Итоговый отчет по гранту РФФИ №[16-06-00578](https://kias.rfbr.ru/index.php)

Морфосинтаксический анализатор текстов на тибетском языке

Руководитель: Смирнова Мария Олеговна

Реферат

Проект направлен на создание последовательного формально-грамматического описания (формальной грамматики) тибетского языка, включающего в себя все грамматические уровни языковой системы — от морфосинтаксиса (синтактики морфем) до синтаксиса сложных предложений и сверхфразовых единств на базе размеченной коллекции (корпуса) тибетских грамматических сочинений. Тибетский синтаксически размеченный корпус и тибетский синтаксический анализатор разрабатываются впервые в рамках данного проекта.

На материале пилотной версии электронного корпуса тибетских грамматических сочинений, разработанного в рамках гранта РФФИ 13-06-00621 А, создана версия разметки данного корпуса, в которой отражены структуры непосредственных составляющих и зависимостей, соответствующие единицам, ранее выделенным в качестве токенов. На основании размеченного таким образом корпуса создана комбинированная грамматика непосредственных составляющих и зависимостей, отражающая все выделенные в ходе разметки классы моделей, требуемые ими способы линеаризации дочерних составляющих и их грамматические свойства.

Была произведена оценка качества разработанной грамматики, а также корректность разметки корпуса. Для этого грамматика была подключена в качестве языкового модуля к открытому лингвистическому процессору AIIRE, которому были поданы на вход исходные тексты корпуса в неразмеченном виде. Результаты лингвистической обработки были представлены в виде вариантов древовидных структур, аналогичных структурам разметки корпуса.

Результаты автоматической обработки текстов были загружены в корпус-менеджер AIIRE в качестве разметки, для которой в корпус-менеджере выполняется автоматический поиск нераспознанных фрагментов, разрывов между синтаксическими структурами или их перекрытий, свидетельствующих о неполноте разметки, и комбинаторных взрывов, свидетельствующих о её неточности. Данный инструментарий позволил одновременно работать над разметкой корпуса и совершенствовать формальную модель, стоящую за используемым лингвистическим обеспечением.

Таким образом, в данном проекте использованы новые методы компьютерного синтаксического анализа тибетского текста, новые подходы к компьютерному моделированию ряда синтаксических явлений на материале тибетского языка, новые методы корпусного представления неоднозначных синтаксических структур на материале синтаксически размеченного корпуса тибетских текстов.

Всё это позволяет говорить о полномасштабном формальном компьютерном описании тибетского морфосинтаксиса, базирующегося на комплексном корпусном подходе, заключающемся в том, что компьютерные инструменты являются средством создания корпуса, а сами корпуса становятся мощным и гибким способом изучения языка, что в свою очередь положительно влияет на качество программных продуктов.

Разработанные материалы размещены на корпусном портале СПбГУ <http://corpora.spbu.ru>

Содержание

1. Введение . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .4

2. Основная часть отчета о НИР «Морфосинтаксический анализатор текстов на тибетском языке» . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .6

3. Заключение . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .14

2

Введение

На сегодняшний день создано множество прикладных систем, работа которых ограничена морфологическим анализом (морфологических анализаторов). Систем, в которых реализован синтаксический анализ текстов (парсинг), на сегодняшний день на несколько порядков меньше. К числу таких систем на сегодняшний день можно отнести AIIRE, Helis, ABBYY Compreno, ЭТАП-3, DictaScope, Link Grammar Parser, RASP, OpenNLP, Gate, Tomita Parser, АОТ, AGFL, NLTK и некоторые другие системы, однако абсолютное большинство этих систем не поддерживают работу с языками, отличающимися от тех, для которых они были созданы.

Из перечисленных систем только AIIRE, Helis и NLTK позволяют работать с произвольными формальными грамматиками, из них только AIIRE позволяет выполнять синтаксический анализ на неоднозначно выделяемых единицах, отличающихся от традиционных “токенов”, т.е. на морфемах, что необходимо для любого языка, письменность которого не предусматривает разделителей между словоформами, в том числе и для тибетского. Кроме того, для тибетского языка в данном исследовании такой анализатор создается впервые.

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

На последних четырёх конференциях Международной тибетологической ассоциации (IATS) проводилась секция «Тибетские информационные технологии» («Tibetan Information Technology»), на которой были представлены проекты применения информационных технологий в области тибетологии, в том числе и создания корпусов тибетского языка. Участники проекта располагают материалами данных конференций и поддерживают научные контакты с их авторами: Пол Дж. Хэкетт «Автоматическая сегментация и частеречная разметка для тибетского языка: первый шаг к машинному переводу», «Анализ Юникод-кодировки для тибетского языка на основе энтропии», Роберт Р. Чилтон «Сортировка тибетских слов в формате Юникод с помощью алгоритма многоразового повтора», «Сортировка данных на тибетском языке в формате Юникод: от алгоритма повтора к реализации», Церинг Гья и Ванчуг Церинг «Исследование стандарта для частеречной разметки текстов на современном тибетском языке». В 2012 г. Нейтан Хилл и Ульрих Пагель из Колледжа востоковедения и африканистики при Лондонском университете получили грант на разработку тибетского корпуса. Объем данного корпуса составляет 1 миллион слогов, а тексты относятся к классическому тибетскому языку. Корпус с полуавтоматической морфологической разметкой, проверенной вручную, был опубликован в открытом доступе только в прошлом году.

Помимо отсутствия синтаксических анализаторов, следует отметить также и отсутствие достоверных подробных описаний тибетского синтаксиса как такового. Синтаксис тибетского языка изучен и описан крайне мало и недостаточно, попытки описывать его с помощью средств компьютерной лингвистики также отсутствуют. Единственное имеющееся подробное описание тибетского синтаксиса в монографии "Классический тибетский язык" американского тибетолога Стивена Бейера (1994) содержит ряд неточностей и упущений, а также носит в ряде формулировок умозрительный характер, не будучи основанным на достаточном текстовом материале, не говоря уже о корпусе тибетских текстов.

Актуальность данного исследования продиктована назревшей как в общей лингвистике, так и, тем более, в тибетологии, где количество последовательных грамматических описаний исчисляется единицами, необходимостью в создании синтаксически размеченных корпусов текстов и формально-грамматических моделей, способных не только описывать, но и объяснять наблюдаемые языковые явления. Множество дискуссионных вопросов современной тибетологии до сих пор не могут найти своего решения, поскольку аргументы в пользу тех или иных решений часто носят умозрительный, не подкреплённый языковым материалом характер. В данном исследовании будет осуществлена попытка не только создать такой корпус и инструментарий для работы с ним и автоматической обработки текстов, но и найти решение некоторым наиболее актуальным вопросам тибетской грамматики на материале этого корпуса.

Основная часть отчета о НИР

Целью данного исследования было создание последовательного формально-грамматического описания (формальной грамматики) тибетского языка, включающего в себя все грамматические уровни языковой системы — от морфосинтаксиса (синтактики морфем) до синтаксиса сложных предложений и сверхфразовых единств на базе размеченной коллекции (корпуса) тибетских грамматических сочинений. Проект был направлен на решение следующих задач:

1. Разработать синтаксическую разметку для корпуса тибетских текстов, отражающую информацию об особенностях тибетских структур составляющих и зависимостей, о системе тибетских грамматических категорий, их возможных значений и ограничений на их совмещение;

2. Создать синтаксически размеченный корпус тибетских текстов, в котором будут полностью отражены структуры высказываний, из которых состоят тексты, от уровня морфосинтаксиса до уровня сложных предложений и сверхфразовых единств;

3. Разработать формально-грамматический языковой модуль открытой системы обработки текстов на естественных языках, позволяющий этой системе автоматически анализировать тексты на тибетском языке, выполняя полный морфологический и синтаксический разбор каждого предложения.

Ожидалось, что созданный инструментарий лингвистической обработки текстов позволит автоматизировать процедуру разметки корпусов текстов и в дальнейшем обеспечить пополнение корпуса в полуавтоматическом режиме.

Использование древовидных синтаксических структур на различных языковых уровнях должен пощволить отразить особенности тибетской грамматической системы, которые при иных подходах вызывают существенные затруднения даже в вопросе о словоделении (в частности, вызывают споры вопросы о статусе формантов, образующих нефинитные формы глагола; последние, по мнению участников проекта, носят синтетический характер, а по мнению тибетологов из Школы востоковедения и африканистики - аналитический).

В рамках проекта была предложена система тэгов, разработанная на основе лингвистического описания тибетского языка, предложенного первым руководителем проекта Гроховским П.Л. в теоретических публикациях. Помимо морфологической разметки, в корпусе также осуществлена и синтаксическая разметка. Лингвистический процессор AIIRE, использованный в данном проекте, позволил использовать комбинированную грамматику составляющих и зависимостей, и при этом анализировать текст на любых языковых уровнях, в том числе - на уровне морфосинтаксиса. Данный лингвистический процессор является многоцелевым, несмотря на то, что в штатном режиме синтаксический анализ работает в связке с семантическим компонентом и компьютерной онтологией, позволяющей автоматически разрешать синтаксическую неоднозначность. В данном проекте использовался другой возможный режим лингвопроцессора - позволивший выполнить синтаксический анализ без участия семантического компонента, что дало возможность увидеть все возможные интерпретации, допустимые с точки зрения формальной грамматики.  
Способность анализировать морфосинтаксические структуры словоформ наряду с синтаксическими структурами словосочетаний и предложений сделала вышеуказанный лингвистический процессор в наибольшей степени соответствующим цели и задачам данного исследования.

В данном исследовании были впервые совмещены методы анализа структур непосредственных составляющих и методы грамматики зависимостей на тибетском материале. Идея совмещения этих методов высказывалась в различных работах, и даже была снабжена необходимым математическим обоснованием (Гладкий А.В. «Синтаксические структуры естественного языка в автоматизированных системах общения», 1985), однако долго не могла получить алгоритмической (программной) реализации по причине зависимости критериев выделения синтаксических единиц от недоступной на стадии синтаксического анализа семантической информации. В данном исследовании была использована комбинированная модель грамматики непосредственных составляющих и зависимостей, разработанная в рамках диссертационного исследования одного из участников проекта (Добров А.В. «Автоматическая рубрикация новостных сообщений средствами синтаксической семантики», 2014) и получившая программную реализацию в виде открытого лингвопроцессора AIIRE, основанная на более совместимом с программной реализацией математическом аппарате (см. aiire.org).

Для разметки корпуса текстов, в отличие от других исследований, где преобладает так называемый «вертикальный» формат, был использован формат XML, позволяющий отразить иерархический (древовидный) характер моделируемых структур и обеспечить возможность поиска языкового материала по структурно-синтаксическим, а не только морфологическим и лексическим характеристикам. Формат XML используется в ряде синтаксических корпусов (в частности, в синтаксически размеченном подкорпусе Национального Корпуса Русского Языка — проект «СинТагРус»), однако, как правило, в разметке отражается лишь информация о зависимостях (подчинительных связях) между лексическими единицами.

В разработанном языковом модуле тибетского языка разделение входной строки на морфемы осуществляется с использованием алгоритма Ахо-Корасик (Альфред Ахо и Маргарет Корасик). Данный алгоритм позволяет находить все возможные подстроки входной строки по заданному словарю. Алгоритм предполагает построение дерева, описывающего конечный автомат, терминальными узлами которого являются завершения цепочек символов слов (в данном случае морфем) из входного словаря.  
Языковой модуль содержит словарь морфем, что позволяет строить дерево автомата заранее, на этапе сборки самого языкового модуля. Во время работы лингвопроцессора это дерево загружается в виде части исполняемого модуля, и время на его загрузку минимально.

Алгоритм Ахо-Корасик позволил с максимальной эффективностью выделять в тексте известные системе единицы (токены), однако корректность выделения тех или иных единиц в тексте с точки зрения правил грамматики на уровне данного алгоритма никак не проверялась; более того, выделяемые единицы могли пересекаться, и лишь последующий морфосинтаксический анализ позволял верифицировать ту или иную сегментацию.

Инструментарий корпус-менеджера AIIRE позволяет одновременно работать над разметкой корпуса и совершенствовать формальную модель, стоящую за используемым лингвистическим обеспечением, что представляет собой новый подход к разработке модулей лингвистического процессора, обеспечивающий постоянную верифицируемость формальной модели и её соответствие корпусному материалу. При создании формальных грамматик изначально учитываются, как правило, лишь наиболее типичные и понятные разработчику явления языковой грамматики, однако при работе с корпусом текстов обнаруживается множество неочевидных, но частотных конструкций, не учтенных при создании модели. К числу таких явлений в тибетском корпусе относилось употребление различных сочетаний с цифрами и числительными, имён и именованных сущностей, окказионализмов и специфических конструкций, в том числе - металингвистические употребления экспонентов языковых единиц (например, суффикс -*pa*-/-*ba*-).

Работа с данным корпусом велась последовательно. Устраняя нераспознанные фрагменты, разрывы в разметке, перекрытия и комбинаторные взрывы путём усовершенствования лингвистического обеспечения, разработчик в итоге добивается не только полной разметки корпуса, но и такого состояния формальной модели, при котором она объясняла бы все наблюдаемые в корпусе явления.

На материале пилотной версии электронного корпуса тибетских грамматических сочинений, разработанного в рамках гранта РФФИ 13-06-00621 А, была создана версия разметки данного корпуса, в которой были отражены структуры непосредственных составляющих и зависимостей, соответствующие единицам, ранее выделенным в качестве токенов. В этой версии разметки корпуса использовался формат XML, позволяющий единообразно отразить иерархические структуры составляющих словообразовательных, формообразовательных и словоизменительных основ и структуры текстовых словоформ, с указанием роли каждой единицы и типа деривационной модели (используемого класса непосредственно составляющих (КНС)).

На обработанном пилотном материале удалось выделить следующие основные типы токенов: v\_suff (глагольный суффикс), p\_ind\_root (корень указательного местоимения), n\_root (именной корень), p\_pers\_root (корень личного местоимения), case\_marker (показатель падежа), v\_root (глагольный корень), num\_root (корень числительного), p\_def\_root (определительное местоимение), fin (показатель конца высказывания), n\_suff (суффикс существительного), num\_suff (аффикс числительных), conj\_root (корень союза), posessive\_root (корень притяжательного местоимения), adj\_root (корень прилагательного), adv (наречие), p\_int\_root (корень вопросительного местоимения), p\_emph\_root (эмфатическая частица), quot (показатель цитирования); а также пунктуационные типы токенов: punct (цег), square\_bracket (квадратные скобки), round\_bracket(круглые скобки), tibetan\_bracket (тибетские скобки), double\_quotes (двойные кавычки) и др.

Для всех перечисленных типов токенов в файле grammarDefines.py были указаны их возможные морфологические и морфонологические признаки. Например, для глагольного корня расписаны такие возможные свойства, как наличие наклонения (true / false), наличие времени (true / false), наклонение (индикатив, императив) и время (настоящее, прошедшее, будущее), тип финальной фонемы, определяющий сочетаемость глагольного корня с суффиксальными алломорфами, трансформативность и переходность. Выделенные типы токенов и их грамматические признаки легли в основу разрабатываемой формальной грамматики, позволяющей лингвопроцессору строить древовидные синтаксические структуры различных видов.

На основании размеченного таким образом корпуса была создана комбинированная грамматика непосредственных составляющих и зависимостей, отражающая все выделенные в ходе разметки классы моделей, требуемые ими способы линеаризации дочерних составляющих и их грамматические свойства. В общей сложности в созданной грамматике за три года разработано 500 различных классов непосредственных составляющих, для которых указана информация о допустимых главных и зависимых дочерних составляющих, об их грамматических свойствах, допустимых способах линеаризации, возможностях эллипсиса, о типах отношений зависимости и иных свойствах, используемых морфосинтаксическим анализатором при обработке текста.

При классификации деривационных моделей был использован опыт разработки системы правил морфологического анализа, полученный британскими исследователями Нейтеном Хиллом, Эдвардом Гарреттом и Ульрихом Пагелем (Школа востоковедения и африканистики при Лондонском университете) и при создании их морфологического анализатора, а также семантико-синтаксические характеристики каждой модели, определяющие валентностные характеристики порождаемых единиц.

В качестве отдельного (базового) грамматического модуля был создан модуль системы тибетских грамматических категорий, их возможных значений и ограничений на их совмещение. Для работы с морфологией и морфонологией тибетского языка были созданы несколько специализированных файлов: файл grammarDefines.py для определения типов выделенных токенов, их свойств и ограничений, и несколько файлов, содержащих разные типы атомов (словари алломорфов), где для каждого алломорфа указана морфема, тип токена и свойства в соответствии с файлом grammarDefines.py.

Была построена разметка структур более высокого уровня — словосочетаний и простых предложений, сложных предложений и некоторых типов сверхфразовых единств. В процессе этой работы древовидные структуры синтаксической разметки были объединены в более крупные древовидные комплексы, с учётом информации о способах линеаризации дочерних составляющих и их грамматических свойствах, а также о классах синтаксических единиц, в том числе — коннекторов и иных средств объединения простых предложений.

Сложное предложение в тибетском языке образуется, прежде всего, при помощи групп конвербов: клауза, от которой образуется конверб, выступает в качестве простого компонента в составе сложного предложения. Поскольку группы конвербов в тибетском языке одновременно являются простыми компонентами в составе сложного предложения и адвербиальными формами глагола, классы непосредственных составляющих для конвербов были частично проработаны на начальных этапах работы. Также были созданы классы, необходимые для функционирования координации клауз в составе бессоюзных предложений (механизм союзного сочинения в тибетском языке отсутствует). На заключительном этапе работы были добавлены некоторые классы конвербов, которые образованы при помощи падежных показателей – создан класс TerminativeConverb (пример: སློབ་ཏུ (slob tu) ‘чтобы изучать’). Все классы конвербов были встроены в сложные предложения (КНС ClosedStatement). Например, རྗེས་འཇུག་ཡི་གེ་བཅུ་པོ་ལ། །དྲུག་པ་སྦྱར་ན་འབྱེད་སྡུད་ཡིན། (rjes ‘jug bcu yi ge bcu po la/ drug pa sbyar na ‘byed sdud yin/) ‘если добавить к десяти конечным фонемам шестую, [полученный показатель] будет являться соединением и разделением’.  
Были созданы классы для заголовков (КНС Header), сочетаний заголовок+подзаголовок (КНС TitledChapter); разные вариации классов для повествовательных предложений (клаузы/группы клауз) с финальной частицей (КНС StatementFin и КНС StatementFinNoSep) и без финальной частицы (КНС StatementNoFin) и класс КНС ClosedStatement – для повествовательных предложений с необходимыми пунктуационными знаками в конце.

В рамках заключительного этапа данной работы были также выделены классы сверхфразовых единств. С формально-грамматической точки зрения, в тибетском языке возможно выделить только некоторые типы сверхфразовых единств. Класс Text объединяет несколько предложений с заголовком. Был создан общий класс для сверхфразового единства, типичного для всех тибетских текстов, содержащего информацию об авторе, издательстве, месте и времени публикации: WorkTitleWithAuthorAndPublisherWithDateAndPlace, например: བོད་ཀྱི་བརྡ་སྤྲོད་རིག་པའི་ཁྲིད་རྒྱུན་རབ་གསལ་མེ་ལོང༌། སྐལ་བཟང་འགྱུར་མེད་ཀྱིས་བརྩམས། སི་ཁྲོན་མི་རིགས་དཔེ་སྐྲུན་ཁང༌། 1981 ལོ་རུ་ཁྲེང་ཏུའུ། (bod kyi brda sprod rig pa’i khrid rgyun rab gsal me long/ skal bzang ‘gyur med kyis brtsams/ si khron mi rigs dpe skrun khang/ 1981 lo ru khreng tu’u/) ‘Поток наставлений по тибетской грамматике, ясное зерцало. Сочинил Келзанг Гьюрме. Сычуаньское издательство национальностей. В 1981 году, Чэнду’; включающий следующие классы: WorkTitleWithAuthor – название текста и автор; WorkTitleWithAuthorAndPublisher – название текста, автор и издательство; PublicationDateAndPlace – дата и места публикации.

Был создан класс для элементов нумерованного списка NumberedElement, например, (1) བྱེད་པ་པོ་སྟོན་པ ((1)byed pa po ston pa) ‘1. Выражение значения деятеля’.

Сверхфразовые единства, в которых несколько предложений объединены общностью темы, на уровне морфосинтаксиса неотличимы от текста. Для таких типов сверхфразовых единств необходимо тематическое моделирование текста, которое возможно осуществить только с использованием семантического анализа, представляющего собой задачу для отдельного исследования.

В процессе работы над разметкой корпуса на уровне сложных предложений и сверхфразовых единств был выявлен ряд явлений, потребовавших доработки грамматики, созданной на предыдущих этапах работы, на различных уровнях текста от морфонологии и синтаксиса словосочетания до синтаксиса простого предложения.  
В грамматику были добавлены классы конструкций с вспомогательными глаголами (в т.ч. модальными), способными присоединять слева группу знаменательного глагола без служебных морфем, например སྐྱེད་ནུས (skyed nus) ‘может породить’. Для этого был создан признак “aux” у свойства транзитивности глагольного корня. Создан соответствующий класс AuxVPNoTenseNoMood, например ལས་བྱེད་ནུས (las byed nus) ‘может делать дело’.

Отдельной обработки потребовало явление топикализации в составе сложного предложения. Помимо обычной группы темы, маркированной показателем ནི (ni), были созданы следующие КНС, распространенные в поэтических произведениях: группа темы в аблативе TopicalizedAblativeNP (пример: སྡེ་བདུན་པ་ལས་ནི (sde bdun pa las ni)’что касается [той, которая] из седьмого ряда’); группа темы в генитиве TopicalizedGenitiveNP (пример: མིང་གི་ནི (ming gi ni) ‘что касается [того, что относится к] имени’); группа темы в терминативе TopicalizedTerminativeNP (пример: བདག་པོའི་ས་རུ་ནི (bdag po'i sa ru ni) ‘что касается [известного] в качестве показателя обладателя’), группа темы в дативе TopicalizedDativeNP; группа темы в эргативе TopicalizedErgativeNP.

В процессе анализа этих разрывов был обнаружен и обработан ряд синтаксических явлений, не зафиксированных на предыдущих этапах работы, в частности – явлений, объясняющихся спецификой текстов корпуса, являющихся грамматическими сочинениями. Так, отдельные классы потребовалось создавать для атомов классов Letter (класс, содержащий буквы тибетского алфавита) и Exponent (класс, содержащий экспоненты морфем). Для сочетаний “Letter/Exponent + слоговой разделитель” были созданы отдельные классы EntityArg и EntityRightArg, которые в свою очередь встраиваются в качестве аргумента в классы TransitiveVP и NPGen соответственно, что позволило добиться связывания предложений типа སུ་ཡི་ཨུ་ཕྱིས་ནས། །དེ་ལ་གསུམ་པའི་དང་པོ་སྦྱར། (su yi u phyis nas/ de la gsum pa’i dang po sbyar) ‘удалив u, [из показателя] su, к ней добавляют первую [букву] третьего [ряда алфавита]’.

В рамках выполнения проекта был обнаружен ряд фундаментальных проблем, препятствующих реализации морфосинтаксического анализа тибетского текста в отрыве от семантики. В связи с этим для снятия морфосинтаксической неоднозначности при разборе идиоматизированных именных групп и групп прилагательного, а также многосложных именных и глагольных корней, которые также являются идиомами, использовалась компьютерная онтология AIIRE. В данном исследовании, как и во многих других, под онтологией понимается база данных, содержащая информацию о концептах и отношениях между ними и, фактически, являющаяся эксплицитной спецификацией концептуализации. Концепты представляют собой формальные (математические, компьютерные) модели понятий. Концепты онтологий включают в себя атрибуты и связаны между собой отношениями. Атрибуты и отношения взаимосвязаны: участие концепта в некотором отношении может трактоваться как атрибут и наоборот. Отношения между концептами характеризуются бинарностью и направленностью и могут быть представлены в виде логических формул, определенных в терминах некоторого исчисления, задающего правила логического вывода. В настоящий момент онтология, разрабатываемая для данного исследования, содержит 4335 концептов, являющихся значениями 3943 тибетских выражений.

Семантическая разметка выполняется вышеупомянутым лингвистическим процессором в процессе анализа текста путём семантической интерпретации синтаксических структур. Данный лингвистический процессор реализует так называемый метод межуровневого взаимодействия: неоднозначность анализа на нижестоящих уровнях разрешается за счёт отсутствия возможности интерпретации на вышестоящем уровне. Таким образом, в частности, отсутствие возможности семантической интерпретации позволяет исключать семантически некорректные версии анализа синтаксических структур. Каждая версия семантической разметки представляет собой так называемый семантический граф.

Прежде всего, использование онтологии было необходимо для разрешения морфосинтаксической неоднозначности, обусловленной множеством грамматически корректных, но семантически бессмысленных вариантов разбиения сложных предложений на компоненты; выбор корректного варианта сегментации чаще всего обусловлен идиоматикой тибетского языка, прежде всего, корректным выделением идиоматизированных морфокомплексов – тибетских композитов.

Все тибетские композиты образованы соположением морфем или словоформ и, таким образом, относятся к атематическому типу (отсутствует соединительный элемент). В некоторых случаях при словосложении возможно выпадение служебных морфем. В зависимости от частеречной принадлежности сложного слова в тибетском языке принято выделять многосложные именные и глагольные корни.

По характеру синтаксических связей между компонентами при осуществлении данного исследования были выделены следующие типы именных композитов: имя-композит однородных именных корней, имя-композит однородных определений, имя-композит именной группы с генитивом, имя-композит сигнификативной именной группы, имя-композит субъектно-предикатной группы, имя-композит именованной сущности, имя-композит с приложением; и глагольных композитов: глагол-композит группы переходного глагола, глагол-композит группы однородных глагольных корней.  
Изначально онтология позволяла установить идиоматический статус для многосложных выражений и единый тип токена для всех имен-композитов.

Вместе с тем, поскольку существенное количество случаев комбинаторного взрыва было вызвано некорректными версиями разборов именно многосложных слов (в основном, композитов, а также некоторых дериватов), было принято решение расширить количество типов токенов в онтологии в соответствии с выявленными в формальной грамматике классами непосредственных составляющих для именных и глагольных композитов, а также некоторых дериватов. Для всех типов таких многоморфемных сочетаний для классов непосредственных составляющих в формальной грамматике был указан признак ‘only\_idiom=True’ с целью исключения не-идиоматических версий разбора.

Помимо моделирования идиоматики, необходимого для разрешения многих случаев синтаксической неоднозначности, онтология позволила ввести ряд ограничений, позволивших задействовать механизм семантических валентностей. В настоящее время в онтологии используется четыре типа семантических ограничений – ограничение генитивных отношений; ограничение приложений; ограничение субъекта, прямого объекта глагола, адресата и терминативного объекта.

Ограничение базового генитивного отношения «обладать любым объектом или процессом» осуществлялось посредством установления подкласса данного отношения между базовыми классами онтологии. Так, наличие номинализованного глагола གྲགས (grags) ‘быть прославленным’ в генитивной конструкции གྲགས་པའི་ལུས (grags pa’i lus) приводило к двум версиям разбора – ‘тело славы’ и ‘тело прославленного’. Для исключения первой версии разбора концепт ལུས (lus) ‘тело’ был связан генитивным отношением с концептом ‘любое существо’, что позволило исключить версию разбора, в которой слава могла обладать телом.

Тибетское приложение присоединяются к именной группе справа. Поскольку в тибетской графической системе отсутствует пробел или другие разделители слов, приложения на письме неотличимы от частей многосложных именных и глагольных корней. Так, например, многосложное слово ཕ་མ (pha ma) ‘мать [и] отец’ (имя-композит однородных именных корней) может быть неверно интерпретировано как «отец-мать» (отец, который является матерью, где мать интерпретируется как приложение) и ‘мать отца’ (имя-композит именной группы с генитивом). Вторая версия исключается установлением для композита ཕ་མ (pha ma) верного типа токена в онтологии. Первая же версия может быть исключена только посредством семантического ограничения возможности эквивалентности базовых классов для частей данного композита. Ограничение субъектов и прямых объектов глаголов использовались не только для верного анализа многосложных именных и глагольных корней и идиоматизированных именных групп и групп прилагательного, но для исключениях лишних синтаксический версий при возникновении комбинаторных взрывов.

В программном обеспечении корпус-менеджера была доработана расширенная функциональность поиска по синтаксическим деревьям с целью поиска сложных предложений и сверхфразовых единств по их синтаксическим структурам. Для этого была доработана функциональность формирования шаблоны структуры составляющих и зависимостей в пользовательском интерфейсе с указанием свойств терминальных вершин, а также ограничений, описанных в формальной грамматике. Результатом работы поиска является список цитат из документов, которые хотя бы в одной интерпретации соответствуют поисковому шаблону синтаксического дерева. При этом реализована возможность просмотра найденной интерпретации, а также всех остальных возможных интерпретаций найденного предложения или сверхфразового единства. Реализованный способ хранения морфосинтаксической и синтаксической разметки предполагает использование одинаковых представлений результатов поиска как для сверхфразовых единств, предложений или словосочетаний, так и для токенов, что позволяет осуществлять поиск не только по структурам словосочетаний, предложений и сверхфразовых единств, но и по разметке внутренних структур единиц, ранее считавшихся токенами.

3

Заключение

Таким образом, в рамках проекта была разработана синтаксическая разметка для корпуса тибетских текстов, отражающая информацию об особенностях тибетских структур составляющих и зависимостей, о системе тибетских грамматических категорий, их возможных значений и ограничений на их совмещение; создан синтаксически размеченный корпус тибетских текстов, в котором отражены структуры высказываний, из которых состоят тексты, от уровня морфосинтаксиса до уровня сложных предложений и сверхфразовых единств.

Был разработан формально-грамматический языковой модуль открытой системы обработки текстов на естественных языках, позволяющий этой системе автоматически анализировать тексты на тибетском языке, выполняя полный морфологический и синтаксический разбор каждого предложения. В общей сложности в созданной комбинированной грамматике непосредственных составляющих и зависимостей разработано 500 различных классов непосредственных составляющих, для которых указана информация о допустимых главных и зависимых дочерних составляющих, об их грамматических свойствах, допустимых способах линеаризации, возможностях эллипсиса, о типах отношений зависимости и иных свойствах, используемых морфосинтаксическим анализатором при обработке текста.

В рамках выполнения проекта был обнаружен ряд фундаментальных проблем, препятствующих реализации морфосинтаксического анализа тибетского текста в отрыве от семантики. Для разрешения морфосинтаксической неоднозначности, обусловленной множеством грамматически корректных, но семантически бессмысленных вариантов разбиения сложных предложений на компоненты; использовалась компьютерная онтология AIIRE. Помимо моделирования идиоматики, необходимого для разрешения многих случаев синтаксической неоднозначности, онтология позволила ввести ряд ограничений, позволивших задействовать механизм семантических валентностей.  
В настоящее время в онтологии используется четыре типа семантических ограничений – ограничение генитивных отношений; ограничение приложений; ограничение субъекта, прямого объекта глагола, адресата и терминативного объекта. В настоящий момент онтология, разрабатываемая для данного исследования, содержит 4335 концептов, являющихся значениями 3943 тибетских выражений.

Семантическая разметка выполняется вышеупомянутым лингвистическим процессором в процессе анализа текста путём семантической интерпретации синтаксических структур. Данный лингвистический процессор реализует так называемый метод межуровневого взаимодействия: неоднозначность анализа на нижестоящих уровнях разрешается за счёт отсутствия возможности интерпретации на вышестоящем уровне. Таким образом, в частности, отсутствие возможности семантической интерпретации позволяет исключать семантически некорректные версии анализа синтаксических структур. Каждая версия семантической разметки представляет собой так называемый семантический граф.

Реализованный способ хранения морфосинтаксической и синтаксической разметки предполагает использование одинаковых представлений результатов поиска как для сверхфразовых единств, предложений или словосочетаний, так и для токенов, что позволяет осуществлять поиск не только по структурам словосочетаний, предложений и сверхфразовых единств, но и по разметке внутренних структур единиц, ранее считавшихся токенами. Для этих единиц на текущий момент мера покрытия составляет в среднем более 99.5%.

Оценка меры покрытия осуществлялась, как и на предыдущем этапе работы, путём расчёта отношения средней длины максимальных выделенных НС к длине предложения. Вместе с тем, так как на данном этапе работы возникла необходимость в подключении семантического анализа, благодаря которому были устранены многие некорректные версии синтаксического разбора, в разметке возникли дополнительные разрывы, обусловленные неполнотой используемой онтологии в сочетании с действующими семантическими ограничениями: некоторые семантически корректные версии синтаксического разбора были исключены наряду с некорректными в силу действия этих ограничений. Поэтому прямая оценка данной величины дала 55%: такой является средняя относительная длина фрагментов текста, получивших не только синтаксическую, но и семантическую разметку.

К сожалению, выполнить аналогичный расчёт для синтаксической разметки, выполняемой без семантики, оказалось невозможным по техническим причинам: количество версий синтаксического разбора сложных предложений и сверхфразовых единств и объём занимаемой ими памяти превысили доступные команде проекта машинные ресурсы. В ходе дальнейших исследований планируется разработка полномасштабного семантического анализатора тибетских текстов на материале данного корпуса, в рамках которой планируется обеспечить полноту покрытия не только для синтаксических, но и семантических интерпретаций. Для оценки меры неоднозначности было рассчитано среднее количество версий синтаксической разметки каждого фрагмента, обладающего какой-либо синтаксической интерпретацией. Благодаря семантическим ограничениям оно не только не выросло с переходом от синтаксиса простого предложения к синтаксису сложных предложений и сверхфразовых единств, но и сократилось по сравнению с результатами 2017 года, составив в среднем 58.13 версий. Вместе с тем, в процессе дальнейшей работы над семантикой планируется дальнейшее сокращение данного показателя.

В программном обеспечении корпус-менеджера была доработана расширенная функциональность поиска по синтаксическим деревьям с целью поиска сложных предложений и сверхфразовых единств по их синтаксическим структурам. Для этого была доработана функциональность формирования шаблоны структуры составляющих и зависимостей в пользовательском интерфейсе с указанием свойств терминальных вершин, а также ограничений, описанных в формальной грамматике. Результатом работы поиска является список цитат из документов, которые хотя бы в одной интерпретации соответствуют поисковому шаблону синтаксического дерева. При этом реализована возможность просмотра найденной интерпретации, а также всех остальных возможных интерпретаций найденного предложения или сверхфразового единства. Реализованный способ хранения морфосинтаксической и синтаксической разметки предполагает использование одинаковых представлений результатов поиска как для сверхфразовых единств, предложений или словосочетаний, так и для токенов, что позволяет осуществлять поиск не только по структурам словосочетаний, предложений и сверхфразовых единств, но и по разметке внутренних структур единиц, ранее считавшихся токенами. Для этих единиц на текущий момент мера покрытия составляет в среднем более 99.5%.

Полученные результаты были представлены научному сообществу в виде выступлений на научных конференциях (в т.ч. Text, Speech, and Dialogue, Брно, "Корпусная лингвистика-2017", "Интернет и современное общество" и "Азия и Африка: наследие и современность") и отражены в восьми научных публикациях:

1. Гроховский П.Л., Добров А.В., Доброва А.Е., Захаров В.П., Михайлова М.О., Смирнова М.О., Сомс Н.Л. Средства автоматической обработки текстов на тибетском языке/ Азия и Африка: наследие и современность. Материалы конгресса. Санкт-Петербургский государственный университет, Восточный факультет. 2017. С. 220-221.

2. Гроховский П.Л., Добров А.В., Доброва А.Е., Сомс Н.Л. Корпус-менеджер для морфосинтаксической разметки: опыт разработки корпуса тибетских грамматических сочинений/ Корпусная лингвистика ‒ 2017. Труды международной конференции. 2017. С. 157-161.

3. Dobrov A., Dobrova A., Grokhovskiy P., Soms N. Morphosyntactic Parser and Textual Corpora: Processing Uncommon Phenomena of Tibetan Language/ Proceedings of the International Conference IMS-2017. P. 143-153 (2017).

4. Dobrov, A., Dobrova, A., Grokhovskiy, P., Soms, N., Zakharov, V. Morphosyntactic analyzer for the Tibetan language: aspects of structural ambiguity/ International Conference on Text, Speech, and Dialogue, 2016. P. 215-222.

5. Grokhovskiy, P., Dobrov, A., Dobrova, A., Smirnova, M., Soms, N. Idioms modeling in a computer ontology as a morphosyntactic disambiguation strategy: The case of tibetan corpus of grammar treatises/ Text, Speech, and Dialogue. 21st International Conference, TSD 2018. Brno, Czech Republic, September 11–14, 2018. Proceedings. - Vol. 11107, LNAI, 2018, P. 76-83.

6. Grokhovskiy, P., Dobrov, A., Dobrova, A., Smirnova, M., Soms, N. Computer ontology of Tibetan for morphosyntactic disambiguation/ Communications in Computer and Information Science. Vol. 859. International Conference on Digital Transformation and Global Society, 2018, P. 336-349.

7. Kramskova, A. Treatise on Writing, Treatise on Music: Comparing Terminology/ FIFTH INTERNATIONAL SEMINAR OF YOUNG TIBETOLOGISTS. Book of Abstracts. Saint-Petersburg State University. September 3–7, 2018, P. 30-31.

8. Smirnova, M. Study of the Tibetan Linguistic Picture of the World Using Computer Ontology/ FIFTH INTERNATIONAL SEMINAR OF YOUNG TIBETOLOGISTS. Book of Abstracts. Saint-Petersburg State University. September 3–7, 2018, P. 50-51.  
  
Разработанные материалы размещены на корпусном портале СПбГУ http://corpora.spbu.ru