Секция «Лексикология и лексикография»

Создание корпуса научных статей Soil

Форафонов Иван Алексеевич[1], Яшин Алексей Игоревич[2]
1 - Орловский государственный университет, Орел, Россия; 2 - Орловский государственный университет, Орел, Россия
E-mail: ifor.osau@yahoo.com

Введение

Лингвистические корпусы представляют собой большие, сбалансированные, репрезентативные и аннотированные коллекции текстов, которые хранятся и обрабатываются в электронном виде при помощи специализированных программ. При этом и сбалансированность, и представительность, и размер - весьма относительные и дискутируемые понятия.

Исследователи часто создают отдельный для определённого типа исследования - скажем, корпус детской речи подходит для изучения механизмов усвоения языка, а исторический корпус - для диахронических исследований. Если предположить, что лингвист решит создать корпус из текстов блюзовых песен то, несмотря на очевидную ограниченность, корпус лирики блюзовых композиций будет в своём роде и репрезентативным, и сбалансированным - ведь он позволит изучить особенности языка этого музыкального жанра.

В процессе анализа научных работ по восстановлению нарушенных почв возникла проблема, заключающая в различных интерпретациях некоторых основополагающих понятий в русско- и англоязычной литературе. В частности, замечено отсутствие унифицированных дефиниций для таких понятий, как *soil remediation, soil restoration, soil recultivation* и многих других. Проблема усугубляется тем, что ни у русско-, ни у англоязычных учёных нет консолидированного мнения о том, что обозначают вышеназванные термины. Так, например, термин *reclamation* может обозначать одновременно и восстановление почвы, и создание новых участков суши. В то же время, русскоязычный аналог "рекламация" не относится к сфере восстановления почвы и обозначает "претензию покупателя или заказчика".

Для разграничения дефиниций и анализа функционирования терминов в тексте был создан лингвистический корпус Soil.

Источники и сбор данных

К качестве источников данных были выбраны научные статьи, находящиеся в свободном доступе в поисковой системе Академия Google (Google Scholar). Индекс поисковой системы включает в себя большинство рецензируемых онлайн журналов крупнейших научных издательств Европы и Америки.

Несмотря на то, что поисковый запрос в системе Академии Google показывает только 1000 результатов, количество статей, доступных для скачивания, исчисляется сотнями. Загрузка всех статей вручную заняла бы огромное количество времени, потому для сбора данных была разработана программа Google Scholar Downloader.

Программа выполнена в виде расширения для веб-браузера Google Chrome. После выполнения запроса в поисковой системе Google Scholar пользователь запускает расширения при помощи иконки на панели инструментов веб-браузера, после чего начинается просмотр результатов запроса для выявления статей находящихся в открытом доступе. Затем доступные статьи автоматически скачиваются для дальнейшей работы.

Анализ результатов поиска продолжается до тех пор, пока пользователь не окончит работу расширения, нажав на иконку на панели задач, либо пока не будет достигнут заданный лимит на количество найденных статей.

Расширение Google Chrome выполнено в виде двух сценариев на языке программирования JavaScript. Основной модуль расширения выполняет работу в контексте веб-браузера и ведет с ним работу, другой выполняется в контексте содержимого веб-страницы и осуществляет ее анализ для выявления ссылок на доступные статьи.

Взаимодействие между сценариями осуществляется путем обмена сообщениями: «готов к использованию», «получить ссылки на статьи», «перейти к следующей странице» и т.п.

Основной модуль содержит команды для создания иконки на панели задач, загрузки статей по найденным ссылкам, отправки и обработки сообщений для другого модуля. Вспомогательный модуль содержит команды для поиска статей, перемещения по страницам результатов поиска, обработки и отправки сообщений основному модулю.

Характеристики и компиляция лингвистического корпуса Soil

Корпус, который был назван Soil (от англ. *soil* - почва), состоит из 2537 научных статей 2000-2015 годов издания. Корпус состоит из 45424906 словоформ или 25545924 слов (лексем).

Корпус Soil размещён на платформе Sketch Engine. Данная платформа предоставляет доступ к практически ста корпусам, тринадцать из которых, не считая параллельных, на английском языке. Sketch Engine обладает разнообразным инструментарием для поиска, сортировки и анализа данных. Некоторые функции поиска в Sketch Engine уникальны, и позволяют значительно облегчить исследование. Особенно хотелось бы отметить Word Sketches - тип поиска, при котором выводятся данные о частотности и значимости конструкций, содержащих искомую языковую единицу.

К недостаткам корпуса Soil можно отнести его ограниченную доступность для нелингвистов, так как, во-первых, характер исследований, проводимых в корпусах, подразумевает, что исследователь знаком с корпусами; во-вторых, корпус Soil размещён на платформе Sketch Engine, работа с которой требует определённых навыков и времени для их освоения.

Другим недостатком является отсутствие метаданных: за исключением названия оригинального файла, документы не содержат метаинформацию об авторах, годах издания, журналах, в которых статьи были опубликованы и т.п.

Отсутствие метаданных могло бы стать колоссальным недостатком при, например, диахронических или социолингвистических исследованиях, однако для выполнения текущих задач их наличие или отсутствие не является основополагающим фактором.

Заключение и перспективы использования

Результаты пробных исследований в корпусе Soil показали, что работа при помощи поисковых систем платформы Sketch Engine поможет дать более точные дефиниции и разграничить значения терминов. В дальнейшем планируется создание аналогичного корпуса статей на русском языке для компаративного изучения функционирования терминов в научных текстах.

Источники и литература

1) Adam Kilgarriff, Pavel Rychly, Pavel Smrz, David Tugwell. The Sketch Engine. Proc EURALEX 2004, Lorient, France; Pp 105-116, and also give the web address http://www.sketchengine.co.uk
2) Baker P. A Glossary of Corpus Linguistics / P.Baker, A. Hardie, T. McEnery. – Edinburgh. : Edinburgh University Press Ltd, 2006
3) Hardie, A., McEnery, T. 2012. Corpus Linguistics: Method, Theory and Practice, New York: Cambridge University Press