Лексическая база корпуса тибетских грамматических сочинений ♣

П.Л. Гроховский, В.П. Захаров, А.М. Попов, М.О. Смирнова, М.В. Хохлова

Санкт-Петербургский государственный университет
plgr@mail.ru, vz1311@yandex.ru, hedgeonline@gmail.com,
2321781@mail.ru, khokhlova.marie@gmail.com

Аннотация

Настоящий доклад посвящен описанию работ по созданию параллельного тибетско-русского корпуса. Корпус представляет собой свод памятников тибетской грамматической традиции с VII по XX вв. н.э. и их переводов на русский язык. Данная традиция во многом основывается на буддийских грамматиках, составленных индийскими учеными, и таким образом восходит к индийской грамматической традиции. Одна из задач проекта – формирование на основе корпуса специальной лексической базы, представляющей интерес как для тибетологов, так и для специалистов по общему языкознанию.

1. О корпусе тибетских грамматических сочинений

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 проекта создается корпус тибетских грамматических сочинений, которые наиболее ценятся в тибетской грамматической традиции: двух первых трактатов «Сумчупа» и «Тагкичжугпа» (VII–VIII вв. н.э.), авторство которых традиционно приписывается создателю тибетской письменности Тхонми Самбхоте, и комментариев к ним [1]. В данный момент корпус насчитывает 24960 словоупотреблений (слогов).

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.

2. Специальная разметка грамматической терминологии

2.1. Тибетская грамматическая терминология

Как и в Индии, научное знание в Тибете формировалось в процессе развития, изучения и адаптации религиозной доктрины. Заимствование буддизма оказало сильное влияние на развитие тибетской лингвистической традиции и прочих областей протонаучного знания.

В Тибете выделяли десять традиционных наук, которые в свою очередь подразделялись на пять великих (тиб. rig gnas che ba) и пять малых наук (тиб. rig gnas chung ba). Лингвистика входила в число пяти великих наук наряду с буддийской религиозной доктриной (санскр. adhyatmavidyā, тиб. nang gi rig pa), логикой (санскр. hetuvidyā, тиб. gtan tshigs kyi rig pa), медициной (санскр. cikitsavidya, тиб. gso ba'i rig pa) и ремеслом (санскр. silpakarmasthanavidyā, тиб. bzo gnas kyi rig pa). К малым наукам относили поэтику (санскр. kāvya, тиб. snyan ngag), синонимику (санскр. abhidhāna, тиб. mngon brjod), стихосложение (санскр. chandam, тиб. sdeb sbyor), астрологию (санскр. jyotiṣa, тиб. skar rtsis) и драму (санскр. naṭa, тиб. zlos gar) [4].

Истоки терминологии и моделей описания тибетской лингвистической традиции, так же как и прочих областей протонаучного знания в Тибете, восходили к буддийской религиозной доктрине. В тибетских грамматических сочинениях, как и в трудах по остальным традиционным наукам, широко использовалась терминология, общая для всех индотибетских традиционных наук, преимущественно религиозного и философского характера.

Современное научное языкознание предполагает разграничение обычной лексики и специальной грамматической терминологии. В рамках лингвистических традиций древности в качестве грамматических терминов использовались, как правило, общеупотребительные слова в специальном значении.

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

Часть тибетской грамматической терминологии сформировалась в результате заимствования или калькирования. Начавшаяся в VII веке работа по переводу индийских религиозных текстов привела к заимствованию индийских моделей описания и терминологии. Однако случаи прямого заимствования санскритских терминов встречаются редко, в основном термины переводились на тибетский язык. Активная деятельность по переводу индийских текстов привела к формированию религиозной и философской специальной лексики. В начале IX в. в Тибете были разработаны первые списки буддийской терминологии, включавшие тибетские эквиваленты санскритских терминов. Некоторые исследователи отмечают, что точность тибетских переводов и последовательность в использовании религиозных и философских терминов могут позволить реконструировать утраченные оригиналы санскритских текстов. Таким образом, главным источником терминологии тибетской грамматической традиции была санскритская терминология.

Для тибетской грамматической традиции не характерно выделение таких традиционных для западной лингвистики разделов, как фонология, морфология и синтаксис. Базовые термины тибетской грамматической традиции представляют собой обозначения первичных единиц разных языковых уровней [4].

Большинство тибетских грамматических сочинений начинают грамматическое повествование с описания тибетского алфавита, различных категорий графем, изложения правил образования слога и сочетания графем, а также правил морфологической сочетаемости фонем.

Также тибетские грамматики содержат описание служебных морфем и лексем.

Тибетская грамматическая традиция заимствовала индийскую концепцию семи падежей. В индийской традиции категория падежа связана с категорией *карака* (санскр. kāraka). Последняя представляет собой промежуточный уровень между семантикой и морфологией. Описание данных категорий относится к семантике (определяются семантически; передают свойство участников групп определенных ситуаций), но выражаются они средствами морфологии – в санскрите это в основном конкретные падежные суффиксы (морфологический уровень). Тибетские грамматисты заимствовали у индийцев данную систему категорий.

2.2. Тэги для грамматической терминологии

В рамках проекта было принято решение выработать специальные тэги, маркирующие тибетскую грамматическую терминологию.

Специальная грамматическая разметка в корпусе (см. Табл. 1) позволяет разделить терминологические поля: грамматическую терминологию (тэг Gram) и термины традиционных наук (тэг GenScien).

Отдельные тэги обозначают модели происхождения термина – посредством терминологизации слов общей лексики (тэг GenLex) или путем заимствования (тэг L).

Базовая грамматическая терминология отмечается тэгом TBas. Поскольку тибетским базовым терминам и в целом тибетской терминологии свойственна полисемия, основной задачей было разграничение фонологической терминологии (тэг TPhon) и терминов-названий графем (тэг TGra).

Также было принято решение разделить термины-названия служебных морфем и лексем (тэг TGrMark) и термины падежной грамматики (тэг TCGr).

2.3. Использование специальной терминологической разметки

Использование специальных тэгов в разметке корпуса позволяет импортировать из корпуса терминологическую лексику и создать базу данных тибетской лингвистической терминологии, включающую дополнительные комментарии к терминам: указание языка заимствования для заимствованных терминов, приведение эквивалентов на других языках, уточнение способа заимствования (фонетическое заимствование, калькирование, семантическое заимствование, гибридные термины) и др.

Последующее создание частотного словаря лексических единиц тибетской грамматической терминологии и семантический анализ лексической базы позволят создать своего рода лингвистическую онтологию: выделить гипонимы и гиперонимы, многозначные и синонимичные термины, проанализировать изменение знаний о языке, состав грамматического терминологического поля, оценить степень терминологизации слов общей лексики.

Таблица 1. Тэги для маркировки терминов

Признак классификации	Тэг	Значение
Терминологическое поле	Gram	термин тибетской грамматической традиции
	GenScien	общенаучный термин, термин традиционных наук
Происхождение термина	GenLex	термин тибетского происхождения (образован посредством терминологизации слов общей лексики)
	L	заимствованный термин
Тип терминологии	TBas	базовый грамматический термин
	TPhon	фонологический термин
	TGra	термин-название типа графемы
	TGrMark	термин-название служебных морфем и лексем
	TCGr	термин-падежной грамматики

3. Обработка размеченных текстов

Разметка текстов осуществляется вручную путем приписывания каждому слову следующей информации: словоформа в тибетском написании, словоформа в латинской транслитерации, лемма (лексема) в тибетском написании, лемма в латинской транслитерации, тэг частеречной разметки, терминологический тэг (см. табл. 2).

Таблица 2. Фрагмент размеченного текста

Словоформа в тибетском написании	Словоформа в латинской транслитерации	Лемма в тибетском написании	Лемма в латинской транслитерации	Тэг частеречной разметки	Терминологический тэг
<s>					
<align>					
དེ་	de	དེ་	de	P	
ཉི་	ni	ཉི་	ni	Top	
སྡུད་	sdud	སྡུད་	sdud	VN	Gram L TGrMark
དང་	dang	དང་	dang	Cj	
འབྱེད་པ་	'byed pa	བྱེད་	byed	VN	Gram L TGrMark
དང	dang	དང	dang	Cj	
‖	//	‖	//	Punct	
རྒྱུ་མཚན་	rgyu mtshan	རྒྱུ་མཚན་	rgyu mtshan	N	Gram GenLex TGrMark
ཚེ་སྐབས་	tshe skabs	ཚེ་སྐབས་	tshe skabs	N	Gram L TGrMark
གདམས་ངག་	gdams ngag	གདམས་ངག་	gdams ngag	N	Gram GenLex TGrMark
ལྔ	lnga	ལྔ	lnga	Num	
འོ	'o	འོ	'o	Fin	
‖	//	‖	//	Punct	
</align>					
</s>					

При работе с корпусом возможно находить контексты для заданных слов или словосочетаний, искать переводы в параллельных текстах, работать со статистическими данными и т.п. В размеченных текстах поиск может осуществляться для разных единиц: словоформ или лемм в тибетском написании, словоформ или лемм в латинской транслитерации, грамматических или терминологических тэгов. Для расширенного поиска используется язык регулярных выражений, позволяющий задавать более сложные запросы. Например, результатом для выражения [lemmalat="bye.*"] будет выдача всех контекстов, в которых встречается соответствующая некоторой словоформе лемма на латинице, начинающаяся с последовательности “bye” или равная “bye”.

Система позволяет получать частотные списки на основе загруженных текстов, при этом атрибута-

ми могут выступать различные единицы (как и в вышеописанном случае поиска) и даже комбинации единиц. Так, для данных текстов наиболее частотными грамматическими тэгами являются N, V, Gen и VN. Ниже приведен список наиболее частотных лемм с их транслитерацией (см. табл. 3).

Таблица 3. Частотный список лемм

Лемма в тибетском написании	Лемма в латинской транслитерации
ཀྱི་	kyi
ཀྱིས་	kyis
ལ་	la
ཏུ་	tu
དང་	dang
ཀྱང་	kyang
དེ་	de
སྟེ་	byed
འོ	'o

4. Структура лексической базы

Наполнение лексической базы тибетской грамматической терминологии составляют лексические единицы, отобранные из тибетской части корпуса по заданным тэгам, а именно по тэгам для маркировки терминов (лингвистических метаданных, см. Табл. 1). Также планируются эксперименты по автоматическому извлечению терминов из текстов с использованием дистрибутивно-статистического метода [3].

В качестве методологической основы наполнения базы и промежуточного формата предлагается рассматривать рекомендации проекта TEI (Text Encoding Initiative) [2]. Разметка TEI основана на синтаксисе языка XML. Подробная информация о TEI представлена в Руководстве TEI P5 [5]. При разработке формата базы принималось во внимание, что TEI содержит метки для перекрестных ссылок, позволяющие фактически отображать сетевое представление данных в линейном XML-файле.

В разметке TEI описание элементов данных лексической базы представляет собой последовательность тэгов верхнего уровня – лингвистических терминов, содержание которых уточняется внутренними тэгами.

Опишем шаблон представления данных лексической базы в формате XML по спецификации TEI в соответствии с требуемой структурой данных. Данный шаблон имеет несколько блоков и уровней представления.

4.1. Уровень лексической единицы

Лексическая единица базы в XML-представлении начинается с записи вида:

```
<entry n="1" type="lex">
<term>གླེ་ཝི་</term>
<pron>A li</pron>
```

где указывается порядковый номер лексической единицы (ЛЕ) (n="1"), ее тип (type="lex" - лексема), а также собственно заголовочное слово A li в тибетице (тэг <term>) и транскрипция (тэг <pron>).

Далее идёт блок грамматической информации (тэг <gramGrp>):

```
<gramGrp>
        <pos> N </pos>
</gramGrp>
```

где указывается часть речи (тэг <pos>), и, возможно, дополнительная грамматическая информация (тэги <gen>, <flex> и др.).

Так же к уровню лексической единицы относится блок этимологической информации (тэг <etym>), их может быть несколько:

```
<etym n="1">
        <lbl> Сумчупа </lbl>
        <date> VIII вв. н.э.</date>
</etym>
```

Атрибут n — порядковый номер блока этимологии, тэг <lbl> описывает источник информации, тэг <date> — дату или год фиксации. Также могут использоваться тэг <mentioned> — форму фиксации, <lang> — язык, из которого предположительно было заимствовано слово.

4.2. Уровень связей

Связей может быть несколько в рамках одной ЛЕ, каждая из них должна начинаться с тэга <rel> и в качестве атрибута должен указываться тип связи (напр., type="synonym"). Так, для описанной выше ЛЕ A li ‘гласная’ приводится связанная с ней единица.

```
<rel type=" synonym ">
        <target>ཡི་གེ་</target>
<'rel:
```

4.3. Уровень примеров

К каждой ЛЕ может быть приведено несколько примеров из корпуса:

```
<cit n="1" type="example">
    <quote>ཡི་གེ་ཐུ་ལི་ཀུ་ལི་གཉིས་॥</quote>
<‘cit>
```

Блок с примером заключается в тэг <cit>, где аргумент n — это номер примера, а type — его тип. Сам текст заключается в тэг <quote>.

5. Система ведения лексической базы

Система ведения базы данных предназначена для хранения всех ЛЕ в виде, удобном для быстрой и оперативной работы. Было принято решение использовать реляционную СУБД Microsoft SQL. Данный выбор был продиктован как широкой распространённостью данной СУБД, так и хорошей интеграцией с используемой для разработки системы платформой Microsoft.NET.

5.1. Структура базы

Для упрощения перевода данных XML в базу и наоборот, структура данных и таблиц внутри базы данных соответствует структуре формата TEI. Для достижения этих целей в базе создано 5 таблиц. Иерархически, таблицы можно разделить на несколько уровней (как и уровни XML). Самый высший, первый — уровень входов (ЛЕ) — строки в этой таблице ни на что не ссылаются, но на них ссылаются строки из других таблиц:

1. **Таблица "Entries"** для хранения данных о ЛЕ. Содержит следующие поля:

id — уникальный идентификатор в таблице; тип: целочисленный

idLocal — уникальный идентификатор в группе; тип: строка в юникоде длиной в один символ

headWord — орфографическая запись ЛЕ; тип: строка в юникоде максимальной длины

entryType — тип ЛЕ (оставлен для совместимости с TEI); тип: строка в юникоде длиной до 5 символов

pron — транскрипция; тип: строка в юникоде максимальной длины

comment — комментарий; тип: строка в юникоде максимальной длины.

Далее, второй уровень, таблицы, строки которых ссылаются на строки таблицы "Entries":

2. **Таблица "Grammar"** для хранения грамматической характеристики слова. Возможен только один блок данного типа на одну ЛЕ. Содержит следующие поля:

id — уникальный идентификатор в таблице; тип: целочисленный

idEntry — ссылка на идентификатор ЛЕ, к которой относится данный блок информации; тип: целочисленный

pos — часть речи; тип: строка в юникоде длиной до 5 символов

flex — парадигма; тип — строка в юникоде длиной до 20 символов

gram — прочие грамматические характеристики для разных частей речи (например, род для существительных); тип: строка в юникоде длиной до 5 символов.

3. **Таблица "Etymology"** для хранения этимологической информации. Возможно несколько блоков данного типа на одну ЛЕ. Содержит следующие поля:

id — уникальный идентификатор в таблице; тип: целочисленный

idEntry — ссылка на идентификатор словарной статьи, к которой относится данный блок информации; тип: целочисленный

idLocal — уникальный идентификатор в группе, если для словарной статьи имеется несколько блоков этимологической информации; тип: строка в юникоде длиной в один символ

source — источник фиксации; тип: строка в юникоде длиной до 20 символов

date — год (или дата) первого упоминания (фиксации); тип — строка в юникоде длиной до 20 символов

mentioned — графическая запись слова; тип: — строка в юникоде длиной до 50 символов

language — язык упоминания; тип: — строка в юникоде длиной до 20 символов.

4. **Таблица "Relations"** для хранения данных о связях. Возможно несколько блоков данного типа на одну ЛЕ. Содержит следующие поля:

id — уникальный идентификатор в таблице; тип: целочисленный

idEntry — ссылка на идентификатор сЛЕ, к которой относится данный блок информации; тип: целочисленный

idLocal — уникальный идентификатор в группе, если для ЛЕ имеется несколько связей; тип: строка в юникоде длиной в один символ

definition — собственно отсылка; тип — строка в юникоде максимальной длины.

5. **Таблица "Examples"** для хранения примеров употреблений. Возможно несколько блоков данного типа на одно значение. Содержит следующие поля:

id — уникальный идентификатор в таблице; тип: целочисленный

example — собственно пример употребления; тип — строка в юникоде максимальной длины.

5.2. Объектная модель представления данных

Центральный и наиболее важный компонент системы, ядро — это программный модуль, являющийся связующим звеном между тремя способами представления данных в разрабатываемой системе — XML-представлением, базой данных и пользовательским интерфейсом. Для обеспечения максимальной совместимости между компонентами ядро и пользовательский

интерфейс разрабатываются на платформе Microsoft.NET на языке программирования C#.

5.3. Пользовательский интерфейс

Управляющим компонентом системы является интерфейс пользователя для взаимодействия с ядром. Интерфейс представляет собой оконное приложение, разработанный на платформе Microsoft.NET и тесно интегрированное с ядром системы. При разработке интерфейса были поставлены следующие требования:

1. интерфейс должен позволять осуществлять поиск по всем ЛЕ в базе;
2. интерфейс должен позволять просматривать ЛЕ в удобной форме;
3. интерфейс должен позволять вручную создавать ЛЕ;
4. интерфейс должен позволять редактировать уже существующие ЛЕ;
5. интерфейс должен позволять загружать в базу (производить импорт) ЛЕ из XML-файлов в формате TEI;
6. интерфейс должен позволять сохранять в файл (производить экспорт) в формате TEI ЛЕ из базы.

Было принято решение ввести два режима работы — для поиска-просмотра и для редактирования. Первый режим предполагается как предпочтительный вариант для работы с системой обычному пользователю, для которого важно в первую очередь возможность быстрого поиска нужной информации. Второй режим предполагается как профессиональный, для использования специалистами-тибетологами при пополнении лексической базы.

Литература

[1] Гроховский П. Л., Захаров В. П., Лебедева Ю. Н., Смирнова М. О., Хохлова М. В. Корпус памятников тибетской грамматической традиции. // Труды международной конференции «Корпусная лингвистика–2013». – СПб.: С.-Петербургский гос. университет, Филологический факультет, 2013. – С. 258-265.

[2] Захаров В.П. Электронный обменный формат для словарей проекта TEI (Text Encoding Initiative): Учебное пособие. – СПб.: СПбГУ. РИО. Филологический факультет, 2013.

[3] Захаров В.П., Хохлова М.В. Автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода как инструмент создания тезаурусов // Структурная и прикладная лингвистика. Выпуск 9. СПб., 2012. С. 222-233.

[4] Смирнова М.О. Базовые термины тибетской грамматической традиции // Вестник Санкт-Петербургского университета. Серия 13. Востоковедение. Африканистика. Выпуск 1. СПб., 2014. С. 23-34.

[5] TEI P5: Guidelines for Electronic Text Encoding and Interchange / Eds. L. Burnard, S. Bauman. S. l., 2010.

Lexical database of the corpus of the Tibetan traditional grammar treatises

P.L. Grokhovskiy, V.P. Zakharov, A.M. Popov, M.O. Smirnova, M.V. Khokhlova

The paper describes the process of developing a parallel Tibetan-Russian corpus. The corpus represents the collection of Tibetan grammar treatises from VII to XX cc. and their translation into Russian. The Tibetan linguistic tradition is mainly based on Buddhist grammars composed by Indian scholars. Thus it is closely connected with the Indian grammatical tradition. The project particularly aims at the creation of specific grammatical lexical database on the basis of corpus. The database will be useful both to tibetologists and general linguistics specialists.

♣ Исследование по разработке корпуса тибетских грамматических сочинений и лексической базы тибетской грамматической терминологии выполнено при поддержке РФФИ в рамках научно-исследовательского проекта РФФИ «Пилотная версия электронного корпуса тибетских грамматических сочинений» (13-06-00621).

Исследование памятников тибетской грамматической традиции выполнено при поддержке гранта СПбГУ 2.38.293.2014 "Тибетская письменная традиция и современность".