ПРИКЛАДНОЙ СЕМАНТИЧЕСКИЙ СЛОВАРЬ РУСЛАН: ОСНОВНАЯ КОНЦЕПЦИЯ И ОБНОВЛЕННЫЙ ПОДХОД

Леонтьева Н. Н. (leont-nn@yandex.ru)
НИВЦ МГУ, Москва, Россия

Ермаков М. В. (mermakov@gmail.com)
РГГУ, Москва, Россия

Крылов С. А. (krylov-58@mail.ru)
ИВ РАН, Москва, Россия

Семенова С. Ю. (sonya_sem@mail.ru)
ИНИОН РАН, Москва, Россия

Соколова Е. Г. (minegot@rambler.ru)
РГГУ, Москва, Россия

Ключевые слова: формальная традиция в компьютерной лексикографии, прикладной семантический словарь, корпусные данные в лексикографии, формальные модели лексического значения

DOI: 10.28995/2075-7182-2020-19-1049-1064

ON TRADITIONAL CONCEPTION AND UPGRADING OF ONE APPLIED SEMANTIC DICTIONARY

Leontyeva N. N. (leont-nn@yandex.ru)
SRCC MSU, Moscow, Russia

Ermakov M. V. (mermsov@gmail.com)
RSUH, Moscow, Russia,

Krylov S. A. (krylov-58@mail.ru)
Institute of Oriental studies RAS, Moscow, Russia,

Semenova S. Yu. (sonya_sem@mail.ru)
INIION RAS, RSUH, Moscow, Russia

Sokolova E. G. (minegot@rambler.ru)
RSUH, Moscow, Russia

The paper deals with upgrading of an electronic semantic dictionary of RUSLAN for automatic processing of Russian texts. The previous versions of the dictionary were created in the 1990-es and early 2000-es mainly for automatic processing of the Russian Federation's state papers. Now the Authors inherit the basic formalism of the Dictionary, including the meta-language and the structure of the dictionary entry. The current version is revised and enlarged in a number of ways. While the initial versions mostly predate the advent of corpus linguistics, the current version is based on corpus data. The Russian National Corpus was used as a source of sample sentences, as well as for determining statistically and empirically which linguistic information is pragmatically relevant. A structural representation for the sample sentences was designed, and a procedure for selecting lexical units from the corpus to use in a pragmatic description of polysemy. A formal representation of situations, previously outlined in the works of Nina N. Leontyeva, has also been detailed and largely realized. Among the lexicon, verbs in particular have received a more flexible description compared to the previous versions, and aspectual meanings are reflected with more nuance.

Keywords: traditions in computational lexicography, NLP-aimed semantic dictionary, corpora data in lexicography, formal models of meaning

Введение

Формализованные семантические словари важны и как ресурсы для прикладных задач, и для фиксации найденных закономерностей в теоретических исследованиях. Наиболее известны англоязычные WordNet — лексическая БД с заданными отношениями синонимии (наборами синсетов) и гиперонимии, и FrameNet — семантический словарь предикатов, в основном глаголов, с описаниями в виде фреймов — наборов бинарных семантических отношений, присоединяющих актанты к предикату.

В России с конца 1950-х гг. создавались семантические словари, которые впоследствии использовались в системах машинного перевода. В них закладывались основы формальной семантической интерпретации лексики. В частности, Н. Н. Леонтьева в [Леонтьева, 1967] предложила бинарные смысловые отношения на год раньше Ч. Филлмора — основателя FrameNet [Fillmore,1968]. Смысловые (или семантические) отношения затем стали основой формализма UNL, который сменил (или дополнил) в практических системах поверхностные и глубинно-семантические отношения, предложенные Н. Хомским. Несмотря на богатую историю нам неизвестен подобный WordNet и FrameNet представительный по объему и общедоступный ресурс для Русского языка (РЯ), хотя несколько подходов к созданию подобных словарей для РЯ предпринимались.

Определенный всплеск в развитии формализованных семантических словарных систем в России возник на рубеже XX и XXI вв. Одной из значительных разработок стала система «Лексикограф», в том числе база данных (БД) «Русский глагол» [Кустова, Падучева, 1994 и др.]. В целом на основе проекта «Лексикограф» была спроектирована семантическая разметка для НКРЯ, проведены глубокие исследования по семантике глаголов, предметных имен, наречий.

В ряду формализованных лексикографических систем находится и словарь РУСЛАН, рассматриваемый в данной работе. Первые версии создавались с 1990-х до второй половины 2000-х гг. Начиная с 2017 г. благодаря поддержке РФФИ (Проект 17-04-00594-ОГН «Автоматический словарь РУСЛАН: обновленная концепция, новая лексика» на 2017–2019 годы) работы по развитию словаря были возобновлены. В его структуре изначально заложен ряд решений, которые не только не устарели, но были повторены в вышеназванных англоязычных базах. В настоящее время авторы видят свою задачу как в поддержке и расширении, так и в модернизации словаря.

1. К предыстории вопроса. Начальные этапы формирования словаря РУСЛАН

Прикладной семантический словарь РУСЛАН, предназначенный для автоматического анализа текста, относится к формальному направлению в отечественной компьютерной лексикографии. Он прямо связан с поиском Языка-посредника для систем машинного перевода (МП), получивших начиная с конца 1950-хх гг. бурное развитие. К тому времени в США уже существовали информационно-поисковые системы (ИПС), работавшие на основе словарей

тезаурусного типа. Подобные работы начались и в отечественном главном информационном центре — ВИНИТИ АН СССР, который задал единую методику создания тезаурусов для многочисленных отраслевых НИИ. Хотя каждый из НИИ имел свою специфическую тематику и, соответственно, лексику (терминологию), структура тезауруса стремилась к единообразию типов отношений между единицами. Отношения были в основном иерархические, но потом появились их разновидности, которым было дано пока общее имя «ассоциативные связи». Каждая предметная область (ПО) стремилась к большей точности в получении нужной информации и добавляла свои семантические пометы. При этом связи в ИПС описывали только парадигматику, но зато Целого текста. Вот тут и понадобилась Лингвистика, которая описывает синтагматику текстовых единиц. Но она локальна, работает только в пределах предложения и стремится построить правильную структуру дерева, с которой можно работать формально. А задачи МП и других типов обработки текста требуют работать с текстом в целом. Казалось бы, почему не иметь инструмент, совмещающий парадигматику и синтагматику? Ведь отношения между единицами ИПС с самого начала были более семантичными, чем чисто синтаксические в составе предложения. Но и лингвистические отношения между словами, выделенными как единицы текста, стремятся к тому, чтобы «обогатиться» семантическими пометами и выйти скорее к уровню Семантики. Лингвистика оказалась не готовой сразу принять этот вызов. А Информатика не готова была погрузиться в тонкие различия значений одной лексемы или конкретной синтаксической связи: ведь она уже работала, обслуживая худо-бедно пользователей. Стали создаваться смешанные системы, выискивающие «нужную» информацию, по заданному пользователем образцу (Information Extraction Systems), но с лингвистическим анализом. Некоторые надежды вселили системы типа FrameWork. Они позволили собирать сложные семантические единицы типа Ситуаций, которые можно достроить до единиц следующего уровня — Текстовых Событий, способных оцениваться с точки зрения истинности/ложности.

В постсоветской России тоже возникло много систем, чаще «доморощенных» (для одной темы, одной задачи и т. п., синтаксически или лексически ориентированных и т. д.). Но Задача создания единой и доступной всем словарной базы для анализа широкого круга деловых текстов осталась нерешённой. Такая задача и была поставлена перед словарём РУСЛАН.

Основной массив словаря вводился в словарную базу в 1990-е и первую половину 2000-х гг. под руководством Н. Н. Леонтьевой. Сначала, в 1990-е гг., словарь создавался как ресурс для системы ПОЛИТекст, нацеленной на информационный анализ официальных документов РФ в Институте США и Канады РАН. Затем, на рубеже веков, работы были перебазированы в НИВЦ МГУ, где словарь получил теперешнее название (РУСЛАН). При этом словник был пополнен общелитературной лексикой, и словарь стал развиваться как общелексический. Формализм словаря, включающий способ представления лингвистических сведений, близкий к алгебраическому (с помощью функциональных зависимостей и бинарных отношений), круг отражаемых сведений и Метаязык (последний задан пока неполной системой конструкций /«синтаксис»/

и структурно-семантической классификацией лексики) разработала Н. Н. Леонтьева. Формализм был построен в соответствии с предложенной ею же моделью понимания текста [Леонтьеваи др., 2001]. В свое время этот формализм прошёл несколько апробаций, ручных (в работе со студентами, с привлечением разных тематик), нескольких ИПС, системы французско-русского перевода ФРАП [Леонтьева и др., 1979], системы ПОЛИТекст, системы русско-английского перевода ДИАЛИНГ, разрабатывавшейся в конце 1990-х гг. и эволюционировавшей потом в Автоматизированное рабочее место прикладного лингвиста, бета-версия.

Словарь РУСЛАН является коллективным продуктом. В разработке статей на разных этапах вместе с авторами данного доклада принимали участие Е. В. Горелик, Е. Р. Иоанесян, А. С. Панина, Е. М. Сморгунова, М. Г. Шаталова, О. А. Штернова. Различные аспекты словарной работы неоднократно отражались в сообщениях на конференции Диалог, напр., [Семенова, 2003а].

В 2015–2016 гг. практическая разработка словаря была приостановлена из-за отсутствия финансирования. В этот период Н. Н. Леонтьева и программист словарной базы передали всё сделанное «народу», т. е. всем тем, кто готов продолжать данный словарь или использовать его для опытов анализа текстов. В 2017 г. благодаря поддержке РФФИ словарные работы были возобновлены остальными авторами (в проекте также участвовала А. С. Панина), и с этого момента работы нацелены не только на поддержку, но и на модернизацию словаря. Кроме проблемы массовых семантических описаний, осталось много трудностей теоретического характера (напр., правила адаптации ситуаций «навстречу» ПО, развитие зоны прагматики и др.).

2. Разработка обновленной версии: конкретные задачи и методики

Обрисуем вкратце круг сведений, отражаемых в основном, «традиционном», разделе словаря РУСЛАН (кроме основного, в процессе модернизации образованы два новых раздела — экспериментальных; см. ниже пп. 3 и 4).

В традиционном разделе описание слова (или, чаще, лексемы) включает информацию нескольких типов: грамматические характеристики; онтологический (или семантический) класс (ЯВЛЕНие , ДЕЙСТВие и др.); структурированные сведения о грамматико-семантической сочетаемости (зона валентностей); тезаурусные связи; словообразовательные и синтагматические лексические функции; термины и словосочетания с описываемым словом; некоторые энциклопедические функции; английские эквиваленты и нек. др. данные.

В общем виде статья основной части словаря включает 11 зон (семантическую, энциклопедическую, лексическую; зону иллюстраций и др.); для каждой зоны предусмотрен свой набор полей и единиц метаязыка. В процессе обновления словаря потребовалась балансировка по каждому типу словарной информации.

Кроме того, в задачи модернизации вошли:

- увеличение словника;
- более подробное (чем в прежних версиях) представление полисемии (при традиционном для данного продукта эмпирическом ограничении — не более 5 лексем для слова);
- ввод свежих текстовых иллюстраций.

Словарь аккумулирует разноплановую лексикографическую информацию, и при его ведении задействован целый ряд методических установок, определяемых характером информации.

В целом, существенную роль при получении обновленной версии играет опора на аргументы статистического характера.

Так, статистическое обоснование положено в основу пополнения словника новыми (для данного словаря!) лексическими единицами. Словник теперь соотносится с данными из Частотного словаря[1]: в отборе новых вокабул основной тактикой стало заполнение статистических лакун (прежнего словника) с опорой на указанный Частотный словарь. Условным порогом для включения вокабулы в обновляемый словник РУСЛАНа сейчас служит значение ее частотной характеристики (ipm), превышающее 1.

Статистика учитывается и при выборе прагматически важных лексем многозначных слов в условиях ограничения, принятого в словаре — не более 5 лексем на слово; ограничение обусловлено потенциальными трудностями дизамбигуации при применении словаря в NLP [Леонтьева, Семенова, 2002]. Если у слова толковые словари выделяют более 5 лексем, то для отбора наиболее частотных привлекается Национальный корпус русского языка (НКРЯ). Частотность лексемы оценивается по первой сотне вхождений слова в корпус, причем сотня первых вхождений рассматривается как в основном подкорпусе, так и в газетном. Тем самым, первая сотня вхождений слова в каждый из двух подкорпусов (преимущественно отражающих современный узус), рассматривается как минимальная выборка для прагматического определения полисемии. Реальная глубина просмотра корпусных вхождений, как правило, бывает большей, но основные приближенные оценки делаются именно на первой сотне [Семенова, Панина, 2019]. Опора на первую сотню вхождений при отборе лексем обусловлена необходимостью ввести ограничение на глубину просмотра подкорпусов; ведь лексика, размещаемая в РУСЛАНе, весьма частотна, и в подкорпусах содержится до нескольких тысяч вхождений рассматриваемых полисемичных слов.

При модернизации большое внимание уделено иллюстрированию. В прежних вариантах иллюстрации были, в основном, модельными [Семенова, 2003а]; зона иллюстраций состояла из одного поля (в котором и фиксировались модельные примеры).

В новой версии было решено развить и структурировать зону иллюстраций, при этом задействовать материал НКРЯ, со ссылками на корпус и на цитируемые в нем источники [Семенова, 2017]. Структуризация зоны иллюстраций

[1] Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

и широкое использование корпусных материалов стали новыми моментами, отличительными от прежних реализаций.

Теперь зона иллюстраций включает как поле модельных примеров (ИЛЛ_МОД), так и ряд полей для примеров корпусных: поле иллюстраций общего характера (ИЛЛ_ОБЩ); поля для «прицельного» иллюстрирования актантов (ИЛЛ_A_i, i = 1, 2, …); поле для показа в тексте фразеологических единиц (ИЛЛ_ФРАЗ) и поле для особых употреблений слова (ИЛЛ_ОСОБ).

Поле модельных примеров решено сохранить, поскольку такие примеры лаконичны и непосредственно отражают ход мыслей лексикографа, его умозрительное представление о типовых (часто минимальных) контекстах лексемы. В ряде случаев в поле ИЛЛ_МОД введены модельные иллюстрации из «Малого академического словаря»[2] (МАС), в силу особого доверия к лексикографической интуиции составителей МАС.

Опыт показал, что для поля корпусных примеров ИЛЛ_ОБЩ целесообразно, с одной стороны, подбирать примеры с полным (или представительным) набором валентностей (как бы не акцентируя никакие из них), а с другой стороны, желательно также размещать примеры, в которых валентности не насыщены — напр., такие, в которых слово репрезентирует свой денотат «как таковой»; ср.:

> *Вероятность стала физическим понятием …, когда Максвелл открыл … первый закон статистической физики ….* [«Знание — сила», 2010 /НКРЯ]

(описываемое слово — *вероятность*; здесь и далее корпусные примеры сокращены, а в словаре Руслан они приводятся полностью).

Поля ИЛЛ_A_i могут заполняться по комбинаторному принципу — с показом (в сочетаниях) тех грамматических моделей (чаще всего предложно-падежных) и семантических классов актанта, которые указаны в зоне валентностей.

Поле ИЛЛ_ФРАЗ иллюстрирует поведение несвободных сочетаний заглавного слова, приведенных в полях словосочетаний, терминов и контекстных лексических функций, причем корпус нередко позволяет обнаруживать новые, умозрительно не предвидевшиеся фразеологизмы, например, *логика вещей* для слова *логика* (более 100 вхождений в основной подкорпус НКРЯ и более 150 вхождений — в газетный; дата обращения 5.05.2020).

Поле ИЛЛ_ОСОБ собирает нестандартные употребления заглавного слова — напр., те, которые не соответствуют (или не вполне соответствуют) описываемой лексеме и которые полезно зафиксировать для будущих уточнений словарных описаний. Могут отражаться и другие «нестандартности». Напр., употребление не самого слова, а его деривата-композита, не включенного в словник (*государство-победитель* для статьи слова *победитель*); нестандартная тональность контекста (например, фрагмент *падение Трои* для того же слова *победитель*) и проч. Систематизация «нестандартностей», встречающихся в корпусном материале (как обусловленных форматом и контентом данного словаря, так и более объективных), может быть предметом отдельного анализа.

---

[2] Словарь русского языка в 4-х томах. М., «Русский язык», 1999. тт. 1–4.

Иллюстрирование направлено не только на показ контекстов, но и на потенциальное использование иллюстраций в NLP — прежде всего, для машинного обучения в целях дизамбигуации; исключение может составлять поле ИЛЛ_ОСОБ — в силу нестандартного характера материала это поле адресовано, гл. образом, самим лексикографам.

Электронная форма словаря не накладывает ограничений на объем иллюстраций, и это дает составителям определенную свободу при отборе примеров (нет ограничений на число примеров, показавшихся чем-либо примечательными) и позволяет накапливать релевантный текстовый материал.

Основной принцип отбора корпусных примеров — близость к «идеальному» прототипическому контексту, расположенность в рамках некоторого фрейма (или фрагмента тезауруса, или семантического поля), образующегося вокруг заглавного слова; так, для имени *победитель* соответствующими прототипу представляются концепты ПРОТИВОСТОЯНИЕ, СОПЕРНИК, НАГРАДА.

Отобранные корпусные предложения заносятся в зону иллюстраций без купюр. Эта стратегия соответствует одной из общих идей корпусной лингвистики (а ранее машинных фондов) — обеспечить целостность и сохранность текста. Можно отметить, что и разработчики НКРЯ сейчас стремятся к более развернутой выдаче, новая поисковая программа выдает целый абзац. При ведении словаря Х составители ограничиваются одним (но целостным) предложением.

Для цитат из корпуса в зоне иллюстраций предусмотрена возможность некоторой внутренней разметки. Можно, с одной стороны, выделять сегменты, представляющиеся особо релевантными для иллюстрирования заглавного слова, с другой стороны, определенные сегменты помечать как балласт.

В качестве балласта могут, к примеру, выступать сегменты, выражающие несущественные (с точки зрения иллюстрирования заглавного слова) фактографические данные: имена персон, хронологию, мало релевантные исторические реалии.

Как балластные могут интерпретироваться также именования сущностей, окказиональных по отношению к мысленному фрейму заглавного слова; так, в следующем примере таковыми по отношению к слову *победитель* видятся *вагоны*:

… победитель должен локализовать производство своих вагонов в России. // «Эксперт», 2014 / НКРЯ].

Конечно, внутренняя разметка корпусных примеров не лишена субъективизма, но она (как дополнительное средство) может учитываться, напр., при разрешении полисемии.

Как уже отмечено, в словаре РУСЛАН возникли два новых раздела. Один из них отражает опыт описания ситуаций, связанных с предикатными лексемами. Этот раздел ведет М. В. Ермаков. Другой раздел содержит статьи глаголов и их адъективных дериватов; в нем более детально, чем в словаре в целом, отражены видовые значения глаголов. Эскиз этого раздела принадлежит Е. Г. Соколовой.

3. Экспериментальный раздел формализованного описания ситуаций

Данный экспериментальный раздел словаря РУСЛАН отражает результаты моделирования семантики предикатных слов при помощи аппарата элементарных ситуаций (ЭСит). Идея разработки и применения такого аппарата намечена в [Леонтьева, 2001]; тем самым, можно считать, что данный раздел словаря построен в развитие названной работы.

В прежних версиях словаря ЭСит использовались, в основном, для описания связей между актантами и другими семантическими элементами. Например, в полях валентностей ВАЛ лексемы *переводить 1* (*Юноша перевёл младшую сестру с одной стороны улицы на другую*) в наличии агент (АГЕНТ, он же первый актант А1), пациенс (ПАЦИЕН, А2), исходная и конечная точка (ИСХ-Т и КОН-Т, А3 и А4). Между этими четырьмя актантами существуют семантические связи: так, после осуществления действия *переводить* пациенс А2 оказывается в конечной точке А4, что выражено в поле ДОП (дополнительных отношений) отношением ЛОК /локализация/ (А4, А2). Однако ЭСит, являясь семантическими отношениями между парой единиц, позволяют записывать более подробные логические презумпции и следствия, вытекающие из значения ситуации. В том же примере можно указать, что до совершения действия А2 находился в точке А3: Элементарная ситуация №1 (ЭСит1) описывает нахождение А2 в точке А3, или ЛОК (А3, А2) и предшествует ЭСит2, когда А2 находится в точке А4 (ЛОК (А4, А2)). Это может быть записано следующим образом в поле ЭСит:

Эсит1. ПРЕДШ (ЭСит2, ЭСит3)
Эсит2. ЛОК (А3, А2)
ЭСит3. ЛОК (А4, А2)

Таким образом, можно сделать вывод, что объект А2 после совершения действия находится в точке А4 .

Благодаря элементарной ситуации ПРЕДШ /«предшествование»/ (А, В) и ее коррелята ПОСЛЕ (А, В) становится возможными моделирование семантики глаголов действия, декомпозиция изменения обозначаемых ситуаций во времени. Такое представление смысла глагольной лексики является более глубоким, чем указание конъюнкций семантических характеристик (что представлено в традиционной части словаря РУСЛАН). Отметим, что близкий (темпоральный) подход был применен в БД «Русский глагол» системы «Лексикограф» [Кустова, Падучева, 1994 и др.]; ср. использованные в этой базе понятия «экспозиция», «момент наблюдения», «импликация»; при этом формализм носил теоретический характер и не был нацелен на NLP.

Бинарная алгебраическая ЭСит представима и в виде графа (в искусственном интеллекте она обычно участвует в интерпретации графа семантической сети). Она может быть применена, например, на стадии семантического анализа предложения — для визуализации семантического представления и экспликации связей между актантами слов.

На предыдущих этапах развития словаря аппарат ЭСит использовался только для описания предметной области «Преступления» [Ермаков, 2007]. Однако данный аппарат можно применять и для других ПО, в т. ч. для представления лексики самой общей тематики.

При разработке данного раздела осуществлялись опыты использования языка ЭСит как универсального языка-посредника для описания общей семантики у слов, связанных по смыслу и входящих в один тематический кластер. Это позволяет создавать тематические «фреймы», благодаря которым слова одной тематики могут описываться сообща, и потенциально наследовать часть описания ситуаций для более общих слов. Напр., семантическое описание и валентности лексемы *продажа* могут частично наследоваться от более общей лексемы (гиперонима) *обмен*. Обе лексемы имеют валентности агента (АГЕНТ) (кто продаёт или меняет), контрагента /покупателя/ (К-АГЕНТ), объекта, который меняется или продаётся (ОБ) и объекта, на который меняется первый объект (ВМЕСТО). В описание четвёртого актанта слова *продажа* будет добавлена семантическая характеристика «финансовое» (ФИН). Отметим, что наследование свойств есть и в других формальных системах, например, во FrameNet [Fillmore, 1968]; о наследовании свойств понятия экземплярами упоминается и в [Шенк, 1980: 170].

В целом аппарат ЭСит использован при описании предикатных слов (действий и ситуаций); такое описание целесообразно (наряду с более традиционным для РУСЛАНа использованием зон валентностей (ВАЛ) и дополнительных отношений (ДОП) между актантами). Аппарат ЭСит позволяет отразить важные изменения состояния, местоположения, принадлежности и т. п. актантов, которые по прототипу происходят в результате обозначаемых действий и в качестве исхода обозначаемых ситуаций. Отражение изменений состояний, локализаций, принадлежности и других следствий может в дальнейшем найти применение в системах с логическим выводом.

В процессе работы над ЭСит проявились сложности как в описании разных семантических узлов и понятий (различие между материальными предметами и информацией о них, между предположениями, реальной или возможной ситуацией), так и при использовании языка-посредника (т. е. самого аппарата ЭСит), который хотя и обладает мощными возможностями, требует дальнейшего упорядочения и уточнения. Проверку формализованных тематических описаний может дать и дальнейшее развитие аппарата ЭСит, и применение словаря РУСЛАН в действующих системах NLP.

4. Экспериментальный раздел описания глаголов с отражением видовых значений

Одна из проблем неадекватности традиционного описания для прикладного словаря связана с представлением в последнем семантики глагольного вида. В традиционном толковом словаре, например, в МАС, указывается видовая форма, а не видовое значение. Но раз есть форма, значение должно существовать. Это потенциальное видовое значение автоматически распространяется

на все лексические значения глагола, описанные в словарной статье, как в МАС. Чаще значения описываются для формы Совершенного вида (СВ), а для формы противоположного вида даётся ссылка на значения, для которых она образуется. В частности, у глагола *угодить–угождать* в МАС приводятся толкования для формы СВ, а для формы НСВ указано, что она образуется от знач. 1 «Удовлетворить кого-л., сделав что-л. приятное, нужное, желаемое» (напр., *угодить начальнику*). Но при переносе этой информации в прикладной словарь и использовании его для анализа текста могут возникнуть проблемы. В частности, любая встреченная словоформа *угождать* будет отнесена к знач. 1, хотя это неверно. В нижеследующем примере в НСВ реализовано знач. 2 «Попасть в какие-л. условия, оказаться в каких-л. обстоятельствах»:

(1) *«В аморальном обществе мафия продолжает процветать даже тогда, когда ее боссы* ***угождают*** *за решетку»* [НКРЯ].

Возможным решением может быть указание в словаре акционального класса[3], и в этом определенным ориентиром могла бы служить БД «Русский глагол» системы «Лексикограф», в которой указывается таксономическая категория (Т-КАТЕГОРИЯ), связанная с акциональным классом. Например, для глагола ПОКЛЯСТЬСЯ 1 указывается «Т-КАТЕГОРИЯ действие: моментальное» и аспектуальная характеристика противоположного вида: «НСВ: КЛЯСТЬСЯ 1: перфектное состояние»; фрагмент статьи ПОКЛЯСТЬСЯ 1 цит. по [Семенова, 2003б].

Определение и состав акциональных классов варьируются у разных исследователей; а в ряде словарных источников напр., в МАС или в «Активном словаре русского языка» акциональный класс не указывается.

Варианты лексического значения глагола одного акционального класса могут вести себя по-разному по отношению к образованию частновидового значения в предложении. В [Соколова, 2019] предложено знач. 2 глагола *угодить-угождать* отнести к акциональному классу «моментальный процесс». Если этим глаголом обозначен социальный процесс, образуется кратное частновидовое значение, как в примере (1), а если обозначен физический процесс, например, *угодить в яму,* то кратного значения не будет (по крайней мере, такие примеры не обнаружены в НКРЯ). В обоих вариантах обозначаемый процесс «*угодить*» происходит случайно, но физ. процесс происходит внезапно, поэтому, скорее всего, единичен, а в основе соц. процесса часто лежит постоянное свойство (или поведение) участника, обусловливающее возможность повторения процесса. В (1) это криминальное занятие боссов. В [Храковский, 2014] высказывается гипотеза о том, что кратное значение создается исключительно контекстом, но, как только что показано, лексическое значение может препятствовать ему (или, наоборот предполагать).

---

[3] «Под акциональным классом глагольной лексемы … понимается ее семантическая характеристика, релевантная для аспектуальных категорий. В первую очередь речь идет о категории „собственно вида“, противопоставляющей перфективное vs. имперфективное значения» [Шлуинский, 2006].

Видовые компоненты в экспериментальном разделе РУСЛАНа вносятся в словарные статьи на основе гипотезы об эшелонировании видовых значений, высказанной в [Соколова, 2019]. Гипотеза предполагает существование «надконтекстных видовых значений», которые и приписываются лексеме в словаре.

В экспериментальном разделе РУСЛАНа продолжен путь описаний, начатый в [Соколова, 2019] с глаголом *угодить-угождать*; этот путь распространен на другие глаголы. В указанном разделе каждой статье глагола приписывается надконтекстное видовое значение, коррелирующее с лексическим значением. Использовались выделенные в [Соколова, 2019] четыре надконтекстных видовых значения: СВ1 — единичное событие, НСВ1 — плавное течение процесса без конечных точек, НСВ2 — кратность или повторяемость и НСВ3 — абстрактное отношение, устанавливаемое говорящим по своей воле между любыми объектами. Надконтекстные значения организованы в систему грамматических оппозиций [Соколова, 2019: 118].

Также в статью глагола вводится категория «сфера восприятия» [Соколова, 2019], отражающая аспекты восприятия ситуации коммуникантами и принимающая четыре значения:

- (физ.) физическая — процесс, физически воспринимаемый участником, например, *я бежал*; *я видел, как он бежал*;
- (мент.) ментальная — процесс, протекающий в сознании участника, например, *я лукавил, он лукавил*;
- (соц.) социальная — процесс, известный говорящему по социальному опыту, напр., *его арестовали*;
- (абстр.) абстрактная — декларируемое говорящим положение вещей, например, *Украина граничит с Россией*.

Отметим, что категория «сфера восприятия» соотносится с параметром «тематический класс» в БД «Русский глагол» системы «Лексикограф» (а также с некоторыми семантическим классами метаязыка основной части словаря РУСЛАН).

Оба признака, надконтекстное видовое значение (соотносимое с акциональным классом) и сфера восприятия, обслуживают различные лексемы многозначного глагола, например:

1. *заходить* (посещать) «событие»: НСВ2 & соц. — *Иван заходит ко мне*; НСВ1 & физ. — *И тут заходит начальник лагеря*;
2. *заходить* (*на посадку, справа* и т. д.) (физ.) «действие, манёвр»: НСВ1 — *самолёт заходит на посадку;* НСВ2- *Пилот Иванов всегда заходит на посадку в восточном коридоре*. Без прямой поддержки контекстом первое образует кратное частновидовое значение, а второе — актуально-длительное.

Более подробно результаты лексикографического эксперимента будут представлены в отдельной статье.

Важной особенностью словаря РУСЛАН является его минимализм. Не ставится задача описать для глагола максимальную парадигму значений или типов контекстов, как, напр., во FrameNet и нек. отечественных толковых словарях.

Это позволяет сосредоточиться на инварианте многозначного глагола, представляемом в предлагаемом подходе как набор, состоящий из нескольких абстрактных сущностей, ассоциированных с актантами глагола в разных его значениях. Объективность описаний повышается за счёт сверки глагольных описаний в словаре РУСЛАН с толкованиями в МАС.

На настоящий момент раздел словаря РУСЛАН, включающий в словарную статью глагола видовые семантические признаки, имеет статус предварительного, исследовательского. Кроме видового значения полезно по возможности формализовать в РУСЛАНе регулярные трансформации, связанные с залоговыми и возвратными формами. Здесь также можно опираться на публикации создателей системы «Лексикограф», в частности, фундаментальные труды Е. В. Падучевой, как и на работы других специалистов по взаимодействию грамматики и семантики русских глаголов.

5. Некоторые выводы

Полезной особенностью охарактеризованного выше словаря РУСЛАН является сочетание в нем разнообразной информации — лингвистической, тезаурусной, энциклопедической. Экспериментальные разделы приближают его к онтологиям, а также показывают, что его формализм удобен для реализации новых (по сравнению с исходными) видов моделирования взаимосвязей грамматической, семантической и энциклопедической информации. В результате модернизации словарь существенно обогатился иллюстративным материалом из НКРЯ. Практика учета корпусных данных, сложившаяся в ходе работ по модернизации, позволяет повысить точность лексикографирования.

Конечно, словарь нуждается в дальнейшей проработке — в наращивании словника, расширении круга отражаемых сведений для конкретных категорий лексических единиц, углублении и лексикографической апробации намеченных формализмов. Авторам очень хотелось бы также, чтобы данный словарь развивался как общедоступный ресурс и находил применение в задачах текстовой обработки. И для достижения этих благих целей требуется поддержка научного сообщества.

Литература

1. *Ермаков М. В.* (2007). Коррекция смысловых отношений как этап семантического анализа (на материале криминальных сводок) // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.). М.: Изд. РГГУ. С. 178–182.
2. *Кустова Г. И., Падучева Е. В.* (1994). Словарь как лексическая база данных // Вопросы языкознания, № 4. С. 96–106.
3. *Леонтьева Н. Н.* (1967). Об одном способе представления смысла текста // Труды 3-й Всесоюз. конф. по информационно-поисковым системам и автоматизир. обработке научно-техн. информации. Т. 2. М., 1967. С. 192–204.

4. *Леонтьева Н. Н. и др.* (1979). Семантическая словарная статья в системе ФРАП. Предварительные публикации / Леонтьева Н. Н., Кудряшова И. М., Соколова Е. Г. / М.
5. *Леонтьева Н. Н.* (2001). К теории автоматического понимания естественных текстов. Часть 2. Семантические словари: состав, структура, методика создания. М.: Изд-во МГУ.
6. *Леонтьева Н. Н., Семенова С. Ю.* (2002). Об отражении полисемии в прикладном семантическом словаре // Компьютерная лингвистика и интеллектуальные технологии. Тр. Междунар. семинара Диалог'2002. Протвино, 6–11 июня 2002 г. М.: Наука. Т. 2. С. 489–496.
7. *Семенова С. Ю.* (2003, а). Примеры в компьютерном семантическом словаре: некоторые наблюдения над процессом подбора // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. конф. Диалог'2003 (Протвино, 11–16 июня 2003 г.). М.: Наука. С. 593–598.
8. *Семенова С. Ю.* (2003, б). О компьютерной лексикографии: семантика и тезаурусные связи прикладных словарях // Теория и практика общественно-научной информации: Ежегодник. М.: РАН. ИНИОН. Вып. 18. С. 88–108.
9. *Семенова* (2017). Об использовании данных Национального корпуса русского языка для иллюстрирования статей компьютерного семантического словаря // Труды международной конференции «Корпусная лингвистика — 2017». СПб. С. 321–324.
10. *Семенова С. Ю., Панина А. С.* (2019). Опыт использования данных НКРЯ при описании полисемии в прикладном семантическом словаре // Труды Международной конференции «Корпусная лингвистика-2019» — СПб.: Изд-во С.-Петерб. ун-та. С.234–240.
11. *Соколова Е. Г.* (2019). Эшелонирование значений категории вида и их взаимодействие с лексическими значениями (на примере глагола угодить–угождать) // Вестник РГГУ. Серия «Литературоведение. Языкознание. Культурология». № 7. С. 101–139.
12. *Храковский В. С.* (2014). Есть ли у несовершенного вида в русском языке повторительное (неограниченно-кратное / многократное / итеративное / хабитуальное) значение? // ВЯ, №4, С. 3–12.
13. *Шенк Р.* (1980). Обработка концептуальной информации / Пер. с англ. М.: Энергия.
14. *Шлуинский А. Б.* (2006). Акциональные классы глаголов в хакасском языке // Третья конференция по типологии и грамматике для молодых исследователей: материалы конф. СПб.: Нестор — История. С.158–164.
15. *Fillmore Charles* (1968). The Case for Case. // «Universals in Linguistic Theory», ed. by E. Bach, R. T. Harms, New York etc.

References

1. *Ermakov M. V.* (2007), Correction of semantyc relations as a stage of semantic analysis (based on the material of criminal reports) [Korrekciya smyslovyh otnoshenij kak etap semanticheskogo analiza (na materiale kriminal'nyh svodok)], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2007" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2007"], Bekasovo, Moscow, pp. 178–182.
2. *Fillmore Charles* (1968), The Case for Case.— In: «Universals in Linguistic Theory», ed. by E. Bach, R. T. Harms, New York etc.
3. *Khrakovsky V. S.* (2014), Does an imperfect form in Russian have a repeated (unlimited-multiple / multiple / iterative / habitual) value? [Est' li u nesovershennogo vida v russkom yazyke povtoritel'noe (neogranichenno-kratnoe / mnogokratnoe / iterativnoe / habitual'noe) znachenie?], Bulletin of linguistics [Voprosy Yazykoznaniya], Moscow, no. 4, pp. 3–12.
4. *Kustova G. I., Paducheva E. V.* (1994), Dictionary as a lexical database [Slovar' kak leksicheskaya baza dannyx] , Questions of linguistics [Voprosy' yazy'koznaniya], no. 4., pp. 96–106.
5. *Leontyeva N. N.* (1967), On one way to represent the meaning of a text [Ob odnom sposobe predstavleniya smysla teksta], Proceedings of the 3rd all-Union. Conf. for search systems and automatic processing of science and technology information [Trudy 3-j Vsesoyuznoj konferencii. po informacionno-poiskovym sistemam i avtomatizir. obrabotke nauchno-tekhnicheskoj Informacii], Moscow, Vol. 2. pp. 192–204.
6. *Leontyeva N. N., Kudryashova I. M., Sokolova E. G.* (1979), Semantic dictionary entry in the FRAT (french-russian automatic translation) system. Pre-publication [Semanticheskaya slovarnaya stat'ya v sisteme FRAP. Predvaritel'nye publikacii], Moscow.
7. *Leontyeva N. N.* (2001), On the theory of automatic understanding of natural texts. Part 2. Semantic dictionaries: composition, structure, and method of creation [K teorii avtomaticheskogo ponimaniya estestvennyh tekstov. Chast' 2. Semanticheskie slovari: sostav, struktura, metodika sozdaniya.], MSU publishing, Moscow.
8. *Leontyeva N. N., Semenova S. Yu.* (2002), On the description of polysemy in the applied semantic dictionary [Ob otrazhenii polisemii v prikladnom semanticheskom slovare], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2002" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2002"], Protvyno, Moscow, Vol. 2. pp. 489–496.
9. *Schank R. C.* (1975), Conceptual information processing, New York.
10. *Semenova S. Yu.* (2003a), Examples in the computer semantic dictionary: some observations on the selection process [Primery v komp'yuternom semanticheskom slovare: nekotorye nablyudeniya nad processom podbora], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2003" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2003"], Moscow, pp. 593–598.

11. *Semenova S. U.* (2003b), On computational lexicography: semantics and thesaurus relations in applied dictionaries [O komp'yuternoj leksikografii: semantika i tezaurusny'e svyazi v prikladny'x slovaryax], Theory and practice of social and scientific information [Teoriya i praktika obshhestvenno-nauchnoj informacii]: Yearbook, Issue 18, pp. 88–108.
12. *Semenova S. Yu.* (2017), On using data from the National corpus of the Russian language to illustrate articles in a computer semantic dictionary [Ob ispol'zovanii dannyh Nacional'nogo korpusa russkogo yazyka dlya illyustrirovaniya statej komp'yuternogo semanticheskogo slovarya], Proceedings Of the international conference "Corpus linguistics-2017» [Trudy Mezhdunarodnoj konferencii «Korpusnaya lingvistika-2017»], Sankt-Petersburg, pp.321–324.
13. *Semenova S. Yu., Panina A. S.* (2019), An experience in the use of NCRL (National Corpus of the Russian Language) data, describing polysemy in an applied semantic dictinary [Opyt ispol'zovaniya dannyh NKRYA pri opisanii polisemii v prikladnom semanticheskom slovare], Proceedings Of the international conference "Corpus linguistics-2019» [Trudy Mezhdunarodnoj konferencii «Korpusnaya lingvistika-2019»], Sankt-Petersburg university publishing, pp.234–240.
14. *Shluinsky A. B.* (2006), Actional classes of verbs in the Khakass language [Akcional'nye klassy glagolov v hakasskom yazyke], Third conference on typology and grammar for young researchers: proceedings of the Conf. [Tret'ya konferenciya po tipologii i grammatike dlya molodyh issledovatelej: materialy konf], Spb., pp. 158–164.
15. *Sokolova E. G.* (2019), Two echelons in the semantics of the Russian verbal aspect and their interaction with lexical meanings (illustrated by the verb ugodit' — ugozhdat') [Eshelonirovanie znachenij kategorii vida i ih vzaimodejstvie s leksicheskimi znacheniyami (na primere glagola ugodit' — ugozhdat')] , RSUH/RGGU bulletin, "Philology. Linguistics. Culturology" Series, [Vestnik RGGU. Seriya «Literaturovedenie. YAzykoznanie. Kul'turologiya»], Moscow, no. 7, pp. 101–139.