КОРПУС КЕТСКИХ И ЭВЕНКИЙСКИХ ТЕКСТОВ

Камаева Елизавета Михайловна (deer95@ya.ru)

Российский государственный гуманитарный университет, Москва, Россия

Abstract

Corpora of Ket and Evenki Texts are electronic texts collected during fieldwork with speakers of the aforesaid languages, with possibility of searching. The corpora are sited here: http://corket.ru/. The texts are provided with vast linguistic information. They are referred to different genres and introduce data of the recent 20 years.

During the report there will be introduced a new type of searching corpus adopted for agglutinative languages, and possibilities of advanced user interface.

Creation corpora for these two languages which are, especially Ket, in a rather disastrous condition, can become a unique instrument in work of researchers of the Northern languages and lighten process of analyzing collected data.

KEYWORDS: Ket, Evenki, Siberian languages, corpus linguistics, morphemic parsing.

Краткая справка о языках

Кетский язык – единственный оставшийся в живых язык из енисейской семьи. Он распространен в верховьях бассейна реки Енисей. У него три диалекта, самый распространенный из которых, южный, считается литературным. На данный момент число носителей, бегло говорящих по-кетски, не превышает 50 человек [Казакевич 2006].

Эвенкийский язык – язык тунгусо-маньчжурской группы алтайской семьи. Он распространен на территории Восточной Сибири от Енисея до Сахалина. Он делится на три наречия, диалект одного из которых, южного, а именно - подкаменно-тунгусский диалект, считается литературным. Общее число говорящих – 13 800 человек [Эвенкийский язык 2016]. На эвенкийском издается газета и учебная литература.

Материал

Выбор языков обусловлен тем, что на данный момент размеченные тексты в электронном формате существуют только на кетском и эвенкийском.

Для наполнения корпуса были взяты обработанные, морфологически размеченные кетские и эвенкийские тексты, собранные в течение 1937 – 2014 гг. Данные тексты любезно предоставлены коллективом лаборатории автоматизированных лексикографических систем НИВЦ МГУ.

Материалы размещены на сайте http://minlang.srcc.msu.ru/ , созданном в рамках проекта «Создание Интернет-ресурса “Малые языки Сибири: наше культурное наследие” (на материале языков бассейна Среднего Енисея и Среднего и Верхнего Таза)» в 2012 – 2014 гг. при поддержке Российского гуманитарного научного фонда. Руководитель проекта – Казакевич Ольга Анатольевна (ЛАЛС НИВЦ МГУ) [SiberianLang. О проекте].

Массив данных включает в себя тексты различных жанров.

Жанр	Кетский	Эвенкийский
мифологические сказки	+	+
песни	+	
случаи из жизни	+	+
внешнее описание камланий	+	+
волшебные сказки		+
истории жизни	+	+
бытовые сказки		+
охотничьи истории		+
истории жизни шаманов		+
сказки о животных	+	+
легенды		+
былички		+
шаманские тексты	+	+

Временной разброс сбора текстов покрывает годы от 1937 до 2014 [SiberianLang. Тексты].

Кетский: ≈1937-1938 гг. (3 текста), 2004 г. (3 текста), 2005 г. (6 текстов), 2006 г. (1 текст), 2009 г. (2 текста). В сумме – 15 текстов.

Эвенкийский: 1952 г. (5 текстов), 2005 г. (3 текста), 2006 г. (2 текста), 2007 г. (27 текстов), 2008 г. (12 текстов), 2009 г. (1 текст), 2010 г. (2 текста), 2011 г. (6 текстов), 2014 г. (2 текста). В сумме – 60 текстов.

Разница в объеме текстов на этих двух языках объясняется ограниченностью проанализированного материала в рамках проекта «Создание Интернет-ресурса “Малые языки Сибири: наше культурное наследие” (на материале языков бассейна Среднего Енисея и Среднего и Верхнего Таза)».

Кетская часть корпуса состоит из 767 предложений и имеет объем 3 450 словоупотреблений; эвенкийская – состоит из 4 618 предложений, содержащих 20 018 словоупотреблений. В сумме – 23 469 словоупотреблений в 5 384 предложениях. Средняя длина предложения – 6 словоформ.

Для кетского языка представлены все три существующих диалекта: северный, центральный и южный. Для эвенкийского – северное и южное наречие, включая подкаменно-тунгусский диалект (свистящие (секающие) говоры южной сибилянтной группы южного наречия), который послужил основой для создания литературного эвенкийского языка [Василевич 1948] [Bulatova, Grenoble 1999: 3]. Для кетского базой литературного языка является южный диалект [Georg 2007: 31].

Техническая реализация

Вся работа по созданию корпуса, выходящая за рамки разметки текстов, выполнена автором доклада.

При разработке корпуса использовались материалы о создании других корпусов, например, [Arkhangelskiy 2016]. Поскольку корпус адыгейского языка не находится в открытом доступе, то были проанализированы рабочие, находящиеся в открытом доступе корпуса, созданные для агглютинативных языков. Например, KORP [KORP by META-SHARE] для финского языка; портал Web-corpora [Лингвистические корпуса и сервисы] для 14 языков, находящихся преимущественно в России; портал БалтоСлав [БалтоСлав: языковые инструменты] для абхазского и аварского языков.

Использованные файлы

Всего было использовано 15 файлов, содержащих тексты на кетском языке, и 60 файлов, содержащих тексты на эвенкийском. Все они записаны в формате *.eaf и имеют расширение XML.

Файлы могут быть отредактированы с помощью программы Elan. В интерфейсе данной программы информация представима в виде слоев. Для файлов, использованных при создании корпуса, их шесть, а именно:

1. слой с элементом-предложением иностранном языке в кириллической записи;
2. слой с элементом-переводом предложения на русский язык;
3. слой с элементами, представляющими собой слова из предложения в фонетической записи;
4. слой с элементами, являющимися фрагментами исходного предложения, разбитого на части по знаку пробела;

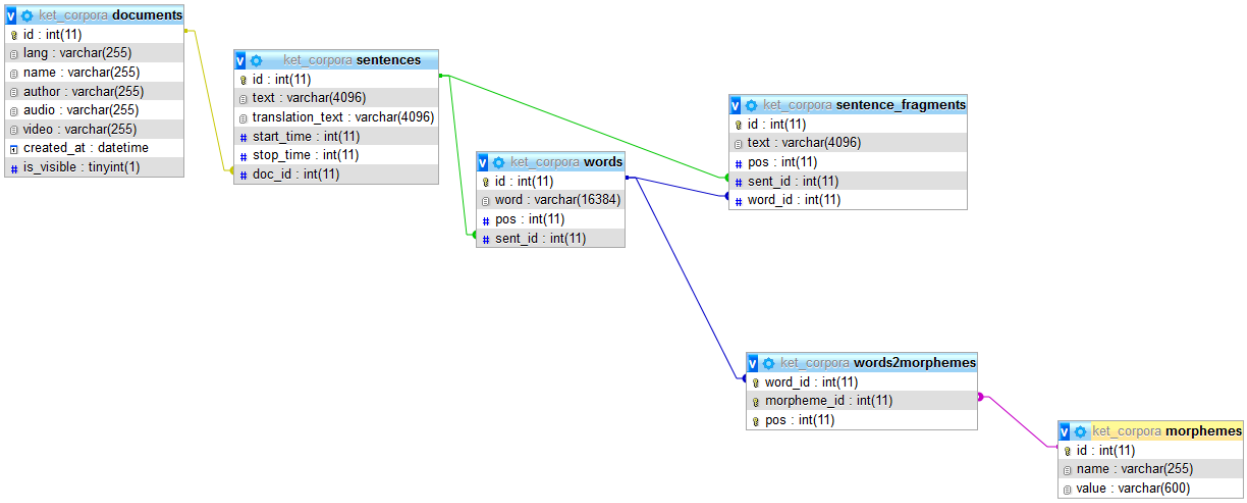
5. слой с элементами, представленный морфемами в фонологической записи;
6. слой с элементами, представляющими собой глоссы морфем.

Инструменты реализации

Для написания алгоритмов извлечения информации из XML-файлов, загрузки их в базу данных и обработки запросов использовался язык Python 3.5.1.

Для синхронизации объектов Python и записей реляционной базы данных была использована программная библиотека SQLAlchemy.

В качестве реляционной СУБД была выбрана MySQL. В качестве программного каркаса использован Pyramid. Для создания сайта использована библиотека JavaScript jQuery и набор инструментов Bootstrap. Все инструменты разработки бесплатны и открыты.



Блок-схема 1. ER-модель СУБД в MySQL

Таблица documents:
id: уникальный код документа;
lang: язык документа;
name: название документа;
author: автор документа;
audio: аудио, сопровождающее текст (пока отсутствует);
video: видео, сопровождающее текст (пока отсутствует);
createdat: дата создания документа.

Таблица sentences
id: уникальный код предложения;
text: само предложение, извлеченное из слоя original;
translation_text: перевод предложения, извлеченный из слоя *rus*;
start_time: время начала предложения в медиафайле;
stop_time: время конца предложения в медиафайле;
doc_id: код документа, в котором присутствует данное предложение.

Таблица sentence_fragments
id: уникальный код фрагмента предложения ;
text: само слово, полученное в результате разбиения предложения на куски по знаку пробела. Извлечен из слоя *sent-fragments*;
pos: позиция слова в предложении;
sent_id: код предложения, в котором присутствует данное слово;
word_id: код слова, соотносящегося с данным фрагментом.

Таблица words2morphemes
word_id: код слова, в котором присутствует морфема;
morpheme_id: код морфемы;
pos: позиция морфемы в слове.

Таблица morphemes
id: уникальный номер морфемы;
name: фонетическое выражение морфемы, извлеченное из слоя *fon*;
value: означаемое морфемы, извлеченное из слоя *gl*.

Таблица words
id: уникальный код слова;
word: само слово, извлеченное из слоя *fonWord*;
pos: позиция слова в предложении (начиная с нуля);
sent_id: код предложения, в котором присутствует данное слово.

При поиске по морфеме выполняется перебор значений в таблице *morphemes* в поле *name*. Далее алгоритм определяет номер предложения, используя уникальные id морфем, слов и предложений, и выводит текст предложения. При поиске по значению морфемы выполняется та же последовательность действий, за исключением одного: перебор значений производится в таблице *morphemes* поле *value*. Когда выполняется операция поиска по переводу, машина просматривает все элементы таблицы *sentences* поля *translation_text* и ищет введенное пользователем слово.

Пользовательский интерфейс

Пользователь работает в браузере. Вверху находятся элементы для выбора языка поиска: кетский или эвенкийский. На ней присутствует два поля для ввода: для поиска фонетическому представлению и глоссам и для полнотекстового поиска по переводу..

Сайт корпуса находится по адресу: http://corket.ru/

Синтаксис запроса

При поиске по морфемам у пользователя имеется в наличии широкий спектр возможностей. Во-первых, можно задавать последовательность морфем в слове. Во-вторых, можно искать морфему как по ее фонетической оболочке (означающему), так и по ее означаемому. Синтаксис запроса таков:

(1) поиск по фонетической оболочке – “[PHON]”;
(2) поиск по означаемому – “MORPH”;
(3) поиск последовательности морфем “ELT > ELT”,

где PHON – фонетическое выражение некой морфемы, MORPH – ее значение, ELT – либо означаемое, либо означающее морфемы.

Любая одиночная морфема обозначается знаком “?”, любая последовательность нескольких морфем – знаком “*”.

Например, запрос “* > PST > [tn]” в кетском языке выдает все формы прошедшего времени глагола «идти», присутствующие в базе.

Подобная организация поискового запроса в случае агглютинативных языков (к которым как раз и относятся кетский и эвенкийский) исключительно важна, поскольку морфемам этих языков свойственна омонимия, и лишь позиция в словоформе различает их значения. Ср. примеры из кетского подкорпуса:

*(1) ət qade d-daq-o-**l**-ej-i-n*

1PL потом S1.1-смеяться-T.PST-**PST**-LV-EP-PL
‘Мы потом смеялись’

(2) *k-a-l-do-n* *eq-**l**-aq-n* *tin*
DET-EP-IMPER-смотреть-S1.PL MOD-**IMPER**-LV-S1.PL это
‘Посмотрите, послушайте это’

В случае уникальной, ничему не омонимичной морфемы или при поиске по значению такой проблемы не встает. Рассмотрим один из результатов запроса “* > DATLOC” по эвенкийскому подкорпусу:

(3) *həgdi-gi-tin* *bi-šo* *śipkā-kā-r-**dū*** *tarə*
большой-SUPERL-PS3PL быть-PANT птица-ATTEN-PL-**DATLOC**
тот.ACC
‘Был самым главным у птиц’

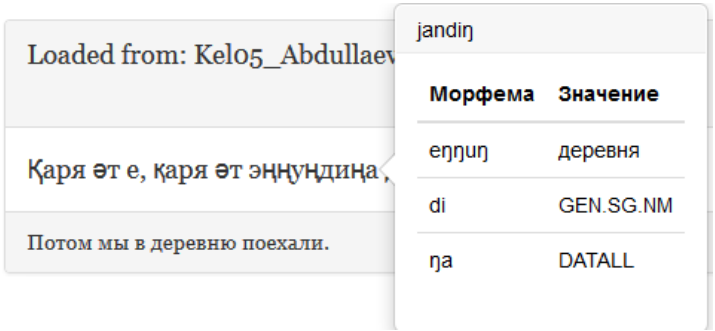
Синтаксический поиск в корпусе реализован через морфемную разметку корпуса. Поскольку аффиксы по большей части могут маркировать только строго определенные части речи, то достаточно ввести в поле поиска морфему или ее значение, специфичные для некой части речи, чтобы найти интересующие лексемы. Например, аффиксы времени или модификатора свойственны глаголу, а падежа – именным частям речи и пред-послелогам.

Важно отметить, что при поиске по фонетическому представлению возможно найти морфемы, представленные лишь в одном диалекте языка. Для поиска по всем диалектам предпочтительнее использовать семантические значения морфем.

Возможности интерфейса

При выводе пользователю демонстрируется 15 предложений. При нажатии кнопки далее открываются следующие 15.

При предварительном просмотре пользователю показывается запись текста на иностранном языке в кириллической записи и перевод под ним. При наведении мышки на иностранное слово выводится таблица с заголовком-записью словоформы в МФА (Международном фонетическом алфавите) и таблицей со столбцами Морфема-Значение.



При нажатии кнопки «Полная информация» выводится отглоссированное предложение, которое поддается копированию. Первая строка – фонетическая запись, вторая – фонематическая запись, третья – глоссы к каждому слову, соединенные дефисом.

Полная информация

ətə	kəlləgbəs'	ʌγa	d'es'kalaan
ət	kellog-bes	əga	d-eska-l-aq-n
1PL	елогуй-PROL	сюда	S1.1-вверх.по.течению-PST-ходить-S1.PL

При нажатии на любое иностранное слово в предпросмотре во всплывающем выводится текст с подчеркнутым предложением, слово которого было нажато. Внизу появляется отглосссированная версия этого предложения и перевод. При нажатии на любое предложение текста во всплывающем окне внизу показывается его отглоссированная версия и перевод.

Туре укон. А баат хы уль тие десьтэйт. Деэсяғот. Қаря қиб баат хай окон хай. Қаря окон ле'сдиңа. Уська бинь ага дә= дътонақ. Окон. Э әт қаря хай усь палаткадиңа диимэсин. « Къма», – қаря обдаңа нима. Әт ситат бәнь тольдамин. Э хънынсин-то тольдамин, Андрей, Максим и я. И әтә около әтн богдот кокдиңтен конхънән. Қае қоноксяң: « Те, къма, е сукдиңтен, е сукдиңтен эңңуңдаңтен». Қаря әт е, қаря әт эңңуңдиңа даңкон. Сюга туде диимэсин. Сюк диимэсин. А қаря э... Туруханскдиңаль ириң диимэсь, что у нас Сережка родился. Вот то'н туре. Это было четырнадцатого июля, в семьдесят четвертом году.

Полная информация

s'uγa	tidə	diiməs'ən
suka	tude	d-ik-i-n-bes-i-n
назад	этот	S1.1-сюда-T.PST-PST-LV-EP-S1.PL

Перевод

Приехали домой.

Общая оценка и перспективы

В качестве положительных черт данной системы нужно указать, что скорость его работы в разы больше, чем в программе Elan, располагающей примерно такими же инструментами поиска. Также корпус предоставляет удобную запись данных для написания статей и научных работ. Недостатком данного корпуса является малый объем данных. Однако потенциально его можно легко расширить, добавив в базу данных новые размеченные тексты. Количество языков также легко может быть увеличено без потери производительности. В будущем возможно сопровождение текстов медиафайлами: видео- и аудиозаписями.

Литература

Arkhangelskiy T. A. et al. Developing a polysynthetic language corpus: problems and solutions //Компьютерная лингвистика и интеллектуальные технологии. – Dialogue, 2016. – №. 15 (22). – С. 38-47.

Bulatova N., Grenoble L. Evenki. / N.Bulatova, L. Grenoble. – Lincom Europa, 1999. – 65.

Georg S. A Descriptive Grammar of Ket (Yenisei-Ostyak): Part 1: Introduction, Phonology and Morphology / S.Georg. – Global Oriental, 2007. – 462.

KORP by META-SHARE. URL: https://korp.csc.fi/#?prequery_within=sentence&cqp=[]&corpus=ftb3europarl,ftb3_jrcacquis,ftb2,reittidemo&lang=en

SiberialLang. Тексты. URL: http://minlang.srcc.msu.ru/ru/textspage

SiberianLang. О проекте. URL: http://minlang.srcc.msu.ru/ru/o-proekte

Василевич Г.М. Очерки по диалектологии эвенкийского (тунгусского) языка / Г.М. Василевич. – Ленинград, 1948. – 343.

БалтоСлав. Языковые инструменты. URL: http://baltoslav.eu/?mova=ru

Казакевич О.А. Кетский язык // Язык и общество. Энциклопедия / Отв. ред. В.Ю. Михальченко М.: Азбуковник, 2016. С. 2012-2017.

Лингвистические корпуса и сервисы. URL: http://web-corpora.net/

Сводный список глосс. URL: https://yadi.sk/i/Lpi1Hg5E3GfTRg

Эвенкийский язык // Язык и общество. Энциклопедия / Отв. ред. В.Ю. Михальченко М.: Азбуковник, 2016. С/ 576-583/.