Словарь и корпус текстов вепсского языка в виде компьютерной онлайн-системы (технический отчёт за 2013 г.)

Veps dictionary and text corpus as an online computer system (technical report 2013)

А. А. Крижановский
Andrew.Krizhanovsky at Gmail

Институт прикладных математических исследований КарНЦ РАН,
Петрозаводск, Россия

Ключевые слова: лексикография, машиночитаемый словарь, корпусная лингвистика

Аннотация. Структура реляционной базы данных корпуса текстов и словаря вепсского языка модифицирована. Новая база данных обеспечивает хранение и поиск информации, связанной с информантами (имя информанта, место рождения, место записи и т.д.). 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).

Введение. Большой трудностью для тех, кто изучает языки малых народностей, является недостаток текстовых материалов в бумажном виде и, тем более, на электронных носителях. Лингвистам, лексикографам и заинтересованным читателям для полноценного изучения языка необходим свободный доступ к текстам на этих языках.

Развитие компьютерных технологий позволяет удовлетворить такую потребность и обеспечить удобный доступ к корпусу текстов и словарю вепсского языка, представленных в сети Интернет.

В настоящей работе описаны ключевые особенности разрабатываемой компьютерной онлайн-системы, включающей словарь и корпус текстов вепсского языка.

Изменение функциональности и интерфейса сайта. В ходе работ за 2013 год были выполнены следующих задачи:

1. База данных (БД) расширена новыми таблицами и интерфейс сайта изменён с тем, чтобы обеспечить работу с метаданными текстов, а именно добавлена возможность редактировать и привязывать к текстам:
   - 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).
2. 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.
3. Добавлена возможность редактировать, добавлять, удалять тексты и переводы.
4. Для сохранения целостности базы данных введён ряд правил, препятствующих вводу противоречивой информации, а именно:
   a. Нельзя удалить "Информанта", пока есть текст, к которому привязан данный информант.
   b. нельзя удалить "Имя" информанта, пока есть информант, который привязан к этому имени.
   c. Нельзя удалить "Географическое местоположение", пока есть информант, который привязан к этому местоположению.

d. Нельзя удалить "Название деревни", пока в базе данных есть запись "Географическое местоположение" с этой деревней.
e. Нельзя удалить запись "Район, область, республику", пока есть запись "географическое местоположение" с этим районом.

5. Добавлены страницы, позволяющие увидеть списки новых или изменённых текстов, новых лемм и словоформ.
6. Расширены возможности поиска, а именно: на страницу "Тексты" добавлена возможность выбора списка текстов (i) по корпусу, (ii) по задаваемым свойствам (параметрам) текста.

Архитектура базы данных корпуса текстов и словаря. База данных корпуса текстов и словаря реализована в единой реляционной базе данных. База данных содержит как таблицы, относящиеся к словарю, так и таблицы, связанные с текстами. Все эти таблицы взаимосвязаны и представляют единое целое. В ходе работы над более полным представлением о текстах была расширена база данных для хранения метаданных текста, а именно: были добавлены следующие таблицы, описывающие информантов (рис. 1):

- *informant* — ключевая таблица об информантах, которые наговорили какой-либо текст, который, в свою очередь, был записан научным сотрудником. Известны дата и место записи. В таблице *text* в поле *informant_id* указан номер записи в таблицы *informant*, что позволяет связать информанта и текст в БД.
- *informant_name* — имя и фамилия информанта;
- *recorder_name* — имя и фамилия сотрудника, записавшего текст информанта;
- *informant_place* — географическое местоположение (место рождения информанта, либо место записи); эта таблица связывает воедино таблицы об информанте (*informant*) и

таблицы с географической информацией (*informant_village*, *informant_region*);

- *informant_village* — название деревни / села;
- *informant_region* — название района / области / республики.

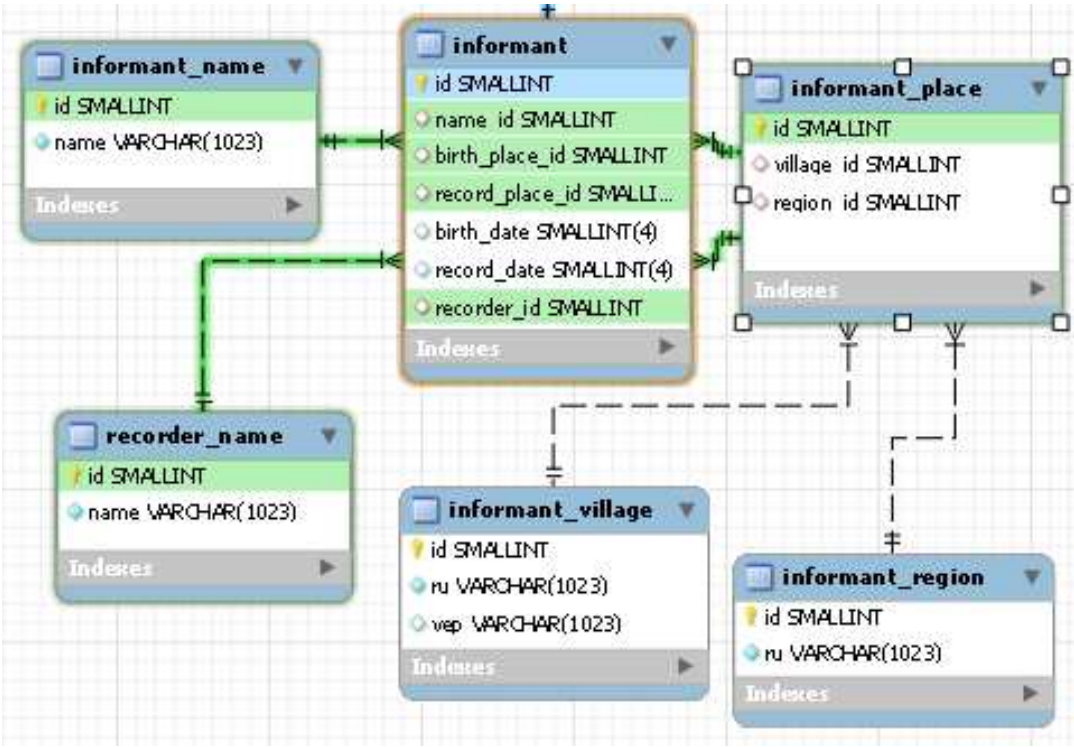

Рис. 1. Группа таблиц в базе данных корпуса и словаря, содержащих информацию об информантах: имя информанта, место рождения, место записи, год рождения, год записи.

Введено понятие корпуса, каждый текст теперь относится к одному из заранее заданных корпусов. На странице «тексты» можно выбрать все тексты одного или нескольких корпусов, возможна фильтрация списка текстов по следующим параметрам: диалект, говор, причитание.

Интерфейс программирования приложений. На языке программирования PHP разработана объектно-ориентированная библиотека, как часть компьютерной онлайн-системы, включающей словарь и корпус вепсского языка. В этой библиотеке для каждой из таблиц, представленной в предыдущем разделе, создан интерфейсный класс типа POPO (Plain Old PHP Object). Для повышения надёжности разрабатываемой компьютерной системы

начато создание Unit-тестов, обеспечивающих проверку правильной работы функций.

Задачи на 2014 год. В планы на будущий год входят следующие работы по изменению структуры базы данных и функциональности сайта, а именно:
работы по словарю:

1. Решение задачи разграничения омонимов и значений слов с помощью изменения структуры базы данных.
2. Расширение списка морфологических признаков (род, число, падеж, время) в описании словарной единицы. Для этого потребуется расширить и изменить структуру базы данных, изменить программный код.

работы по корпусу текстов:

1. Реализация поиска по словарю с учётом новой структуры базы данных.
2. Расширение функциональности сайта с тем, чтобы редактор мог сохранить всю базу данных на свой компьютер по нажатию одной кнопки. Это обеспечить большую сохранность и защиту от сбоев системы.
3. Расширение функциональности поиска по текстам, добавление возможности ограничивать область поиска одним корпусом.

Благодарности. Работа выполнена при финансовой поддержке Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» 2012-2014, направление 3 «Создание и развитие корпусных ресурсов по языкам народов России» (проект «Корпус вепсского языка: пополнение и развитие электронного ресурса», рук. проекта доктор филол. наук, зав. сектором языкознания ИЯЛИ КарНЦ РАН Зайцева Н.Г.).

Публикации

- Смирнов А.В., Круглов В.М., Крижановский А. А., Крижановская Н.Б. Автоматическое извлечение словарных

помет из Русского Викисловаря // Математическое моделирование и информационные технологии (Труды Карельского научного центра РАН), 2014 (принято к печати), *указана поддержка Программой фундаментальных исследований Президиума РАН «Корпусная лингвистика»*

- Браславский П.И., Мухин М.Ю., Ляшевская О.Н., Бонч-Осмоловская А.А., Крижановский А.А., Егоров П.В. YARN: начало // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Электронная публикация: http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BraslavskiyP_YARN.pdf