ISSN 0548-0027

НАУЧНО-ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

СЕРИЯ 2

Информационные процессы и системы

СОДЕРЖАНИЕ

№ 11
1993

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 801.25

Г. И. Кустова, Е. В. Падучева, Е. В. Рахилина, Р. И. Розина, М. В. Филипенко, Н. М. Якубова, Т. Е. Янко

СЛОВАРЬ КАК ЛЕКСИЧЕСКАЯ БАЗА ДАННЫХ: ОБ ЭКСПЕРТНОЙ СИСТЕМЕ «ЛЕКСИКОГРАФ»

Приводится и иллюстрируется идея зависимости семантических и синтаксических свойств глагольной лексемы (семантической структуры, наличия и типа видовой пары, частных видовых значений, сочетаемости с сирконстантами) от ее таксономической категории (действие, процесс, происшествие, состояние). Рассматриваются способы представления этих зависимостей в лексической базе данных реляционного типа.

В данной статье дается описание и лингвистическое обоснование разрабатываемой в ВИНИТИ экспертной системы «Лексикограф», программное обеспечение которой осуществляется И. С. Красильщиком и Е. Н. Хасиной, см. нашу первую публикацию [1]. В основе системы лежит идея представления словарной информации в виде лексической базы данных (БД) реляционного типа: информация о лексеме распределяется по признакам (зонам), имеющим фиксированный набор значений; в результате пользователь может осуществлять поиск любой степени глубины в заданном множестве признаков. Система состоит из двух компонентов: собственно БД и экспертная часть. Основное внимание обращается на семантическую информацию о слове. Таким образом, речь идет, в сущности, о семантическом словаре, представленном в виде реляционной базы данных*.

Для удобства разработчиков ДБ разделена на несколько частей. В работе речь идет о разделе «Глагол», который находится в стадии разработки; разработка же раздела «Предметные имена», в которой принимал участие В. А. Плунгян, близка к завершению [2].

СЛОВАРНАЯ СТАТЬЯ словаря имеет следующие зоны: (1) Заглавная лексема — с примерами ее употребления в предложении; (2) Морфологическая характеристика (по [3]); (3) Актанты; (4) Таксономическая категория; (5) Толкование; (6) Аспектуальная характеристика; (7) Производные значения. Зона Актанты и зона Толкование в свою очередь делятся на поля.

Узловым семантическим признаком слова является его **таксономическая категория** (Т-категория): Т-категория играет такую же роль для семантики слова, как часть речи для грамматики. Простейшие Т-категории глагола — действие, процесс, состояние, происшествие (приблизительно по Вендлеру).

Специфика толкований, применяемых в нашей БД, состоит в том, что они имеют определенный **формат**. Предпосылкой форматирования является расчленение толкования на отдельные синтаксически независимые компоненты (признаки) — аналогично толкованиям в [4] и особенно в [5] (и в противоположность синтаксически связным толкованиям в модели «Смысл↔Текст»). Компоненты имеют предикативную форму, например: 'Субъект действует'; 'Идет процесс в Объекте'; 'Объект перемещается' и под. Форматирование толкования достигается за счет того, что каждый его компонент является значением некоторого **параметра** (один из вариантов идеи форматированного толкования реализован в [5], где компоненты толкования глаголов речи идентифицируются как исходные предпосылки, диктум и иллокутивная цель; ср. также частично форматированные толкования в [6], [7]). Параметры служат названиями зон толкования, например — деятельность, каузация, исходное состояние, конечное состояние, процесс, результат, предел и под. Толкование — это набор семантических компонентов (как значений параметров) и их линейный порядок. Формат характеризуется обязательным наличием определенных параметров на определенных местах толкования.

Формат толкования предопределен принадлежностью глагола той или иной Т-категории. Так, для глагола действия обязательным (категориальным) будет набор параметров «деятельность» и «результат (соответствующий цели Субъекта)»; для глагола происшествия — «исходное состояние» и «конечное состояние». Система Т-категорий имеет иерархическую структуру; например, различные значения параметров «предел» и «каузация» порождают из Т-категории процесс субкатегорию процесс предельный (*таять*) и процесс непредельный (*кипеть*); из Т-категории происшествие — субкатегорию происшествие обычное, с Субъектом (например, *упасть*) и происшествие каузированное, с Объектом (например, *загородить;* так, *Камень загородил вход в пещеру* = 'Камень переместился, в результате вход оказался загорожен').

Строя толкование, мы не стремимся к предельной полноте: обязательно должны быть отражены только **повторяющиеся** семантические противопоставления; т. е. обязательно отмечается только то, что **объединяет** слова друг с другом: различия, если они носят частный характер, могут быть опущены. Так, глаголы *наполнить*

* В 1993 г. работа поддерживалась Российским фондом фундаментальных исследований.

и *заполнить*, в своих основных значениях, имеют одинаковые толкования. Имеющееся между этими словами семантическое различие не эксплицируется, поскольку на данном этапе мы считаем это различие уникальным для данной пары глаголов.

В принципе, информация во всех зонах словарной статьи должна предсказываться из толкования. В самом деле, Т-категория, как следует из уже сказанного, выводима из формата толкования; аспектуальная информация в зоне (2) Морфология тоже задается толкованием. Аспектуальная информация в зоне (6) — наличие видовой пары и допустимый набор частно-видовых значений — также во многих случаях предсказывается толкованием. Например, глаголы состояния (типа *знать*), как известно, не имеют актуально-длительного значения; глаголы постоянного свойства/соотношения (*весить, стоить*) не входят в стандартные видовые пары, и т. д. Есть, однако, и ограничения, не выводимые из семантики; например, отсутствие парного НСВ у глаголов *очнуться, встрепенуться* — это морфологический каприз.

Информация в зоне (3) Актанты предопределена толкованием лишь частично, а частично содержит независимую семантическую и коммуникативную информацию. Каждый актант характеризуется по трем параметрам:

1) синтаксическая характеристика, т. е. поверхностный падеж (различаются: Субъект — подлежащее; Объект — прямое дополнение; и Периферийный актант — предложно-падежные формы косвенных дополнений и обстоятельств не фиксируются, поскольку эти различия не имеют непосредственной семантической значимости);

2) семантическая, или ролевая характеристика, т. е. глубинный падеж по Филлмору (различаются: Агенс, Пациенс, Место и под. роли);

3) таксономическая характеристика ('лицо', 'материальный предмет', 'вещество', 'числовой параметр', 'ситуация' и под.). Избыточен только глубинный падеж (например, наличие в толковании компонента 'Субъект действует' означает, что Субъект данного глагола — Агенс).

Синтаксическая характеристика и Т-категория актанта непосредственно из толкования невыводимы. В толковании актант может быть назван своим поверхностным или глубинным падежом, а иногда и Т-категорией. Переменных для называния актантов мы не используем.

Сопоставление поверхностного падежа актанта с глубинным позволяет охарактеризовать глагол с точки зрения коммуникативной перспективы в смысле Якобсона — Филлмора. Например, глаголы **полного охвата**, типа *наполнять, заваливать* [8], имеют сдвинутую перспективу — Пациенс выражен у них периферийным актантом, а семантическая роль центрального актанта, Объекта, — Место (в норме Пациенс выражен Объектом и входит в перспективу). Сдвиг перспективы и порождает характерный для этих глаголов семантический привесок — семантику «полноты охвата», ср. известные примеры: *загрузить картошку в машину* не равно *загрузить машину картошкой*, поскольку во втором случае, когда *загрузить* употребляется как глагол полного охвата, явно сказано, что машина загружена полностью.

Пользование БД предполагается в нескольких режимах. Помимо семантической информации о каждом отдельном слове, пользователь может получать перечни всевозможных семантических классов глагольных лексем — в принципе, классы задаются любым признаком и любым набором или конфигурацией признаков, например: глаголы СВ, не входящие в видовую пару; физические действия, не допускающие инструмента (например, *полоть, рвать* можно только руками, в противоположность *пахать, резать*). Выразимы на языке БД и традиционные семантические классы; например, глагол движения — это глагол, в толкование которого обязательно входит компонент 'X перемещается в Место' или набор 'в момент t_i X не находится в Месте'; 'в момент t_j X находится в Месте'.

Среди множества (семантических) классов, задаваемых БД, для лингвиста представляют интерес так называемые **релевантные классы.** Релевантный класс — это такой класс (задаваемый набором или конфигурацией признаков в БД), который характеризуется определенной общей особенностью поверхностного поведения входящих в него лексем. Например, релевантным классом является Т-категория: слова одной и той же Т-категории характеризуются одинаковой сочетаемостью с обстоятельствами цели, времени и многими другими общими свойствами: обычно эта сочетаемость считается свободной; на самом же деле она предопределена Т-категорией. Релевантный класс образуют также глаголы полного охвата; импликативные глаголы по Карттунену и мн. др.

В оптимальном случае нахождение релевантного класса позволяет не только получить список глаголов с определенной особенностью поверхностного поведения, но и **объяснить** наличие у данной группы глаголов данного свойства. Например, презумптивный статус компонента «деятельность» у конативов, типа *решить (задачу), уговорить* (факт, отмеченный Ю. Д. Апресяном: *не решил* предполагает 'решал'; *не уговорил* предполагает 'уговаривал'), объясняется наличием в семантике этих глаголов компонента 'удалось', который, в свою очередь, имплицирует компонент 'пытался'.

Базой для формулирования запроса пользователя может быть не только информация о классах слов, но и о том или ином параметре — толкования или актанта — как о совокупности признаков. Например, возможны запросы: полный список Т-категорий; возможные значения параметра «каузация»; Т-категория актанта Агенс и под.

Еще один режим использования БД — проверка всевозможных гипотез о связи семантических характеристик слова с его поверхностными свойствами. Так, база данных дает возможность выяснить, верно ли, что в семантику глагола СВ всегда входит компонент 'наступление нового состояния', как было предложено А. Вежбицкой. (Ответ отрицательный, ср. такие глаголы, как *посветить, защитить, произойти,* которые предполагают скорее конец старого состояния: *плащ защитил меня от дождя* = 'защищал, пока дождь не кончился.)

Второй компонент системы, ее экспертная часть, — это своего рода **грамматика лексикона,** лексическая База знаний (БЗ). БЗ включает семантические правила и общие закономерности, которые позволяют манипулировать семантической и грамматической информацией, помещенной в БД. Мы исходим (опираясь, прежде всего, на работу [9]) из идеи о том, что, в принципе, любая особенность поверхностного поведения слова может быть выведена из его семантики. БЗ должна содержать всевозможные обобщения такого рода, а также формальные правила, позволяющие делать такого рода выводы.

На данном этапе БЗ включает следующие разделы.

1. **Т-категории глагола.** Идея о связи видо-временных характеристик глагола с его семантикой высказывалась многими авторами; самую широкую популярность приобрела классификация Вендлера (ср. также подчиненную аспектуальным задачам семантическую классификацию глаголов Ю. С. Маслова). Однако классификация Вендлера не охватывает всей глагольной лексики и нуждается в уточнениях. Необходимо выявить полный набор Т-категорий русских глаголов,

а также охарактеризовать, исчерпывающим образом, особенности поведения лексемы, вытекающие из ее принадлежности данной Т-категории.

К настоящему моменту установлено, что принадлежность глагола той или иной Т-категории предопределяет: набор основных актантов; возможность образования видовой пары и ее семантический тип; ограничения на дополнительный набор видо-временных значений; сочетаемость с разными видами обстоятельств (цели, времени и под.); способность мотивировать маркированные способы действия. Так, глаголы, относящиеся к Т-категории деятельность (например, *играть*), способны мотивировать делимитатив (*поиграть*) и производный инхоатив (*заиграть*).

Особую проблему составляет разграничение базовых и производных Т-категорий. Так, Т-категория делимитативов (обсуждавшаяся М. Флаером) производная, и потому не случайно, что ей не нашлось места в классификации Вендлера, рассчитанной на базовые категории.

2. **Семантическая сочетаемость.** Ограничения сочетаемости предопределяются не только Т-категорией, но и другими более частными компонентами толкования. Например, у абстрактных глаголов физического действия, типа *расширить* (с компонентами 'способ действия не специфицирован'), затруднена сочетаемость с инструментом [10]. Так, недопустимо **расширить яму лопатой*, при правильном *копать землю лопатой*, хотя действие — расширение ямы — производится, скорее всего, с участием инструмента. Предполагается, что все такого рода связи могут быть исчислены.

3. **Частные видо-временные значения глагола.** В основном ограничения на набор частно-видовых значений выводится из Т-категории, но могут играть роль и другие компоненты толкования; например, невозможность или затрудненность употребления глагола в актуально-длительном значении обусловливается компонентами 'процесс в Объекте сверхкраткий', как у *ударять* или 'процесс в Объекте несинхронный деятельности Субъекта', как у глаголов *стрелять, взрывать, отравлять, убивать*.

4. **Семантическая деривация.** Многим глаголам в нашей БД свойственна регулярная многозначность (в смысле Ю. Д. Апресяна). Ставится задача исчислить регулярные типы семантической деривации, а также выявить **предрасположение** тех или иных классов слов тому или иному типу семантической деривации.

Основной источник регулярной многозначности — метонимический перенос; ср. диатетические соотношения вроде *Сторож наполняет бассейн водой — Вода наполняет бассейн; В своей комедии он высмеивает интеллигенцию — Его комедия высмеивает интеллигенцию,* а также актантные перемещения вроде *полоть сорняки — полоть грядки.* Что касается метафорических переносов, то они обычно считаются непредсказуемыми, что не совсем точно. Так, Т-категория потенциального метафорического деривата иногда может быть предсказана. Например, у глагола действия следует ожидать в первую очередь образования значения, относящегося к Т-категории происшествие (*порезал хлеб — порезал палец*). Что же касается Т-категории актантов семантического деривата, то она, действительно, в значительной степени случайна, ср. *мальчик подпрыгнул* и *цены подпрыгнули.* Случайный характер носит также закрепление метафорического употребления в языке.

Таковы задачи, которые ставит перед собой «Лексикограф». Решение даже части этих задач позволит выйти на новый уровень системности в описании лексики.

СПИСОК ЛИТЕРАТУРЫ

1. Paducheva E. V., Rakhilina E. V. Predicting co-occurrence restrictions by using semantic classifications in the lexicon // COLING-90. Papers presented to the 13-th International conference on computational linguistics. Vol. 3. — Helsinki, 1990.
2. Красильщик И. С., Рахилина Е. В. Предметные имена в системе «Лексикограф» // НТИ. Сер. 2.— 1992.— №. 9.
3. Грамматический словарь русского языка.— М.: Русский язык, 1977.
4. Wierzbicka A. Lingua mentalis. — Sydney: Academic press, 1980.
5. Wierzbicka A. English speech act words. — Sydney: Academic press, 1987.
6. Иорданская Л. Н. Попытка лексикографического толкования группы русских слов со значением чувства // Машинный перевод и прикладная лингвистика. — 1970.— Вып. 12
7. Зализняк Анна А. Семантика глагола «бояться» в русском языке // Изв. АН СССР. Сер. лит. и языка. — 1983. — № 1.
8. Падучева Е. В., Розина Р. И. Семантический класс глаголов полного охвата: толкование и лексико-синтаксические свойства // Вопр. языкознания. — 1993. — № 6.
9. Wierzbicka A. The semantics of grammar. — Amsterdam: Benjamins, 1988.
10. Плунгян В. А., Рахилина Е. В. Сирконстанты в толковании? // Metody formalne w opisie jezykow słowiańskich. — Bialystok, 1990.

Материал поступил в редакцию 06.12.93.

2. Шевелев В. М. К систематизации дифференциальных семантических признаков русских наречий времени // Лингвистическое наследие Т. П. Ломтева и вопросы русистики.— Днепропетровск, 1984.— С. 128—133.
3. Семчинская Н. С. Функционально-семантические особенности наречий времени и их функционирование в русском языке: Автореф. дис.... канд. филол. наук. — М., 1989.
4. Леонтьева Н. Н. Описание слов со значением времени // Машинный перевод и прикладная лингвистика.— 1964.— № 8.
5. Богуславский И. М. Сферы действия лексических единиц: Дис.... д-ра филол. наук.— М.: Ин-т языкознания, 1993.
6. Veyrenc J. L'agregat 'tak i' en Russe contemporain //Les particules énonciatives en Russe contemporain. —Université de Paris 7, 1986.— P. 13—52.
7. Firbas J. Functional sentence perspective in written and spoken communication.— Cambridge University Press, 1992.

Материал поступил в редакцию 03.12.93.

Редактор Т. Н. Лаппалайнен

Технический редактор Л. В. Кутакова

Сдано в набор 18.01.94 Подписано в печать 23.02.94 ЛР № 040228 от 22.01.92
Формат бумаги 84×108¹/₁₆ Бум. типографская Литературная гарнитура Высокая печать
Усл. печ. л. 3,36 Усл. кр.-отт. 4,02 Уч.-изд. л. 4,88 Тираж 1142 экз. Заказ 284 Цена 64 р.
Адрес редакции: 125219, Москва, А-219, ул. Усиевича, 20а. Тел. 152-66-71
Производственно-издательский комбинат ВИНИТИ,
140010, Люберцы, 10, Московской обл., Октябрьский пр., 403