А. М. Галиева, Д. Д. Якубова
A. M. Galieva, D. D. Yakubova

СПЕЦИФИКА СЕМАНТИЧЕСКОЙ АННОТАЦИИ ГЛАГОЛОВ В НАЦИОНАЛЬНОМ КОРПУСЕ ТАТАРСКОГО ЯЗЫКА[1]

SPECIFIC FEATURES OF SEMANTIC ANNOTATION OF VERBS IN THE NATIONAL CORPUS OF THE TATAR LANGUAGE

Аннотация. В статье предлагается подход к семантической аннотации глагольной лексики в Корпусе татарского языка, основанный на разделении таксономии и строевых компонентов значения. В качестве базовых строевых (категориальных) сем выделяются каузативность, интерперсональность, отрицание.

Abstract. The article offers an approach to semantic annotation of verbal lexis in the Corpus of the Tatar language based on the differentiation of taxonomy and structural components of the meaning. Causativity, interpersonality and negation are singled out as basic categorial semes.

Введение

Татарский национальный корпус[2] можно рассматривать как совокупность концептуально-функциональных моделей различных уровней татарского языка. Корпус содержит тексты различных жанров и стилей современного татарского литературного языка. В Корпусе представлены тексты художественной, учебной и научной литературы, предоставленные различными издательствами, тексты Интернет-публикаций, относящиеся к новостной или общественно-политической тематике, а также тексты официальных

[1] Работа выполнена в рамках Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» 2012-2014.

[2] **http://web-corpora.net/TatarCorpus/search/?interface_language=ru**

документов. Объем текстовой коллекции составляет примерно 20 млн. словоупотреблений. К настоящему времени в Корпусе выполнено грамматическое аннотирование, актуальной задачей остается разработка оптимального инструментария для семантической аннотации и классификации массива лексики. В статье рассматриваются общие принципы к построению семантической классификации, а также специфические аспекты классификации глагольной лексики для татарского языка.

1. Общие принципы

Смысловая структура слова как продукт мыслительной деятельности человека, связанная с компрессией информации человеческим сознанием, с такими видами мыслительных процессов, как сравнение, классификация, обобщение, в некоторой степени предопределяется грамматическим строем языка. Базой для семантической разметки лексики обычно является морфологическая (частеречная) разметка, позволяющая выделить основные лексико-грамматические классы слов, которые затем получают различные пометы в зависимости от задач разработчиков.

В языках агглютинативного типа, к которым относится татарский, не только слово- и формообразующие основы слов, но и аффиксы, используемые в каждой словоформе, оказываются значительно более самостоятельными и психологически более «весомыми» языковыми элементами, чем в языках флективных, что накладывает свой отпечаток на системную организацию лексики в языке.

Семантическая разметка имеет существенные отличия от грамматической, которая строится на строго определенном, фиксированном количестве грамматических категорий. Количество семантических признаков, хотя теоретически и обозримо, но неопределенно, поскольку зависит от степени генерализации. К значительным трудностям при семантическом описании лексики приводит отсутствие во многих случаях ясной границы между таксонами, необходимость оперировать

огромным количеством признаков и свойств, сложность разграничения компонентов значения в лексической единице и часто невозможность однозначного определения их характера и др.

Лексико-семантические группы (ЛСГ) слов представляют собой одно из фундаментальных проявлений системности в языке. Говоря о семантическом поле или лексико-семантической группе, лингвисты имеют в виду не просто набор слов, но и характер семантических отношений между этими словами.

Формы семантического описания в значительной мере зависят от свойств описываемых языковых единиц. Полное семантическое описание должно, с одной стороны, основываться на определении места лексемы в системе языка, установлении его парадигматических отношений с другими словами, с другой стороны, – учитывать синтаксический потенциал и синтагматику слова. Для различных типов слов предлагаются различные форматы семантического значения.

При работе над семантической разметкой слов в Национальном корпусе русского языка (НКРЯ) были использованы сведения о значении слов и структуре семантических классов из имеющихся семантических словарей русского языка (Русский семантический словарь под ред. Н. Ю. Шведовой, Толковый словарь русских глаголов под ред. Л. Г. Бабенко, Системный семантический словарь русского языка Л. М. Васильева)[3]. Поскольку к настоящему времени не создано идеографических словарей татарского языка, семантическая разметка Корпуса татарского языка осуществляется по имеющимся толковым и двуязычным словарям.

Таким образом, перед создателями Корпуса[4] стояла задача не только разработать общие принципы семантической аннотации

---

[3] О лексико-семантической информации в Корпусе // Национальный корпус русского языка [Офиц. сайт]. URL: http://**www.ruscorpora.ru/corpora-sem.html**

[4] *Гильмуллин Р.А., Невзорова О.А., Хакимов Б.Э.* Корпус татарских текстов: проблема репрезентативности // Труды международной

лексикографической базы данных, но и произвести реальную классификацию словарного массива татарской лексики, то есть из алфавитного списка слов получить реальное наполнение выделяемых ЛСГ. При разработке критериев классификации использовались материалы имеющихся идеографических словарей других языков, в первую очередь, русского.

Схема семантической разметки, как и любой другой разметки, предполагает наличие набора тэгов, описания того, что каждый из них означает, а также правил присвоения тэгов единицам текста или словаря. В настоящее время стандарты по созданию семантической разметки отсутствуют. Количественные и качественные характеристики наборов тэгов, применяемых в различных идеографических словарях, электронных корпусах и лексикографических базах, варьируются. Очевидно, что чем большим является набор тэгов, тем более детальный анализ языкового материала может быть произведен с его помощью. Однако упрощенная система кодировки также имеет свои достоинства, так как она дает возможность избежать большого количества ошибок, непоследовательности при аннотировании, большого объема ручной (неавтоматизированной) работы. Поэтому одной из задач при разработке системы разметки является создание необходимого баланса между степенью детализированности разметки и ее простотой и прозрачностью, удобство для разработчика и пользователя.

При разработке семантической разметки для Корпуса татарского языка нами производилась ориентация на семантическую разметку, реализованную в НКРЯ, с некоторыми поправками и уточнениями с учетом специфики лексической и словообразовательной системы татарского языка. При этом слова получили более детализированное описание. Базовые сведения об

конференции «Корпусная лингвистика–2011». 27–29 июня 2011 г., Санкт-Петербург. СПб., С.-Петербургский гос. университет, Филологический факультет, 2011. С. 125–130.

основных таксономических (тематических) классах татарских существительных и глаголов были представлены в статье «Семантико-грамматическая аннотация в русско-татарской лексикографической базе данных»[5].

2. Строевые компоненты значения

В Татарском национальном корпусе, помимо таксономии, для глагола определены базовые категориальные семы (строевые компоненты значения). В ходе анализа научной литературы нами не было найдено специальных работ по татарскому языкознанию, где выделяются строевые компоненты значения глаголов, но имеются подходы по их выделению на материале русского языка. Так, Е.В. Падучева говорит: «Строевые компоненты имеют максимально широкую сочетаемость и не являются темообразующими: не они определяют специфическую лексическую семантику слова. Чтобы выявить принадлежность слова к тематическому классу, надо «очистить» его семантическую формулу от строевых компонентов. Строевые компоненты меняют лексическое значение, сохраняя тему»[6]. По мнению Е.В. Падучевой, к строевым компонентам значения у русского глагола могут быть отнесены каузация, начинательность, отрицание, оценка, многие аспектуальные значения (в частности, предельность, узуальность, итерация и т.п.), модальность и некоторые другие[7].

Для татарского глагола нами предлагаются следующие строевые (категориальные) пометы: каузативность/некаузатив-

---

[5] *Невзорова О.А., Салимов Ф.И., Хакимов Б.Э., Гатиатуллин А.Р., Гильмуллин Р.А., Галиева А.М., Якубова Д.Д., Аюпов М.М*. Семантико-грамматическая аннотация в русско-татарской лексикографической базе данных // Филологические науки. Вопросы теории и практики. Тамбов, Грамота, 2012. №7 (18): в 2-х ч. Ч. I. С. 141–146.

[6] *Падучева Е.В*. Динамические модели в семантике лексики. М. Языки славянской культуры, 2004. С. 45–46.

[7] *Там же,* С. 46–48.

ность – **cat:caus/cat:noncaus;** интерперсональность (совместность/взаимность – **cat:inter;** отрицание – **cat:neg**.

Строевые компоненты значения отчасти коррелируют с залоговыми показателями глагола (понудительный и взаимно-совместный залоги), но не могут быть сведены только к ним. Так, в татарском языке может быть выделено значительное количество глаголов с каузативным значением без аффикса понудительного залога, типа *көчләү*, *ирексезләү* (неволить, принуждать), *ихтиярсызлау* (заставлять, вынуждать), *илтү* (нести, относить, вести), *ташу* (таскать), *ватy* (ломать), *җибәрү* (отправлять, посылать), *түгү* (проливать).

Присутствие в структуре значительного количества лексем, относящихся к разным тематических группам, компонента значения 'совместность' или 'взаимность' (ср.: *дуслашу* – подружиться, *дошманлашу* – стать врагами, *вәгъдәләшү* – договориться, условиться, *киңәшләшү* – советоваться, совещаться, *хәбәрләшү* – извещать друг друга и т.п.) обусловило появление специальной пометы **cat:inter**, которая также отнесена нами к категориальным пометам.

Еще одна строевая сема, выделяемая нами, – отрицание. Отрицание – универсальная метакатегория, представленная во всех языках и имеющая различные средства выражения на разных уровнях. Для татарского языка характерно синтаксическое и морфологическое выражение отрицания: частица *түгел*, глагольный аффикс *-ма/-мә*, аффикс для прилагательных *-сыз/-сез.* Нас интересуют преимущественно случаи, когда сема отрицания содержится имплицитно, в связанном виде в лексеме, не выражаясь при помощи формальных средств (аффиксов).

В ходе разработки системы аннотации глаголы с отрицательной семантикой (типа *молчать* в русском языке) было решено отнести к тому же классу, что и антонимы этих глаголов с «положительным» значением. Так, глагол *тыну* (стихнуть, смолкнуть) мы относим к глаголам звучания. При этом появляется необходимость в дополнительной помете для выражения отрицания **cat:neg**, общего для таких глаголов, как

туктау (остановиться), *югалту* (терять) и некоторых других; в данных глаголах сема отрицания синкретично содержится в самом значении глагола, а отрицание не является словоизменительной категорией (аспектом) глагола, выражаемой специальным аффиксом -ма/-мә (в отличие от случаев типа: *әйтү* – сказать, *әйтмәү* – не сказать). Отрицательный аспект глаголов выделяет морфологический анализатор, что маркируется при грамматической аннотации.

Было принято решение вывести помету «Отрицание» из рубрики «Таксономия», основанной на тематической классификации, в разряд строевых компонентов значения (cat:neg). Сема отрицания присуща словам разных частей речи вне зависимости от тематической отнесенности слова.

Помета для выражения отрицания в списке лексико-семантических помет НКРЯ нами не обнаружена, но она реально представлена при описании семантических признаков ряда лексем, например, *тишина*: семантика основная – r:abstr, t:fam, **t:neg**, t:sound[8]. Но помета t:neg, в НКРЯ используется непоследовательно, она отсутствует при описании ряда других слов с компонентом значения «отрицание», например:

молчание: семантика основная – der:v, r:abstr, t:speech;

остановиться: семантика основная – ca:noncaus, d:pref, der:v, t:move.

Использование пометы **t:neg** (с буквой **t**) свидетельствует о том, что отрицание разработчиками НКРЯ отнесено к таксономии, то есть к классификации по тематическому признаку. Мы, как и Е.В. Падучева, считаем его строевым (категориальным) признаком.

[8] Национальный корпус русского языка [Офиц. сайт]. URL: **http://www.ruscorpora.ru/search-main.html**

Заключение

Таким образом, для семантической аннотации татарского глагола нами предлагаются следующие строевые (категориальные) пометы:

- каузация: **cat:noncaus / cat:caus**;
- совместность / взаимность: **cat:inter**;
- отрицание: **cat:neg**.

Каузативность и интерперсональность характеризуют множественность субъекта, выполнение коллективного действия, характер распределения действия между тем, кто вызывает это действие и тем, кто его выполняет. Отрицание как универсальная метакатегория является важной составляющей семантической структуры слова, в ряде случаев являясь формально немаркированным членом класса. Выделение строевых компонентов в Корпусе дает возможности для исследования в системном плане корреляции между тематическими классами глаголов и нетематическими аспектами внутренней структуры значения глагола.