Электронный корпус мансийских текстов: проблемы разработки и перспективы использования

Дарья Олеговна Жорник, МГУ имени М. В. Ломоносова
Фёдор Олегович Сизов, Институт языкознания РАН

В докладе мы представим электронный корпус литературных и диалектных текстов на мансийском языке (http://digital-mansi.com/corpus), а также обсудим некоторые перспективы его использования. Мансийский язык, как и родственный ему хантыйский, демонстрирует высокий уровень диалектной вариативности (см. [Honti 1988]). Хотя на сегодняшний день живыми остаются лишь некоторые северные говоры манси, существует значительное количество текстов и на других диалектах, которые также могут и должны быть включены в корпус. Самые ранние из этих текстов датируются 1840-ми годами, таким образом, мы полагаем, что полный диалектный корпус будет способен отражать изменения в мансийском языке, происходившие на протяжении последних 170 лет. Это позволит использовать его для исследований диахронической эволюции и диалектной дивергенции мансийского языка.

Корпус включает многочисленные коллекции литературных и диалектных мансийских текстов, на базе которых могут быть созданы корпусные словари, распределенные по подкорпусам, каждый из которых соответствует определённой диалектной группе. Тексты XIX и первой половины XX века, представленные в корпусе, были записаны венгерскими, финскими и русскими исследователями (А. Регули, А. Альквистом, В. Н. Чернецовым и др.). Современные литературные тексты представляют собой печатные издания, опубликованные в советский и постсоветский период (они были оцифрованы посредством сканирования и распознавания при помощи программы Tesseract OCR). Другим важным источником литературных мансийских текстов является газета “Лӯима Сэ̄рипос” (http://www.khanty-yasang.ru/luima-seripos), выпускаемая с 1989 года до настоящего времени. Не менее важен подкорпус текстов, записанных Д. О. Жорник и С. В. Покровской в рамках полевых исследований слабо документированного верхнелозьвинского диалекта [Жорник, Покровская 2017] летом 2017 и зимой 2018 года. Эти тексты включаются в корпус вместе с соответствующими аудиофайлами.

При работе над корпусом мы в значительной степени опираемся на опыт предшественников. Особенно важной для нас является универсальная платформа UniParser, разработанная Т. А. Архангельским [Arkhangelskiy et al. 2012]: в нашем корпусе морфологический анализ производится универсальным парсером AmpEngine, который был разработан Ф. О. Сизовым с использованием некоторых моделей из UniParser. С целью анализа текстов на мансийском языке для AmpEngine был создан “мансийский модуль” на основе грамматики [Ромбандеева 1973]. Он поддерживает обработку текстов с высоким уровнем языковой вариативности — последнее, как известно, является одной из основных проблем морфологической разметки слабо стандартизованных языков (см., например, [Gerstenberger et al. 2017]). Транскрипционные системы, используемые в текстах, могут различаться в зависимости от их диалектной принадлежности. AmpEngine имеет интерфейс для свободного переключения между различными системами записи (как транскрипционными, так и орфографическими).

Представленный корпус может быть, в частности, использован для исследования различных грамматических категорий. К примеру, залог и система дифференцированного маркирования объекта в обско-угорских языках подвержены сильному влиянию со стороны информационной структуры и могут быть эффективно лишь на основе большого объёма текстов (ср. [Kulonen 1989], [Толдова, Сердобольская 2012]). В докладе будут представлены примеры таких корпусных исследований.

Список литературы

1. Жорник Д. О., Покровская С. В. *Документация верхнелозьвинского диалекта мансийского языка.* Постерный доклад на конференции “Малые языки в большой лингвистике”, МГУ им. Ломоносова, Москва, 2-3 ноября 2017.
2. Ромбандеева Е. И. *Мансийский (вогульский) язык.* Москва: Наука, 1973.
3. Толдова С. Ю., Сердобольская Н. В. *Дифференцированное маркирование прямого дополнения в финно-угорских языках* // Финно-угорские языки: фрагменты грамматического описания. Формальный и функциональный подходы. М.: Языки славянских культур, 2012, с. 59-142.

4. Arkhangelskiy, T.; Belyaev, O.; Vydrin, A. *The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform* // Proceedings of COLING 2012: Posters. Mumbai: The COLING 2012 Organizing Committee, 2012. Ch. 9. P. 83–91.
5. Gerstenberger C., Partanen N., Rießler M., Wilbur J. *Utilizing Language Technology in the Documentation of Endangered Uralic Languages* // The Northern European Journal of Language Technology 4, 2017, pp. 29-47.
6. Honti, L. *Die ob-ugrischen Sprachen – Die wogulische Sprache.* In The Uralic Languages: Description, History and Foreign Influences, Denis Sinor (ed.), 1988, 147–171.
7. Kulonen, U-M. *The passive in Ob-Ugrian.* SUS 203, Helsinki, 1989.