Л.Д. Бадмаева, Ж.Б. Бадагаров, Б.З. Цыдыпов

ОБЩИЕ ПРОБЛЕМЫ ФОРМИРОВАНИЯ КОРПУСА БУРЯТСКОГО ЯЗЫКА[1]

При уже разработанных или интенсивно разрабатываемых лингвистических корпусах (**http://www.natcorp.ox.ac.uk/**, **http://ruscorpora.ru/**, **http://www.narusco.ru/** и др.) в настоящее время представляется уже естественным почти любое обращение к постановке подобной задачи на материале того языка, который еще не вовлечен в орбиту корпусно-ориентированных исследований и проектов. В нашем случае бурятский язык как один из языков народов Сибири также становится как бы «главным действующим лицом» в смысле формирования на его материале лингвистического корпуса. При всех сложностях наших целей и задач в данном направлении, нам, тем не менее, представляется безотлагательным их реализация, вернее начало их реализации, в силу бесспорных преимуществ самих корпусных технологий для языковедов в первую очередь. Уточнение насчет начала реализации сказано выше в связи с тем, что впредь подобные задачи по бурятскому языку и вообще в бурятоведении не ставились в принципе. Новаторство корпусно-ориентированных начинаний в бурятоведении закономерно влечет за собой цепь определенных проблем, которые, как правило, появляются непосредственно в процессе решения поставленных целей и задач (т.е. элементарно если каким-то делом не заниматься, то и проблем с ним связанных появляться не будет).

Ясно, что при организации любого корпуса в первую очередь встает проблема электронных версий текстов, включаемых в него. Нами определены основные пути накопления архива текстовых материалов. Это в первую очередь сканирование, затем

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 08-06-00151а) и ACLS.

договор с официально признанными бурятскими издательствами (Издательский дом «Буряад Үнэн», издательство «Бэлиг», редакция журнала на бурятском языке «Байгал», Бурятское государственное книжное издательство) и с организаторами сайта[1] бурятской литературы (**http://www.nomoihan.org**), давшими свое согласие на предоставление нам электронных версий, опубликованных бурятских текстов для их корректного использования в научных целях.

Вместе с тем, при сотрудничестве с издательствами в плане получения от них электронных версий текстов возникают свои специфические проблемы, заключающиеся в основном в разности компьютерных программ, использующихся как нами (Word), так и издательствами (PageMaker, InDesign). Данное обстоятельство требует осуществления реформатирования, например, издательских текстов для их включения в наш корпус. Кроме этого могут возникать проблемы, связанные с разными бурятскими шрифтами, использующимися издательствами. Таким образом, перевод текстов с бумажных носителей на электронные влечет за собой некоторый «шлейф» проблем, требующий специального и профессионального подхода для их разрешения. Не всегда подобные проблемы могут быть решены самими лингвистами, тем более традиционными (здесь мы имеем в виду традиционных в противовес компьютерным). В подобных случаях необходимо бывает сотрудничество со специалистами-программистами. Также в случае получения текстов от издательств в некоторых случаях могут быть как бы «потеряны» данные, важные для паспортизации текстов корпуса. Данное обстоятельство может диктовать сверку электронного варианта текста с его бумажным. Вообще, как

[1] Организаторы названного сайта (инициатор – Ж.Б. Бадагаров) выполняют подготовку электронных версий опубликованных текстов в сотрудничестве с Центром информатизации Минобразования Республики Бурятия. На данное время количество бурятоязычных сайтов исчисляется числом не более десяти.

показывает наш опыт подготовки электронного архива текстовых материалов, на первых порах работы бумажные версии текстов необходимо сохранять под рукой для справки, уточнения и т.п.

Уже в начале подготовки текстов корпуса представляется целесообразным подготовить инструкцию, по которой удобно было бы брать на заметку различные параметры для паспортизации текстов. Понятно, что данная информация в дальнейшем будет материалом для внешней, т.е. экстралингвистической разметки корпуса. На данное время нами сформирован определенный список параметров по выходным данным текстовых материалов, имеющий 17 пунктов, и по жанровой классификации текстов из трех основных пунктов, каждый с несколькими подпунктами:

- **Выходные данные**

1. автор (пол)
2. составитель (пол)
3. редколлегия (пол)
4. редактор (пол)
5. художник (пол)
6. технический редактор (пол)
7. корректор (пол)
8. название
9. место издания
10. издательство
11. типография
12. год издания
13. объем издания (количество страниц)
14. аннотация
15. шифры ББК, УДК, ISBN
16. тираж
17. тип издания (научное, литературное / художественное, учебное, популярное и т.д.);

- **Жанровая классификация текстов**

1. художественная литература

проза
поэзия
фольклор
драма

2. общественно-публицистическая

газета
журнал
брошюра

3. учебно-научная

учебник
пособие
монография
брошюра
сборник
словарь
справочник.

Основные принципы организации корпуса бурятского языка в целом не расходятся с принятыми в корпусной лингвистике[1]. В соответствии со сказанным мы нацелены на литературный бурятский язык, на котором созданы тексты, начиная со второй половины XX века по настоящее время. При нашей нацеленности на включение в корпус текстов трех основных стилей бурятского языка, таких как художественный, общественно-публицистический и учебно-научный преобладающее распределение происходит в сторону первого, около 70%. Далее идет учебно-научный стиль – 20% и общественно-публицистический, соответственно – 10%. Процентное соотношение текстов разных стилей в готовящемся корпусе бурятского языка выглядит следующим образом:

– художественный стиль (проза – 40%, поэзия – 10%, фольклор – 10%, драма – 10 %) – 70%;
– учебно-научный стиль (учебные пособия, научные труды – 5%, словари – 15%) – 20%;
– общественно-публицистический стиль (пресса, журналы; летописи, исторические сочинения, документация, делопроизводство) – 10%.

Данные пропорции нельзя принимать как окончательные. Обзор соответствующей литературы показывает, что пропорции текстов могут варьировать для их максимальной репрезентативности[2]. Говоря о репрезентативности текстов корпуса, здесь можно заметить два основных подхода. По одному подходу, разработчики того или иного корпуса придерживаются равной / пропорциональной представленности текстов разных стилей и жанров (например, **http://www.narusco.ru/**). По другому подходу,

---

[1] *Гарабик Р.* Словацкий национальный корпус // Труды международной конференции «Корпусная лингвистика – 2004». СПб., 2004. С. 99–121.

[2] *Герд А.С., Захаров В.П.* Национальный корпус русского языка в свете проблем современной филологии // Труды международной конференции «Корпусная лингвистика – 2004». СПб, 2004. С. 122–130.

наряду с вышеописанным требованием еще учитывают распространенность самих форм языка и стилей в нем и соответственно текстов данных форм и стилей[1]. Применительно к нашим целям и задачам второй подход представляется нам более приемлемым. Вероятно, здесь сказывается то, что бурятский язык относится к миноритарным или региональным языкам в силу чего его функции и сферы распространения значительно сужены по сравнению, например, с русским языком и поэтому не все стили и формы бурятского функционируют одинаково активно и пропорционально относительно друг друга. В свою очередь подобное обстоятельство уже само может заведомо предопределять в создающемся корпусе различающуюся в пропорциональном отношении представленность разных стилей и жанров языка.

При подготовке текстовых материалов возникают также проблемы, связанные с алфавитом, который используется данным языком. Например, бурятский алфавит основан на стандартном русском с добавлением трех дополнительных букв для обозначения звуков, специфичных для бурятского языка. Три бурятские буквы содержатся в универсальной кодировке Юникод в диапазоне расширенной кириллицы, как и многие буквы кириллических алфавитов, использовавшихся и до сих пор использующихся многими языками на территории бывшего СССР.

На сегодняшний день языки, использующие расширенную кириллицу, используют самые разные региональные и национальные стандарты. К сожалению, на данный момент не представляется возможным как-то унифицировать это разнообразие, хотя положительные стороны этого очевидны. Единственной унифицированной кодировкой такого рода, охватывающей символы большого количества различных языков, является кодировка Cyrillic Asian компании «Паратайп». Однако эта кодировка не смогла стать региональным стандартом, видимо, из-за того, что

[1] *Гарабик Р.* Словацкий национальный корпус…

ее создатели не ставили перед собой такой цели. Поскольку на данный момент большинство языков, использующих кирилличес-кие алфавиты, находится на территории Российской Федерации, самым удобным решением для обеспечения их полной поддерж-кой в операционных системах наряду с русским является одно-байтовая кодировка, основанная на кириллической кодировке Windows-1251.

Для того чтобы использовать программы, работающие в однобайтовой кодировке, мы использовали однобайтовую кодировку для монгольских языков[1]. В некоторых случаях требовалось конвертировать тексты из однобайтовой в двухбайтовую (юникод) кодировку и наоборот. Для этой цели были написаны два макроса в текстовом редакторе MS Word.

Так, сканирование и распознавание бурятских текстов осуществляется программой ABBYY Fine Reader, которая работает в кодировке Юникод. Для облегчения правки неправильно распознанных символов использовалась упрощенная клавиатурная раскладка для бурятского языка[2].

Вычитка больших объемов отсканированных и распознанных текстов позволила выявить ряд типичных ошибок распознавания. Например, наиболее распространенными ошибками были следующие: 1) распознавание буквы *э* как *з* – *тзрз* вместо *тэрэ* «тот»; 2) распознавание *э* как *а* при плохом качестве оригинала – *хаэна* вместо *хаана* «где»; 3) распознавание *ү* как *у* и наоборот – *уүдэн* вместо *үүдэн* «дверь» и т.п. Для автоматизации их поиска был написан набор макросов, значительно ускоряющий работу по нахождению ошибок и их исправлению.

---

[1] *Бадагаров Ж.Б.* Бурятский язык в эру высоких технологий. Однобайтовая кодировка для монгольских языков // Газета «Буряад Үнэн». Улан-Удэ, 19.08.2004. С. 10–11.

[2] Разработана Ж.Б. Бадагаровым.

В ближайшей перспективе находятся проблемы разработки основных принципов лемматизации слов бурятского языка. В решении данной проблемы нам представляется полезным опыт разработки принципов лемматизации слов современного монгольского языка[1] близкородственного бурятскому.

Из сказанного выше можно видеть, что на данном этапе в целях создания соответствующего корпуса мы заняты в основном подготовительными работами, которые, представляются как неизбежными, так и необходимыми.

[1] *Badam-Osor Khaltar, Atsushi Fujii*. A Lemmatization Method for Modern Mongolian and its application to information retrieval // URL: **http://if-lab.slis.tsukuba.ac.jp/fujii/paper/ijcnlp2008khab.pdf**