Корпус кетских и эвенкийских текстов

В докладе будет рассмотрен проект "Корпус кетских и эвенкийских текстов", представляющий собой реляционную базу данных с возможностью поиска по текстам на обоих языках.

Для написания программного обеспечения использован язык Python 3. В качестве системы управления базами данных взят MySQL, для работы с СУБД использована программная библиотека SQLAlchemy. При создании веб-приложения, которое является связующим звеном между пользователем и базой данных, использован программный каркас Pyramid.

Тексты, использованные при создании, являются обработанным полевым материалом на разных говорах кетского и эвенкийского языков с морфологической аннотацией, взятым с сайта «Малые языки Сибири: наше культурное наследие», созданного и поддерживаемого ЛАЛС НИВЦ МГУ им. М.В. Ломоносова, общим объемом около 40000 словоупотреблений, в формате *.eaf.

Поиск осуществляется по фонетической оболочке морфемы, по ее означаемому и по словоформе, использованной в тексте. При запросе можно задать позицию морфемы в слове, а также искать несколько морфем в одной словоформе, указав их порядок следования друг за другом.

При выводе пользователю демонстрируется предложение с выделенным словом, в котором находится интересующая морфема или которое является искомым словом, и перевод. При наведении указателя мышки на каждое слова предложения показывается его морфемный состав. Нажатием кнопки можно открыть поле, в котором написан отглоссированный текст, представленный в виде таблицы с количеством столбцов, равным числу слов в предложении. В каждом из столбцов три строки: словоформа, записанная латиницей; морфемы, соединенные дефисом, и их значения, тоже через дефис. Также можно открыть текст, в котором присутствует это предложение, и в диалоговом окне посмотреть любое отглоссированное предложение оттуда.

Новизна данного проекта состоит в том, что на текущий момент это единственный корпус текстов с возможностью поиска по морфемам, их значению и позиции в словоформе. Также это единственный корпус кетских и эвенкийских текстов, что позволит исследователям языков Сибири быстро получать интересующие их языковые данные.

В качестве положительных черт данного проекта нужно указать, что скорость его работы в разы больше, чем в программе Elan, располагающей примерно такими же инструментами поиска. Более того, для получения интересующей информации вовсе не нужно устанавливать на устройство новое программное обеспечение - пользователь взаимодействует с базой данных через HTML-

страницу. Также корпус предоставляет очень удобную запись данных для написания статей и научных работ: отглоссированные предложения выводятся в записи, готовой для вставки в текст для иллюстрации языковых данных.

Недостатком данного корпуса является малый объем данных. Однако потенциально его можно легко расширить, добавив в базу данных новые размеченные тексты.

Литература:

User Guide for ELAN Linguistic Annotator. URL: http://www.mpi.nl/corpus/html/elan_ug/index.html
HeidiSQL. URL: http://www.heidisql.com/
MySQL. URL: https://www.mysql.com/
Pyramid. URL: https://trypyramid.com/
Siberian Lang. URL: http://siberian-lang.srcc.msu.ru/
SQLAlchemy. URL: http://www.sqlalchemy.org/