ST PETERSBURG STATE UNIVERSITY
INSTITUTE FOR LINGUISTIC STUDIES (RAS)
HERZEN STATE PEDAGOGICAL UNIVERSITY OF RUSSIA

# PROCEEDINGS
## OF THE INTERNATIONAL CONFERENCE
## «CORPUS LINGUISTICS–2017»

June 27–30, 2017, St. Petersburg

SAINT PETERSBURG UNIVERSITY PRESS
2017

*Организационный комитет конференции*
*«Корпусная лингвистика–2017»*

В. П. Захаров (председатель), Е. Л. Алексеева,
Л. Н. Беляева (зам. председателя), А. О. Гребенников,
О. Н. Камшилова, О. Н. Крылова, О. А. Митрофанова,
И. С. Николаев (зам. председателя), Я. К. Харапет, М. В. Хохлова

*Программный комитет конференции*
*«Корпусная лингвистика–2017»*

В. П. Захаров (председатель), И. В. Азарова, Е. Л. Алексеева,
Л. Н. Беляева, В. Бенко (Словакия), Н. В. Борисов, В. В. Бочаров,
Р. Вальденфельс (Польша), Л. А. Вербицкая, Р. Гарабик (Словакия),
А. С. Герд, Л. Л. Иомдин, Н. Н. Казанский, В. Б. Касевич,
М. В. Копотев (Финляндия), Д. А. Кочаров, О. Н. Ляшевская,
В. Матоушек (Чехия), О. А. Митрофанова, К. Пала (Чехия),
В. Петкевич (Чехия), В. А. Плунгян, Л. В. Рычкова (Беларусь),
С. О. Савчук, В. П. Селегей, Д. В. Сичинава, М. В. Хохлова,
А. Я. Шайкевич, С. А. Шаров (Великобритания), Т. Ю. Шерстинова

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ РАН
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ
ИМ. А. И. ГЕРЦЕНА

# ТРУДЫ
## МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
# «КОРПУСНАЯ ЛИНГВИСТИКА–2017»

27–30 июня 2017 г., Санкт-Петербург

ИЗДАТЕЛЬСТВО С.-ПЕТЕРБУРГСКОГО УНИВЕРСИТЕТА
2017

ББК 81.1
Т78

*Ответственный редактор издания*
В. П. Захаров

**Труды международной конференции «Корпусная лингвистика–2017».** — СПб.: Изд-во С.-Петерб. ун-та, 2017. — 340 с.

**Т78**

Сборник содержит материалы докладов, представленных на научной конференции «Корпусная лингвистика–2017» 27–30 июня 2017 г. в Санкт-Петербурге.

Создание корпусов текстов является одним из приоритетных направлений в современной компьютерной лингвистике. Проведение конференции по данной тематике знакомит ученых с современными разработками и новыми технологическими решениями в этой области, а также способствует обобщению опыта научных исследований по корпусной лингвистике.

**ББК 81.1**

*Программный комитет конференции выражает искреннюю благодарность Российскому фонду фундаментальных исследований за финансовую поддержку, грант No 17-06-20225 Г*

*П.Л.Гроховский, А.В.Добров, А.Е.Доброва, Н.Л.Сомс*
*P.L.Grokhovsky, A.V.Dobrov, A.E.Dobrova, N.L.Soms*

# КОРПУС-МЕНЕДЖЕР ДЛЯ МОРФОСИНТАКСИЧЕСКОЙ РАЗМЕТКИ: ОПЫТ РАЗРАБОТКИ КОРПУСА ТИБЕТСКИХ ГРАММАТИЧЕСКИХ СОЧИНЕНИЙ[1]

# CORPUS MANAGER FOR MORPHOSYNTACTIC ANNOTATION: EXPERIENCE OF DEVELOPMENT OF CORPUS OF INDIGINEOUS TIBETAN GRAMMAR TREATISES

**Аннотация.** В данной статье представлен опыт разработки корпус-менеджера для морфосинтаксической разметки на материале корпуса тибетских грамматических сочинений. Рассматриваются проблемы токенизации и вертикальной разметки тибетского текста, обусловленные особенностями синтактики тибетских алломорфов. Предлагается новый подход к организации разметки корпуса, не требующий разбиения текста на словоформы и основанный на синтаксической разметке. Описывается созданная технология отладки морфосинтаксической разметки, объединяющая корпус-менеджер, формальную грамматику и лингвистический процессор и позволяющая эффективно дорабатывать языковые модули лингвистического процессора так, чтобы формальная модель объясняла все, а не только некоторые явления в корпусе.
**Ключевые слова.** Корпусный менеджер; тибетский язык; морфосинтаксическая разметка; токенизация; лингвистический процессор.
**Abstract.** The article presents the experience of developing a corpus manager for morphosyntactic annotation on the basis of The Corpus of Indigenous Tibetan Grammar Treatises. The problems of tokenization and vertical markup of Tibetan texts are considered, which are conditioned by Tibetan allomorph syntactics features. A new approach to organization of the corpus annotation is proposed, which does not require segmentation of text into word forms and is based on syntactic annotation. The developed technology of debugging morphosyntactic markup is described, which integrates the corpus manager, the formal grammar and the natural language processor and allows to effectively refine linguistic modules of the natural language processor so that the formal model can explain not just some, but all the phenomena in the corpus.
**Keywords.** corpus manager; Tibetan language; morphosyntactic annotation; tokenization; natural language processor.

В рамках проекта РФФИ «Морфосинтаксический анализатор текстов на тибетском языке» был создан корпусный менеджер (далее — КМ) для ранее подготовленного корпуса тибетских грамматических сочинений. Токенизация и морфологическая разметка были первоначально выполнены вручную: искусственное разделение тибетских текстов на токены характеризовалось некоторой произвольностью,

[1] Исследование выполнено в рамках научно-исследовательского проекта РФФИ «Морфосинтаксический анализатор текстов на тибетском языке» (16-06-00578 А).

так как графематическое или какое-либо иное деление на словоформы в тибетском языке отсутствует; поэтому было принято решение ограничиться инвентарём более чётко определяемых атомарных единиц морфосинтаксической структуры: алломорфов, знаков пунктуации, разделителей, цифр.

Синтактика алломорфов в тибетском языке переплетена с синтаксисом предложения, поэтому формальная грамматика должна моделировать все уровни грамматической системы от алломорфов до высказываний и текстов.

Существующие КМ ориентированы на языки иного строя и работают с токенизацией и морфологической разметкой, поэтому было принято решение разработать иной КМ, который бы позволил: 1) работать с синтаксической (морфосинтаксической) разметкой и 2) находить в ней места, требующие усовершенствования лингвистического процессора, её порождающего.

КМ позволяет загружать неразмеченные тексты или тексты в "вертикальном" формате для их дальнейшей автоматической разметки, отражающей структуры непосредственных составляющих и зависимостей, при этом единицы, ранее считавшиеся токенами, разбиваются на алломорфы, которые далее объединяются в древовидные структуры.

Поиск, организованный в КМ, позволяет находить морфосинтаксические структуры по заданным моделям. На данный момент доступен поиск по тибетским моделям слово- и формообразования, однако может быть реализован поиск по синтаксическим структурам любой сложности (в разметке других корпусов в КМ представлены структуры предложений и текстов, и расширение границ разметки тибетского корпуса планируется в будущем). Поскольку поиск осуществляется не по словам, а по морфосинтаксическим деревьям, результатом поиска являются фрагменты синтаксических структур (морфосинтаксические деревья с грамматическими характеристиками и морфемным наполнением).

КМ включает в себя поддержку для корпусов на различных языках документов корпусов и включает ряд инструментов для автоматической разметки корпуса и обнаружения фрагментов этой разметки, требующих усовершенствований лингвистического обеспечения («ошибок» разметки).

КМ предоставляет возможность просмотра разметки полностью размеченных фрагментов текста. Для частично размеченных фрагментов отображается три дополнительных вида помет: нераспознанные

единицы, разрывы и перекрытия синтаксических деревьев. Нераспознанными считаются фрагменты, для которых в разметке отсутствуют синтаксические деревья; разрывами — позиции, в которых дерево не может быть связано с соседним; перекрытиями — фрагменты текста, в которых пересекаются синтаксические деревья, не полностью покрывающие текст: фрагмент, покрытый одним деревом, включает позицию начала фрагмента, покрытого вторым деревом, но не позицию его конца.

Данный инструментарий 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. 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 объясняет все наблюдаемые в корпусе явления.

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 осуществлялась именно методом последовательного устранения недостатков разметки путем последовательного пополнения и исправления словарей и формальной грамматики; при этом было устранено около 700 разрывов, 100 перекрытий и 200 комбинаторных взрывов.

При разработке КМ использовалась гибкая архитектура, включающая в себя СУБД PostgreSQL, серверное приложение (скрипты на языке Python, подключённые к веб-серверу Apache через mod_wsgi) для обработки JSON-RPC запросов и статический веб-интерфейс с адаптивной верстkой на основе технологии Bootstrap: КМ доступен как в десктопных, так и в мобильных браузерах. Лингвистический про-

цессор также подключён к КМ: серверный компонент КМ отправляет запросы gsoap-серверу AIIRE на обработку и автоматическую разметку текстов; разметка передаётся в формате XML. КМ обеспечивает сохранение разметки в СУБД; на стороне СУБД средствами хранимых процедур производится парсинг XML и сохранение разметки в таблицы непосредственных составляющих и грамматических признаков. Дальнейшая обработка разметки производится средствами серверного компонента КМ: производится поиск и сохранение в СУБД нераспознанных фрагментов, разрывов, перекрытий и комбинаторных взрывов. Кроме того, серверный компонент КМ обладает API для выдачи компонентов разметки в виде JSON-объектов, отражающих НС-структуры с грамматической информацией, информацией о зависимостях и морфемном наполнении. Отрисовка древовидных структур реализована на клиентской стороне на языке Javascript: НС-структуры изображаются в виде SVG-файлов. При существенном объеме данных этот процесс может занимать длительное время. Исходный код КМ, как и исходный код лингвистического процессора, является открытым и доступен по адресу: http://svn.aiire.org/repos/tproc/trunk/t/corpus_manager. КМ допускает вертикальное и горизонтальное масштабирование стандартными средствами Apache и PostgreSQL. В КМ реализован механизм аутентификации и распределения прав доступа; гостевой доступ обеспечивает возможность просмотра опубликованных корпусов и их разметки, а также поиска по ним; доступ разработчика и административный доступ позволяют работать с разметкой и инструментарием её отладки.

В ходе морфосинтаксической разметки корпуса 86 192 единицы размечены как атомарные, что почти в 2 раза превосходит количество единиц, ранее считавшихся токенами (48 166). При этом 44 837 из них получили автоматическую разметку в виде морфосинтаксических деревьев, полностью покрывающих исходные токены. Таким образом, суммарное покрытие автоматической разметки составило 97 %. КМ доступен по адресу http://corpora.spbu.ru/corman/.

**Литература**


1. *Beyer S.* (1992), The Classical Tibetan language. New York.
2. *Grokhovskii P. L., Zakharov V. P., Smirnova M. O., Khokhlova M. V.* (2015), The Corpus of Tibetan Grammatical Works // Automatic Documentation and Mathematical Linguistics. Vol. 49, no. 5, pp. 182–191.
3. *Haspelmath M.* (2011), The indeterminacy of word segmentation and the nature of

morphology and syntax // Folia Linguistica. Vol. 45, iss. 1, pp. 31–80.
4. *Захаров В. П.* (2016), Корпусная лингвистика // Прикладная и компьютерная лингвистика / под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. М.: УРСС. 320 с.; с. 138–155.

## References

1. *Beyer S.* (1992), The Classical Tibetan language. New York.
2. *Grokhovskii P. L., Zakharov V. P., Smirnova M. O., Khokhlova M. V.* (2015), The Corpus of Tibetan Grammatical Works. In: Automatic Documentation and Mathematical Linguistics, vol. 49, no. 5, pp. 182–191.
3. *Haspelmath M.* (2011), The indeterminacy of word segmentation and the nature of morphology and syntax. In: Folia Linguistica, vol. 45, iss. 1, pp. 31–80.
4. *Zakharov V. P.* (2016), Corpus Linguistics. In: Applied and Computer Linguistics. Eds. I. S. Nikolaev, O. V. Mitrenina, T. M. Lando. Moscow, URSS, 320 p.; pp. 138–155.

---

**Гроховский Павел Леонович**
**Grokhovskiy Pavel**
***E-mail: p.grokhovskiy@spbu.ru***

**Добров Алексей Владимирович**
**Dobrov Aleksei**
***E-mail: a.dobrov@spbu.ru***

Санкт-Петербургский государственный университет (Россия)
Saint Petersburg State University (Russia)

**Доброва Анастасия Евгеньевна**
**Dobrova Anastasia**
***E-mail: adobrova@aiire.org***

**Сомс Николай Леонидович**
**Soms Nikolai**
***E-mail: nsoms@aiire.org***

ООО "АИРЕ" (Россия)
AIIRE LLC (Russia)