Тезисы
всероссийской конференции
«От языковых машинных фондов к лингвистическим корпусам: памяти В.М. Андрющенко»

Лаборатория автоматизированных лексикографических систем
НИВЦ МГУ имени М.В. Ломоносова
Институт русского языка имени В.В. Виноградова

Москва, 28 – 29 сентября 2018 г.

словарями на этом диалекте или соединять с другими с помощью этимологических связей. В ЛингвоДоке есть возможность вывода всех употреблений той или иной морфемы в любом из словарей или корпусов. Это дает позволяет в будущем объединить весьма большие массивы текстов в единой лаборатории и дать им комплексное осмысление с помощью морфологических и лексических аудиословарей, которые будут объединены этимологическими связями.

Помимо этого в ЛингвоДоке существует возможность сколь угодно сложных запросов поиска и отражения их **на карте** мира http://lingvodoc.ispras.ru/map_search. В настоящее время в системе зарегистрированы и работает около 100 ученых лингвистов из большинства крупных городов России и 6 стран Европы (Германии, Австрии, Финляндии, Швеции, Эстонии, Венгрии).

Корпусная экосистема Школы лингвистики НИУ ВШЭ

Орехов Б. В.

НИУ ВШЭ
(Москва)

Коллектив Школы лингвистики факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» начал формироваться в 2011 году. С самого начала в нем состояли специалисты по созданию корпусов, в т. ч. принимающие активное участие в разработке НКРЯ Е. В. Рахилина, А. А. Бонч-Осмоловская, С. Ю. Толдова, О. Н. Ляшевская, Т. А. Архангельский. В 2011 году заметный толчок развитию корпусов различных языков дала Программа фундаментальных исследований Президиума РАН «Корпусная лингвистика», активное участие в осуществлении которой принимали названные лингвисты.

К этому моменту в распоряжении коллектива был корпусный менеджер, разработанный в 2000-х годах для Восточно-армянского национального корпуса: http://eanc.net/ На платформе этого менеджера ещё в рамках Программы Президиума РАН были размещены корпуса бурятского, калмыцкого, татарского, казахского, албанского, лувийского, монгольского, цыганского языков (все размещены на ресурсе http://web-corpora.net/)

[Архангельский 2012]. По окончании Программы все они поддерживаются сотрудниками ШЛ НИУ ВШЭ.

Использование единообразной платформы позволило выстроить последовательную работу по сбору коллекций текстов и по подготовке инструментов разметки для других языков. В ходе учебно-научной деятельности в рамках сотрудничество преподавателей и студентов на домене http://web-corpora.net/ в течение 2012-2017 годов размещены корпуса удмуртского, новогреческого, тайского, амхарского языков, языка идиш, а также башкирский поэтический корпус [Arkhangelskiy 2014]. Подготовленные для этих корпусов текстовые коллекции не только доступны в интерфейсе поиска корпуса, но и служат полигоном для осуществления учебной проектной работы: студентами разработаны инструменты для снятия морфологической неоднозначности для новогреческого и языка идиш, морфологический анализатор для амхарского языка, основанный на технологии машинного обучения, апробированы методы автоматического поиска когнатов в близкородственных языках.

Ключевая работа по поддержке корпусной инфраструктуры проделана Т. А. Архангельским, который, являясь экспертом по работе с корпусным менеджером также является разработчиком платформы универсального языконезависимого морфологического анализатора UniParser [Архангельский 2014], используемого для разметки удмуртской, албанской, казахской и других текстовых коллекций.

Унифицированный характер интерфейса корпусной платформы позволил создать универсальную веб-страницу для одновременного доступа ко всем развернутым на сервере http://web-corpora.net/ корпусам. Индексальная страница сайта позволяет не только перейти к нужному корпусу, но и осуществить быстрый поиск по точной форме в любом интересующем нас корпусе.

В рамках проекта «Языки России» (http://web-corpora.net/minorlangs/) собраны коллекции текстов для будущих корпусов, которые будут обладать богатой социолингвистической разметкой.

Одновременно с созданием корпусов для не охваченных языков идет работа по совершенствованию НКРЯ. Разработаны (пока не внедрены по организационным причинам) современные средства визуализации выдачи, составлены скетчи (http://linghub.ru/RNC_sketches/), идет работа над алгоритмами снятия морфологической неоднозначности в русском языке.

В ноябре 2017 года Т. А. Архангельским представлена первая версия нового корпусного менеджера, «Цакорпус», при разработке которой учтены все недостатки использовавшейся ранее платформы. Развертывание корпусов с помощью нового менеджера должно происходить быстрее и проще. Новые корпуса будут размещаться на новом домене ШЛ НИУ ВШЭ, linghub.ru, который, как следует из названия, будет аккумулировать различные ресурсы компьютерно-лингвистической природы, не ограничиваясь корпусами.

Литература

Т. А. Архангельский. Электронные корпуса албанского, калмыцкого, лезгинского и осетинского языков // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2012. № 4. С. 24-29

T. Arkhangelskiy, M. Medvedeva. An online annotated corpus of Udmurt language // 4th Mikola Conference on Lexicology and Lexicography of the Uralic and Siberian languages. Szeged, Hungary, November 14–15, 2014

Т. Архангельский. Система морфологического анализа текстов UniParser // Конференция по компьютерной и когнитивной лингвистике ``TEL-2014''. Казань, 6–9 февраля 2014

Мультимедийный корпус языка идиш // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2015. № 3. С. 18–25

T. Arkhangelskiy, E. Kuzmenko. Composing the Corpus of Modern Greek: features and methods // 12th International Conference on Greek Linguistics. Berlin, Germany, September 16–19, 2015

Т. Архангельский, Т. Панова. Мультимедийный корпус языка идиш // XX ежегодная международная конференция по иудаике "Сэфер". Москва, 12 июня 2015

T. Arkhangelskiy. Digital corpora at web-corpora.net: Features and development // 8th Tsakonian conference. Leonidio, Greece, September 9–11, 2016

T. Arkhangelskiy, M. Medvedeva. Developing Morphologically Annotated Corpora for Minority Languages of Russia // CLiF 2016, Bloomington, IN, USA, June 6–10, 2016

T. Arkhangelskiy, N. Serdobolskaya, M. Usacheva. Corpus-oriented lexicographic database for Beserman Udmurt. In: Acta Linguistica Academica, Vol. 64, No. 3, 2017, P. 397-415.

Т. Архангельский. Выявление диалектных особенностей удмуртского языка при помощи интернет-корпуса // VI международная молодёжная научно-практическая конференция Института социально-гуманитарных наук Тюменского государственного университета “Множественность интерпретаций: цифровая перезагрузка”. Тюмень, 14–17 февраля 2018

О проекте создания факсимильно-транскрипционного корпуса рукописей Пушкина

Перцов Николай Викторович
ИРЯ РАН
(Москва)

Факсимильно-транскрипционный корпус рукописей (ФТК; иначе – рукописный факсимильно-транскрипционный корпус) – это корпус, единицами которого служат факсимильно-транскрипционные представления (ФТП), т.е. пары вида «страница рукописи, транскрипция этой страницы». Кроме этих необходимых компонентов в составе ФТП могут быть даны и другие, «факультативные», «поля», часто представленные в архивных описаниях или в публикациях писателей: характер текстов (черновой, беловой с поправками, беловой), отнесение текста к тому или другому произведению, особенности бумаги, пишущих средств, пометы посторонних лиц, идентификация рисунков, датировка, послойное представление вариантов текста и другие сведения.

Будет охарактеризован проект создания ФТК рукописей Пушкина и рассказано о проведённой работе по транскрибированию и компьютеризации его рукописей. В пушкинской текстологии накоплено довольно обширное множество транскрипций черновиков, разбросанных в многочисленных работах. Прежде всего, желательно собрать и «компьютеризовать» эти транскрипции (проведя по возможности их проверку).