

ПРИНЦИПЫ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ НАЦИОНАЛЬНОГО КОРПУСА КАЛМЫЦКОГО ЯЗЫКА

Principles of Semantic Annotation in the National Corpus of the Kalmyk Language

*В. В. Куканова (V. Kukanova)*¹

¹ кандидат филологических наук, заведующий Лабораторией прикладной и экспериментальной лингвистики Калмыцкого института гуманитарных исследований Российской академии наук (Ph.D. of Philology, Head of Laboratory of Applied and Experimental Linguistics at the Kalmyk Institute for Humanities of the Russian Academy of Sciences). E-mail: vika.kukanova@gmail.com.

В статье дается описание принципов семантической разметки в Национальном корпусе калмыцкого языка, разработанных на основе Национального корпуса русского языка. В созданном корпусе калмыцкого языка имеется морфологическая разметка (www.kalmscorpora.ru). Семантическое аннотирование для будущих исследований структуры и семантики языка, в частности вопросов сочетаемости лексических единиц, является важным шагом в развитии корпуса как информационно-справочного ресурса.

Ключевые слова: корпусная лингвистика, калмыцкий язык, Национальный корпус калмыцкого языка, семантическое аннотирование, фасетная классификация, древесная классификация.

This paper presents description of the semantic annotation principles in the National Corpus of the Kalmyk language (www.kalmscorpora.ru) which is agglutinative with rich morphology. The Kalmyk language belongs to the Mongolian language family and is used by the Oirats in Xinjiang (China) and the Kalmyks living in the Lower Volga region of Russia. The corpus of the Kalmyk language is open data of the Kalmyk texts of different styles from 1950–2012 but it mainly includes literary works and newspaper articles. The model of morphological analysis is based on the formal description of inflectional types and paradigms without which the corpus could not have automated language processing.

The semantic annotation is a crucial step in the project development because the Kalmyk language belongs to the endangered ones, that is why it is necessary to create conditions for thorough and systematic research of the language facts on the wide range of textual materials with particular word collocations. Children can learn grammatical rules and vocabulary, however, it is difficult to acquire how a certain word “works” in the context, and without this knowledge we are not able to produce natural speech. Owing to the availability of semantically-based computerize queries and the information deriving from semantic annotation with or without combination of morphological description in the Kalmyk corpus, we can receive relatively distinct material for researching various linguistic phenomena.

The work on semantic annotation is based on the list of lexical units from the Kalmyk-Russian dictionary [1977] edited by B. Muniev. In other words, we use dictionary-based approach to annotation. Combining different methods for processing the list of words, we analyze them from four aspects: 1) lexical and grammatical characteristics (revealing categories in the part of speech); 2) thematic characteristics (one list of themes for all parts of speech); 3) word connotation (negative, positive or both of them); 4) information on word derivatives (it is not the main purpose of annotation, however, we try to point out some of them in cases where it is easy to discover them).

The semantic annotation is based on the faceted and tree classification, as a result we do not have a clear ontology of lexica but in the process of work we realize that it is difficult to give unambiguous characteristics because of word polysemy. In some cases, we add some specific operators to the universal taxonomical classification to emphasize the existence of branched systems within some word groups in the Kalmyk language, for instance the system of animals' marking depending on their age and gender. These marks are necessary to convey cultural peculiarities reflected in the language.

We analyze lexical units of all parts of speech, except linking words, which make almost 27 thousand units. Two third of all the words have more than one mark in each group of annotation. The result of this annotation is accessed as a closed database (corpus) but we will have opened and published it by the end of 2014. At the moment, we are searching for and emending mistakes in the program code of the morphological analyzer.

Keywords: Corpus Linguistics, Kalmyk language, the National Corpus of the Kalmyk Language, semantic annotation, faceted classification, tree classification.

В последнее время корпусная лингвистика стремительно развивается, что, несомненно, обусловлено, с одной стороны, появлением «прорывных» разработок в вычислительной технике и, с другой стороны, бурным развитием Интернет и Интернет-технологий, само существование которых можно определить как толчок к прогрессу информационно-коммуникационных систем в целом. Исследования, проводимые на материале корпуса текстов на том или ином языке, позволили во многом скорректировать имеющиеся описания языков. Например, на материале Национального корпуса русского языка созданы новые словари, грамматики, проводятся исследования проблем общей и прикладной лингвистики и частных вопросов [Ляшевская, Шаров 2009; Рахилина 2010; Гришина 2011; и др.].

Не случайно, что проект по созданию и развитию корпуса калмыцкого языка появился в калмыцком языкознании. На данный момент существует два корпуса калмыцкого языка: первый создается аспиранткой РГГУ А. Э. Ванькаевой в рамках работы над диссертационным исследованием (Калмыцкий корпус, далее – КК), второй (Национальный корпус калмыцкого языка, далее – НККЯ) разрабатывается в Калмыцком институте гуманитарных исследований Российской академии наук (одним из разработчиков которого является автор данной статьи). Отметим, что работы по созданию этих справочно-информационных систем начались параллельно и независимо друг от друга, но не на равных условиях и на основе различных подходов.

Во-первых, КК создавался на основе уже имеющихся программных элементов, необходимых для функционирования корпуса¹ [Сай 2012], не хватало лишь формализованного описания морфологии калмыцкого языка. НККЯ создавался без теоретического задела. На начало выполнения проекта имелась лишь большая электронная библиотека художественных произведений на калмыцком языке, которая насчитывала около 2 млн словоупотреблений. Этот факт был единственным преимуществом по сравнению с первым корпусом (КК).

Что касается материала для создания формализованного описания морфологии, то в КК и в НККЯ грамматический словарь основан на базе словника Калмыцко-рус-

ского словаря, единственной лексикографической работы академического характера² [Калмыцко-русский словарь 1977]. Другими словами, материал для создания формализованного описания морфологии калмыцкого языка совпадает. Однако впоследствии словник для грамматического словаря НККЯ был значительно расширен путем оцифровки терминологических и орфографических словарей [Краткий... 1968; Павла 1973; Очир-Гаряев 1990; 1995; 1996; Корсункиев 1992; Орфографические... 2000; Бардаев 2007; Манджикова 2007].

Мы предпочли обрабатывать массив слов на основе обратного (или перевернутого) списка слов, что во многом ускорило работу по обработке лексических единиц с грамматической точки зрения. Программное обеспечение, база данных по метаописанию [Куканова и др. 2012а] разработаны самостоятельно в Институте, что чрезвычайно важно в целях развития корпусной лингвистики и информационно-коммуникационных систем на региональном уровне.

Разработка новых типов разметки и постоянное пополнение корпусов калмыцкого языка новыми материалами важны для простого пользователя, поскольку существование параллельных электронных ресурсов с разными подходами, разными разметками создает благоприятные возможности для более глубокого и разностороннего исследования структуры калмыцкого языка (преимущественно с морфологической и семантической точек зрения).

Лингвистически аннотированный корпус калмыцкого языка, который разрабатывается в рамках проектов КИГИ РАН³, принадлежит к фундаментальному направлению развитию лингвистики, создание которого зависит от решения многих задач, каждую из которых можно приравнять к отдельному исследованию. Конечно, отправной точкой в создании НККЯ стала разработка обратного словаря, который

² Электронную копию Калмыцко-русского словаря автор данной статьи получил от Э. В. Шарманджиева, за что выражает ему искреннюю благодарность и признательность.

³ Проектов РГНФ, Программы фундаментальных исследований Президиума Российской академии наук и научно-исследовательской работы согласно государственному заданию КИГИ РАН.

¹ Речь идет о морфологическом анализаторе и «движке» сайта (или корпусном менеджере).

стал основой для создания формализованного описания морфологии. Аннотирование по морфологическим критериям осуществлено в 2012 г. [Куканова и др. 2012б; Куканова 2012а; 2012б], естественным ходом расширения лингвистического аннотирования стала семантическая разметка, которая проведена в 2013 г. В течение этого периода также разрабатывалось программное обеспечение, без которого корпус не может функционировать. Коллекция текстов КИГИ РАН, постоянно пополняющаяся новыми материалами, стала бы только «корпусом первого порядка», т. е. без автоматически проведенной разметки. В разработку программного обеспечения входили создание графического и морфологического анализатора, корпусного менеджера⁴. Отдельной задачей, не связанной непосредственно с созданием программного обеспечения, но не менее важной, была разработка сайта, поисковых запросов, базы данных, в которую «выгружаются» уже проанализированные с точки зрения морфологии, семантики тексты⁵.

В разработке системы помет для семантического аннотирования мы опирались на разметку для Национального корпуса русского языка [Кустова и др. 2005; Кустова, Толдова 2009; Кустова 2014; Ляшевская 2008; Рахилина и др. 2006], корпуса башкирского языка [Бускунбаева, Сиразетдинов 2014] и др., а также на труды, посвященные таксономическим группам, на Русские семантические словари под ред. С. Г. Бархударова [1983] и Н. Ю. Шведовой [1998; 2000; 2003; 2007] и работу А. И. Кретова [2009]. Ориентиром в решении многих вопросов, возникающих при создании корпуса калмыцкого языка, служит прежде всего Национальный корпус русского языка, первый российский опыт разработки информационно-справочной системы с семантической разметкой. Сами разработчики НКРЯ основывались в свою очередь на системе «Лексикограф», которую, конечно, переработали с учетом особенностей корпуса [Кустова и др. 2005: 156–158].

⁴ Программное обеспечение корпуса разрабатывает инженер-исследователь КИГИ РАН А. Ю. Каджиев в рамках темы НИР.

⁵ Данный тип работы осуществляется в рамках проекта РГНФ «Национальный корпус калмыцкого языка: создание и разработка» (№ 12-04-12047).

Для чего нужна семантическая разметка? Ответим на этот, казалось бы, простой вопрос словами С. Гиндина: «Возможны два подхода к описанию языка – „от формы к значению“ и „от значения к форме“». В первом случае задача описывающего состоит в возможно более точной и полной характеристике всех существующих значений и способов употребления некоторого языкового элемента. При втором подходе, напротив, перечисляются и описываются все имеющиеся в данном языке способы воплощения некоторого содержания, некоторого смыслового комплекса. <...> Оба эти подхода взаимно дополняют друг друга и в равной степени необходимы в языковедении и в преподавании языка. Да и «пользователь», адресат лингвистической продукции, будь то грамматики и словари его родного языка или языка, который он еще только изучает, нуждается в описаниях обоих типов. Если вам встретилось незнакомое слово или неизвестная конструкция, вы тянетесь за словарем или перечнем грамматических элементов, в котором будут перечислены значения озадачившего вас незнакомца. Но когда вам самим предстоит выступать или писать, как часто вы оказываетесь в положении немого: что хотите сказать, вроде бы знаете или по крайней мере чувствуете, а нужное слово или грамматическую форму найти не можете» [Гиндин 2001].

Действительно, семантическая разметка очень важна для решения задач исследования лексики, в частности проблем сочетаемости слова, его синтактики. Например, в калмыцком языке прилагательное *хуучн* ‘старый’ сочетается только с неодушевленными существительными, *көгшин* ‘старый’ – только с одушевленными. Это свидетельствует об ограничениях в сочетаемости прилагательных с разными типами существительных. Создание разных пользовательских запросов с учетом семантики позволит на большом массиве примеров уточнить, выявить правила сочетаемости тех или иных лексических единиц

Методика и материал работы

Существует два подхода в создании семантического аннотирования – фасетная и древесная классификации. Последняя реализована в Русском семантическом словаре [1998; 2000; 2003; 2007] в виде древа, где каждая помета является производной от главного. При проведении фасетной клас-

сификации исследователь отталкивается от мысли, что существуют несколько независимых классификаций, которые могут быть использованы одновременно по различным основаниям. Та или иная лексическая единица, которая может обладать множеством признаков, имеет потенциальную возможность пересечения этих признаков.

При анализе использовалась и та, и другая классификация, так как, на наш взгляд, достаточно трудно выбрать один тип классифицирования материала⁶. Слово может обладать более чем одним значением, следовательно, для наших целей не всегда подходит древесная классификация, в некоторых случаях было целесообразно вводить семантику в виде фасетов (или граней). Например, слово *ноха* 'собака', кроме пометы *t: animal*, может иметь и помету *t: hum*, поскольку обладает пейоративным значением и употребляется по отношению к человеку (на что указывает помета *ev: neg*).

Если лексическая единица имеет несколько характеристик, которые, как правило, зависят от контекста, то указывался набор возможных таксонов, к которым может принадлежать то или иное слово. Такие пометы давались через точку с запятой. Первая запись является основной, а вторая и последующие — дополнительными (тем самым были зафиксированы переносные значения слова, насколько это было возможно).

Пометы давались через двоеточие, причем второй знак указывает на оператора, который направлен на конкретизацию значения. Например, *t: size* может иметь конкретизацию *t: size: max* (*боржн* 'широкий, массивный'), *t: size: neut* (*тегг* 'среднего роста, коренастый') и *t: size: min* (*ахр* 'короткий'), если такая имела.

Работа строилась и по частеречному принципу, и по списку слов в алфавитном порядке. Первый подход использовался при определении лексико-грамматических разрядов именных частей речи, второй — при

⁶ При анализе словаря мы столкнулись с теми же самыми проблемами, с которыми встретились разработчики семантической разметки в НКРЯ. Такая «непоследовательность в использовании древесного или фасетного принципов классификации», по мнению исследователей, не создает условий для онтологического описания лексической системы, но, тем не менее, интуитивно понятна обычным пользователям [Рахилина и др. 2009: 223].

тематической классификации, так как однокоренные слова располагались рядом, то их было легче отнести к одной тематической группе, к тому же наличие однородного материала со сходной тематикой создавало благоприятные условия для выявления дифференциальных сем и их обозначения. Обрабатывались только знаменательные части речи. Значения служебных частей речи во многом зависят от контекста, что создает условия для появления большого количества вариативных (омонимичных) разборов. По этой причине служебные части речи не обрабатывались. Так, например, послелог *deer* может выражать пространственные и временные значения. И только после ручного просмотра можно решить, какое значение несет тот или иной послелог⁷.

Как уже говорилось выше, в качестве словаря выступил Калмыцко-русский словарь [1977], который был дополнен лексическими единицами из других словарей. Аннотирование происходило в тех же самых таблицах, в которых содержатся морфологические характеристики лексических единиц. Были дополнительно введены четыре столбца, в которых и давалась характеристика слова по четырем возможным группам.

I. Лексико-грамматическая разметка: определение разряда внутри именных частей речи. Определение разряда внутри каждой части речи проводилось на основе лексико-семантической классификации знаменательных частей речи, представленной в Грамматике калмыцкого языка [1983].

Имя существительное

1. *r: proper* — имя собственное⁸:

1.1. *name* — имя;

1.2. *sur* — фамилия;

1.3. *patr* — отчество⁹;

⁷ Оговоримся, что не все послелоги имеют несколько значений, но, тем не менее, было принято решение пока не включать служебные части речи в фокус анализа с точки зрения семантики.

⁸ Желательно введение классифицирования онимов (гидронимов, ойконимов, оронимов, микропонимов и др.) для будущих исследований собственных имен.

⁹ В калмыцком языке нет отчеств, но в художественных текстах можно найти примеры, когда они употребляются при обращении или при назывании имени героя. Использование отчеств

1.4. *geox* — топоним;

1.5. *orgn* — организация;

2. *r:appel* — нарицательное существительное:

2.1. *concr* — предметное имя;

2.2. *abstr* — непердметное имя.

Имя прилагательное

1. *t:qual* — качественное прилагательное;

2. *t:rel* — относительное прилагательное.

Числительное

1. *ord* — порядковое числительное;

2. *card* — количественное числительное;

3. *par* — разделительное числительное;

4. *age* — возрастное числительное;

5. *col* — собирательное числительное.

Местоимение

1. *pers* — личное местоимение;

2. *dem* — указательное местоимение;

3. *qua* — определительное местоимение;

4. *refl* — возвратное местоимение;

5. *inter* — вопросительное местоимение;

6. *ind* — неопределенное местоимение.

II. Лексико-тематическая разметка: выделение таксономических групп, мерологии, топологии, если можно было определить принадлежность к определенной тематаксону.

Вслед за А. А. Кретовым, мы считаем, что тематические группы одинаковы для всех частей речи, семантика слова, на наш взгляд, является константой языка (и, шире, языков). Например, группа слов, связанных совсем не характерно для калмыцкой речи, но в связи с влиянием русской культурной традиции общения в речи калмыков это также стало традицией.

(1) *Очр Бадмаевич көвүндэн цэглһв*. [Балакан Алексей. Алтн булгин зөөр].

(2) — *Михаил Михайлович, мендвт. Минь ода ирүв*. [Инжин Лиж, Мартгдшго нерд].

(3) — *Мукөвүн Борлыкович, су өөрэн. Ода хамдан көдлх болжанавидн, — гинэд начальник келв*. [Бадмин Алексей. Му көвүн].

Примеры даны из Национального корпуса калмыцкого языка (www.kalmscorpota.ru).

с темой животных, есть в каждом языке и может быть семой в лексических единицах разных частей речи. К тому же в калмыцком языке распространены так называемые изафетные конструкции (например *модн гер* ‘деревянный дом’), т. е. если в линейной последовательности встречаются два существительных, то первое из них будет прилагательным по своей частеречной принадлежности (определением)¹⁰. Таким образом, *модн* ‘деревянный’ имеет помету *t:plant* в любом случае (т. е. когда оно может выступать в роли прилагательного или существительного).

III. Оценочная характеристика: определение коннотации — отрицательной и/или положительной, если таковая имеется (*ev:posit*; *ev:neg*).

IV. Деривационная характеристика: указание, если возможно, на производную основу, выделение морфосемантических словообразовательных групп (диминутивов, аугментивов и т. д.). Лексико-семантическая характеристика слова тесно связана со словообразовательными пометами, но, тем не менее, не являлась главной целью. Давались наиболее важные и легко обнаруживаемые деривационные характеристики. Например, если слово заканчивается на *-го*, то данную лексическую единицу относили к каритивам.

Семантическая разметка НКРЯ и НККЯ

На основе достижений в области аннотирования словарного материала по семантическим критериям была выработана система разметки, которая уточнялась по мере обработки материала, дополнялась семантическими пометами применительно к калмыцкому языку. Например, в калмыцком языке существует разветвленная система обозначения домашних животных в зависимости от их возраста и пола. По этой причине были введены следующие пометы: *t:animal:age*, *t:animal:gender* и при совмещении сем *t:animal:age&gender*. На наш взгляд, опущение данного аспекта при семантическом аннотировании может привести к потере важной информации о лингвокультурологических особенностях, содержащихся в языке. В отличие от семантического аннотирования в НКРЯ мы исключили для глаголов отметку о наличии или отсутствии каузации, поскольку в калмыцком языке каузация является морфоло-

¹⁰ Ср. с английским языком.

гически выраженной категорией. Мы также не стали выделять для каждой части речи набор таксонов лексико-тематического аннотирования, хотя в начале анализа нашего словника мы пошли по пути НКРЯ, но через некоторое время осознали, что таксоны не носят грамматический характер и, по сути, небольшая часть из них уникальна или специфична, т.е. присутствует только в одной части речи.

Проблемы семантической разметки НККЯ

Работа над лексико-семантическим аннотированием не обошлась и без проблем, таких как словообразовательная нерелевантность, субъективизм в определении, возможность отнесения к двум разным группам одного и того же слова, зависимость от контекста и т. п. Систематизация лексических единиц внутри одной части речи по

определенным критериям позволит анализировать лексический состав не от формы к значению, а наоборот — от значения к форме. Это важный шаг в исследовании калмыцкого языка, возможность по-другому взглянуть на объект анализа создаст условия для построения синтактики языка, что необходимо в ситуации утраты языка как средства мышления, познания и общения. Ребенок интуитивно, на бессознательном уровне, не впитывает законы языка, среди которых самым сложным в восприятии является не грамматика, а лексическая сочетаемость единиц языка. Изучающему язык необходимы формально выраженные правила, которыми он будет оперировать при продуцировании высказываний. Таким образом, важность разработки семантической разметки несомненна для развития калмыцкого языкознания.

Литература

- Бардаев Э. Ч.* Материалы к калмыцко-русскому и русско-калмыцкому словарю лингвистических терминов. Элиста, 2007. 102 с.
- Бархударов* — Караулов Ю. Н., Молчанов В. И., Афанасьев В. А., Михалев Н. В. Русский семантический словарь: Опыт автоматического построения тезауруса: от понятия к слову / отв. ред. С. Г. Бархударов. М.: Наука, 1983. 566 с.
- Бускунбаева Л. А., Сиразетдинов З. А.* Система разметок в Национальном корпусе башкирского языка [электронный ресурс] // URL: <http://www.marlamuter.ru/forum/index.php?topic=21.0> (дата обращения: 11.04.2013).
- Гиндин С.* Семантические словари – карты языкового мира // Русский язык. 2001. № 5.
- Грамматика калмыцкого языка: фонетика и морфология.* Элиста: Калм. кн. изд-во, 1983. 336 с.
- Гришина Е.А.* Да в русском устном диалоге // Russian Linguistics. Vol. 35. P. 169–207.
- Калмыцко-русский словарь* / под ред. Б. Д. Муниева. М.: Изд-во «Русский язык», 1977. 768 с.
- Корсункиев Ц. К.* Калмыцко-русский и русско-калмыцкий терминологический словарь: Медицина. Элиста, 1992. 190 с.
- Краткий словарь общественно-политических терминов калмыцкого языка.* Элиста, 1968. 88 с.
- Кретов А. А.* Анализ семантических помет в НКРЯ // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / отв. ред. В. А. Плулунян. СПб.: Нестор-История, 2009. С. 240–257.
- Куканова В. В.* Словоизменительные типы в калмыцком языке в свете автоматической обработки текстов (на примере имени существительного) // Вестник Калмыцкого института гуманитарных исследований РАН. 2012. № 2. С. 168–177.
- Куканова В. В.* Словоизменительные типы в калмыцком языке в свете автоматической обработки текстов (на примере имени существительного) – II // Вестник Калмыцкого института гуманитарных исследований РАН. 2012. № 3. С. 151–161.
- Куканова В. В., Бембеев Е. В., Мулаева Н. М., Очирова Н. Ч.* Метаразметка в Национальном корпусе калмыцкого языка // Вестник Калмыцкого государственного университета. 2012а. № 3. С. 67–72.
- Куканова В. В., Бембеев Е. В., Мулаева Н. М., Очирова Н. Ч.* Национальный корпус калмыцкого языка: архитектура и возможности использования // Вестник Калмыцкого института гуманитарных исследований РАН. 2012б. № 3. С. 138–150.
- Кустова Г. И., Толдова С. Ю.* НКРЯ: семантические фильтры для разрешения многозначности глаголов // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / отв. ред. В. А. Плулунян. СПб.: Нестор-История, 2009. С. 258–278.

- Кустова Г. И.* Семантическая разметка в электронных корпусах и электронных словарях [электронный ресурс] // URL: http://lexicograph.ruslang.ru/TextKust/Kustova_sem_razm_el_sl_SPb_11.pdf (01.04.2014).
- Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В.* Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005. С. 155–174.
- Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. 1090 с.
- Ляшевская О. Н.* Топологические классы имен в семантической разметке Национального корпуса русского языка // Труды международной конференции «Корпусная лингвистика–2008» (Санкт-Петербург, 6–10 октября 2008 г.). СПб.: С.-Петербургский гос. университет, Факультет филологии и искусств, 2008. С. 276–284.
- Манджикова Б. Б.* Калмыцко-русский терминологический словарь: флора и фауна. Элиста, 2007. 95 с.
- Орфографические правила и Орфографический словарь калмыцкого языка* / отв. ред. Г. Ц. Пюрбеев, Э. У. Омакаева. Элиста: Калм. кн. изд-во, 2000. 480 с.
- Очир-Гаряев В. Э.* Калмыцко-русский и русско-калмыцкий словарь терминологический словарь: Агрономия. Элиста, 1990. 85 с.
- Очир-Гаряев В. Э.* Калмыцко-русский, русско-калмыцкий терминологический словарь: Народное образование. Элиста, 1996. 91 с.
- Очир-Гаряев В. Э.* Калмыцко-русский и русско-калмыцкий терминологический словарь: Рыбное хозяйство. Элиста, 1995. 64 с.
- Павда Дорж.* Чикэр бичлһнэ толь. 2-гч һарц. Элст: Хальмг дегтр һарһач, 1973. 240 с.
- Рахилина Е. В.* Лингвистика конструкций / отв. ред. Е. В. Рахилина. М.: Азбуковник, 2010. 584 с.
- Рахилина Е. В., Кобрицов Б. П., Кустова Г. И., Ляшевская О. Н., Шеманова О. Ю.* Многозначность как прикладная проблема: семантическая разметка в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2006». М., 2006. С. 445–450.
- Рахилина Е. В., Кустова Г. И., Ляшевская О. Н., Резникова Т. И., Шеманова О. Ю.* Задачи и принципы семантической разметки лексики в НКРЯ // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / отв. ред. В. А. Плунгян. СПб.: Нестор-История, 2009. С. 215–239.
- Сай С. С.* Корпус калмыцкого языка [электронный ресурс] // URL: <http://www.corplingran.ru/otchety.html> (дата обращения 06.03.2012).
- Шведова — Русский семантический словарь.* Толковый словарь, систематизированный по классам слов и значений / Ин-т рус. яз. им. В. В. Виноградова РАН; под общ. ред. Н. Ю. Шведовой. Т. 1. М.: Азбуковник, 1998. XXV, 807с. Т. 2. М.: Азбуковник, 2000. XXXII, 762с. Т. 3. М.: Азбуковник, 2003. 720 с. Т. 4. М.: ИРЯ РАН, 2007. 952 с.