Особенности обработки тибетских композитов-существительных в лексической базе данных

М.О. Смирнова, П.Л. Гроховский

Санкт-Петербургский государственный университет

2321781@mail.ru, p.grokhovskiy@spbu.ru

Аннотация

Настоящая статья посвящена исследованию композитов-существительных корпуса текстов на современном тибетском языке с использованием реляционной лексической базы данных, содержащей единую непротиворечивую классификацию значений тибетских лексических единиц, с установлением между этими значениями различных семантических отношений. В статье описываются структура базы данных и принципы обработки тибетских композитов-существительных; типы семантических структур композитов, которые позволила выявить обработка. Поскольку большая часть композитов представляют собой грамматические, религиозно-философские и общенаучные термины, в статье анализируются особенности обработки концептов предметных областей знаний, в том числе, обусловленные буддийской языковой картиной мира.

Ключевые слова: лексическая база данных, тибетский корпус, тибетские композиты-существительные, тибетская языковая картина мира

1. Введение

Данная работа является продолжением ряда исследовательских проектов («Базовый корпус тибетского классического языка с русским переводом и лексической базой данных», «Пилотная версия электронного корпуса тибетских грамматических сочинений»), направленных на разработку методов создания параллельного тибетско-русского корпуса.

Технологический процесс создания корпуса включал несколько этапов, главными из которых были токенизация, т.е. разделение входного текста на составные элементы (словоформы, знаки препинания, цифры и т. п.), разметка текста, импорт размеченных текстов в структуру корпусного менеджера. Изначально токенизация осуществлялась вручную.

Процесс создания корпуса текстов современного тибетского языка в рамках проектов «Программные средства автоматической обработки текста на современном тибетском языке (морфологический уровень)», «Морфосинтаксический анализатор текстов на тибетском языке» был автоматизирован. Было обработано 9 текстов длиной от 2170 до 8900 токенов, общий объём составил 55533 токена, из которых 52539 (94%) токенов было размечено автоматически [1].

В результате разметки тибетских текстов по границам словоформ (традиционная тибетская орфография маркирует лишь границы слогов) был сформирован список из 1466 многосложных слов (преимущественно двухкомпонентных), включающий композиты-существительные, именные группы, многосложные именные корни. Многосложные слова требовали дополнительной обработки для корректного распознавания их в текстах морфо-синтаксическим анализатором. Для решения данной проблемы помимо «технических» лексических баз данных, необходимых для работы разрабатываемых программных средств автоматической обработки текстов на тибетском языке (деления на словоформы, приписывания словоформам частеречных тегов и идентификации лемм для словоформ), было принято решение о необходимости разработки реляционной лексической базы данных, содержащей единую непротиворечивую классификацию значений тибетских лексических единиц, с установлением между этими значениями различных семантических отношений.

Задачей начального этапа работы с базой данных стало внесение и обработка композитов-существительных из числа многосложных слов и их составных элементов.

2. Структура лексической базы данных и принципы обработки тибетских композитов-существительных

Лексическая база данных в рамках данного исследования создается на основе онтологии AIIRE и представляет собой единую систему понятий на тибетском и русском языках, для которых описаны различные семантические отношения [2].

Помимо имён-композитов, допускается размещение в базе данных именных групп, групп прилагательного и наречия. Во всех этих случаях кроме самого выражения в базе данных моделируются и его компоненты, также подлежащие частеречной классификации.

Для каждого выражения в базе данных приводится перевод и описание его значения или значений, включающее толкование на русском языке (предпочтительный тип толкования — генетическое определение понятия) и полное толкование из тибетского «Большого толкового словаря» [3] на тибетском языке с указанием номера словарной статьи. Если в соответствии с толковым словарем у выражения имеется несколько значений, подходящее переводится на русский язык (за исключением случаев, когда в словаре выражение определяется через синонимы).

Значения понятий в лексической базе данных делятся на классы и экземпляры. Экземпляры — это концепты, которые соответствуют

представлениям о конкретном объекте, явлении или отношении, существующем в реальной или вымышленной действительности. Классы — это концепты, которые соответствуют обобщенным представлениям о системе атрибутов объектов, явлений или отношений, существующих в реальной или вымышленной действительности.

Экземпляры всегда относятся к одному или нескольким классам (между экземплярами и классами устанавливается отношение наследования. Например, «*smra sgo* — «Врата речи»; тибетский грамматический трактат Смритиджнянакирти» является экземпляром класса «грамматический текст — литературное произведение по тибетской лингвистике»).

Между понятиями лексической базы данных устанавливаются различные отношения. Отношение синонимии носит характер абсолютной синонимии (полное совпадение денотатов при возможных различиях в сигнификатах). Концепты-синонимы образуют синонимический ряд, каждый концепт которого обладает одними и теми же атрибутами, то есть одинаковыми отношениями и объектами этих отношений.

Отношение гипо-гиперонимии (родо-видовое отношение или отношение наследования) устанавливается между классами и подклассами, либо классами и экземплярами в тех случаях, когда одно из понятий (гипоним) является частным случаем другого (гиперонима). Например, класс *bu* ‘мальчик’ является подклассом *pho gsar* ‘особа мужского пола юного возраста’, который, в свою очередь, является подклассом класса *pho* ‘мужчина, конкретная особь мужского пола’, являющегося подклассом класса *mi* ‘человек, конкретная особь вида homo sapiens’. Конечными гипонимами для классов являются экземпляры.

В случае если лексическая база данных не содержит подходящего тибетского гиперонима для вводимого концепта, и работающий с системой лингвист затрудняется его подобрать, указывается гипероним на русском языке. Например, последовательность гиперонимов для концепта *mgo* ‘голова’: *yan lag* ‘часть тела’ > часть организма > *cha* ‘объект являющийся частью чего-либо в противопоставление самостоятельному объекту’. По мере пополнения системы данные пробелы заполняются. Русский язык также используется для описания глагольной семантики и отношений между концептами.

Подбор синонима и гиперонима для композита осуществляется, основываясь на его толковании в тибетском толковом словаре и употреблении в тексте корпуса. Синонимы в словаре отмечены специальными терминами — *don gcig* ‘одно значение’, *ming gi rnam grangs* ‘синоним’. Гиперонимы определяются по генетическому толкованию в словаре (при его наличии): *X* — «*Y*, который...», «*Y*, расположенный в...», где *Y* — гипероним.

Указание гиперонима каждого понятия является обязательным для построения иерархии понятий, восходящей к базовым классам (верхние классы в иерархии понятий, а также важнейшие классы вроде класса «человек»). Базовые классы, как правило, обладают значительным числом отношений разных типов (то есть не только отношениями синонимии и гипо-гиперонимии, в отличие от небазовых классов): генитивными, глагольными, предложными отношениями и другими.

Данный вариант лексической базы данных находится в данный момент на стадии пилотной разработки (проведено проектирование и развёртывание в

пилотном режиме базы данных и интерфейса редактирования, создание пользователей, адаптация инструментария для тибетского языка).

3. Типы семантической структуры тибетских композитов-существительных

Все тибетские композиты образованы соположением морфем или словоформ и, таким образом, относятся к атематическому типу (отсутствует соединительный элемент) [4, с. 469]. В некоторых случаев при словосложении возможно выпадение служебных морфем, например, *blo* ‘ум’ + *bzang po* ‘благой’ → *blo bzang* ‘мудрец’; *drang po* ‘добродетельный’ + *srang po* ‘идущий вперед’ → *drang srong* ‘риши, мудрец’.

В зависимости от частеречной принадлежности компонентов и образованного сложного слова в тибетском языке принято выделять несколько моделей образований композитов. Следующие пять моделей являются исконно тибетскими:

1. существительное + существительное → существительное;
2. существительное + прилагательное → существительное;
3. прилагательное + существительное → существительное;
4. прилагательное + прилагательное → существительное;
5. существительное + глагол → глагол [5, p. 103–105].

Также С. Бейер выделяет две модели, заимствованные тибетцами при осуществлении переводов с санскрита:

1. существительное + глагол → существительное;
2. усиливающее слово + глагол → глагол [5, p. 108–110].

По характеру синтаксических связей между компонентами принято выделять сложные слова с подчинительными отношениями и сложные слова с равноправными, или сочинительными отношениями [4, с. 468]. Подчинительный и сочинительный типы также обнаруживаются среди тибетских композитов.

При описании в лексической базе данных отношений синонимии и гипо-гиперонимии для подчинительных композитов и их компонентов, удалось выявить набор типичных семантических отношений между значениями компонентов сложных слов.

3.1. Семантические структуры сочинительных композитов

Сочинительные композиты можно разделить на подвиды в зависимости от типов лежащих в их основе семантических структур. Сочинительные композиты первого подвида представляют собой сочетание слов, образующее одно понятие (slob sbyong ‘обучение’, букв. ‘изучение [и] тренировка’; drang srong ‘риши, мудрец’, букв. ‘добродетельный [и] идущий вперед’).

К данному подвиду относятся также сложные слова, образованные двумя глаголами, указывающими на признак, как, например, *ring thung* ‘длина’ (букв. ‘быть длинным — быть коротким’). Специфика обработки таких композитов в лексической базе данных заключается в том, что они, как правило, обозначают признак (свойство) объекта, а их компоненты — атрибуты (ring po ‘длинный,

имеющий большую протяжённость', thung thung 'короткий, имеющий небольшую протяженность'). В таких случаях сложное слово связывается с компонентами соответствующими отношениями — в приведенным выше примере это отношения «быть размером любого объекта (о длине)» и обратное «обладать длиной (о любом объекте)».

В сочинительных композитах второго подвида оба компонента могут выступать в качестве родового понятия по отношению к понятию, выраженному сложным словом (например, *rba rlabs* 'водная волна', где *rba* и *rlabs* обладают схожим значением 'складка, вздутие'; *sa gnas* 'район, регион', где *sa* и *gnas* соответствуют концепту 'место; область физического пространства').

Семантическая структура композитов третьего вида представляет собой пару концептов, объединенных сочинительной связью (gser dngul 'золото [и] серебро'). В некоторых случаях компоненты, образующие композит, обладают противоположными значениями (legs 'быть хорошим' + nyes 'быть плохим' → legs nyes '[то, что является] хорошим [и то, что является] плохим') или соответствуют разным базовым классам (lam lugs 'традиция [и] путь [следования ей]'). Особенности семантической структуры сочинительных композитов второго и третьего подвида затрудняют их обработку в соответствии с правилами лексической базы данных (например, построение иерархии понятий).

3.2. Семантические структуры подчинительных композитов

Подчинительные композиты (преимущественно двухкомпонентные) также обладают разными типами семантической структуры, обуславливающими особенности их обработки в базе данных.

В первом типе композитов второй компонент является гиперонимом, который:

1. обозначает совокупность или родовое понятие, например:

bdud ***rtsi*** 'нектар; амрита; субстанция, разрушающая смерть'

> ***rtsi*** 'субстанция, оказывающая влияние на цвет и вкус';

2. имеет значение «категория, вид, класс, тип»; может встречаться в качестве компонента нескольких композитов одной иерархии понятий, например:

rgyal ***rigs*** 'каста воинов; кшатрии; одна из каст в Индии' >

mi ***rigs*** 'вид людей; каста' >

rigs 'тип; класс; вид; однородная группа'.

Употребление некоторых существительных в качестве второго элемента композита соответствует устойчивым словообразовательным моделям, описанным С. Бейером: например, *sa* 'место' (lha sa 'Лхаса', букв. 'земля богов') [5, p. 122, p. 129]. Статус других слов, выступающих в аналогичной роли, требует дальнейшего исследования.

Отдельного внимания требуют случаи, когда второй компонент композита-существительного является субстантивированным причастием и заносится в лексическую базу данных как глагол, но при этом одновременно указывает на родовое понятие по отношению к композиту. Несмотря на то, что в большинстве случаев подобрать гипероним-существительное не составляет

труда, при обработке тибетской терминологии возникает опасность утраты некоторых терминологических значений (подробнее см. пункт 4 данной статьи).

Во втором типе подчинительных композитов гиперонимом является первый компонент. В основном данные композиты образованы существительным и прилагательным с выпадением служебной морфемы, например: ***gos*** *hrul* ‘обноски’ (*hrul po* ‘рваный’) > **gos** ‘одежда; совокупность предметов, изделий, которыми покрывают тело или которые надевают на него’; *rdo ring* ‘стела’ (ring po ‘длинный’) > *rdo* ‘камень’.

Третий тип подчинительных композитов объединяет случаи, когда второй элемент является синонимом или сокращенной формой композита (например, tshig rkang ‘стихотворная строка’, где rkang имеет такое же значение ‘строка’).

Четвертый тип представлен композитами-метафорами (*mkha’ nor* ‘небесная драгоценность’, *bshes gnyen* ‘духовный друг’ — синонимы с метафорическим значениям слова *nyi ma* ‘солнце’). Для корректного подбора гиперонима и описания различных отношений для таких композитов в лексической базе данных возникает необходимость установления отношений синонимии между концептами, соответствующими различным типам токенов (метафорическим композитом-существительным и именным корнем / корнем-морфемой, обозначающими соответствующий неметафорический синоним).

4. Тибетские композиты-существительные, относящиеся к предметным областям знания

Тибетские композиты, подлежащие обработке в лексической базе данных, представляют несколько предметных областей.

Часть композитов относится к терминам общенаучного характера. Большинство общенаучных терминов представляют собой обозначение различных частей или разделов текста. Так на данный момент выражение «структурная единица текста» в базе данных уже насчитывает десять гипонимов-композитов: *nang gses* ‘поразд ел текста’, *mtha’ dpyod* ‘полное исследование’, *re’u mig* ‘таблица’, *sa bcad* ‘раздел текста’, *skabs don* ‘часть текста, раскрывающая главную тему’, *sdom tshig* ‘краткий вывод’; и три гипонима имени (компоненты композитов): *bam po* ‘раздел или глава текста’, *skabs* ‘глава’, *sdom* ‘вывод’, *mchod brjod* ‘выражение почтение Будде и богам’.

Ряд композитов являются названием типов научных, художественных или религиозных текстов. В частности, для тибетской научной традиции характерно выделение базовых текстов *gzhung* и многочисленных комментариев к ним — *rgyas bshad* ‘детальный комментарий’, *rnam bshad* ‘полный комментарий’ и др.

Большим количеством композитов представлен общий для всех индо-тибетских наук класс *don* ‘смысл текста’, который включает такие важные для тибетской научной словесности концепты-гипонимы как *gnad don* ‘ключевое значение’, *gzhung don* ‘главный смысл текста’, *dgongs don* ‘подразумеваемый смысл’, *brjod don* ‘тема’, *go don* ‘суть смысла, значение’ и др.

Поскольку рассматриваемый тибетско-русский корпус включает, в том числе, тексты по одной из тибетских традиционных наук — лингвистике (тиб. *sgra’i rig pa*), большая часть обработанных композитов относится к грамматическим терминам и специальной лексике теории письма. Кроме

собственно лингвистических терминов, в тибетских грамматических сочинениях также широко использовались религиозные и философские термины.

Для концептов, относящихся к тибетским традиционным наукам — лингвистике, буддийской религиозной доктрине и др. в лексической базе данных указывается предметная область.

Родо-видовые отношения для концептов традиционных областей знания описываются в соответствии с тибетскими представлениями. Например, гиперонимом тибетского термина *'phul rten* 'добавочная графема' является *gsal byed* 'любая графема тибетского алфавита, обозначающая согласную фонему', который, в свою очередь, является подклассом класса *yi ge* 'графема'.

Терминология тибетских средневековых областей знания не отвечает всем критериям научного термина, в связи с чем, уместнее говорить о предтерминах — лексемах, используемых в качестве терминов в предметных областях для называния новых сформировавшихся понятий, но не отвечающие основным требованиям, предъявляемым к термину [6, с. 44].

Образование терминов тибетских средневековых наук в основном происходило двумя способами: путем терминологизации слов общей лексики и заимствования. В некоторых случаях узкие терминологические значения приобретали именные глагольные формы — упомянутые в пункте 3 субстантивированные причастия, а также формы, образованные посредством добавления к глаголу суффиксов-номинализаторов со значением 'способ, место, путь' и др. В случае с общей лексикой, такие формы не обрабатываются — в базу данных заносится только соответствующий глагол. Однако когда они приобретают специальное значение в терминологических полях: например, *klog tshul* 'транскрипция' (букв. 'способ чтения'), такие термины было решено включать в систему как композиты-существительные.

Незавершенность процесса терминологизации, близость тибетской специальной лексики к общеразговорному языку, наличие большого количества консубстанциональных терминов — встречающихся как в обыденной, так и в профессиональной речи [6, с. 25], часто приводят к возникновению отношений гипо-гиперонии между значениями одного выражения. В случаях, когда сферы употребления гипонима и гиперонима или их валентности различаются, в лексической базе данных использовалось отношение «обладать типичным представителем о классе объектов» (обратное «быть типичным представителем класса»). Поскольку по-тибетски класс и типичный представитель выражены одним словом, в базе выражение, обозначающее типичного представителя, указывалось на русском языке. Так, для тибетского слова *rtags* 'знак; локализуемый в любом пространстве инструмент передачи и получения информации' в качестве типичного представителя указан «грамматический знак — знак, маркирующий различные грамматические значения».

Ряд тибетских терминов образован посредством добавления к имени или имени-композиту числительного, обозначая, таким образом, совокупности (например, *dus gsum* 'три времени глагола', *'byung ba lnga* 'пять элементов, стихий'). Для связи совокупностей и их элементов в лексической базе данных использовалось отношение «включать объекты класса» (и обратное «быть объектом класса»): *dus gsum* 'три времени глагола' > *включать объекты класса*

> *da lta ba* ‘настоящее время’, *ma ‘ongs ba* ‘будущее время’, *‘das ba* ‘прошедшее время’.

Одной из особенностей тибетских грамматических терминов является их многозначность и отсутствие у терминов уникальных значений в рамках лингвистического терминологического поля. Не любой контекст может демонстрировать использование термина в конкретном значении. Например, один из базовых терминов тибетской грамматической традиции *yi ge* соответствует понятию фонемы, которая может быть выражена графически, а иногда — слогу или силлабографеме; то есть понятия фонемы, слога и его компонентов в тибетской традиции не разделялись. В одном понятии, обозначаемом термином *yi ge*, также объединялись минимальные единицы системы выражения звукового уровня языка и минимальные единицы графической системы языка.

В этом и аналогичных случаях в лексической базе данных было использовано отношение «означать понятие» (и обратное «обозначаться знаком»). Для тибетского термина-композита описывался базовый концепт, который связывался отношением с выражением на русском языке, например: *yi ge* ‘графема (силлабографема); языковой знак, обозначающий фонему’ > *означать понятие* > *любая фонема*. Параллельные иерархии были выстроены для всех типов тибетских графем/фонем.

5. Тибетские композиты-существительные и языковая картина мира

Лексическую базу данных, в которой выражения связаны различными типами отношений, можно рассматривать как формализованную модель представлений о реальном мире, отраженном в тибетском лексиконе и тибетской грамматике. Совокупность представлений о мире, исторически сложившаяся в сознании языкового коллектива и зафиксированная в языке, называется языковой картиной мира. Различия в языковых картинах мира проявляются в фундаментальных особенностях как лексической, так и грамматической семантики и наоборот.

5.1. Особенности тибетской грамматической и лексической семантики

Характеристика тибетской грамматической семантики на основании обработки исключительно композитов-существительных затруднительна. Одним из главных принципов работы с базой данных, является стремление моделировать только те значения выражения, которые могут быть подтверждены текстами. Если при описании значений композитов есть возможность опираться на текст, то при описании компонентов сложных слов – только на семантическую и синтаксическую структуру композита.

В первую очередь сложности возникли при описании лексико-грамматической семантики глаголов. Тибетские глаголы разделяются С. Бейером на переходные, непереходные и глаголы сущности (эквативы). Глаголы данных видов различаются не только в семантическом плане, но и в синтаксическом – набором обязательных валентностей [5, p. 254].

Грамматическая категория вида в тибетском языке выражается только аналитическими морфологическими формами глагола, поэтому русские аналоги при толковании значений тибетских глаголов приводятся в лексической базе данных как в совершенном, так и в несовершенном виде (sbyin ‘давать / дать; осуществлять / осуществить передачу чего-либо’). Без опоры на текст в частности сложно отделить динамические значения «становиться/стать каким-либо» от статических «быть каким-либо», глаголы действия от глаголов деятельности.

В тибетской грамматике отсутствует четкая граница между состоянием и свойством, поскольку ряд атрибутов, являющихся прилагательными в русском языке, в тибетском передаются именными формами глаголов состояния. В лексической базе данных такие глаголы описаны как гипонимы выражения «пребывать в неадресованном ненаправленном состоянии кого-либо»; например, *gtsang* ‘быть чистым’, *gsar* ‘быть новым’, *nges* ‘быть определенным’, *rgan* ‘быть достигшим старости’ и др. В ряде случаев удается обнаружить прилагательные, выражающие соответствующие свойства. Следует отметить, что в тибетском толковом словаре такие прилагательные и именные формы глаголов связаны отношением синонимии, например, *legs pa* ‘хороший’ — *bzang po* ‘хороший’; *ngan pa* ‘плохой’ — *sdug po* ‘плохой’.

Число в тибетском языке при необходимости может также выражаться наречиями или прилагательными, чей признак выражен через отношение к числу: *mang po* ‘множественный, многочисленный’, *nyag ma* ‘единичный’ и др.

Специфика тибетской языковой картины мира проявляется особенно ярко в сфере лексической семантики. Основные лингвоспецифические концепты, ключевые для тибетской культуры и одновременно плохо переводимые на большинство языков, являются терминами буддийской религиозной доктрины. Это касается и сложных для обработки парных понятий, упомянутых выше (*legs nyes* ‘все плохое и хорошее’), и собственно категорий имен существительных.

Для тибетской картины мира актуально различие между животными и людьми; живыми существами, обладающими дуалистическим сознанием и находящимися в сансаре и Буддой. Данные особенности требуют построения в лексической базе данных нескольких сложных иерархий понятий, в некоторых из которых будут объединены люди и животные (*sems can* ‘существо, обладающие рациональным сознанием’, *‘gro ba* ‘существо, находящееся в сансаре’, *skyes ldan* ‘существо, обладающее рождением’), в других — люди, боги и будды (*blo bzang* ‘мудрец; обладающий благим умом’). Таким образом, буддизм во многом определяет особенности тибетской языковой картины мира.

Взаимосвязь буддизма и языковой картины мира проявляется в основополагающих категориях. Так, говоря об одной из базисных категорий — пространстве — в его восприятии человеком, различают проксимальные и дистальные пространственные параметры. Обычно выделяемые оппозиции-примитивы «верх / низ», «передний / задний» относятся к описанию проксимальной пространственности — пространства, которое непосредственно «прилегает» к человеку. Первоначальный смысл оппозиции — в формировании координат относительно человеческого тела. Другие объекты для архаичного человека оцениваются, в том числе, с точки зрения дружественности / враждебности, поэтому и оппозиция, важнейшая для дистального пространства,

— это «свой / чужой». Таким образом, дистальное пространство определяется человеком как социальным организмом, тогда как проксимальное — человеком как определенным образом устроенным организмом [7, с. 133-134].

При описании восприятия дистального пространства в тибетской языковой картине мира наибольшее значение приобретает отношение объектов пространства к буддийской религиозной доктрине. Вместо характерного для проксимальных пространственных характеристик социоцентризма, точкой отчета становится Будда и его Учение. Например, тибетский композит *mtha' 'khob* может обозначать как просто любой пограничный регион или окраину, так и дикие места, на знакомые с учением Будды или даже человека, не исповедующего буддизм или не принадлежащего к членам буддийской монашеской общины. Связь пространственных представлений с буддизмом демонстрирует и оппозиция тибетских терминов *nang pa* 'буддист'/ *phyi rol pa* 'не буддист', буквально означающих «находящийся внутри» / «находящийся снаружи».

5.2. Регистры вежливости в тибетском языке

В письменном тибетском языке, а также в некоторых разговорных вариантах существует несколько «регистров вежливости». Как правило, выделяют два основных регистра: *zhe sa* 'вежливая речь' и *skad dkyus ma* 'обычная речь'. Разница между ними лежит в области лексики. Среди личных местоимений, а также многих существительных, глаголов и служебных частей речи в тибетском языке встречаются обычные и вежливые пары (используемые при обращении к уважаемому лицу или рассказе о нем) с одинаковым значением. Некоторые значения могут быть выражены не двумя, а более вариантами, между которыми существуют иерархические отношения, то есть одно слово используется, если речь идет о Будде, другое — о высоком ламе, третье о монахе, и так далее. В связи с этим некоторые исследователи разделяют до пяти регистров вежливости в тибетском языке [8, p. 446–449].

Фактически, в зависимости от того, о ком идет речь или кто является субъектом действия, два предложения, описывающие один и тот же набор объектов и одинаковое действие, лексически могут быть выражены по-разному: *dmag mi rta la zhon* 'военный скачет на лошади' / *dmag dpon chibs la 'chibs* 'генерал скачет на лошади'; *khol pos cgu la kha 'khrud* 'слуга умыл лицо водой' / *rgyal pos chab la zhal bsil* 'царь умыл лицо водой' [5, p. 152]. В приведенных примерах вежливые и невежливые формы обладают одними и теми же денотатами, однако коннотативный компонент у них различен.

Не всегда вежливые и обычные формы обладают общим денотатом. Некоторые первичные вежливые формы приобрели специфичные узкие значения — преобразовались в лингвоспецифические концепты. Так, из пары концептов *zas / bshos* 'еда', вежливая форма *bshos* часто использовалась в ритуале, вследствие чего приобрела узкое значение еды-подношения божеству [5, p. 155].

Тибетская лексика различных предметных областей часто демонстрирует асимметрию в использовании вежливых форм. В частности, такая асимметрия наблюдается в терминах родства, терминах теории письма. Так, для названия элементов тибетских графем используются в основном обычные формы,

обозначающие части тела: *mig* ‘глаз’, *mgo bo* ‘голова’, *rkang pa* ‘нога’, однако для описания элемента графемы, соединяющего круглый элемент (глаз) с верхней горизонтальной чертой (головой) используется как обычная, так и вежливая форма с буквальным значением шея (*ske / mgul ba*).

Вышеперечисленные особенности тибетских регистров вежливости не позволяют связывать концепты, выраженные вежливой и обычной формами, отношениями синонимии в лексической базе данных. На данный момент в системе их объединяют только общие родо-видовые иерархии.

6. Заключение

В общей сложности в рамках данного семантического исследования в лексической базе данных было заведено 1888 выражений, обработано 554 композита-существительных, 1334 компонентов и технических выражений на русском языке. Данная работа является частью более широко исследования. Впоследствии планируется расширение лексической базы данных, введение дополнительных и уточнение имеющихся отношений; добавление выражений, что позволит преобразовать систему в компьютерную онтологию, которую можно будет использовать для моделирования лексической семантики текстов тибетско-русского корпуса.

Работа выполнена при поддержке проекта РГНФ № 16–04–12016 «Программные средства автоматической обработки текста на современном тибетском языке (морфологический уровень)».

Литература

[1] О проекте «Программные средства автоматической обработки текста на современном тибетском языке (морфологический уровень)». http://corpora.spbu.ru/svn/newtibet/index_newtibet.html.
[2] Dobrov A. Semantic and ontological relations in AIIRE natural language processor. In: Computational Models for Business and Engineering Domains, ITHEA, RzeszowSofia, 2014, pp. 147-157.
[3] Bu chung. Mngon brjod tshig mdzod. Lhasa, 1997.
[4] Большой энциклопедический словарь. Языкознание. 2-е (репринтное) изд. лингвистического энциклопедического словаря. М.: Научное издательство «Большая Российская энциклопедия», 1998.
[5] Beyer S.V. The classical Tibetan language. New York, 1992.
[6] Гринев-Гриневич С.В. Терминоведение: учеб. пособие для студ. высш. учеб. заведений. М.: Издательский центр «Академия», 2008.
[7] Касевич В.Б. Буддизм. Картина мира. Язык. СПб.: Центр «Петербургское Востоковедение», 1996.
[8] Tournadre N., Sangda Dorje. Manual of standard Tibetan: language and civilization. Ithaca, New-York, Boulder, Colorado, 2003.

Principles of Tibetan compounds processing in lexical database

M.O. Smirnova, P.L. Grokhovskiy

Saint-Petersburg State University

The paper is devoted to the study of noun-compounds from modern Tibetan corpus with the use of relational lexical database. The lexical database represents a consistent classification of concepts of Tibetan lexical units with different relations between them. The paper describes the structure of the database; principles of processing Tibetan compounds; recognized types of compounds semantic structure.

Most Tibetan compounds belong to grammatical, religious philosophical and general scientific terms. Therefore the paper specifies the processing principles of subject area compounds, including areas, identified by Buddhist linguistics picture of the world.

Keywords: lexical database, Tibetan corpus, Tibetan noun-compounds, linguistic picture of the world.