Фёдор Олегович Сизов

Институт языкознания РАН, Москва

Коротких Григорий Викторович

ТРОО «Союз КМНС Томской области», Северск

К созданию электронного портала селькупского языка

В данном докладе мы представим недавно открытый для свободного доступа первый электронный портал селькупского языка (http://selkup.org), включающий эффективные корпусные и лексикографические инструменты для исследования селькупского языка и содействия его ревитализации. Несмотря на то, что в настоящее время число носителей селькупского языка составляет (даже по официальным данным) всего около 1 000 человек, многие из них активно участвуют в деятельности по ревитализации и внесли существенный вклад в создание и пополнение корпуса.

Ключевым этапом развития проекта должна стать готовящаяся публикация большого объёма аннотированных текстов с удобной поисковой системой. Одной из важных проблем развития проекта является вариативность систем записи в текстах на селькупском языке, характерная для многих бесписьменных языков. Поддержка такой вариативности требует учитывать данные о происхождении текста -- не только его диалектную принадлежность, но также место, время и автора записи, и соответствующим образом дифференцировать их обработку. Это, в свою очередь, влечёт за собой необходимость создания и применения специального программного обеспечения (проблема достаточно широко обсуждалась в специальной литературе, ср., например, [Gerstenberger et al. 2017]).

При создании корпуса селькупского языка используются печатные материалы преимущественно советского периода, учебные пособия, а также тексты носителей, некоторые из которых публикуются по их личной инициативе. Эти данные проходят оптическое распознавание и последующую обработку.

Кроме того, в корпус включаются тексты носителей обского шёшкупского и нарымского диалектов, публикуемые в социальных сетях. Опыт обработки подобных текстов уже имеется для удмуртского корпуса, где их доля достигает 6% (см. http://web-corpora.net/UdmurtCorpus/search/index.php). В нынешних условиях существования селькупского языка и при небольшом числе носителей такой тип корпусных материалов представляется ценным.

Портал также включает в себя коллекцию электронных словарей, созданных на основе печатных публикаций [Быконя 2005], [Быконя, Ким, Купер 1994], [Григоровский 2007]. Вместе с тем, одной из целей проекта является интеграция корпусных и лексикографических данных. В частности, это должно способствовать улучшению поиска в словарях и корпусе, использованию естественных примеров употребления в словарных статьях, повышению качества морфологической разметки корпусных документов. Интеграция в случае селькупского языка осложняется тем, что в настоящее время не существует в достаточной мере универсальных решений в области корпусной лексикографии, которые были бы эффективны в условиях высокой диалектной вариативности и сравнительно небольшого объёма (100-150 тыс. токенов) текстов корпуса. В нашей работе мы опираемся на частично аналогичный опыт создания электронного словаря удмуртского языка [Arkhangelskij et al. 2015] и Открытого корпуса вепсского и карельского языков (ВепКар) (http://dictorpus.krc.karelia.ru/ru).

Список литературы

1. Arkhangelskij T. A., Idrisov R. I., Serdobolskaya N. V., Usacheva M. N. Designing a lexicographic database for an online dictionary // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции Диалог (Москва, 27 — 30 мая 2015 г.). — Vol. 14. — Изд-во РГГУ Москва, 2015.
2. Gerstenberger C., Partanen N., Rießler M., Wilbur J. (2017), Utilizing Language Technology in the Documentation of Endangered Uralic Languages, The Northern European Journal of Language Technology 4: pp. 29-47
3. Быконя В. В. Селькупско-русский диалектный словарь. Томский гос. педагогический университет. Томск, 2005.
4. Быконя В.В., Ким А.А., Купер Ш.Ц. Словарь селькупско-русский и русско-селькупский. Томск, 1994.
5. Григоровский Н.П. Южноселькупский словарь Н.П. Григоровского // Обраб. и изд. Евгения Хелимского. Hamburg: [Inst. für Finnougristik/ Uralistik der Univ. Hamburg], 2007.