КОРПУСНАЯ ЛИНГВИСТИКА В РОССИИ

Виктор Павлович Захаров

кандидат филологических наук,
доцент кафедры математической лингвистики СПбГУ
Санкт-Петербург, Россия

Email: vz1311@yandex.ru

1. Введение

В последние годы создание корпусов и корпусно-ориентированные исследования стали неотъемлемой частью деятельности лингвистов. В мире корпусная лингвистика как особое направление сложилась к началу 1990-х годов. Россия встала на этот «корпусный» путь с некоторым опозданием, но движется по нему очень быстро. Безусловно, ключевым моментом стало создание Национального корпус русского языка (2004г., http://ruscorpora.ru). О востребованности корпуса свидетельствуют многочисленные публикации, подготовленные на корпусе. Часть из них находится на сайте корпуса в разделе Studiorum. С корпусной проблематикой во всем ее многообразии можно ознакомиться по материалам конференции «Корпусная лингвистика» (см. на сайте http://corpora.phil.spbu.ru), проводимой кафедрой математической лингвистики СПбГУ (2002, 2004, 2006, 2008, 2011, 2013 гг.). Начинают выходить учебные пособия по корпусной лингвистике - см. Литературу.

2. Первые корпусы русского языка

Первый русскоязычный корпус был создан в 1980-е гг. в Университете Уппсалы, Швеция. Однако еще до первых русскоязычных корпусов в 1960–1970-е гг. был создан ***Частотный словарь русского языка*** под рук. Л.Н. Засориной, построенный на основе

примитивных текстовых файлов объемом в 1 млн словоупотреблений, включавшей в себя лексику 4 жанров в примерно равной пропорции: общественно-политические тексты, художественную литературу, научные и научно-популярные тексты из разных областей и драматургию.

В процессе создания этих файлов – сегодня мы бы сказали, корпуса – обсуждались все проблемы современной корпусной лингвистики:

- репрезентативность,
- графематический анализ,
- нормализация,
- лемматизация.

Так что фактически это был первый корпус русского языка, не дошедший до наших дней.

В 1985 г. в СССР по инициативе академика А.П. Ершова были начаты работы по созданию ***Машинного фонда русского языка***. Это был грандиозный проект. В создании фонда принимали участие более 40 организаций-соисполнителей, среди них Институт русского языка, Московский, Санкт-Петербургский, Харьковский, Гродненский, Сыктывкарский и Саратовский университеты и др. В задачи фонда входило накопление на машинных носителях и в базах данных текстовых, лексикографических и грамматических источников, необходимых для научного изучения русского языка и для осуществления прикладных разработок. Одновременно велось создание программных средств для проведения лингвистических исследований. В 1985–1992 гг. были осуществлены разработка концепции и архитектуры Машинного фонда русского языка, разработка концепции терминологического банка данных, введены в компьютер текстовые источники русской литературы XIX–XX вв., главнейшие словари русского языка, краткая академическая грамматика, созданы текстовые корпусы поэзии, художественной прозы, общественно-политических и технических текстов. Однако после 1991 г. в новых экономических условиях работы по созданию фонда постепенно стали сокращаться и наконец совсем прекратились.

Уппсальский корпус русского языка (Uppsal'skij korpus russkix tekstov) состоит из 600 текстов, его объем составляет 1 млн словоупотреблений, поровну распределенных между образцами специальной и художественной литературы. По замыслу создателей, корпус должен был отражать современное состояние русского языка. Цель формирования корпуса заключалась в том, чтобы представить, в первую очередь, литературный язык, поэтому в массиве нет образцов разговорной речи.

В корпус отбирались специальные тексты (включались не фрагменты текстов, а целые тексты) с 1985 по 1989 г. и художественные тексты с 1960 по 1988 г. В аннотации к корпусу отмечается, что среди специальных текстов особое внимание уделено более важным, с точки зрения создателей корпуса, темам, а среди художественных текстов предпочтение отдано более известным авторам. Тексты в корпусе записывались латиницей. Фрагмент корпуса выглядит следующим образом:

&Perestrojka vse glubhe zatragivaet hiznennye interesy millionov, obqestva v celom. Estestvenno, l~di xot„t lu†we u„snit' sut' i nazna†enie processov obnovleni„, blihnie i dal'nie celi preobrazovanij, opredelit' svoe otnowenie k nim

Уппсальский корпус сейчас входит в так называемые ***«Тюбингенские корпусы русских текстов»***, созданные в рамках работы специального научно-исследовательского сектора SFB 441 Тюбингенского университета в 1990–2000-е гг. с возможностью поиска on-line (http://www. sfb441.uni-tuebingen.de/b1/rus/korpora.html#uppsalakorpus).

Корпусы размечены тегами морфологической аннотации. Разметка была осуществлена при помощи статистического теггера (TnT). Поиск может производиться как по словоформам, так и по морфологическим тегам. Возможен вывод текста вместе с разметкой. Для ввода поискового выражения и вывода найденного текста можно выбрать одну из следующих кодировок: кириллицу (KOI8 или Windows-1251) или транслитерацию латинскими буквами. Поиск осуществляется при помощи программы CQP, представляющей собой систему для управления большими корпусами, разработанную Институтом машинной обработки языка Штутгартского университета.

Компьютерный корпус текстов русских газет конца XX в. был создан на Филологическом факультете МГУ в 2000–2002 гг. в Лаборатории общей и компьютерной лексикологии и лексикографии (http://www.philol.msu.ru/~lex/corpus/). Подбор обширного газетного материала для корпуса (тексты общим объемом более 1 млн словоупотреблений) был осуществлен на основе принципов включения в него полных номеров 13 российских газет на русском языке за отдельные даты 1994–1997 гг. (23110 текстов), представленности в нем ежедневных и неежедневных («МН», «Новая газета»), «левых» («Завтра», Правда», «Правда-5») и «правых», центральных и местных, общих и профессионально ориентированных газет (например, «Литературная газета»). Эти принципы позволяют получить относительно объективную и надежную картину соотношения в газетном материале текстов различного типа (например, различных жанров и жанровых типов), их единиц и отношений между ними.

Корпус создан, анализируется и управляется на основе системы Диктум-1, разработанной в Лаборатории общей и компьютерной лексикологии и лексикографии МГУ. С помощью этой системы тексты и единицы корпуса автоматически и полуавтоматически маркируются различного рода маркерами: тексты – маркерами газеты-источника, объема текста, его жанра, даты публикации и т. п.; словоупотребления – маркерами грамматических, лексических, морфемных и иных категорий.

При подготовке демонстрационного варианта корпуса для Интернета был выделен фрагмент корпуса общим объемом более 200 тыс. словоупотреблений, проведена автоматическая лемматизация и морфологическая квалификация словоупотреблений корпуса (с последующими контролирующими процедурами), а также морфемная сегментация словоформ и лексем.

3. Современные корпусы русского языка

Национальный корпус русского языка (НКРЯ) (http:// ruscorpora.ru/ (с апреля 2004 г.)) отвечает критерию репрезентативности и другим требованиям, предъявляемым к современным корпусам, о чем свидетельствуют следующие его характеристики (см. статистику http:// ruscorpora.ru/corpora-stat.html):

1) объем НКРЯ, который в сумме составляет около 500 млн словоупотреблений (по данным сайта на март 2013 г.);

2) жанровое разнообразие составляющих его текстов, которые относятся ко всем основным сферам использования русского языка (научной, официально-деловой, публицистической, церковно-богословской, художественной, разговорно-бытовой, включая устную и электронную коммуникацию).

Этот корпус хорошо известен, поэтому в числе целого ряда его особенностей, отличающих его от других национальных корпусов, отметим всего лишь две: мультимедийный подкорпус (http://ruscorpora.ru/search-murco.html) и семантическая разметка (http://ruscorpora.ru/corpora-sem.html).

Мультимедийный русский корпус (МУРКО), образованный фрагментами кинофильмов 1930–2000-х гг., представленных в виде параллельных видеоряда, аудиоряда и текстовой расшифровки звучащей речи, а также наблюдаемых в кадре жестов. Возможен поиск не только по произносимому тексту, но и по жестам (кивание головой, похлопывание по плечу и т. п.) и типу речевого действия (согласие, ирония и т. п.). В поисковой выдаче видеофрагменты доступны для просмотра и прослушивания.

Семантическая разметка приписывает единицам текста один или несколько семантических и словообразовательных признаков, например, 'лицо', 'вещество',

'пространство', 'скорость', 'движение', 'обладание', 'свойство человека', 'диминутив', 'отглагольное имя' и т.п. Используется фасетная классификация, при которой одно слово может попадать в несколько классов. В основу семантической разметки положена система классификации русской лексики, принятая в базе данных "Лексикограф", которая разрабатывалась с 1992 г. в Отделе лингвистических исследований ВИНИТИ РАН под рук. Е. В. Падучевой и Е. В. Рахилиной.

И по указанным признакам, разбитым по нескольким категориям: разряд, таксономия, мереология, оценка, словообразование и т. п. – в корпусе осуществляется поиск. Можно искать контексты по основному значению слова или по вспомогательному. Можно найти все приставочные глаголы движения (отметить признак «движение» в категории «Таксономия» и признак «приставочные глаголы» в категории «Словообразование»). И т.д.

Хельсинский аннотированный корпус (ХАНКО) создан в Хельсинском университете в начале 2000-х гг. как часть проекта «Функциональный синтаксис русского языка» (рук. проф. А. Мустайоки) и постоянно развивается. Объем корпуса – 100 тыс. словоформ. Доступен в Интернете по адресу: http:// http:// www.ling.helsinki.fi/projects/hanco/. Корпус создан на основе статей из журнала «Итоги» за 2001 г. В корпусе реализованы морфологическая и синтаксическая разметки и, соответственно, морфологический и семантический поиск. Особенность корпуса – тщательно проработанный формат лингвистического описания данных и полная визуальная (ручная?) проверка результатов автоматической разметки, имеющая следствием полное снятие грамматической омонимии. Синтаксическая разметка корпуса должна совмещать разметку в терминах членов предложения (уже реализована) и в терминах деревьев зависимостей. Кроме того, размечены многословные устойчивые обороты (примерно 2000 единиц). Планируется семантическая разметка в терминах семантических категорий. Корпус преследует в первую очередь учебные цели.

Корпусы университета г. Лидс (Великобритания). В 2000-е гг. в университете г. Лидс в Центре переводческих исследований С.А. Шаровым создано большое количество корпусов для разных языков (английский, арабский, китайский, французский, немецкий, итальянский, японский, испанский, польский и др.) (http:// corpus.leeds.ac.uk/). Среди них имеются корпусы и русского языка (http://corpus.leeds.ac.uk/ruscorpora.html). Это версия Национального корпуса русского языка объемом в 116 млн словоупотреблений (на основе ее был создан Частотный словарь русского языка (авторы О.Н. Ляшевская, С.А. Шаров). Кроме того, на этом сайте представлены Корпус русских газет (2001–2004 гг., 76 млн словоупотреблений), корпус русских текстов из Интернета (160 млн словоупотреблений),

бизнес-корпус (12 млн словоупотреблений) и объединенный корпус, составленный из всех вышеназванных.

Поисковый интерфейс Leeds CQP базируется на корпусном менеджере IMS Corpus Workbench и предоставляет интересные возможности. Он позволяет вести очень точный лексико-грамматический поиск, поскольку дает возможность использовать специальный язык запросов, в том числе и с использованием языка регулярных выражений. Имеются способы управления выходным интерфейсом, формой представления результатов поиска. Можно также получить списки коллокаций, вычисленных и упорядоченных на основе ассоциативных мер MI, T-score, Log-likelihood. Там же имеется коллекция различных программных средств для обработки корпусных текстовых данных (http://corpus.leeds.ac.uk/tools/).

Корпус Библиотеки Мошкова. На сайте группы АОТ (http://aot.ru/search1.html) имеется большой корпус русских текстов объемом 680 млн слов, созданный А. Сокирко по текстам из библиотеки Мошкова. Можно искать по лексическим единицам с учетом частей речи и морфологических характеристик, используя мощный язык запросов корпусного менеджера DDC. Там же имеется сервис поиска биграмм (54 млн), вычисленных по мере MI.

Корпусы в системе Sketch Engine. Английская лингвистическая служба Lexical Computing Ltd. (A. Kilgarriff) предоставляет на коммерческой основе доступ более чем к 40 корпусам различных языков. Среди них имеется ряд корпусов русского языка и прежде всего корпус, созданный из текстов Интернета по технологии Wacky объемом 20 млрд словоупотреблений. Английские исследователи совместно с чешскими разработчиками из Университета им. Масарика разработали корпусный менеджер Sketch Engine (http://sketchengine.co.uk/).

Менеджер обладает многими уникальными возможностями. Помимо стандартного поиска с выдачей конкорданса он выдает списки коллокаций по отдельным синтаксическим моделям, формирует частотный словарь, группирует лексические единицы в лексико-семантические поля с внутренней кластеризацией и указанием силы связи между лексемами.

4. Звуковые корпусы русского языка

Это в первую очередь хорошо известный ***Корпус устной речи НКРЯ*** объемом более 10 млн словоупотреблений (http://ruscorpora.ru/sbornik2008/07.pdf), уже упоминавшийся *мультимедийный*, а также *диалектный, поэтический и акцентологический* корпусы в составе НКРЯ.

В качестве еще характерного примера рассмотрим речевой разрабатываемый в Институте филологических исследований Филологического факультета СПбГУ. ***Один речевой день*** (ОРД) – звуковой корпус современного русского языка повседневного общения. Корпус создается с целью изучения реальной речи носителей языка в естественных условиях коммуникации, и в этом его отличие от абсолютного большинства речевых корпусов, записанных в лабораторных и других специальных условиях.

Первая серия звукозаписей осуществлена осенью 2007 г. Для этого была отобрана группа информантов из 30 человек, представляющих разные социальные и возрастные слои населения Санкт-Петербурга и давших согласие прожить один день с «диктофоном на шее». Информанты получили подробный инструктаж о методике проведения звукозаписи своих речевых контактов в течение суток, заполнили социологические анкеты и прошли психологическое тестирование. Помимо речи информантов, в корпусе представлены записи их коммуникантов (родственников, друзей, коллег, знакомых и незнакомых), среди которых были люди самого разного возраста и разных специальностей. Общая длительность записанного материала – более 500 часов. Расшифровке и многоуровневой разметке подвергнуто 45 часов (по данным на март 2013 г.).

Данный звуковой корпус позволяет изучать лингвистическую динамику записанного материала: исследовать временны*е* ряды количественных переменных с помощью стандартных статистических методов и анализировать частотные ряды (лексики, грамматических и, в частности, синтаксических структур, семантики или разговорных тем, тех или иных акустических явлений или просодических контуров) в зависимости от времени суток и условий коммуникации в самом широком понимании этого термина, а также решает множество других задач, таких как анализ влияния профессии на бытовую жизнь человека, получение информации о среднем артикуляционном темпе спонтанной речи носителей русского языка.

Мультимодальные корпусы включают видеозапись участников коммуникации, поэтому с их помощью можно исследовать эмоции.

Русскоязычный эмоциональный корпус (REC), размеченный с учетом данных о мимике, движениях рук, бровей и т. д., позволяет изучить стратегии эмоционального взаимодействия и конфликта, непрерывное коммуникативное поведение, хезитации и речевые сбои и др. (http://www.harpia.ru/rec). Он может также использоваться как материал для обучения работников клиентских служб или как база данных эмоциональных реакций для мультипликаторов и режиссеров.

В Иркутском государственном лингвистическом университете идет работа по созданию *Учебного Мультимодального Корпуса* (УМКО) видеозаписей неподготовленных учебных диалогов носителей и «не носителей» русского и китайского языков по определенным темам, размеченных в программе ELAN и представленных также в виде параллельных корпусов, выровненных по смысловым блокам внутри диалогов. Например, диалог носителей русского языка на русском языке сопоставляется с диалогом на ту же тему («Знакомство», «Регистрация в аэропорту» и др.) китайцев, изучающих русский язык, на русском языке. Данный корпус предназначен, в первую очередь, для лингводидактических целей, так как позволяет выявить типичные ошибки и найти пути их устранения в ходе учебных занятий и самостоятельной работы студентов.

Назовем еще *Корпус устных рассказов* на русском языке, стимулом для которых послужил 6-минутный видеосюжет, так называемый «Фильм о грушах» (“Pear film”). Об этом фильме было записано 8 рассказов, сделанных студентами МГУ, общей продолжительностью около 20 минут. Всего в корпусе было 595 элементарных дискурсивных единиц, которые обычно совпадают с простым предложением, и 327 иллюстративных жестов, которые, в соответствии с подходом Г.Е. Крейдлина, понимаются как носители информации, выступая в качестве знаковых кинетических единиц выражения и передачи информации. На примере из корпуса устных рассказов исследователям удалось показать, как отдельные признаки жестов и положения рук могут добавлять дополнительную информацию касательно организации дискурса, состояния говорящего и процесса коммуникации. Так, изменение положения покоя рук между жестами достаточно последовательно указывает на границу между сегментами нарратива. Данный пример демонстрирует предоставляемые мультимедийным корпусом возможности изучения связи структуры устного нарратива и иллюстративных жестов.

5. Специальные корпусы текстов

Специальные корпусы текстов – это сбалансированные корпусы, как правило, небольшие по размеру, подчиненные определенной исследовательской задаче и предназначенные для использования преимущественно в целях, соответствующих замыслу составителя.

Примером может быть *Санкт-Петербургский учебный корпус текстов школьников, изучающих английский язык (SPbEFLLC)*, созданный на кафедре прикладной лингвистики РГПУ им. А.И. Герцена. Основной целью его создания было исследование особенностей английских текстов, порождаемых русскими школьниками. Аутентичный текстовый материал был собран в школах Санкт-Петербурга в период с ноября по декабрь 2007 г.

Авторами текстов являются 78 учеников 911 классов, предварительно прошедших тестирование. Уровень владения английским языком был определен как средний / intermediate (26 %) и выше среднего / upper-intermediate (74 %). Размер данного корпуса составляет около 50 тыс. словоупотреблений.

Исследование на базе корпуса показало, что систематическое предпочтение максимально простых структур развернутым и более естественным моделям стандартного английского языка приводит к так называемой «структурной бедности» речевых произведений не-носителей языка. В репертуаре грамматических структур, обнаруженных в SPbEFLLC, есть такие, которые представляют собой случаи «переходной грамматики» (интеръязыка), выражающиеся, например, в нарушении правил наполнения компонентов базовых структур. Так формируется ядро грамматики EFL (English as a Foreign Language), которое не совпадает с базовыми грамматическими структурами литературного английского языка. На основании корпусных данных авторы высказывают предположение о том, что складывающиеся нормы «глобального английского» во многом опираются на «окаменевшие» модели интеръязыка.

Сложным объектом с точки зрения создания и стандартизации являются исторические корпусы, такие как *Санкт-Петербургский Корпус Агиографических*[1] *Текстов XV–XVII вв.* (СКАТ), доступный на сайте http://project.phil.pu.ru/skat. СКАТ – это электронный корпус текстов по памятникам древнерусской агиографической литературы, созданный на кафедре математической лингвистики филологического факультета СПбГУ. Язык агиографических произведений во многом обусловил судьбу и характер русского литературного языка XV–XVII вв. Отображение этого языка является первостепенной задачей создаваемого корпуса текстов русских житий того времени, что достигается, в частности, за счет широкого географического охвата территорий, где в разное время создавались памятники русской агиографии. В 2011 г. объем корпуса составил 500 тыс. словоупотреблений.

Специальные корпусы текстов могут быть востребованы не менее, чем национальные. Любой *отраслевой* специальный корпус текстов может пригодиться и в данной конкретной отрасли, и в смежных областях (кораблестроение, металлы, экология, навигация и т. д.), поскольку он дает специалисту самое главное – термины в их профессиональном конкретном окружении (что тот или иной автор имеет в виду под

---

1 Агиогра́фия (от греч. ἅγιος «святой» и γράφω «пишу») – научная дисциплина, занимающаяся изучением житий святых, богословскими и историко-церковными аспектами святости. Жития святых могут изучаться с лингвистической, историко-богословской, исторической, социально-культурной и литературной точек зрения.

данным термином, какое понятие за ним стоит), позволяет отследить изменения в терминологии, включая появление новых терминов.

В числе других специальных корпусов следует упомянуть Регенсбургский диахронический корпус русского языка (древнерусские тексты), корпус рукописных памятников Древней Руси (берестяные грамоты, летописи, рукописные книги), параллельный корпус переводов «Слова о полку Игореве», корпус русского электронного наследия «Манускрипт».

6. Сервис НКРЯ «Графики»

Кратко остановимся еще на одном виде работы с корпусами – это диахронические исследования, основанные на статистике. Первой из таких систем был ***Google Books NGram Viewer*** (см. подробнее *А.Ц. Масевич.* Проблема старой орфографии в текстах корпусов русского языка). Затем эта же база данных была представлена М. Дэвисом в составе его корпусов на сайте Brigham Young University (http://corpus.byu.edu/).

С 2012 г. аналогичный сервис появился и в России. Он работает на текстах НКРЯ и называется «Графики». Сервис «Графики» функционально подобен сервису Google Books N Gramm Viewer. Он показывает хронологическое распределение заданных и найденных лексических единиц (словоформ, словосочетаний) в основном корпусе НКРЯ. Вход в этот сервис возможен как со страницы с результатами по произвольному запросу к основному корпусу (ссылка *Распределение по годам*), так и из главного меню (рис. 1).

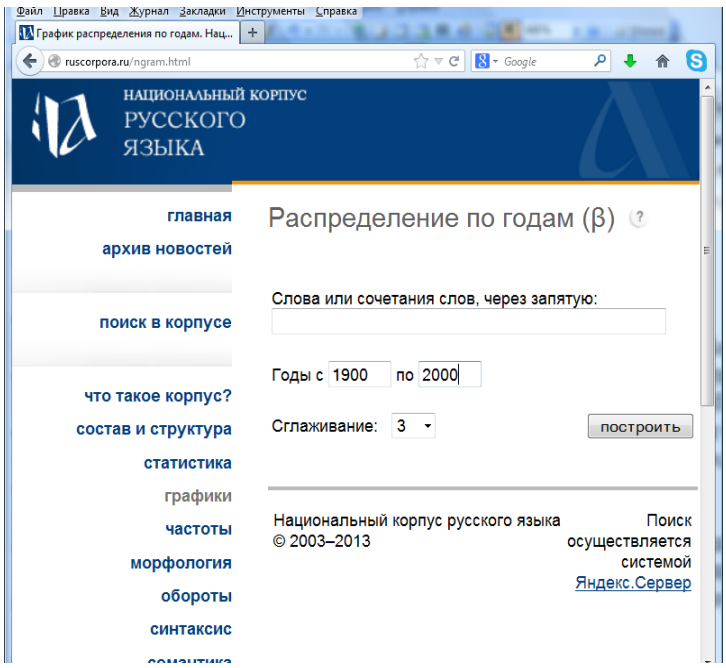

Рис. 1. Интерфейс сервиса «Графики»

Можно задать временные границы, например, с 1930 по 1960 г. Нажав на кнопку «Построить», мы получим график (рис. 2), где каждому элементу сравнения (в нашем примере Черчилль, Рузвельт, Франко) соответствует линия своего цвета (см. легенду в правом верхнем углу).

По вертикальной оси графика откладывается относительная частота употреблений данной лексической единицы (в ipm). Подведя мышку к любой точке на линии, мы увидим относительную частоту употребления за определенный год (ipm). Сглаживание графиков позволяет увидеть общую тенденцию за случайными колебаниями частот. Например, сглаживание в 10 лет усредняет частоту слова с учетом предшествующих и последующих 5 лет, т.е. для данного года берется средняя величина за 11 лет.

Имеется возможность показать таблицы с абсолютными частотами употреблений за каждый год. Из таблиц гиперссылки позволяют перейти к просмотру примеров из корпуса по годам.

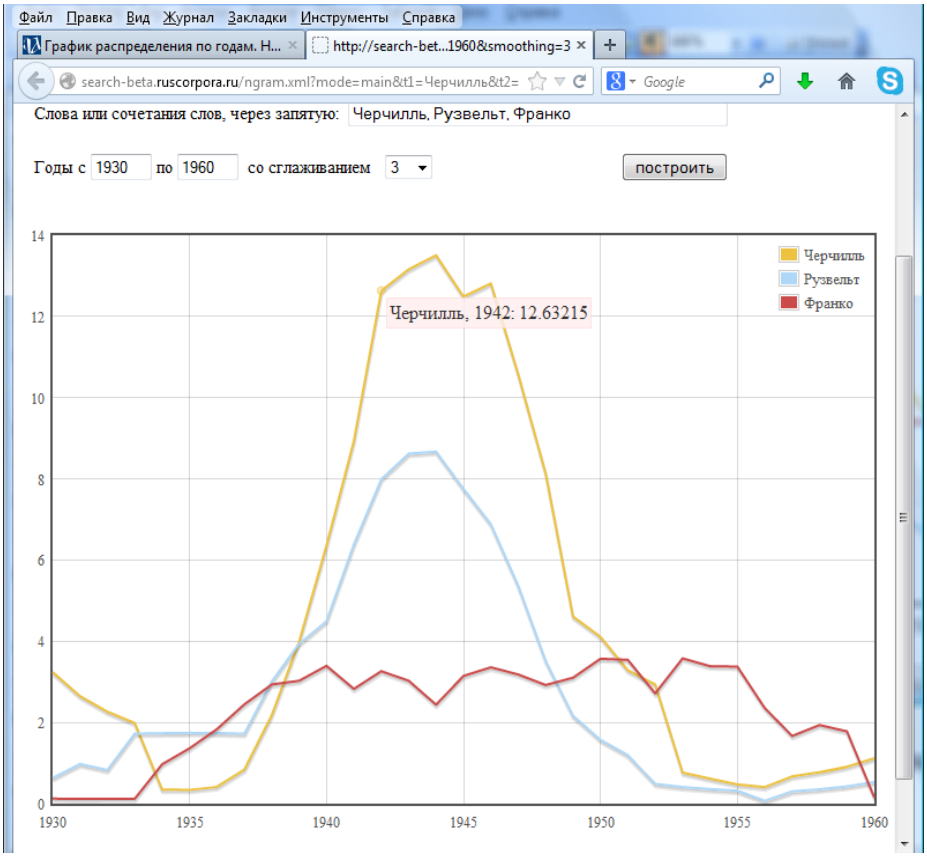

Рис. 2. График встречаемости имен Черчилль, Рузвельт, Франко в текстах, опубликованных с 1930 по 1960 гг.

Из сервиса «Графики» можно перейти на сервис Google Ngram Viewer, работающий на русскоязычной коллекции текстов Google Books. Однако при сходной идеологии, формулы подсчета относительной частоты в сервисах Национального корпуса и Google Ngram Viewer отличаются.

ЛИТЕРАТУРА

Баранов А.Н. Введение в прикладную лингвистику. – М., 2007.

Гвишиани Н.Б. Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. – М.: Высшая школа, 2008.

Грудева Е.В. Корпусная лингвистика: Учебное пособие. 2-е изд., стереотип. - Москва: Флинта: 2012.

Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика» (1-е изд., Иркутск: ИГЛУ, 2011; 2-е изд., перераб., 2012, URL: http://www.twirpx.com/file/1098705 , 3-е изд., перераб. и дополн., СПб.: СПбГУ, Филологический факультет, 2013).

Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005.

Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: Учеб. пособие. – М.: Издательский центр «Академия», 2004.

Камшилова О.Н. Учебный корпус текстов: потенциал, состав, структура. – СПб.: ООО «Книжный Дом», 2012.

Национальный корпус русского языка: 2003—2005. Сборник статей. М.: Индрик, 2005.

Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009.

Труды международной конференции «Корпусная лингвистика» (2004, 2006, 2008, 2011, 2013). Санкт-Петербург. – СПб.: СПбГУ. Филологический факультет.