ATLANTIS PRESS *Advances in Social Science, Education and Humanities Research (ASSEHR), volume 122*
45th International Philological Conference (IPC 2016)

К вопросу об отличиях традиционных и веб-корпусов на основе анализа частотных существительных

Хохлова Мария Владимировна

Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия

On the Differences between Traditional and Web-Corpora based on the Analysis of High-Frequency Nouns

Khokhlova Maria

Saint Petersburg State University (SPSU), St. Petersburg, Russia

Аннотация

В последнее время появляется все больше корпусов текстов, создаваемых автоматическими методами. В статье обсуждаются корпусы русского языка ruTenTen и Araneum Russicum Maximum, а также сравниваются результаты, полученные на их основе, с данными из Частотного словаря современного русского языка.

Abstract

The paper gives a survey of corpora and analyzes a number of Russian nouns across the following corpora: ruTenTen (18.3 bln tokens) and Araneum Russicum Maximum (13.7 bln tokens). The research focuses on the discussion on these corpora, their comparison and the study of frequency properties for the high-frequency Russian nouns comparing them with data published in the Frequency Dictionary.

Ключевые слова: корпус текстов, Интернет-корпус, частотный словарь, существительные.

Keywords: text corpus, web corpus, frequency dictionary, nouns.

Введение

В последнее время появляется все больше корпусов текстов, создаваемых автоматическими методами. От традиционных текстовых коллекций они отличаются как по объему, так и по содержанию. Связано это с ростом технических возможностей и постепенным уходом от «классического» составления корпусов к автоматическому. Под «классическим», или традиционным, подходом понимается разработка корпусов по заранее описанной технологии: предварительный отбор

Copyright © 2017, the Authors. Published by Atlantis Press.

This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

текстов с учетом репрезентативности и сбалансированности, их вычитка, разметка и загрузка. Новые корпусы могут быть названы «сверхбольшими» [Benko, Zakharov 2016] и в основном содержат тексты, собранные при помощи специальных программ из Интернета. Для русского языка можно упомянуть ряд проектов, активно развивающихся в последнее время (например, [Беликов, Селегей, Шаров 2012; Benko 2014]).

Материал

Целью нашего исследования является сравнение результатов, получаемых при работе с двумя корпусами разных объемов, и данных из словаря [Ляшевская, Шаров 2009], который был создан на основе выборки из Национального корпуса русского языка. Рассматриваемый словарь представляет современный русский язык 1950–2007 гг. и содержит тексты разных функциональных стилей: художественная литература, публицистика, нехудожественная литература (учено-научная, официально-деловая, электронная коммуникация, церковно-богословская, реклама, бытовая, производственно-техническая) и устная непубличная речь.

Для экспериментов в качестве материала для исследования нами были отобраны два русскоязычных корпуса: ruTenTen и Araneum Russicum Maximum. Первый корпус относится к семейству корпусов TenTen, включающему в себя корпуса разных языков [Jakubíček, Kilgarriff, Kovář, Rychlý, Suchomel 2013], объем каждого из которых превышает 1 млрд. слов. Корпус русского языка является одним из самых больших наряду с английским, немецким, французским и испанским языками. В качестве второго корпуса был использован корпус Araneum Russicum Maximum, который является представителем семейства Aranea. Корпуса Aranea [Benko 2014] также были созданы для полутора десятков разных языков, среди которых есть и русский. Объемы корпусов следующие: ruTenTen (18,3 млн словоупотреблений), Araneum Russicum Maximum (13,7 млрд словоупотреблений).

Отбор данных

В основном тексты, входящие в состав Интернет-корпусов русского языка, представляют собой материалы новостных ресурсов, блогов, рекламных сайтов, групп социальных сетей и др. Художественные тексты представлены не так широко, поэтому было решено обратиться к спискам частотной лексики, которые отражают именно данные функциональные стили, — нехудожественные и публицистические тексты. Нами был сформированы два списка слов, в который были отобраны 20 наиболее частотных существительных по словарю [Ляшевская, Шаров 2009]. Список, характерный для нехудожественной литературы (в скобках приведены частоты в ipm): *год* (4624,2), *время* (2080,5), *человек* (1945,3), *система* (1798,0), *работа* (1766,4), *статья* (1363,0), *дело* (1339,5), *случай* (1259,0), *процесс* (1221,8), *вопрос* (1180,9), *лицо* (1175,9), *суд* (1153,9), *часть* (1153,8), *вид* (1147,9), *решение* (1122,3), *право* (1117,6), *ребенок* (1078,4), *отношение* (1077,5), *развитие* (1059,6), *федерация* (1003,1).Список, характерный для публицистических текстов: *год* (5589,50), *человек* (2950,10), *время* (2364,60), *жизнь* (1548,40), *дело* (1482,00), *день* (1397,80), *работа* (1272,40), *страна* (1203,90), *вопрос*

(992,00), *слово* (989,70), *место* (976,10), *мир* (887,80), *дом* (879,70), *друг* (850,90), *случай* (744,30), *город* (738,50), *рука* (713,00), *власть* (711,80), *конец* (710,80), *сила* (709,80).

Далее нами были рассмотрены ранговые распределения данных русских существительных в обоих корпусах.

Результаты

Лексемы *сайт* и *компания*, которые оказались наиболее частотными в обоих корпусах ruTenTen и Araneum Russicum Maximum и не вошли в список 20 частотных существительных, отобранных из словаря, отражают специфику текстов, взятых из Интернета, во-первых, из-за большого количества новостных ресурсов, во-вторых, ввиду направленности на описание содержания веб-страниц. В корпусе ruTenTen в отличие от Araneum Russicum Maximum наиболее частотными также являются слова *возможность*, *качество*, *программа*, *услуга*, что может свидетельствовать о том, что данный корпус даже в большей степени ориентирован на описание веб-ресурсов.

Для обоих списков существительных, представленных как в словаре другой нехудожественной литературы, так и в словаре публицистики [Ляшевская, Шаров 2009], частотные характеристики в корпусах ruTenTen и Araneum Russicum Maximum оказываются наиболее согласованными. Для первого списка существительных коэффициент корреляции Спирмена между ранговыми распределениями по обоим корпусам равен 0,89 (что говорит о достаточно высокой тесноте корреляционной связи), а по словарю и корпусу ruTenTen — 0,63, по словарю и корпусу Araneum Russicum Maximum — 0,76. Для второго списка существительных коэффициент корреляции Спирмена между ранговыми распределениями по обоим корпусам равен 1 (что свидетельствует о полной корреляции), а по словарю и корпусу ruTenTen — 0,84, по словарю и корпусу Araneum Russicum Maximum — 0,82. Можно высказать предположение, что оба корпуса по своему наполнению оказываются в большей степени схожи с публицистическими текстами (хотя для корпуса Araneum Russicum Maximum характерна высокая корреляция с данными, полученными на основе нехудожественных текстов), и более широко, с Национальным корпусом русского языка, поскольку словарь [Ляшевская, Шаров 2009] создавался на его материале.

Выводы

Общий вывод, который можно сделать на основе полученных данных, свидетельствует о том, что тексты больших корпусов отражают язык Сети и по своему составу схожи с публицистикой. Для рассмотренных частотных существительных наблюдается очень тесная связь между данными двух корпусов текстов, можно предположить, что в случае высокочастотных слов не существует разницы между корпусами, созданными автоматически.

Оба корпуса показывают достаточно большую согласованность с частотным словарем. Данные, приведенные в частотном словаре, были основаны на Национальном корпусе русского языка, поэтому можно сделать вывод, что полученные результаты отражают корреляцию между традиционными корпусами и веб-корпусами.

Благодарность

Статья подготовлена в рамках работы по гранту Президента Российской Федерации для государственной поддержки молодых российских ученых № МК-5274.2016.6 «Исследование статистических закономерностей сочетаемости лексических единиц».

Литература

Беликов В. И., Селегей В. П., Шаров С. А. Пролегомены к проекту Генерального интернет-корпуса русского языка (ГИКРЯ) 2012 // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). М., 2012. Т. 1. С. 37–49.

Корпусы текстов Aranea. URL: http://sketch.juls.savba.sk/aranea_about/ (Дата обращения: 26.06.2016).

Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). Москва: Азбуковник, 2009. 1087 с.

Национальный корпус русского языка. URL: http://ruscorpora.ru (Дата обращения 26.06.2016).

*Benko V.*Aranea: Yet Another Family of (Comparable) Web Corpora. In *P. Sojka, A. Horák, I. Kopeček* and *K. Pala* (Eds.): Proceedings of Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. LNCS 8655.Springer International Publishing Switzerland, 2014. P. 257–264.

Benko V., Zakharov V. Very Large Russian Corpora: New Opportunities and New Challenges. In Computational linguistics and intellectual technologies. Vol. 15 (22), 79-93. Moscow: Izd-vo RGGU, 2016.

*Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V.*The TenTen Corpus Family.In Proceedings of the 7th International Corpus Linguistics Conference.Lancaster, 2013. P. 125–127.

References