

**Словарь и корпус текстов вепсского языка в виде компьютерной
онлайн-системы (технический отчёт)**

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

А. А. Крижановский
Andrew.Krizhanovsky at Gmail

Институт прикладных математических исследований КарНЦ РАН,
Петрозаводск, Россия

Ключевые слова: лексикография, машиночитаемый словарь, корпусная лингвистика

Аннотация. Представлена структура реляционной базы данных корпуса текстов и словаря вепсского языка. Построены списки многозначных вепсских слов.

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.

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). Несмотря на это деление таблицы взаимосвязаны и представляют единое целое (рис. 2):

1] таблицы словаря:

- *lemma* – таблица лемм. Особенность таблицы в том, что лемма (т.е. каноническая, основная форма слова) не является уникальной в данной таблице, число лемм равно числу значений данного слова. Например слово “leta” имеет два значения «летать» (уточнить: «строить» - это третье значение?) и «поднимать, поднять». В таблице *lemma* будет две одинаковых записи “leta”, однако при редактировании словаря пользователь будет видеть две записи “leta” и “leta [2]”.
- *wordform* – таблица словоформ. Словоформа привязана к лемме с помощью поля *lemma_id*.
- *translation_lemma* – перевод какого-либо значения с главного языка на один из иностранных. «Главным» языком в данном проекте является вепсский, иностранный – это русский или английский. Таблица содержит следующие поля:
 - *lemma* – текст перевода на один из иностранных языков;
 - *lang_id* – идентификатор языка (отсылка к таблице *lang*), указывающий на каком иностранном языке записан перевод в поле “lemma”;
- *translation* – таблица, связывающая лемму с переводом, т.е. связывает таблицы *lemma* и *translation_lemma*.

2] таблицы корпуса текстов:

- *text* – ключевая таблица, описывающая и содержащая текст на каком-либо языке, содержит следующие поля:
 - *text* – сам текст;

- *lang_id* – идентификатор языка, указывающий на каком языке записан текст в поле “*text*”, аутентичный (в данном случае – вепсский) или русский (в будущем, может быть, английский) для текстов с параллельным переводом с вепсского языка;
 - *title* – название текста;
 - *source_id* – идентификатор источника текста, описанного в таблице *source*;
 - *informant* – данные об информанте со слов которого был записан данный текст.
- *text_pair* – таблица, связывающая тексты с параллельным переводом, например, первый текст – источник на вепсском и второй текст – перевод первого на русский язык.
 - *label* – описание характеристик текста:
 - вид причитаний (свадебные или похоронные и поминальные причитания);
 - диалект (северновепсский, средневепсский или южновепсский);
 - говор (восточный или западный);
 - *text_label* – связывает две таблицы: *text* и *label*, т.е. для каждого текста может быть указаны: диалект, говор, вид причитаний. При необходимости в таблицу *label* могут быть добавлены новые характеристики для более подробной классификации текстов.
 - *source* – характеристики источника текста. Таблица содержит следующие поля:
 - *title* – название книги, содержащий данный текст;
 - *author* – имя автора книги;
 - *year* – год издания книги;
 - *ieeh_archive_number1* – первый номер в архиве;
 - *ieeh_archive_number2* – второй номер в архиве;
 - *page_from* – номер первой страницы в книге, с которой начинается данный текст;

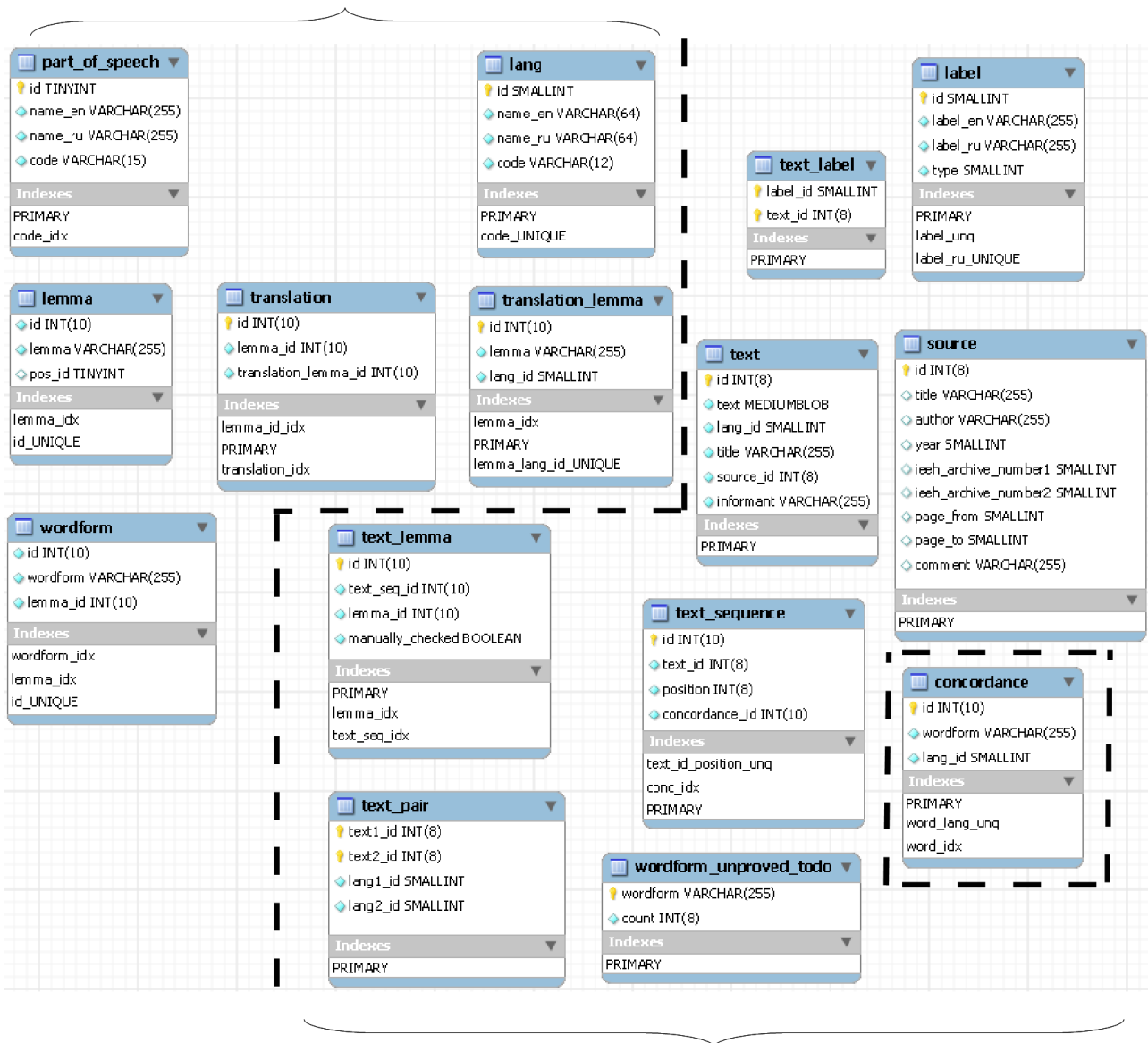
- *page_to* – номер последней страницы книги, содержащей данный текст; (в следующей версии базы данных необходимо будет либо разбить таблицу *source* на две таблицы, либо отказаться от полей *page_from* и *page_to*, т.к. сейчас дублируется информация в полях *title*, *author*, *year* для разных страниц одной и той же книги; в идеале должна быть одна запись в таблице *source* для одной книги);
- *comment* – комментарий к источнику. Эта таблица была создана в ходе реструктуризации базы данных. Сейчас все данные об источнике автоматически записаны в данное поле *comment*. В дальнейшем необходимо будет вручную разнести данные по соответствующим полям (автор, год и т.д.)
- *text_lemma* – таблица, связывающая леммы и текст. Для заданной леммы можно найти тексты, которые её содержат. Для этого необходимо последовательно просмотреть таблицы: *lemma* – *text_lemma* – *text_sequence* – *text*.
- *text_sequence* – содержит упорядоченный список слов для каждого текста с указанием словопозиции в тексте. Обеспечивает быстрый поиск по текстам и связывает три таблицы: *text*, *text_lemma* и *concordance*.
- *concordance* – содержит список всех словоформ, найденных в текстах, с указанием языковой принадлежности. Данные таблицы позволяют выполнять *быстрый поиск* в текстах даже по тем словам, которые ещё не добавлены в словарь. Эта таблица относится к «двум мирам»: и к корпусу, и к словарям (рис. 1), т.к. таблица построена по текстам корпуса, однако данные будут использованы для улучшения словаря.
- *wordform_unproved_todo* – резервная таблица, где будет храниться автоматически созданный список словоформ для последующей проверки этого списка вручную (например, список

наиболее часто встречающихся слов в корпусе, отсутствующих в словаре);

3] В особую группу можно выделить вспомогательные таблицы, содержащие «лексикографические константы», т.е. данные в этих таблицах изменяются очень редко, в результате согласованной работы инженера-программиста и лингвиста:

- *lang* – таблица со списком языков, сейчас это вепсский, русский и английский. В таблице указаны:
 - уникальный идентификатор (поле “*lang_id*”);
 - название языка на русском;
 - название языка на английском;
 - буквенный код языка в соответствии со стандартом ISO 639.
- *part_of_speech* – список частей речи (сущ., прил., и т.д.), всего на данный момент в базе данных – 11 частей речи. Запись в таблице содержит уникальный идентификатор, название части речи на английском, на русском и текстовое сокращение для обозначений этой части речи.

Словарь



Корпус текстов

Рис. 1. Две группы таблиц в базе данных корпуса и словаря, таблица *concordance* на особом положении (см. описание таблиц)

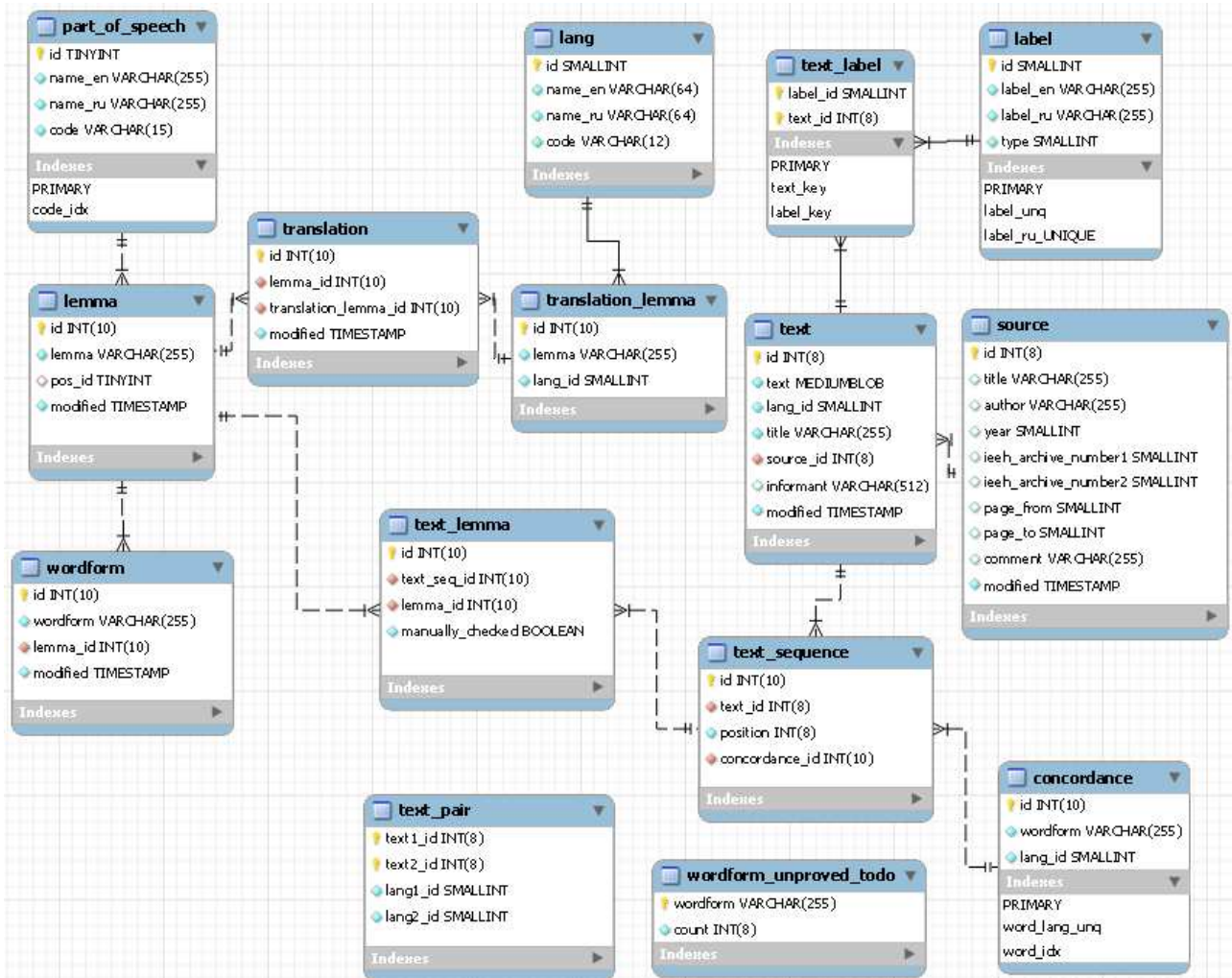


Рис. 2. Таблицы и отношения в базе данных корпуса и словаря

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

Эксперименты. В ходе реструктуризации базы данных были получены следующие интересные списки многозначных вепских слов для последующей работы с ними лексикографов. Необходима ручная проверка правильности преобразования словарных статей для многозначных слов (см. описание таблицы лемм *lemma* выше, т.е. нужно проверить, что при редактировании словаря пользователь будет видеть, например, две записи “leta” и “leta [2]”, если слово имеет два значения).

1. Список из 13 значений вепсских слов, содержащих знак «точка с запятой» в тексте перевода значения на русский язык, и имеющих *перевод на английский язык*.

http://vepsian.krc.karelia.ru/commons/wordlist/2012/13_meanings_column_in_ru_with_en.html

2. Список из 55 значений вепсских слов с *переводом на русский язык*, содержащих знак «точка с запятой» в тексте перевода.

http://vepsian.krc.karelia.ru/commons/wordlist/2012/55_meanings_column_in_ru.html

3. Список из 376 значений вепсских слов, содержащих знак «запятая» в переводе на русский язык.

http://vepsian.krc.karelia.ru/commons/wordlist/2012/376_meanings_comma_in_ru.html