Электронный корпус мансийских текстов

Проблемы создания и перспективы использования

Дарья Жорник, Фёдор Сизов

МГУ имени М. В. Ломоносова, Институт Языкознания РАН

Введение

Мансийский язык

- Мансийский язык < финно-угорские < уральские;
- Перепись 2010г.: 938 носителей, ХМАО и север Свердловской области;
- 4 диалектные группы, однако на сегодняшний день в живых остаются лишь некоторые северные говоры;
- Проект по документации верхнелозьвинского диалекта мансийского языка (`www.digital-mansi.com`), исследование поддержано грантом РФФИ№18-012-00833А;
- Экспедиции в дд. Ушма и Тресколье Ивдельского района Свердловской области в 2017-2018гг. (Д. Жорник, С. Покровская, Л. Козлов);
- Создание мультимедийного подкорпуса текстов на верхнелозьвинском диалекте в рамках многодиалектного корпуса.

Существующие ресурсы

- По сравнению с другими уральскими языками, мансийские тексты в электронном виде представлены мало;
- Сайты газеты ”Лӯима Сэ̄рипос” `www.khanty-yasang.ru/luima-seripos` и Обско-угорского института прикладных исследований и разработок `www.ouipiir.ru`: некоторые мансийские тексты, без перевода и глоссирования;
- 272 частично аннотированных мансийских текста в базе данных ObUgric Database (Мюнхен), о которой см. [Schön, Wisiorek 2016]: `www.babel.gwi.uni-muenchen.de`;
- Проект “Languages under the influence”, представленный в работе [Simon, Mus 2017]; описанный корпус недоступен для использования.

Проблемы существующих ресурсов

- Малое количество текстов: в основном, тексты из газет советского и пост-советского периодов;
- Аудиофайлы отсутствуют во всех существующих базах мансийских текстов;
- Данные современных мансийских диалектов не представлены;
- При глоссировании не учитываются деривационные показатели и аналитические конструкции, отделяются только словоизменительные морфемы;
- Не решается проблема диалектной вариативности (фонетической, лексической, грамматической).

Корпус мансийских текстов

- В рамках нашего проекта разрабатывается корпус мансийского языка `www.digital-mansi.com/corpus`, см. [Zhornik, Sizov 2018];
- Многодиалектность: тексты как на живых, так и на уже мёртвых мансийских диалектах;
- Тексты XIX и первой половины XX века, записанные венгерскими, финскими и русскими исследователями (А. Регули, А. Альквистом, В. Н. Чернецовым и др.);
- Современные литературные тексты: печатные издания, опубликованные в советский и постсоветский период;
- Тексты из мансийской газеты "Лӯима Сэ̄рипос" `http://www.khanty-yasang.ru/luima-seripos`;
- Подкорпус верхнелозьвинских полевых текстов с соответствующими аудиофайлами (5ч текстов, в обработке).

Структура корпуса

Подкорпусы

- Южный мансийский;
- Восточный мансийский;
- Западный мансийский;
- Северный мансийский;
- Литературный мансийский (печатная литература, периодические издания, переводы Библии);
- Верхнелозьвинский мансийский (с аудио);
- Сосьвинский мансийский (современный).

Подкорпусы

- Для аннотирования текстов применяется морфологический движок AmpEngine;
- Каждый подкорпус объединяет на основе метаданных тексты, использующие систему записи, отличную от стандартного сета (литературный мансийский > периодические издания);
- Пример: Северный мансийский – *sus*, западный мансийский – *šuš*, восточный мансийский – *šonš*, южный мансийский – *šoš*;
- Система записи — Σ, $\Sigma^* = \bigcup_{i=0}^{\infty} \Sigma^i$, $\Sigma^+ = \Sigma^* \backslash \{\varepsilon\}$.

Обработка вариативности

- База данных для стемминга на основе [Chernetsov, Chernetsova 1936], [Balandin, Vakhrusheva 1958], [Rombandeeva, Kuzakova 1982], [Munkacsi, Kalman 1986], [Rombandeeva 2005], [Kannisto 2014], [Bakhtiyarova, Dinislamova 2016];
- Словарные статьи и альтернативные формы записи кластеризуются по признаку диалектной принадлежности и используемого алфавита;
- Оптимизация поиска внутри базы данных и стемминга.

Обработка вариативности

- При линейной сегментации токен представляется в виде суффиксного дерева;
- Токен — t, собственный префикс для t существует при условии, что $wv = t, |v| \neq 0$; суффикс – при $vw = t, |v| \neq 0$;
- АЕ-контейнер содержит данные о морфемах, которые представляются при поиске в t как $w, t = wv \vee t = vw$ (представляется в виде правого или левого ветвления);
- По мере построения деревьев разбора избираются только те последовательности подстрок, которые соответствуют морфологическим правилам подкорпуса.

Система данных корпуса

Предварительная обработка текстов

- Сканирование печатных текстов;
- Распознавание отсканированных текстов с помощью Tesseract OCR Engine (с модулем для поддержки манси);
- Исправление ошибок распознавания;
- В случае с аудиофайлами производится их разметка в программе ELAN: записи транскрибируются и сегментируются на клаузы;
- Аудиофайлы для текстов верхнелозьвинского диалекта размечаются носителем диалекта Т. П. Бахтияровой, впоследствии участники проекта производят выверку.

Интеграция с корпусной платформой

- Для работы с размеченными данными используется корпусная платформа Tsakorpus;
- Аннотированные тексты с помощью присвоенных им метаданных распределяются по подкорпусам;
- Кластеризация по подкорпусам используется далее для обработки диалектной вариативности в поиске.

Итоги

- В ходе всех этих шагов мы имеем многочисленные подкорпусы с аннотированными текстами;
- В корпусе будут представлены разнообразные данные, и он хорошо справляется с обработкой вариативности;
- Временной разброс текстов в корпусе может достигать 170 лет, что позволит наблюдать развитие языка в диахронической перспективе.