П. Л.Гроховский, В. П.Захаров, Ю. Н. Лебедева,
М. О. Смирнова, М. В. Хохлова
P. L. Grokhovskiy, V. P. Zakharov, Yu. N. Lebedeva,
M. O. Smirnova, M. V. Khokhlova

КОРПУС ПАМЯТНИКОВ ТИБЕТСКОЙ ГРАММАТИЧЕСКОЙ ТРАДИЦИИ[1]

CORPUS OF THE TIBETAN TRADITIONAL GRAMMAR TREATISES

Аннотация. Проект направлен на разработку модели корпуса памятников тибетской грамматической традиции, предположительно, сформировавшейся в VII–VIII вв. н.э. Корпус полезен исследователям тибетской грамматической традиции, а также может быть использован для лингвистических исследований классического и современного тибетского языка, его изучения и преподавания.

Abstract. The project aims at developing a model of a corpus of Tibetan traditional grammar treatises which is proposed to date back to 7–8th centuries C.E. The corpus will be useful to scholars focusing on Tibetan traditional grammar treatises and as well for linguistic research on classical and modern Tibetan language, its description and teaching.

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научно-исследовательского проекта РФФИ «Пилотная версия электронного корпуса тибетских грамматических сочинений» (13-06-00621). Общие принципы и система морфологической разметки разработаны при поддержке Фонда фундаментальных лингвистических исследований (http://www.ffli.ru) (проект С-47 «Базовый корпус тибетского классического языка с русским переводом и лексической базой данных»). Исследование памятников тибетской грамматической традиции выполнено при финансовой поддержке РГНФ в рамках научно-исследовательского проекта РГНФ «Тибетская грамматическая традиция» (11-34-00227а1).

1. Задачи проекта

Проект направлен на разработку корпуса памятников тибетской грамматической традиции, которая, предположительно, начала свое формирование в VII–VIII вв. н.э., когда были созданы первые дошедшие до нас грамматики «Сумчупа» и «Тагкичжугпа». Данная традиция во многом основывается на буддийских грамматиках, составленных индийскими учеными, и таким образом восходит к индийской грамматической традиции. По методам описания и анализа явлений языка они значительно отличаются от западного языкознания. Современные тибетские лингвисты продолжают поддерживать и развивать традицию классического тибетского языкознания.

В рамках данного проекта планируется создать корпус грамматических сочинений, которые наиболее ценятся в тибетской грамматической традиции: 1) двух первых трактатов «Сумчупа» и «Тагкичжугпа» (VII–VIII вв. н.э.), авторство которых традиционно приписывается создателю тибетской письменности Тхонми Самбхоте, 2) грамматики Смтритиджнянакирти «Врата речи, [подобные] мечу» (XI в. н.э.), 3) комментария к двум первым трактатам Ситу Махапандиты «Прекрасный жемчужный венок – ожерелье мудреца» (XVIII в.), 4) комментария к двум первым трактатам Нгулчу Дхармабхадры «Устные наставления по сочинению великого ученого Ситу» (XIX в.), 5) комментария к двум первым трактатам неизвестного автора под названием «Драгоценный венок благих изречений» (XVIII/XIX вв.), 6) тибетской грамматики Келсанг Гьюрме «Ясное зерцало – введение в тибетскую грамматику» (XX в.).

При работе над проектом участники используют электронные текстовые ресурсы – тибетские тексты и русские переводы следующих тибетских грамматических сочинений разного периода (VII–XX вв.) общим объемом около 10 000 словоформ: «Сумчупа» («Тридцать строф» – один из двух базовых трактатов тибетской грамматической традиции, ок. 1000

словоформ), два наиболее важных комментария к нему «Устные наставления по сочинению великого ученого Ситу» (XIX в., ок. 4000 словоформ) и «Драгоценный венок благих изречений» (XVIII/XIX вв., ок. 5000 словоформ), фрагменты современных лингвистических трудов Келсанг Гьюрме «Ясное зерцало – введение в тибетскую грамматику» (Пекин, 1981, ок. 1000 словоформ) и «Правила написания тибетского письма – Зрелище, желанное для всех» Пари Сангье (Пекин, 1997, ок. 2000 словоформ).

Работа построена на использовании интерактивной электронной базы данных по грамматике тибетского языка «Тибетская грамматическая терминология», созданной в результате работы участников данного проекта в рамках тематического плана Восточного факультета.

2. Современная корпусная лингвистика тибетского языка

Несмотря на то, что над разработкой средств представления тибетских текстов трудятся ученые в разных странах (Германия, Великобритания, КНР, США, Япония), до сих пор не выработан единый стандарт корпусного представления тибетского языкового материала.

На последних четырех конференциях международной тибетологической ассоциации (IATS Seminar) проводилась секция «Тибетские информационные технологии» («Tibetan Information Technology»), на которой были представлены проекты применения компьютерных технологий в области тибетологии, в том числе и проекты создания и использования корпусов тибетского языка[2].

Создание корпусов тибетского языка за рубежом только начинается. Совместным научно-исследовательским проектом 441 в Тюбингенском университете Эберхарда Карла (Тюбинген, Германия) под руководством Б. Цайслер был создан подпроект

[2] См. **http://www.columbia.edu/~ph2046/iats/it/**. Дата обращения 10.04.2012

под номером B11 «Семантические и падежные отношения и межклаузальная референция в тибетском языке», существовавший с 2002 по 2008 год. В результате деятельности этого подпроекта был создан небольшой тибетский корпус, состоящий из четырех текстов – 2 текста на доклассическом тибетском и по одному тексту на классическом тибетском языке и ладакхском диалекте, на котором говорят тибетцы, живущие в Индии и Пакистане. В данном корпусе использовалась частеречная, синтаксическая и семантическая разметки. На сайте проекта 441 доступны фрагменты этого небольшого корпуса, в котором все тексты снабжены переводом и указанием на источник (оригинальный текст). Но основная цель, для которой и создавался данный корпус тибетского языка это заполнение «пробелов» в тибетской грамматике и, в частности, в семантике глаголов, которая изучаются на основе текстов[3].

В 2012 г. Н.Хилл и У.Пагель из Колледжа востоковедения и африканистики при Лондонском университете начали разработку тибетского корпуса объемом 1 миллион слогов, тексты которого будут относится к трем периодам истории тибетского языка – доклассическому, классическому и современному[4].

Первое отличие описываемого здесь проекта от перечисленных выше заключается в том, что в его рамках будет разработана система лингвистической разметки, разработанная на основе теоретического описания тибетского языка, предложенного руководителем проекта в его публикациях[5].

---

[3] См. **http://www.sfb441.uni-tuebingen.de/b11/index-engl.html** Дата обращения 10.04.2012

[4] См. **http://www.soas.ac.uk/news/newsitem73472.html** Дата обращения 10.04.2012

[5] См., например, *Гроховский П.Л.* Категории сказуемости и номинализации действия (субстантивно-адъективные формы) в классическом тибетском языке // Очерки по теоретической грамматике восточных языков: Существительное и глагол./Под редакцией В.Г. Гузева. – СПб.: Издательский дом СПбГУ, 2001, С. 269–288;

Анализ опубликованных материалов и материалов, находящихся в разработке, показывает значительные расхождения исследователей в подходах к лингвистическому анализу грамматических категорий классического тибетского языка. Второе отличие связано с характером используемых материалов – это тексты, посвященные одной из традиционных наук тибетцев – лингвистике.

3. Структура корпуса

Проект направлен на решение двух задач: создание параллельного корпуса тибетских грамматических сочинений с русским переводом, а также создание на его основе лексической базы данных с частотными характеристиками и семантическими отношениями.

Интерфейс корпуса позволяет осуществлять поиск, сортировку и фильтрацию в корпусе по всем элементам аннотации, переход к конкордансу – соответствующим кратким контекстам словоупотребления внутри корпуса, а также к расширенным контекстам.

Тибетские тексты и русские переводы в корпусе будут выровнены по границам предложений тибетской части, в тибетском тексте размечены границы словоформ (что само по себе является нетривиальной задачей, т.к. традиционная орфография маркирует лишь границы слогов, приблизительно в 95 случаях из 100 совпадающие с границами морфем). Существуют различные программы для автоматического выравнивания, например, Hunalign, Vanilla и др. Однако их настройка для тибетского языка непроста и представляет собой отдельную задачу, решение которой не входит в данный проект. Сравнительно небольшой объем корпуса и необходимость разметки границ словоформ, требующая ручного просмотра всего текста корпуса, свидетельствуют в пользу ручного выравнивания.

Гроховский П.Л. Грамматика имени существительного в классическом тибетском языке // Там же, С. 76–91.

4. Разметка корпуса

Тибетский текст будет подвергнут морфологической разметке (лемматизация, частеречная разметка, грамматическая аннотация глагольных форм, снятие грамматической омонимии) в полуавтоматическом (с постредактированием) режиме. Для автоматизации разметки планируется адаптировать для тибетского языка программу TreeTagger. При этом не исключается возможность разработки собственной программы-разметчика.

Создана система тегов для тибетского языка. Приведем список тегов для обозначения наиболее употребительных служебных лексем (см. в табл. 1).

Также текстам корпуса приписываются элементы метаразметки, которые включают информацию о жанре, датировке, авторе текста, его принадлежности к конкретной буддийской школе. Размеченные тексты загружаются в базу данных корпусного менеджера. В качестве корпусного менеджера в тестовом режиме используется система Sketch Engine. Был создан небольшой фрагмент корпуса и загружен в систему Sketch Engine (см. рис. 1).

В состав лексической базы, полученной на основе корпуса, будут добавлены элементы традиционного грамматического метаописания – разметка по терминологическим категориям (фонологическая, синтаксическая терминология), научный комментарий, санскритские эквиваленты (для заимствованных терминов), гиперссылки на синонимы, гиперонимы, гипонимы.

Таблица №1. Список тегов для обозначения служебных лексем в тибетском языке.

№	Тег	Служебная лексема	Пример
1	Cj	союз	dang
2	Pp	послелог	drung du
3	Erg	показатель эргатива	алломорфы kyis, gyis, gis, s, yis
4	Com	показатель комитатива	dang
5	Dat	показатель датива	la
6	Loc	показатель локатива	na
7	Dest	показатель дестинатива	алломорфы tu, du, ra, ru, su
8	Abl	показатель аблатива	las
9	El	показатель элатив	nas
10	Comp	показатель компаратива	алломорфы pas, bas
11	Gen	показатель генитива	алломорфы kyi, gyi, gi, ‘i, yi
12	Fin	конечная частица	алломорфы go, ngo, do, no, bo, mo, ‘o, ro, lo, so, to
13	Top	выделительная частица	ni
14	Ind	неопределенная частица	алломорфы cig, zhig, shig
15	Emph	усилительная частица	алломорфы kyang, yang, 'ang
16	Quant	слова, выражающие количественные значения ('столько', 'именно' и т.п.)	tsam, kho na, 'ba' zhig, snyed
17	Pl	показатель множественного числа	rnams
18	Quot	показатель прямой речи	алломорфы ces, zhes

Рис. 1. Фрагмент выдачи данных для поиска по части леммы «ра»

5. Заключение

Пилотная версия электронного корпуса тибетских грамматических сочинений будет полезна исследователям памятников тибетской грамматической традиции, а также может быть использована для лингвистического исследования тибетского языка, его изучения и преподавания, т.к. в настоящее время отсутствуют общедоступные корпусы размеченных текстов на тибетском языке, тем более с переводом на русский язык.

В дальнейшем предполагается снабдить корпус синтаксической разметкой, увеличить его объем, а также развить его в более масштабный корпус текстов на тибетском языке, в том числе текстов, посвященных другим традиционным наукам тибетцев: буддийской религиозной доктрине, логике, медицине, ремеслу, поэтике, синонимике, стихосложению, астрологии и драме.