Н.Г. Зайцева, М.М. Филатова,
Н.Л. Шибанова, А.А. Крижановский
N.G. Zaitseva, M.M. Filatova,
N.L. Shibanova, A.A. Krizhanovsky

КОРПУС ВЕПССКОГО ЯЗЫКА[1]

THE VEPS CORPUS

Аннотация. В работе описаны ключевые особенности разрабатываемой компьютерной системы, включающей словарь и корпус текстов вепсского языка. К 2015 году Корпус включает более тысячи текстов и более 800 библиографических источников, Словарь содержит более 10 тысяч лемм и словоформ. Корпус и Словарь вепсского языка доступны онлайн: http://vepsian.krc.karelia.ru.

Ключевые слова. Вепсский язык, корпус, словарь.

Abstract. The computer system under development includes the Veps corpus and dictionary. There are more than one thousand of texts and more than 800 bibliographic sources in the corpus, more than 10 000 lemmas and word forms in the dictionary at 2015. The Veps corpus and dictionary are available online at http://vepsian.krc.karelia.ru.

Keywords. Veps, corpus, dictionary.

1. Введение

Тексты на различного рода языках, которые в языкознании называются обычно «образцами речи», становятся все бо-

[1] Работа А.А. Крижановского поддержана грантом РГНФ (проект № 15-04-12006). Работа Н.Г. Зайцевой, М.М. Филатовой, Н.Л. Шибановой выполнена при частичной финансовой поддержке Программы фундаментальных исследований Секции литературы и языка ОИФН РАН «Язык и информационные технологии» 2015–2017 (проект «Корпус вепсского языка: разработка и формирование морфологической базы электронного ресурса»).

лее востребованными не только в исследовательской практике лингвистов, но и этнографов, фольклористов. Историки, введя понятие «история повседневности», также активно стали использовать разного рода тексты на языках исследуемых ими народов. Поэтому появление электронных ресурсов с размещенными в них текстами исключительно востребовано. Что же касается малочисленных народов, языки которых используются в практике жизни небольшим кругом населения и известны узкому количеству специалистов, то параллельные корпусы, включающие в себя переводы текстов на другие языки (на русский, английский) особенно популярны, поскольку из них черпаются разного рода знания не только по языку, но и истории народа, его духовной и материальной культуре. И, таким образом, развитие компьютерных технологий позволяет удовлетворить потребность ученых в сведениях по языку, материальной и духовной культуре народа и обеспечить удобный доступ к корпусу текстов в сети Интернет.

2. Компьютерная реализация

Разработана структура реляционной базы данных корпуса текстов и словаря вепсского языка. База данных включает группы таблиц, обслуживающих словарь, таблицы корпуса текстов, таблицы лексикографических констант (язык, часть речи). Созданы специальные таблицы для хранения и поиска информации, связанной с информантами (имя информанта, место рождения, место записи и т.д.).

2.1. Корпус текстов

На уровне базы данных и объектно-ориентированного языка программирования введено понятие корпуса, каждый текст теперь относится к одному из заранее заданных корпусов, что позволяет пользователю более гибко фильтровать список текстов, выбирая корпуса и какие-либо из параметров

текста (диалект, говор, причитание, стиль текста – публицистический, художественный и др.).

Все тексты (1097) размещены в *5 подкорпусах*:

1) параллельный подкорпус *вепсских диалектных текстов*: 199 текстов разного характера на всех диалектах вепсского языка;
2) параллельный подкорпус *вепсских причитаний* с переводами на русский язык, которые являются продолжением диалектного подкорпуса, поскольку каждая сказка рассказано на своем диалекте или говоре: **47;**
3) параллельный подкорпус *вепсских народных сказок* с переводами на русский язык, которые также являются продолжением диалектного подкорпуса: 55 сказок; **младописьменный подкорпус (43)**;
4) ***библейские тексты*** (переводы Нового Завета) на младописьменном языке вепсов подкорпус библейских текстов (приближен к параллельному подкорпусу, поскольку по номеру стиха всегда можно легко найти соответствующий русский или иной перевод перевод)**: 431;**
5) *младописьменный подкорпус*, который содержит художественные и публицистические тексты разного характера на младописьменном языке вепсов: 89 текстов.

Была проделана кропотливая работа *по вычитке* всех диалектных текстов, помещенных в Корпус вепсского языка, поскольку при передвижении текстов в корпус, при их разбивке на абзацы и т.д. могла выпасть часть слова, могли быть удалены некоторые слова, могли возникнуть отдельные моменты непрочтения латинской графики и специфических графем вепсского алфавита. Поскольку в корпусе предусмотрена *техническая возможность редактировать*, добавлять, удалять тексты и переводы, то работа по сверке и вычитке текстов была проведена, и в настоящее время можно быть уверенным, что все вепсские тексты соответствуют действительности, и

заинтересованные пользователи могут апеллировать к текстам, в которых действительно представлена подлинная вепсская речь без ошибок и искажений.

2.2. Электронный словарь.

Поскольку язык малочисленного вепсского народа, который большую часть своего функционирования находился в бесписьменном состоянии, не настолько богат корневыми словами или леммами, то новые леммы при лемматизации стали встречаться все реже. В настоящее время электронный словарь включает ~ 10 000 лемм и словоформ.

2.3. Архитектура базы данных корпуса текстов и словаря

База данных корпуса текстов и словаря реализована в единой реляционной базе данных. Все таблицы базы данных можно условно разделить на две взаимосвязанные группы: таблицы, обслуживающие словарь, и таблицы корпуса текстов (рис. 1). В свою очередь, из таблиц корпуса можно выделить таблицы, предназначенные для описания паспорта текстов и описания библиографических ссылок (рис. 1).

2.4. Целостность базы данных

Для сохранения целостности базы данных введён ряд правил, препятствующих вводу противоречивой информации. Правила по работе с информантами таковы:

1. Нельзя удалить «Информанта», пока есть текст, к которому привязан данный информант.
2. Нельзя удалить «Имя» информанта, пока есть информант, который привязан к этому имени.

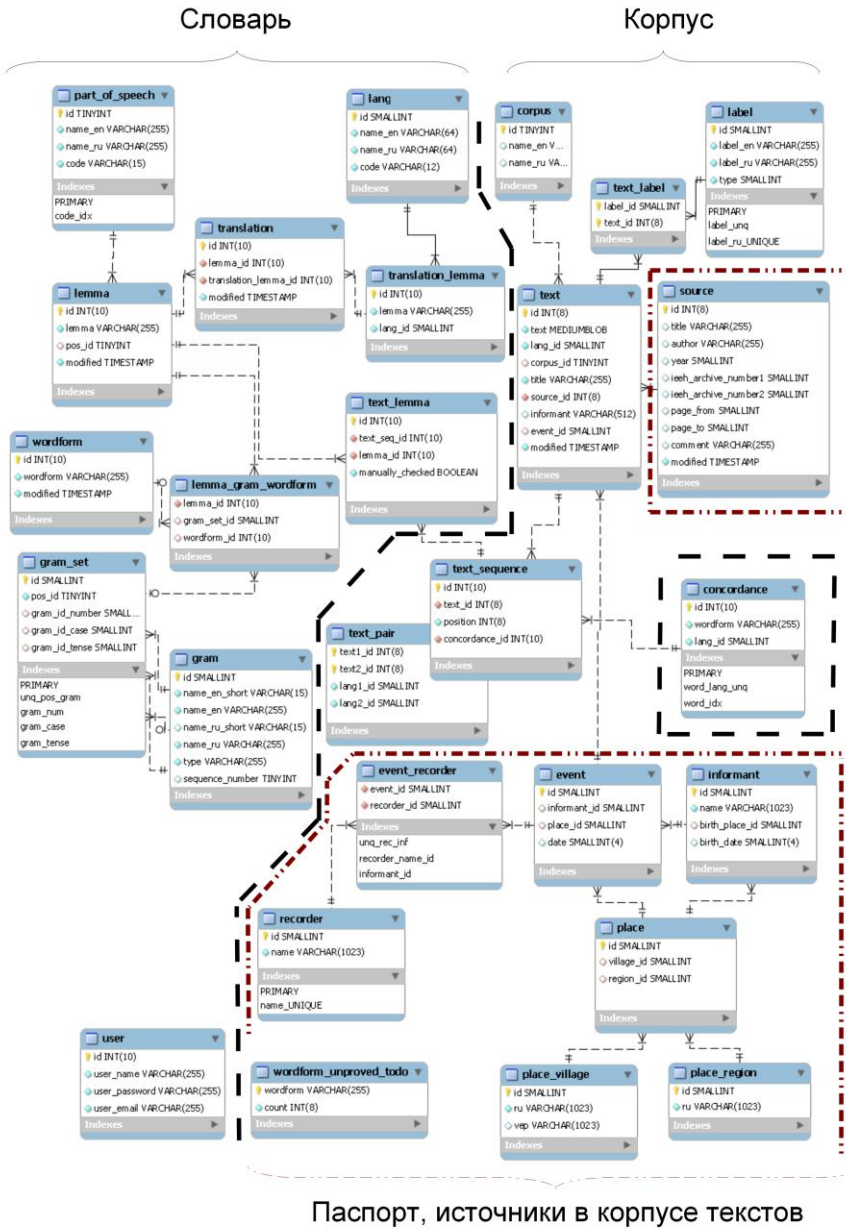

Рис. 1. Три группы таблиц в базе данных корпуса и словаря:

(1) таблицы словаря, (2) таблицы корпуса и (3) таблицы корпуса, обслуживающие паспорт текстов и библиографические ссылки текстов (в этой третьей группе таблицы *source* и *event* являются ключевыми)

3. Нельзя удалить «Географическое местоположение», пока есть информант, который привязан к этому местоположению.
4. Нельзя удалить «Название деревни», пока в базе данных есть запись «Географическое местоположение» с этой деревней.
5. Нельзя удалить запись «Район, область, республику», пока есть запись «географическое местоположение» с этим районом.

Реализован механизм удаления связанных записей для лемм, т.к. леммы связаны посредством таблиц базы данных со словоформами и переводами. Теперь при удалении леммы удаляются связанные с ней словоформы и переводы. Прежде чем выполнить удаление, (1) система выводит на экран удаляемые слова, (2) система выводит предупреждение для тех омонимичных словоформ, которые принадлежат нескольким разным леммам, и не удаляет такие словоформы. Аналогично отрабатывается удаление переводов, связанных с удаляемой леммой. Система выводит предупреждение, что данный перевод используется для перевода других лемм (они выводятся на экран), и не удаляет такие переводы.

2.5. Дополнительные возможности компьютерной системы

Разрабатываемая система обладает богатым функционалом, облегчающим работу читателя и редактора корпуса и словаря:

- Добавлены страницы, позволяющие увидеть списки новых или изменённых текстов, новых лемм и словоформ.

- Расширены возможности поиска, а именно: на страницу "Тексты" добавлена возможность выбора списка текстов (i) по корпусу, (ii) по задаваемым свойствам (параметрам) текста.
- Можно редактировать, добавлять, удалять тексты корпуса и переводы этих текстов.
- К текстам корпуса можно привязывать:

  список информантов (фамилия, имя, отчество, год рождения и записи информанта, место рождения и место записи);

  список людей, выполнявших запись информантов;

  список географических мест (места рождения информантов, места записи).
- Для удобного ввода географического места рождения информантов или места записи разработан интерфейс для редактирования (i) списка деревень (название на русском и вепсском), (ii) списка районов, областей и республик.

Заключение

Разрабатываемый Корпус вепсского языка представляет интерес для ученых разного профиля (лингвистов, этнографов, фольклористов и т.д.), поскольку представляет раритетные материалы языка малочисленного вепсского народа, находящегося под угрозой исчезновения. Результаты проекта также востребованы и в практике жизни. В настоящее время вепсский язык введен как предмет в школьное образование, в соответствующие специализации в вузах, и материалы корпуса востребованы при составлении учебников и учебных пособий на вепсском языке, в преподавательской деятельности, в поддержке работы СМИ как источник по материальной и духовной культуре вепсского народа. Материалы корпуса широко используются при впервые создаваемых правилах орфографии вепсского языка, поскольку системы поиска наглядно демон-

стрируют все возможные диалектные варианты, позволяя выбрать нужный для орфографического словаря вепсского языка. Книга «Орфографический словарь вепсского языка» [1], изданный в практике функционирования вепсского языка впервые, широко использовала возможности корпуса. Находящийся в настоящее время в работе при поддержке РГНФ (2012-2014 гг.; руководитель проекта Зайцева Н.Г.) «Лингвистический атлас вепсского языка» в качестве главного источника опирается на диалектные материалы корпуса, которые четко паспортизированы и в отношении диалектов, и пунктов и времени записи и т.д. и поэтому уникальны в диалектологической работе.

Таким образом, и научный мир, и все заинтересованные пользователи обладают в настоящее время *впервые созданным* в отечественной науке доступным ресурсом, состоящим из электронного словаря и корпуса *диалектных текстов* этнографического и фольклорного содержания. Тексты включают уникальные вепсские причитания, расшифрованные специально для корпуса. Другой уникальный подкорпус – это впервые размещенные в сети Интернет *младописьменные тексты* на вепсском языке (художественные, публицистические, тексты для детей, переводы Библии на младописьменный вепсский язык).

К 2015 году Корпус и Словарь включают более одной тысячи текстов, более 800 библиографических источников, более 10 тысяч лемм и словоформ.

Литература

1. *Зайцева Н.Г., Жукова О.Ю., Харитонова Е.Н.* (составители) (2012), Орфографический словарь вепсского языка. Петрозаводск. 429 с.

References

1. *Zaitseva N.G., Zhukova O.Yu., Kharitonova E.N.* (composers) (2012), Orfograficheskiy slovar' vepsskogo yazyka [Veps language Spelling dictionary]. Petrozavodsk. 429 pp.

Зайцева Нина Григорьевна

Институт языка, литературы и истории Карельского научного центра РАН (Россия).

Zaitseva Nina

Institute of linguistics, history and literature of the Karelian Research Centre of the Russian Academy of Sciences (Russia).

Филатова Мария Михайловна

Институт языка, литературы и истории Карельского научного центра РАН (Россия).

Filatova Maria

Institute of linguistics, history and literature of the Karelian Research Centre of the Russian Academy of Sciences (Russia).

Шибанова Нина Леонидовна

Институт языка, литературы и истории Карельского научного центра РАН (Россия).

Shibanova Nina

Institute of linguistics, history and literature of the Karelian Research Centre of the Russian Academy of Sciences (Russia).

Крижановский Андрей Анатольевич

Институт прикладных математических исследований Карельского научного центра РАН (Россия).

Krizhanovsky Andrew

Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences (Russia).

E-mail: andrew.krizhanovsky@gmail.com