ISSN 2079-9160

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ОТДЕЛЕНИЕ ИСТОРИКО-ФИЛОЛОГИЧЕСКИХ НАУК
РОССИЙСКИЙ КОМИТЕТ ТЮРКОЛОГОВ

Российская ТЮРКОЛОГИЯ

Учредители: Институт языкознания РАН
Российский комитет тюркологов при ОИФН РАН
«**Российская тюркология**» – преемник и продолжатель журнала «Советская тюркология», издававшегося как совместный орган АН СССР и АН АзССР (Баку, 1970–1991 гг.). -

Издатель: От имени РКТ при ОИФН РАН – И.В. Кормушин

№ 2(15)

МОСКВА–КАЗАНЬ 2016

ISSN 2079-9160

COMMITTEE OF RUSSIAN TURCOLOGISTS
RUSSIAN ACADEMY OF SCIENCES
BRANCH OF HISTORY

RUSSIAN TURKOLOGY

Founded by Institute of Linguistics of the Russian Academy of Sciences
The Committee of Russian Turkologists,
Russian Academy of Sciences, Branch of History

«Rossijskaja tjurkologija» (Russian Turcology) is an heir and successorof the journal «Sovetskaja tjurkologija» (Soviet Turcology) that was a joint organ of the Academy of Sciences of the USSR and the Academy of Sciences of the Azerbaijan SSR (Baku, 1970–1991)

Publisher: On behalf of the Committee of Russian Turkologists – I.V. Kormushin

№ 2(15)

MOSCOW–KAZAN 2016

Российская тюркология

№ 2(15) 2016

РАССТАНОВКА СЕМАНТИЧЕСКИХ И ДЕРИВАЦИОННЫХ ТЭГОВ В ЭЛЕКТРОННОМ ХАКАССКО-РУССКОМ СЛОВАРЕ [1]

А.В. Дыбо, А.В. Шеймович, С.А. Крылов, г. Москва

Резюме. Эта статья описывает возможности аннотировать семантически и деривационно языковой корпус хакасского языка. Здесь представлена предварительная версия инвентаря признаков для семантического аннотирования. В отличие от большинства семантических систем классификации этот инвентарь учитывает и парадигматические и синтагматические особенности семантики слова. Этот подход основан на идеях Ч. Филмора о структурной семантике, которая интерпретирует значение слова через семантические структуры (или предикаты, с точки зрения российской лингвистики), содержащие элементы структуры (семантические роли), а так же на идеях Лексической семантики и модели «Смысл<=>Текст» Ю. Апресяна и И. Мельчука. Система семантических признаков является иерархической: ее элементы вступают в отношения включения, перекрывания, объединяясь так же как «аргумент – функция». Мы показываем примеры некоторых задач, для которых может использоваться семантический теговый корпус. Главная из них – это снятие лексической и грамматической омонимии. Статья демонстрирует несколько примеров полуавтоматического снятия омонимии глаголов, используя семантические особенности их актантов, снятие омонимии имен, используя их сочетаемость, и снятие грамматической омонимии посредством соглашения о числе подчиненного предиката.

Работа проводится в рамках корпоративного проекта РАН по развитию корпусов миноритарных языков Российской Федерации, включая тюркские языки миноритарных коренных народов.

Ключевые слова: корпус языка, семантическая маркировка, этимологическая маркировка, инвентарь семантических признаков, лексическая омонимия, грамматическая омонимия, снятие омонимии.

1. Введение

Работа по семантической разметке корпусов миноритарных тюркских языков в настоящий момент проводится в форме расстановки в статьях электронной хакасско-русской словарной базы, созданной на основе Большого хакасско-русского словаря [БХРС 2006], семантических тэгов.

[1] Работа ведется на средства гранта РГНФ № 15-04-12030 «Система автоматического морфологического и синтаксического анализа для корпусов миноритарных тюркских языков России» и программы ОИФН РАН «Евразийское наследие и его современные смыслы». Направление 4. Мультимедийные технологии в филологических исследованиях. (Проекты «Развитие корпусов миноритарных тюркских языков России» и «Создание версии «3» генерального корпуса современного монгольского языка»).

Семантическая разметка словарной базы и текстового корпуса значительно расширяет возможности пользователя при создании поисковых запросов и улучшает качество результатов поиска [Кустова, Толдова 2009]. Она необходима для решения различных задач на множествах лексем, объединенных в семантические поля, или лексико-семантические классы, по признаку обладания одним или несколькими общими семантическими признаками.

Семантическая информация о лексеме представлена в виде набора семантических помет (тэгов) и записана в поле SEMTAG. Ср. рис. 1.

С учетом необходимости унификации семантической маркировки всех имеющихся национальных корпусов, мы стремимся сделать систему семантических помет для корпуса хакасского языка максимально универсальной, для чего используем существующий опыт и наработки в этой области. В частности, нами были использованы следующие классификации лексики: тезаурусные – для предметных имен: [A.L.E. 1973; СИГТЯ; НКРЯ]; для предикатных имен и глаголов: [НКРЯ], частично: [Апресян и др. 2007; Апресян, 1967].

2. Общие подходы к организации инвентаря семантических помет

К настоящему моменту разработана предварительная версия инвентаря семантических тэгов; в отличие от помет в НКРЯ, она учитывает не только парадигматические, но и синтагматические характеристики семантики слова. Здесь мы исходим из следующих соображений.

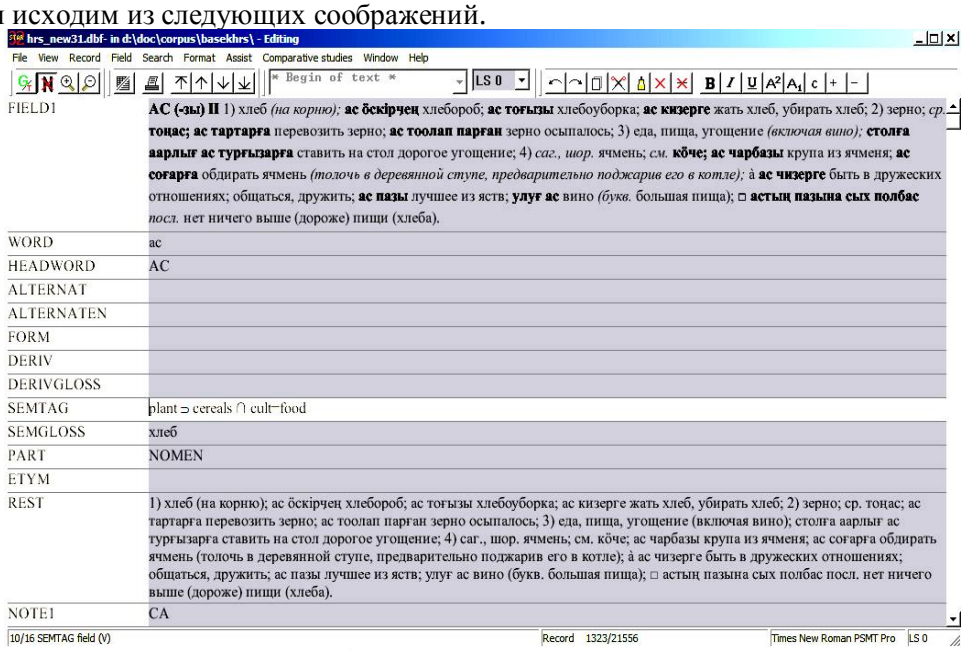

Рис. 1. Семантическая информация о лексеме

Идея представления языковой семантики через фреймы (или, в более традиционной для российской лингвистики терминологии, предикаты) не является чем-то специфически новым. Предикатное представление лексической семантики было принято как в модели «Смысл <=> Текст» (ср. у Ю.Д. Апресяна: «В общем случае толкуемой единицей должно быть не отдельно взятое слово, а содержащее его выражение вида *XPY*, где *P* – толкуемое слово, а *X* и *Y* – переменные, сообщающие данному выражению форму предложения или словосочетания» [Апресян 1974: 97]; еще раньше – в работах А.К. Жолковского,

Н.Н. Леонтьевой и Ю.С. Мартемьянова, с 1960 г.), так и в различных направлениях генеративистики (несколько позже – примерно с 1963 г., поскольку генеративная грамматика далеко не сразу признала необходимость работы с семантическим уровнем языка). Пожалуй, наиболее подробная разработка «фреймовой» семантики сейчас принадлежит Ч. Филлмору и его сотрудникам, см.: Berkeley Framenet Project [BFP] (сам Ч. Филлмор вел исследования в этом направлении с 1982 г.)[2].

Вкратце современные лингвистические представления о «фреймовой» семантике можно очертить следующим образом. Слово естественного языка – знаковая единица, то есть двусторонняя единица, имеющая форму и значение. Отдельные значения многозначного слова рассматриваются при семантическом анализе как составные части отдельных единиц. По крайней мере для значительной части слов – лексических единиц (ниже мы вернемся к вопросу – для какой части) следует считать, что каждое из них описывает стандартизованное представление о некотором типе реальных ситуаций. Типизованная ситуация («фрейм») включает «участников ситуации», они же «семантические роли», связанные элементарными отношениями. Семантические роли делятся на «ядерные» («обязательные») – сущностно важные для ситуации, такие участники, без которых сама ситуация не может иметь места, и которые характеризуют в данном наборе именно этот тип ситуаций; и «неядерные» – не характеризующие специфически данный тип ситуаций (хотя, возможно, и необходимо присутствующие в нем). Пример – ситуация торговли; она включает в качестве ядерных роли покупателя, продавца, акт передачи (сам по себе может быть описан как фрейм с более элементарными участниками), предмет торговли («ресурс» в терминологии [Розенцвейг 1964]), деньги (или иной кодифицированный эквивалент обмена). «Неядерные» роли здесь – время, когда совершается сделка, место, где она совершается (роли, характеризующие любую ситуацию типа события или процесса), цель сделки (в принципе, может отсутствовать); о последнем типе «неядерных» ролей иногда говорят как о «внешних» ролях. Типы ситуаций могут быть расклассифицированы внутри тезауруса, соответственно, вступая в отношения включения, пересечения и подчинения между собой. Торговля – подтип ситуации обмена, в ситуации обмена не специфицирован как кодифицированный эквивалент второй ресурс; купля и продажа – подтипы ситуации торговли, в зависимости от того, покупатель или продавец рассматривается как главный герой ситуации; спекуляция – еще один подтип ситуации торговли, где участник «цель» включен как внутренний – торговля с целью получения неоправданно высокой прибыли (для нас неважно, что реально соответствующая ситуация, возможно, ничем не отличается от просто торговли – здесь речь идет об обычном значении слова).

Предполагается, что словарь каждого языка должен представлять собой базу данных с «фреймовыми» толкованиями для слов, включающую информацию о синтаксической интерпретации «семантических ролей» и (если требует грамматика языка) стандартном морфологическом выражении этой синтаксической интерпретации – то есть, возвращаясь к принятой у нас терминологии, информацию о модели управления слова.

[2] Более подробно о этом классе семантических теорий см. в: Лингвистика конструкций, 2010, 18–75.

Вообще-то, различие между традиционным лексикографическим толкованием и толкованием в предикатной форме носит скорее технический характер; в общем случае из традиционного толкования можно получить «фреймовое» автоматически; для русского языка для этого следует перевести инфинитив правой части толкования в личную форму и приписать справа и слева от нее переменные X и У к ближайшим двум зависимым существительным, далее Z и т.п.; эксперимент был проделан С.А. Крыловым над толкованиями словаря Ожегова с помощью системы автоматического морфологического анализа, встроенной в СУБД Starling (система управления этимологическими базами данных, разработанная С.А. Старостиным, http://starling.rinet.ru) и дал вполне удовлетворительные результаты. Другое дело, что традиционная лексикография не требует в эксплицитной форме подробного описания модели управления слова, почему в словарях языков с менее развитой лексикографической традицией, чем русская, французская, немецкая или английская (а лучше всего – латинская и древнегреческая) зачастую моделью управления слова вообще пренебрегают, а «предикатная» лексикография специально заостряет внимание лексикографа на этой проблеме, эксплицитно рассматривая описание возможных типов управляемых единиц как основную часть толкования управляющего слова.

При «фреймовых» толкованиях слов кроме сказанного даются также для каждого фрейма ссылки на все слова, фреймы которых совпадают с данным или находятся с ним еще в каких-либо отношениях. Имея такой словарь, автоматическая система может производить «аннотирование» предложений любого текста, то есть относительно каждого выбранного слова («мишени») размечать в тексте обозначения «участников» соответствующего фрейма, что и моделирует процесс понимания предложения и, в дальней перспективе, текста. Соответственно, при наличии таких баз для нескольких языков можно смоделировать автоматический перевод с языка на язык.

Из изложенного видно, что легко поддаются «фреймовому» толкованию языковые предикаты, то есть слова, обозначающие действие, состояние или процесс. Со словами других типов возникают сложности, разрешение которых производится различными способами, ad hoc. Наибольшую трудность для осмысленной фреймовой интерпретации представляют так называемые предметные имена, так сказать, термы par excellence.

Мы не будем здесь специально уделять внимания частной проблеме в семантике естественного языка, а именно, связи различных типов значений многозначного предметного имени с его типичной синтаксической функцией, на которую впервые обратил внимание В.В. Виноградов [Виноградов 1977] (у имен типа «свинья» основное значение – конкретное, производное – метафорическое, и второе встречается практически исключительно в синтаксической функции предиката). В дальнейшем Н.Д. Арутюнова [Арутюнова 1980] дала уточненную классификацию типов «таксономического» и «характеризующего» значений; в системе фреймовых толкований у Филлмора «характеризующие» значения также обрабатываются особым образом.

Надо сказать, что их толкование вызывает сложности практически во всех семантических теориях (ср. традиционный для лексикографии вопрос, насколько словарное толкование должно включать энциклопедическую информацию).

Общий теоретически мыслимый подход к толкованию предметных имен таков: во всяком случае, для имен артефактов и иных предметов, регулярно участвующих в человеческой деятельности, явно можно использовать классы типа «материал», «инструмент», «помещение», «сосуд» и под. – каждый из таких

классов легко интерпретируется как фрейм (т.е. предикатно, с более элементарными участниками); понятно, что отнесение предметного имени к одному из таких классов отражает наиболее типичное его использование в человеческой деятельности. По-видимому (как показывают, во всяком случае, историко-семантические исследования), в реальных языках подобным же образом устроена и семантика ряда наименований природных объектов – например, названий растений и животных, частей тела («конопля – травянистое растение, используемое для изготовления веревок»; именно такая семантика приводит к тому, что в истории языков название конопли часто переходит в название крапивы и наоборот). Неудобство такого подхода к фреймовым толкованиям, очевидно, то, что предложения, описывающие ситуацию нетипичного использования объекта и потому не поддающиеся простому автоматическому «аннотированию», будут встречаться значительно чаще и поддаваться дополнительной типизации значительно меньше, чем предложения, описывающие ситуации отступления от словарных фреймов глагольного типа (вроде случаев с разнотипными опущениями участников или наложениями нескольких разных фреймов), просто в силу того, что глагольные предикаты – в сущности, продукт именно человеческой интерпретации ситуаций, деятельности человеческого разума, а множество предметных имен – понятийная сетка, с трудом налезающая на множество объектов действительности, не являющееся само по себе чем-то специально приспособленным для человеческой деятельности, и потому значительно более разнообразное.

Соответственно этому, техническое решение, которое принимает команда Филлмора (и, по-видимому, наиболее естественное) – следующее. Создание фреймовой базы данных производится не «вообще», а для «аннотирования» текстов из определенного корпуса. При толковании предметных имен сначала выявляют наиболее частотные в данном корпусе предикаты, с которыми употребляется данное предметное имя, соответственно, выясняется, в каких фреймах значения этих имен типичным образом заполняют «семантические роли». Соответственно, для имени строится фрейм с данным типом фреймов в качестве «участника», а также, по возможности, с еще одним или несколькими «участниками», представляющими типизованную «качественную» характеристику (вроде «материала», из которого предмет изготовляется, или «цели», для которой он предназначен).

Собственно говоря, сам Филлмор не решается говорить о предикатных (фреймовых) толкованиях предметных слов, утверждая, что при аннотировании текста отдельным образом производится аннотация для мишени-носителя фрейма и мишени – «заполнителя слота», но в действительности толкования «заполнителей слотов», устроенные так, как было описано выше, легко записать как предикатные.

Такой подход, в общем, соответствует теоретико-семантическому представлению о «возможных мирах», в рамках которых только и может быть правильно проинтерпретирован смысл той или иной лингвистической единицы. Соответственно, тематически ограниченный корпус текстов может рассматриваться как реализация своего рода «платоновской идеи» некоторого возможного мира, внутри которого происходит понимание текста и слова текста получают толкование.

Разметка хакасского словаря (и впоследствии автоматических словарей других миноритарных тюркских языков) предназначена для предварительной работы с корпусом текстов, разнообразных по тематике; статистическая обработка текстов с выяснением сочетаемости слов планируется впоследствии, но видится как один из промежуточных результатов, а не исходный пункт работы. Поэтому пока мы

даем синтагматическую информацию в тэгах только для глаголов, а также очевидных операторов, в частности, имен действия, состояния и процесса. В дальнейшем планируется расширить этот подход к пометам; пока же мы ориентировались в целом на подход, декларированный в статье [Апресян и др. 2005: 193–214]. Согласно ему, инвентарь семантических тэгов (или дескрипторов), своеобразный семантический метаязык, должен обеспечивать содержательное и адекватное описание лексики языка, предметной и предикатной[3], а также, в совокупности с морфологической и синтаксической разметкой, давать исследователю достаточную информацию о закономерностях поведения элементов всех лексико-семантических классов в текстах на естественном языке. Предметные тэги членят словарь языка с наивно-энциклопедической точки зрения, отражая некоторым образом систему представлений носителя языка об окружающем мире. Поэтому, к слову сказать, так удобно было использовать для разработки семантического инвентаря 5-й том [СИГТЯ], организованный по принципу тезауруса и содержащий разделы, соответствующие важнейшим сторонам природы, жизни и хозяйственной деятельности носителей древних диалектов пратюркского языка.

3. Об инвентаре семантических помет

Наша система семантических помет выглядит иерархически: тэги вступают в отношения включения, пересечения и объединения, а также отношения «Аргумент – Функция» между собой. Для символизации отношений между признаками мы воспользовались стандартной теоретико-множественной и логической символикой:

- $\supset$ включение, помета, следующая за этим знаком, конкретизирует предыдущую (*hum* $\supset$ *persn* , *hum* $\supset$ *prof*); в словарной статье конкретизируемый признак может опускаться.
- $\cap$ – пересечение признаков (*plant*$\cap$*cult*).
- & «и» – конъюнкция признаков (*period&quant(time)*)
- $\vee$ «или»- дизъюнкция признаков (*animal*$\vee$*hum*).
- Аргументы признаков-операторов ставятся в скобках после операторов (*part(body), part(plant), part(constr)*); скобки же размечают соподчинение связок.
- | – отделяет сведения о валентностях предиката;
- : двоеточие – отделяет заполнения валентностей (|*Ag:hum, Pat:plant*).
- = – разделяет пометы, принадлежащие разным значениям многозначного слова (*plant*$\cap$*cult=food*).

Соответственно, слово может попадать в несколько семантических классов (т.е. одно слово м.б. снабжено несколькими наборами помет – см рис. 1 выше (так же устроены пометы в НКРЯ)). Слова, принадлежащие к разным лексико-грамматическим классам, могут оказаться в одном семантическом классе:

HEADWORD ААЛ I село, селение; населённый пункт; аал; улус // сельский;
SEMTAG settl=aggr(hum)

[3] Предметная лексика – названия живых существ, растений, гор, рек, овощей, фруктов и т.п., предикатная лексика – любые другие валентные лексемы, главным образом обозначающие действие, состояние или процесс, дескрипторами для которых будут ‘действие’, ‘деятельность’, ‘занятие’, ‘воздействие’, ‘свойство’, ‘интерпретация’ и такие их подклассы как ‘начало’ и ‘прекращение’, ‘каузация’ и ‘ликвидация’ [Апресян и др. 2005].

HEADWORD АҒАС (-зы) 1. 1) дерево; ... 2) древесина
SEMTAG plant=stuff

HEADWORD АС (-зы) II 1) хлеб (на корню); ... 2) еда, пища
cereals∩cult=food⊃SEMTAG plant

HEADWORD ЧИДІЛЕРГЕ /чиділ-/ кашлять;
disease|Pat:hum⊃SEMTAG physiol

HEADWORD ЧИДІЛ кашель;
disease⊃SEMTAG physiol

Ниже следует фрагмент инвентаря семантических помет для хакасской словарной базы, составленный с опорой на вышеприведенные источники; при некоторых пунктах даются примеры сложных обозначений:

а) Оппозиция собственные имена *onym* vs предметные имена *concr* vs имена предикатов *abstr*

Onym ⊃ (*persn; patrn; famn; topon*)⊃
имена *hum* ∩ *persn*
отчества *hum* ∩ *patrn*
фамилии *hum* ∩ *famn*
топонимы *landsc* ∩ *topon*

б) Тезаурусная классификация

I. Природные явления и объекты *nature*
1. Небо и небесные тела *astr*
2. Атмосферные явления; ветер, погода *weather*
3. Вода; моря, реки *landsc* ⊃ *water*
...
II. Сверхъестественное *supernat* /мифологическое *myth*
[джинны и пери hum∩supernat; драконы animal∩supernat]

III. Человек
1. Наименования лиц *hum* в том числе:
...
IV. Виды и области человеческой деятельности (*activity*)
///
...
VII. Меронимия
1. Части в том числе: *part()* (напр., *part(hum), part(constr)*)
...
VIII. Топология
...
XIV. Каузативность и под. *Caus*
XV. Пространство, место и пространственные отношения
...
XVII. Качества qual
[*qual(hum)* человеческие качества, *qual(tool)* качества инструментов
...
XXII. Числительные *num*
...

XXIII. Местоимения *pron*
1. личные *pers (я,⊃pron он)*
...

4. Отношения семантической и словообразовательной разметок

Отдельно следует сказать о существенной для нашей базы особенности расстановки частеречных и семантических помет и различиях в их трактовке. Как известно, существуют некоторые отличия в способе выделения частей речи в тюркских языках различными и грамматиками, скажем так, тюркски и европейски ориентированными. Напомним, что одно и то же слово в тюркских языках может трактоваться как существительное, или прилагательное, или наречие в зависимости от синтаксической функции, выполняемой им в предложении (*тас тура* ‘каменный дом’ *кічіглер ойнапчалар* ‘маленькие (малыши) играют’). Поэтому в нашей модели нет разных систем тэгов для традиционно выделяемых частей речи, как, например, в НКРЯ, разработчики которого пишут: «Лексико-семантическая информация имеет различную структуру для разных частей речи» (http://www.ruscorpora.ru/corpora-sem.html). В НКРЯ свою структуру помет имеют собственные, предметные и непредметные существительные, прилагательные, наречия, глаголы. Причем в систему семантических помет включена и грамматическая и словообразовательная информация: каузация и служебный статус у глаголов, диминутивы, аугментативы, сингулятивы и т.п. у имен.

В электронной хакасско-русской словарной базе для информации такого рода предназначено поле DERIVGLOSS, где представлено деление основы на словообразовательные форманты и даны стандартизованные граммемные имена этих формантов, в основном из инвентаря лексических функций Модели «Смысл <=> Текст».

Примеры:

Отметим, что такие операторы, как каузация (caus) используются и как имя словообразовательного типа, и как семантический тэг, т.к., с одной стороны, основы, содержащие словообразовательный каузативный аффикс, не всегда являются каузативами семантически:

HEADWORD аарладарға быть уваженным
DERIVGLOSS тяжелый=Oper=Caus-
SEMTAG *Pass(inter∩posit)*

(семантическое развитие, очевидно, через «вызывать уважение») – а, с другой стороны, семантическая каузативность не всегда формально выражена в основе:

HEADWORD ХЫРАРҒА /хыр-/ I отскабливать что-л.
DERIVGLOSS -
SEMTAG *Caus(disapp)*

HEADWORD ХЫРАРҒА /хыр-/ II уничтожать, истреблять
DERIVGLOSS -
SEMTAG *Caus(disapp)*

Что касается таких традиционно выделяемых в грамматиках лексико-грамматических классов (и соответствующих им частей речи) как «местоимение» и «числительное», то мы предпочитаем не вычленять их из лексико-грамматического класса «имя». Набор их грамматических категорий и синтаксических функций совпадает с именным. Как и другие имена, местоимения и числительные могут употребляться в функциях актантов, атрибутов и предикатов.

Примеры:
ікі алтам **два** шага;
ікі ал салғам [я] получил **двойку**
ікі хоныхтығ **дважды** женатый
*Пастап **ол миннең** хорыххан, ам пызоҷах осхас чабас*
Сначала **он меня** боялся, а теперь как теленок смирный
*Мына, **оларның** пірсі Алексей Кружков*
Вот (этот) **один из них** Алексей Кружков
***Ол** Тойоңзар, пысхан нымырт осхас, хара харахтарынаң хази көрібіскен, **олох** көрізін Арина Петровназар тастаан*
Она пронзительно посмотрела на Тойона [своими] как спелая черемуха черными глазами, **этим же** взглядом окинула Арину Петровну

Зато эти группы имен представляют собой специфические семантические классы (о местоимениях см. классическую работу [Якобсон 1972, 95–99]). Поэтому пометы «местоимение» и «числительное», включая их лексико-грамматические разряды (*pron ⊃ card⊃pers, num*), в хакасской словарной базе относятся к полю SEMTAG:

HEADWORD ОЛ мест. он, она, оно; тот, та, то
SEMTAG *dem⊃ pers=pron ⊃pron*

HEADWORD СИГІС восемь
SEMTAG *ord⊃num*

5. Некоторые примеры использования семантических помет

Самый распространенный тип запроса к семантически размеченной словарной базе – это нахождение по данному семантическому признаку слова или группы слов. Пользователь получает множество лексем, характеризующееся той или иной степенью близости значений (синонимический ряд, гипероним с его гипонимами, антонимы, конверсивы, семантическое поле в версии Ю.Н. Караулова, набор семантических функций от данного слова в версии Мельчука–Жолковского–Апресяна или лексико-семантическую группу в версии Э.В. Кузнецовой).

Второй тип запросов служит решению более сложных научно-исследовательских задач. Это, например, определение лексико-грамматической сочетаемости слов, снятие семантической неоднозначности многозначных слов и т.п.

5.1. Задача полуавтоматического снятия семантической неоднозначности глаголов на основе сем. характеристик актантов

Задачи снятия семантической неоднозначности глаголов исходя из семантической разметки контекста, а именно с помощью данных о глагольных моделях управления, в которых актантам глагола проставлены семантические тэги, т.е. с помощью т.н. «семантических фильтров» [Кустова, Толдова, 2009, 258—276].

В настоящий момент лексические омонимы и просто многозначные слова в словарной базе различаются римскими и арабскими цифрами в поле FIELD 1. Теперь мы собираемся различать их еще и с помощью семантических помет, как самих глаголов (в словаре), так и их актантов (в размеченных текстах корпуса).

Примеры глаголов-омонимов:

КИЗЕРГЕ /кис-/ I 1) резать, разрезать, срезать, отрезать что-л.; іпек кизерге резать хлеб; пычахнаң кизерге резать ножом;

кис- *impact* | Ag: *Pers*; Pat: *concr*; Instr: *tool*

іпек *food*
пычах *instr*

КИЗЕРГЕ /кис-/ II 1) надевать что-л.; ödік кизерге надевать обувь, обуваться; 2) носить что-л.: тон кизерге а) надевать пальто;
кис- *put*| Ag: *pers*; Pat: *cloth*
тон *cloth*
ödік *cloth*

КИЗЕРГЕ /кис-/ III переходить что-л.; переправляться через что-л.; суғ кизерге переправляться через реку; чол кизерге переходить через дорогу; чазағ кизерге переходить вброд.
кис- *move* | Fact: Pers&Anim; Loc: Space
суғ *water⊃landsc*
чол *space∩transp*

Соответственно, автоматический парсер дает форме глагола с основой *кис-* три варианта разбора, отличающихся номером и переводом глагола, взятыми из словаря. Можно построить фильтр, который будет проверять соответствие семантических помет актантов в словарной статье глагола семантическим пометам имен, встретившихся в том же предложении, что и глагол, в тексте. Как мы видим, в примерах предложений, приведенных в словаре при каждом глаголе, такое соответствие соблюдается. Если при каком-то из вариантов анализа совпадения нет, такой вариант анализа отбрасывается. Т.е., может быть выведено правило вида: «Если актант глагола *кис-* выражен именем, принадлежащим к сем. классу “одежда”, то в данном контексте значение этого глагола – ‘надевать’». В дальнейшем правило может быть обобщено для целого класса однотипных глаголов; выделение таких классов для наших корпусов представляется делом будущего.

5.2. Задача полуавтоматического снятия семантической неоднозначности многозначных имен на основе их лексической сочетаемости

Разрешение омонимии в группе имен «час I “слеза” [*physiol∩water*]; II возраст, год, лета [*quant(time)*]; III 1) час // часовой [*period&quant(time)*]; 2) часы [*device*]; IV 1) молодой, зелёный [*age=color(plant)*]; 2) свежий [*qual(food)*]; V сырой [*qual(concr)*]; VI весна [*season*]» не осуществимо автоматически для всех значений многозначного слова (при том, что слово III пополнило эту группу тюркских омонимов, будучи заимствовано из русского – но не обладая формальными признаками русизмов).

По признаку лексической сочетаемости с числительными лучше всего определяются значения ‘год’ и ‘час’ (*ансахха тоғызон час* старику девяносто лет; *пала пис час толдырды* ребёнку исполнилось пять лет), также они встречается в контексте слов с сем. тэгом *time* (*иртен читі часта* утром в семь часов) – зато различие между ними практически невыявимо; значение ‘часы’ (единственное среди них с пометой, входящей в класс *concr*) выявляется, например, в контексте слов с пометой *stuff* (*алтын час* золотые часы). Значение ‘зеленый’ можно присвоить слову *час* в контексте слов с сем. тэгом *plant, plant:part* (*час от* молодая трава; *час от чіли пӳктелче, час хамыс чіли мондылча фольк.* как зелёная поросль изгибается, как молодой камыш качается *(о пластике движения молодой девушки)*), однако значительно сложнее вычленить значения *час* ‘свежий’, т.к. это может быть и пища (*час халас* свежий хлеб), и ребенок (*час пала* новорождённый младенец) – слова, принадлежащие к различным семантическим группам; такая же ситуация и со значением ‘влажный’ – древесина, шкура, облака (*час ағас* сы-

рая древесина *(как строительный материал); час теер* сырая шкура *(только что снятая с животного); час пулуттар* тяжёлые, кучевые облака *(дождевые или снеговые)*).

5.3. Отражение числа в глаголе при согласовании подлежащего и сказуемого

В хакасском языке согласование подлежащего и сказуемого в числе зависит от того, чем они выражены. Если подлежащее в единственном числе, сказуемое всегда с ним согласовано (ед. число не маркировано). Если подлежащее во множественном числе, то его согласование имеет некоторые особенности [ГХЯ 1975, 303-304]. Согласование по числу обязательно лишь в том случае, если подлежащее выражено местоимениями 1 и 2 л. мн.ч. В большинстве остальных случаев, когда подлежащее выражено 3-м лицом, согласование его со сказуемым происходит факультативно: *Олар тоғынча / тоғынчалар* ‘Они работают’; *Піс (2 л. мн.ч.) часкалығбыс, я?* ‘Мы счастливые, да?’ На последнем примере можно наблюдать забавный случай автоматического снятия лексической омонимии: слову *піс* морфологический анализатор автоматом припишет 2 перевода: личное местоимение *мы* и существительное *шило*, а по признаку Pers2 Pl сказуемого можно смело выбрать из двух омонимов местоимение *мы*.

ЛИТЕРАТУРА

Апресян 1967 – Апресян Ю.Д. Экспериментальное исследование семантики русского глагола. – М.: Наука, 1967. – 256 с.

Апресян 1974 – Апресян Ю.Д. Лексическая семантика (синонимические средства языка). – М.: Наука, 1974. – 366 с.

Апресян и др. 2005 – Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003—2005. – М.: Индрик, 2005. – С. 193–214.

Апресян и др. 2007 – Апресян Ю.Д., Дяченко П.В., Лазурский А.В., Цинман Л.Л. О компьютерном учебнике лексики русского языка // Русский язык в научном освещении. – 2007 – № 2 (14). – С. 48–112.

Арутюнова 1980 – Арутюнова Н.Д. К проблеме функциональных типов лексического значения // Аспекты семантических исследований. М.: Наука. – С. 156-249.

БХРС 2006 – Большой хакасско-русский словарь / Под ред. О.В. Субраковой. – Новосибирск: Наука, 2006. – 1115 с.

Виноградов 1977 – Виноградов В.В. Основные типы лексических значений слова // Избранные труды. Лексикология и лексикография. – М., 1977. – С. 162–189.

ГХЯ – Грамматика хакасского языка / Под ред. Н.А.Баскакова. – М., 1975. – 418 с.

Кретов 2009 – Кретов А.А. Анализ_семантических помет в НКРЯ // Национальный корпус русского языка: 2006 – 2008. Новые результаты и перспективы. – СПб.: Нестор-История, 2009. – С. 240–257.

Кустова , Толдова 2009– Кустова Г.И., Толдова *С.Ю.* НКРЯ: семантические фильтры для разрешения многозначности глаголов // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. – СПб.: Нестор-История, 2009. – С. 258–276.

Лингвистика конструкций 2010 – Лингвистика конструкций / Изд. Е.В. Рахилина. – М., 2010. – 584 с.

НКРЯ – Национальный корпус русского языка (http:// www.ruscorpora.ru/ corpora-sem.html)

Розенцвейг 1964 – Розенцвейг В.Ю. Лексика имущественных отношений // Машинный перевод и прикладная лингвистика. – Вып. 8. – М., 1964. – С. 104-108.

СИГТЯ – Сравнительно-историческая грамматика тюркских языков: Лексика. – М.: Наука, 2001. – 822 с.

Якобсон 1972 – Якобсон Р.О. Шифтеры, глагольные категории и русский глагол / пер. с англ. А.К. Жолковского // Принципы типологического анализа языков различного строя. – М.: Наука, 1972. – С. 95–113.

A.L.E. 1973 – Atlas Linguarum Europae sous la rédaction de A. Weijnen, rédacteur-en-chef Mario Alinei, Manuel Alvar, R.I.Avanesov et al. Première questionnaire (onomasiologie, vocabulaire fondamental). Secrétariat de la Rédaction de l'A.L.E., Nimègue 1973.

BFP – Berkeley Framenet Project (http://framenet.icsi.berkeley.edu/ framenet)

Erdal 1991 – Erdal M. Old Turkic Word Formation. A Functional Approach to the Lexicon. Vol. I and II. – Wiesbaden: Otto Harrassowitz, 1991. – 874 p.

Fillmore 1982 – Fillmore Ch. Frame semantics // Linguistic in the Morning Calm. – Seoul-Hanshin, 1982. – P. 111–137.

Dybo A.V., Sheimovich A.V., Krylov S.A. **Creation of semantics and derivation tags in the electronic Khakas-Russian dictionary**

Summary. This paper describes an effort to annotate a Khakass language corpus semantically and derivationally. It presents a preliminary version of the inventory of tags for semantic annotation. Unlike most semantic classification systems, this inventory takes into consideration both paradigmatic and syntagmatic characteristics of the word's semantics. This approach is based on Ch. Fillmore's idea of Frame semantics which interprets the word meaning through semantic frames (or predicates, in terms of Russian linguistics) containing frame elements (semantic roles), as well as on ideas of Lexical semantics and the Sense – To - Text model by Yu. Apresyan and I. Melchuk.

The system of semantic tags is hierarchical: its elements enter the relations of inclusion, overlapping, integration as well as "argument – function" relation. We show some examples of tasks for which a semantically tagged corpus could be used. The main one is disambiguation of lexical and grammatical homonymy. The paper demonstrates several examples of semi-automatic disambiguation of homonymy: disambiguation of homonymic verbs using semantic characteristics of their arguments (actants); disambiguation of homonymic names using their cooccurrences; and disambiguation of grammatical homonymy by means of agreement of subjects and predicate s in number.

The work follows the framework of the RAS corporate project in regards to the development of corpora for languages of the Russian Federation, including Turkic minority languages.

Key words: corpus of a language, semantic tagging, etymologic tagging, inventory of semantic tags, frame semantics, lexical homonymy, grammatical homonymy, disambiguation of homonymy

✦✦✦