Количественные методы исследования сербского тимокского диалектного текста

Теодора Вукович
Университет Цюриха
teodora.vukovic2@uzh.ch

Дарья Владимировна Конёр
Институт лингвистических исследований РАН (Санкт-Петербург)
dsuetina@yandex.ru

Анастасия Леонидовна Макарова
Институт лингвистических исследований РАН (Санкт-Петербург)
abeatina@rambler.ru

Андрей Николаевич Соболев
Институт лингвистических исследований РАН, Санкт-Петербургский государственный университет (Санкт-Петербург)
sobolev@staff.uni-marburg.de

В докладе исследуется частотность дифференциальных фонологических и грамматических диалектных признаков (рефлексы прасл. *tj, *dj; слогообразующее *l*; гласный среднего ряда среднего подъема; прогрессивные палатализации; финальнослоговое *l*; аналитическое выражение косвенного объекта; опущение частицы конъюнктива при модальных глаголах и формах футура; ренарратив; редупликация прямого и косвенного объекта; постпозитивный артикль; формы личных и указательных местоимений и др.) в зависимости от темы наррации (автобиография, история семьи, традиционное хозяйство, календарная обрядность, обряды перехода, былички и др.) в речи Драгини Микич (1906 г.р.), носительницы тимокского говора сербского языка.

Исследование выполнено квантитативным методом на материале базы данных по диалектам Восточной Сербии и Западной Болгарии (подкорпус "Sprachatlas Ostserbiens und Westbulgariens", с. Берчиновац, ок. 5300 словоупотреблений), при финансировании РФФИ (грант № 18-512-76002 "Изучение дивергенции и конвергенции традиций Центральных Балкан: реализация и перцепция"). Материалы диалектологических экспедиций транскрибированы при помощи программы транскрипции и разметки Partitur-Editor из пакета EXMARaLDA (Schmidt 2009). Аннотация частей речи и лемматизация проведены с использованием таггера ReLDI (Ljubešić et al 2016) и набора морфосинтаксических меток, сформированного в рамках проекта MULTEXT-East (Erjavec, Ljubešić 2016). Ради достижения большей точности, в настоящее время продолжается дополнительная ручная обработка данных и создание инструментов, специализирующихся на обработке данного диалекта. Корпус, общий объем которого составляет 700 000 токенов, выполнен в формате XML в соответствии с конвенциями TEI. Поиск по корпусу и извлечение данных для квантитативного анализа осуществляется при помощи специально разработанной программы на языке Python.

Литература

Вуковић, Теодора; Самарџић, Тања. 2018. Просторна расподела фреквенције постпозитивног члана у тимочком говору// Ћирковић Светлана (Ур.). Тимок. Теренска истраживања 2015–2017. Књажевац: Библиотека. – в печати.

Мирић, Мирјана М. 2018. Употреба/изостављање субјунктивног маркера *да* у конструкцији футура првог у тимочким говорима// Ћирковић Светлана (Ур.). Тимок. Теренска истраживања 2015–2017. Књажевац: Библиотека. – в печати.

Erjavec, Tomaž; Ljubešić, Nikola: MULTEXT-East Morphosyntactic Specifications, Version 5 (draft), Croatian Specifications. (http://nl.ijs.si/ME/V5/msd/html/msd-hr.html).

Ljubešić, Nikola; Klubička, Filip; Agić, Željko; Jazbec, Ivo-Pavao: New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. Nicoletta Calzolari, Khalid Choukri,Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis (ur.): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož: European Language Resources Association (ELRA).

Schmidt, Thomas. 2009. Creating and Working with Spoken Language Corpora in EXMARaLDA. in: Verena Lyding (ed.): Lesser Used Languages & Computer Linguistics II. Bolzano: Eurac Research. 151-164.

Sobolev, Andrej N. 1998. Sprachatlas Ostserbiens und Westbulgariens. Bd. III. Texte. Marburg: Biblion Verlag.