КОРПУС КАК ЯЗЫК: ОТ МАСШТАБИРУЕМОСТИ К ДИФФЕРЕНЦИАЛЬНОЙ ПОЛНОТЕ

Беликов В. И. (vibelikov@gmail.com)
РГГУ, Москва, Россия

Копылов Н. Ю. (Nikolay_Ko@abbyy.com)
РГГУ; ABBYY, Москва, Россия

Пиперски А. Ч. (apiperski@gmail.com)
РГГУ, Москва, Россия

Селегей В. П. (Vladimir_S@abbyy.com)
РГГУ; МФТИ; ABBYY, Москва, Россия

Шаров С. А. (s.sharoff@leeds.ac.uk)
РГГУ, Москва, Россия; University of Leeds, Великобритания

Основным вопросом всякого корпусного исследования, будь то эксперименты с Интернетом, работа с НКРЯ или иным корпусом, должен быть вопрос об объекте наблюдения: изучается конкретный корпус, поисковая машина или собственно язык? К сожалению, почти всегда исследователь принимает в качестве не требующего доказательства предположения «масштабируемость» результатов частного корпусного исследования на весь язык. В статье рассматриваются критерии, с помощью которых разработчики корпусов обосновывают возможность такого масштабирования частных корпусных данных и предполагается новый подход к оценке границ действия обнаруженных исследователем фактов, принятый в рамках продолжающегося проекта создания Генерального интернет-корпуса русского языка (ГИКРЯ). Одним из базовых положений этого проекта является идея, что само масштабирование результатов является операцией весьма ограниченного применения. Для большинства лингвистических и лексикографических задач корпусной анализ должен проводиться с точностью до четко определенных жанровых и социолингвистических границ.

CORPUS AS LANGUAGE: FROM SCALABILITY TO REGISTER VARIATION

Belikov V. (vibelikov@gmail.com)
RSUH, Moscow, Russia

Kopylov N. (Nikolay_Ko@abbyy.com)
RSUH, ABBYY, Moscow, Russia

Piperski A. (apiperski@gmail.com)
RSUH, Moscow, Russia

Selegey V. (Vladimir_S@abbyy.com)
RSUH, ABBYY, Moscow, Russia

Sharoff S. (s.sharoff@leeds.ac.uk)
RSUH, Moscow, Russia; University of Leeds, UK

The main research question of any corpus investigation, either while experimenting with the Internet or working with the RNC or any other corpus, should be the question of the object of investigation: do we study a particular corpus, search engine or the language "overall"? Unfortunately, researchers usually accept as self-evident the assumption of "scalability" of the results obtained with a specific corpus study to the whole body of language. The article examines the criteria to justify the possibility to scale specific data and proposes an approach to assessing the limits of discovered facts, as adopted in the framework of an ongoing project to create the General Internet Corpus of Russian (GICR). One of the basic ideas of this project is that scaling the results is a very limited operation. For the majority of linguistic and lexicographical problems, corpus analysis should be carried out within a well-defined genre and sociolinguistic parameters.

О проекте ГИКРЯ[1]

Данная статья является продолжением работы [Беликов, Селегей, Шаров 2012], в которой обосновывалась необходимость запуска еще одного корпусного проекта для русского языка и определялись основные технологические принципы создания сверхбольшого корпуса на основании полностью автоматических методов сбора и лингвистической, и метатекстовой разметки.

[1] Проект ведется при финансовой поддержке Министерства образования и науки Российской Федерации (гос. контракт № 07.514.11.4142) и программы стратегического развития РГГУ.

В данной работе мы не касаемся вопросов технологии корпусного строительства. Рассматриваются только вопросы оценки достоверности корпусных исследований. ГИКРЯ в данный момент находится в фазе сбора первой тестовой версии, позволяющей тем не менее проводить отдельные лингвистические исследования, некоторые результаты которых будут представлены ниже.

Объем этой версии в данный момент составляет около 4 млрд слов, по составу текстовая версия ориентирована на исследования блогосферы (более 20 млн записей и комментарии к ним) и актуальной художественной литературы и публицистики (около 45 тыс. текстов объемом 250 млн слов).

Масштабируемость и сбалансированность

Скрупулезный анализ адекватности полученных результатов не стал еще, к сожалению, частью культуры корпусных исследований. Этот удивительный факт находится в разительном противоречии с популярностью и значимостью этих исследований в современной лингвистике.

Большое количество работ основаны на абсолютном доверии к полученному количественному результату, часто вся выводная часть строится на сопоставлении частот, полученных запросом по всему доступному корпусному материалу, будь то НКРЯ или Интернет. Типичное резюме выглядит следующим образом:

Материалом для исследования стали данные корпуса N. Количество вхождений для каждой из сравниваемых конструкций составило (приводятся несколько цифр, часто одного порядка). Далее следует вывод *о приоритете конструкции с большей частотой в исследуемом языке*. Вопросы, которые обычно остаются без внимания:

- сравнение данных по числу вхождений, документов и авторов;
- анализ временной динамики;
- анализ распределения результатов по типам источников (параметрам метатекстовой разметки);
- наличие дублетов или иных систематических факторов, «накручивающих» счетчик.

Обобщение частных корпусных результатов на весь язык (включая и отрицательные результаты!) является господствующей тенденцией, которую в значительной степени поддерживают сами создатели корпусов.

Если исследователь может найти в корпусе примеры на интересующее его явление, то его выводы во многом зависят от того, как позиционируется этот корпус. В названии некоторых корпусов содержится информация об их составе (например Michigan Corpus of Academic Spoken English), и пользователи таких ресурсов едва ли рискнут масштабировать полученные результаты на весь язык. Однако корпуса, позиционирующие себя как национальные, намного сильнее навязывают своим пользователям представление, что по ним можно делать выводы про язык в целом.

Обычно корпуса пытаются так или иначе обосновать эту претензию. Возможность масштабирования выводов на язык связывается с понятиями сбалансированности и представительности/репрезентативности, которые часто

упоминаются в описаниях корпусов. Иногда эти понятия рассматриваются как эквивалентные. Например, создатели Национального Корпуса Русского Языка (НКРЯ) указывают [НКРЯ, 2012]:

> *Национальный корпус ... характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус* ***по возможности пропорционально их доле в языке соответствующего периода*** [выделение наше].

Едва ли не текстуально совпадающие автохарактеристики можно найти на сайтах других национальных корпусов, например, болгарского и британского[2].

Как создатели корпусов доказывают наличие указанных свойств? Здесь успехи достаточно скромны. Даже НКРЯ, увы, «не защищен» солидными формальными доказательствами сбалансированности и репрезентативности. Имеется лишь одна работа, посвященная статистическому анализу НКРЯ в целом [Шаров, Ляшевская, 2009]. При создании частотного словаря, основанного на текстах этого корпуса, применялись определенные принципы подсчета частот, позволяющие доказать лексическую сбалансированность этого ресурса (сегментирование корпуса и подсчет частот с учетом равномерности распределения по сегментам). Но оценка сбалансированности в целом не была дана, было показано только, что относительные частоты в ядре языка устроены в НКРЯ близко к тому, что показывает Интернет.

Чтобы доказать свою представительность и сбалансированность, некоторые корпуса эксплицитно указывают принципы отбора текстов: например, Корпус современного американского английского языка — COCA за каждый год начиная с 1990-го включает примерно по 4 млн словоформ в устный, художественный, журнальный, газетный и научный подкорпуса. Другие корпуса просто сообщают свой состав как данность. Так поступает, например, Венгерский национальный корпус, который приводит информацию об объеме текстов в зависимости от их жанровой и географической принадлежности[3]. Наконец, есть корпуса, которые ограничиваются лишь общими рассуждениями о типологическом разнообразии и не указывают конкретных цифр (в их число входит НКРЯ).

В целом, апробированных научным сообществом способов обеспечения представительности и сбалансированности корпусов не предложено ([Hunston 2008]; [McEnery, Hardie 2011]). Эта тема активно обсуждается в компьютерной лингвистике (ср. [Biber 1993], [Leech 2007]), но пока что получается, что корпус

---

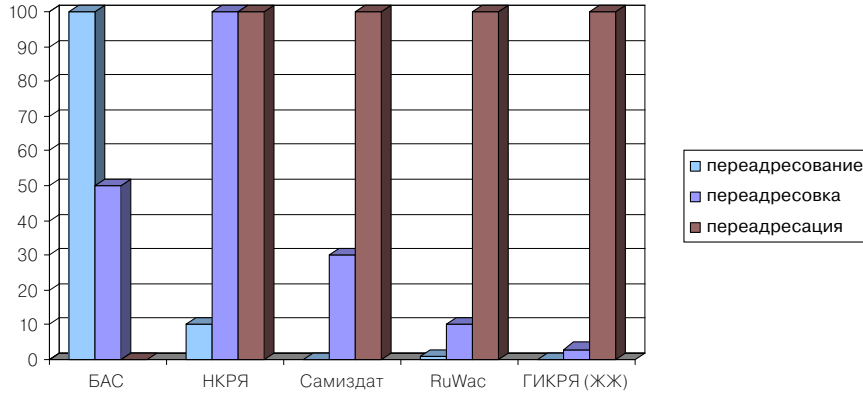
[2] http://www.ibl.bas.bg/BGNC_classific_bg.htm; http://www.natcorp.ox.ac.uk/corpus/index.xml.

[3] Эта информация суммирована в таблице 5×5 (**жанры**: пресса, литература, наука, официальные тексты, частная сфера; **регионы**: Венгрия, Словакия, Прикарпатье, Трансильвания, Воеводина).

считается представительным и сбалансированным тогда, когда на этот счет имеется негласный договор между его создателями и пользователями.

Масштабируемость и состав корпуса

Препятствием для масштабирования исследовательского результата на весь язык (что бы ни понималось под таковым — см. далее) является не только типологическая несбалансированность (несоответствие пресловутой доле в языке), но и случайность подбора текстов каждого типа, неизбежная при ограниченном объеме корпуса, а также лексическая неактуальность корпусного контента. На графике ниже показано примерное соотношение приведенных частот употребления 3-х вариантов именования понятия: *переадресование* (основной вариант по данным словарей[4]) — *переадресовка* — *переадресация*.



Эти данные, к анализу которых мы вернемся несколько позже, показывают, что два корпуса, претендующих на представительность, НКРЯ (ручная сборка) и RuWac (полностью автоматическая сборка), показывают весьма различающиеся данные, что плохо соотносится с идеей масштабирования[5]. Различие картин мира БАС («национального» толкового словаря) и блогов ГИКРЯ не требует комментария.

---

[4] В БАС (т. 15, 2011) и в словаре Ефремовой (2006) *переадресовка* отсылает к *переадресование*; в словаре Шведовой *переадресовка* помечена *разг*. *Переадресация* в толковых словарях отсутствует.

[5] Вообще говоря, количество вхождений слов в НКРЯ (до 20) делает результат в некотором смысле случайным, чувствительным к процедуре подсчета, но что самое важное, никак не демонстрирует процесс ухода варианта *переадресовка*.

Недокументированные особенности интернет-поиска

Максимально возможным приближением к полному *множеству* текстов является Интернет. Однако, как было показано в [Беликов, Селегей, Шаров 2012], Интернет не является *корпусом*, поскольку не предоставляет адекватных инструментов доступа и анализа содержащихся в нем текстов.

Это не означает, что интернет-поисковиками не нужно или тем более — нельзя пользоваться. Необходима аккуратная постановка эксперимента, позволяющая надежно верифицировать полученные результаты. Это возможно только при ясном понимании того, как этот результат получен.

Некритичное использование Интернета для лингвистических целей стало беспокоить и самих разработчиков этого ресурса, по крайней мере тех, кто имеет дело с лингвистами. Так, в лекции о принципах поиска Яндекса в Политехническом музее [Плахов 2012], было в очередной раз, но теперь — уже из самых первых уст сказано, что любые цифры, которые нельзя перепроверить вручную (до 1000 поисковых результатов) являются результатами не прямых подсчетов, а различных аппроксимаций. Не зная алгоритмов аппроксимации и их параметров, используемых на момент поиска, добросовестный лингвист не может полагаться на результаты поиска, которые часто можно рассматривать как случайные.

Мы не станем перегружать текст доказательствами этого тезиса, приведем только один новый пример, полученный авторами в ходе сравнения данных ГИКРЯ и Яндекса. В таблице ниже видны не только обычные нарушения аксиом арифметики и алгебры, но и совершенно неожиданные и не имеющую ничего общего с изучаемым явлением зависимости результата поиска от предполагаемого или явно указанного местонахождения исследователя (все скриншоты сохранены авторами):

	«украину»	**«на украину»**	**«в украину»**
Поиск от 12.08.2011 в Угловке Новгородской обл.			
без ограничения региона	136 млн	310 млн	321 млн
Поиск от 14.03.2013 в Петербурге			
без ограничения региона	3 млн	138 тыс.	196 тыс.
✓ в Санкт-Петербурге	2 млн	951 тыс.	2 млн
Поиск от 15.03.2013 в Москве			
без ограничения региона	5 млн	4 млн	14 млн
✓ в Москве	69 млн	3 млн	6 млн

Работая с поисковиками, исследователь использует в качестве рабочего инструмента то, что программисты назвали бы недокументированными возможностями (не описанными в официальных руководствах и свободными от обязательств поддержки). Это относится как к общим алгоритмам поиска и подсчета частот отдельных словоформ и коллокаций, так и к специальным инструментам «лингвистического» анализа, которые периодически появляются.

Например, в течение некоторого времени Яндекс поддерживал специальный инструмент «Пульс блогосферы». В феврале 2013 г. он был закрыт как

непопулярный. Между тем среди лингвистов он был довольно популярен. В готовящемся в изд. НЛО сборнике «Русский язык как глобальный ресурс и новые технологии» он независимо упоминается двумя авторами. Б. В. Орехов называет его «удобным электронным инструментом» для исследования микроистории лексики. М. А. Кронгауз, основываясь на данных «Пульса блогосферы», пишет о динамике популярности интернет-мемов. Между тем, популярность мема в соответствующей среде может расти, а его частотность в постоянно перестраивающейся в социальном отношении блогосфере — падать. При этом принципы подсчета частоты совершенно скрыты от пользователя[6].

Масштабируемость как цель

Таким образом, в основе представлений о возможности масштабирования результатов поиска в корпусе лежат три идеи о его контенте:

1. о его сбалансированности (остается, как мы показали, совершенно неформальным свойством корпуса);
2. о его актуальности;
3. о его общем количестве, достаточном для генерализации выводов (также совершенно неформальный критерий).

Но что именно получает ведомый идеей масштабируемости исследователь, работая пусть даже с идеальным корпусом, удовлетворяющим этим условиям?

Он полагает, что можно обращаться к корпусу как к самому языку с вопросами, ответы на которые будут относиться также ко всему языку (получая нечто вроде усредненной частоты явления в языке).

Но о каком потенциальном множестве текстов идет речь, когда создатели корпуса говорят о «доле в языке» некоторого типа, представленного в корпусе? Является ли идеалом сбалансированности гипотетический полный корпус текстов, содержащий все написанное на данном языке в некотором временном промежутке (с точностью до фактов устной речи, требующей фиксации иного типа)?

Хорошо известно мнение некоторых лингвистов о том, что «Интернет — это большая помойка, в нем полно неграмотных текстов». В этом случае Интернету противопоставляется некоторый свод текстов на Правильном Русском Языке (ПРЯ) со своей идеальной картиной типового распределения. Получение корпуса этого ПРЯ является задачей, достойной с какой угодно точки зрения, но не с точки зрения лингвистики и лексикографии.

Заметим, что методы недифференцированного анализа корпуса часто применяются в компьютерной лингвистике, где условная сбалансированность позволяет создавать модели, эффективные в среднем (обученные на предполагаемой в деятельности пользователей жанровой и тематической взвеси).

---

[6] Объем статьи не позволяет входить в детали, для интересующихся более полные данные опубликованы в расширенном варианте статьи на сайте конференции Диалог.

На наш взгляд вопрос о «процентном составе языка» является лингвистически совершенно бессмысленным, и любые полученные цифры отражают процессы, имеющие малое отношение к лингвистике и даже социолингвистике. Так, если в практике создания текстов носителями языка в данном историческом периоде на один протокол осмотра происшествия приходится 0,01 некролога, 10 медицинских заключений и 1000 «объявлений о знакомстве», вряд ли разумно требовать от корпуса, чтобы и в нем соблюдалось то же соотношение.

Более «справедливая» модель отбора, известная со времен Ноева ковчега, также порождает при запросах «в среднем» совершенно неадекватную картину.

Кроме того, относительно самого набора «корпусных тварей» пока не наблюдается ни малейшего единодушия [Sharoff, 2010]. Количество категорий классификации текстов варьирует от 15 (Брауновский корпус), 70 (БНК), 181 (НКРЯ), 349, отобранных в исследовании предпочтений пользователей [Crowston, Kwasnik, Rubleske, 2010] до более 4 тысяч [Adamzik, 1995].

О неоднородности корпусных данных

Принятие гипотезы о масштабируемости корпуса порождает исследовательские ошибки не только из-за несбалансированности, недостаточности объема и прочих бед, о которых было сказано. Самой большей проблемой является то, что даже идеально сбалансированный корпус будет приводить исследователя к неверным выводам, если не учитывается принципиальная неоднородность языковых данных. Некоторая типология явлений, требующих дифференциального подхода была дана в работе [Беликов, Селегей, Шаров 2012].

К сожалению, сегодня дифференциальные исследования остаются на периферии интересов лингвистов и лексикографов. Одно из немногих исключений — масштабное исследование региональной вариативности в русском языке в проекте «Языки русских городов» (http://community.lingvo.ru/goroda/, [Беликов 2010]).

Удивляет то, что недифференциальный подход свойственен даже многим работам по собственно вариативности. Исследователи рассматривают дивергенцию в языке как некоторое общеязыковое свойство, не пытаясь анализировать параметры этой вариативности. Например, на сайте НКРЯ объявлен научный проект по изучению вариативности «Проблемы русской стилистики по данным НКРЯ» (http://studiorum.ruscorpora.ru/stylistics/intro.html). Приведены несколько дающих проектные установки работ, остановимся только на одной: М. В. Шкапа «Склонение топонимов на *-ово (-ево), -ыно (-ино)*». Автор подсчитывал количество вхождений в текстах за 2000–2010 гг., работая с корпусом в 50,0 млн слов. Уже то, что автор работает с вхождениями, а не документами, резко снижает ценность исследования. Так, один только появившийся после исследования текст Льва Дурнова «Жизнь врача» на 40% увеличил число склоняемых вхождений топонима *Перово*. Кроме того, газетный корпус НКРЯ составлен

из публикаций семи изданий, в разном объеме и за разные годы; редакционная политика изданий заметно влияет на результаты. Все это указывает на то, что решаются не проблемы *русской стилистики*, а проблемы стилистики *конкретного собрания текстов*.

Но важнее всего то, что многие такие частные проблемы решать «в среднем по корпусу» просто не следует. Так, даже беглый анализ вариативности в склонении топонимов по данным блогосферы показывает, что на нее влияют самые разные факторы. Прежде всего — региональные (см. версию статьи на сайте). Имеют место различия в выборе варианта для своего и чужого топонима, и т.д. Для усмотрения каких-то тенденций и параметров, которые на них влияют, нужен существенно больший материал, чем имеется в НКРЯ.

Мы провели небольшое исследование по сравнительному анализу двух условно сопоставимых по жанровой смеси и объему корпусов: художественных текстов НКРЯ и журнального подкорпуса ГИКРЯ, полученного на основании текстов Журнального зала (magazines.russ.ru).

Помимо упомянутого выше анализа употребления тройки «*передресация* и т. п.», исследовалась структура значения слова *окучивать* по данным этих корпусов. Выяснилось, что на собранном вручную (НКРЯ) и автоматически (ГИКРЯ) корпусах обнаруживаются заметные отличия, связанные, например, с обилием остросюжетных текстов в современной части НКРЯ, представленных в основном романами (при поиске *окучивать* их оказалось более половины) и известной «однобокостью» поэтического корпуса НКРЯ[7]. В результате в ГИКРЯ среди 142 вхождений глагола *окучивать* 5,5 % оказывается в поэзии, на традиционное агротехническое значение падает 65 % всех вхождений, а в НКРЯ они составляют лишь 45 % от 42 вхождений.

Причин этих отличий несколько: от серьезного влияния невыявленных дублетов (к каковым в случае НКРЯ относятся, например, многократные повторы об условиях подписки на издание), до влияния неслучайности отбора (тут могут быть важны и соображения доступности, авторских прав и т. п.).

Выводы просты: создатели корпусов должны добиваться независимости любого подкорпуса, выделенного по какой-либо группе параметров, например, жанровой, от процедуры отбора корпусного материала. То есть, корпус должен обнаруживать устойчивость в результатах анализа по тем свойствам, которые являются объектом исследования. В корпусах небольшого объема добиться такой устойчивости можно только резким сужением круга изучаемых явлений.

От сбалансированности к дифференциальной полноте

Как показано выше, масштабируемость является не универсальным свойством корпуса, а частным следствием из доказанной однородности корпусных данных относительно данного языкового факта.

---

[7] 203 наболее поздних текста, датированные периодом (1981–1995), принадлежат трем авторам 1903–1907 г. р.: И. В. Чиннову (126), Г. А. Глинке (74) и А. А. Штейнбергу (3).

Для того, чтобы вывод о масштабируемости был надежным, нужен корпус с особыми свойствами. Вместо понятия сбалансированности мы предлагаем использовать понятие *дифференциальной полноты* корпуса.

В отличие от понятия сбалансированности-репрезентативности, дифференциальная полнота означает не просто предположительную типологическую полноту корпуса, но и

- явное указание типа за счет метатекстовой разметки и использование типологических данных при обработке запросов;
- доказуемые предположения относительно необходимого и достаточного объема текстов каждого типа в корпусе.

В дифференциально полном корпусе результат обработки любого запроса может быть разложен по типологическим координатам и оценен с точки зрения однородности. Разумеется, при желании исследователь может дать веса типам для того, чтобы получить какие-то обобщенные данные, соответствующие его представлению о составе языка.

Вопрос о формальных критериях дифференциальной полноты корпуса не является решенным. Собственно говоря, это является одной из основных «математических» целей проекта ГИКРЯ.

Мы можем говорить о достаточной дифференциальной полноте корпуса относительно некоторого языкового явления в следующих смыслах:

- при наличии метатекстовой разметки (аналог обучения с учителем, supervised learning), которая сама может быть порождена автоматическими методами с разной степенью надежности;
- при отсутствии такой разметки, как оценка максимального варьирования параметров в рамках автоматически определенных кластеров (аналог обучения без учителя, unsupervised learning).

В первом случае, например, дифференциальная полнота по регионам определяется наличием в корпусе статистически значимого количества текстов, авторы которых происходят из представительного списка регионов; полнота по жанрам предполагает наличие значимого количества текстов по всем жанрам, которые должны быть представлены в корпусе.

С другой стороны, этот вид полноты предполагает, что все категории классификации известны нам заранее. Это явно не имеет места для жанровой классификации, и даже для классификации региональной мы не можем быть априорно уверены в значимости параметра дробности для отдельных категорий. Мера полноты может быть оценена выбором признаков (вся лексика, избранная лексика ключевых слов, частеречные коды и т.п.), кластеризацией корпуса и оценкой различий между различными кластерами [Sharoff, 2007].

При этом существенным для определения значимых различий между сегментами корпуса (либо определенными по метатекстовой разметке, либо кластеризацией) является:

- выбор языкового явления, которое измеряется в числовом выражении и может быть надежно извлечено автоматическими методами на большом корпусе (обычно это частота леммы, словоформы или грамматической конструкции);
- количество примеров этого явления в рамках отдельных текстов каждого из сегментов (частота по документам по сегментам);
- числовое выражение явления в рамках документов в сегменте.

В любом случае, полнота представлена не как абсолют, а как относительное понятие: один корпус полнее другого (или дает более достоверную картину) в рамках того или иного языкового явления, либо два корпуса более похожи по степени своей полноты в сравнении с третьим.

Рассмотрим пример для региональных частот. С точки зрения статистики мы исходим из предположения («основной гипотезы» в терминах статистики), что два региональных сегмента в сравнимых жанрах не отличаются по использованию лексемы X. Мы можем оценить:

1. статистическое распределение частот употребления этой лексемы в текстах каждого региона и
2. определить значимость «альтернативной гипотезы» о различии в этих частотах.

Пункт 1 считается как вероятность по документам (частота в документе, деленная на его размер), в результате мы получаем распределение частот по текстам каждого региона. Если f_d — частота термина в документе d, а N_d — размер этого документа, то:

$$p_d = \frac{f_d}{N_d}$$

Пункт 2 считается стандартными методами аппроксимирования альтернативной гипотезы, например, используя t test распределения вероятностей, либо коэффициент логарифмического правдоподобия отклонений абсолютных частот (a и b) в двух сегментах (размерами соответственно c и d), который учитывает как абсолютные значения, так и относительные размеры каждого сегмента:

$$E1 = c\frac{(a+b)}{(c+d)}$$

$$E2 = d\frac{(a+b)}{(c+d)}$$

$$G^2 = a\log\frac{a}{E1} + b\log\frac{b}{E2}$$

В результате сравнение разных результатов для разных регионов или феноменов может быть оценено с точки зрения вероятности того, что «альтернативная гипотеза» верна не случайным образом. Например, значение G^2 большее 3,84 дает 95 % вероятность того, что разница не случайна [Rayson, Garside, 2000].

Заключение

Основным содержательным результатом проекта ГИКРЯ через год после начала систематической работы авторов над этим проектом можно считать появление первой тестовой версии корпуса, позволяющей верифицировать результаты корпусных исследований (проводимых как в рамках существующих корпусов РЯ, так и в Интернете). ГИКРЯ пока едва достиг 10 % планируемого объема, но уже сейчас дает исследователю результаты, каждый шаг получения которых является совершенно прозрачным. Другим важным результатом является формирование понятия дифференциальной полноты. Что касается ее вычислимых критериев, тут предстоит еще большая работа, требующая нескольких итераций сбора корпуса и коррекции используемых классификаторов. Важным результатом станет возможность тестирования лингвистической устойчивости корпуса при расширении его состава и возможность объективного сравнения свойств разных корпусов.

Литература

1. *Adamzik, K.* 1995. Textsorten — Texttypologie. Eine kommentierte Bibliographie, Nodus, Münster. http://www.unige.ch/lettres/alman/adamzik/akt/naments.pdf
2. *Biber D.* (1993), Representativeness in Corpus Design, Literary and Linguistic Computing, Vol. 8, No. 4, pp. 243–257
3. *Crowston, K., Kwasnik, B., Rubleske, J.* 2010. Problems in the use-centered development of a taxonomy of web genres. // Mehler, A., Sharoff, S., Santini, M., editors, Genres on the Web: Computational Models and Empirical Studies. Springer, Berlin/New York.
4. *Hunston S.* (2008), Collection strategies and design decisions, in A. Lüdeling and M. Kyto (eds.), Corpus linguistics: an international handbook, Walter de Gruyter, Berlin/New York, pp. 154–168.
5. *Leech G.* (2007), New resources, or just better old ones? The Holy Grail of representativeness, in M. Hundt, N. Nesselhauf and C. Biewer (eds.), Corpus Linguistics and the Web, Amsterdam, Rodopi, pp. 133–49.
6. *McEnery T., Hardy A.* (2011), Corpus linguistics, Cambridge University Press, Cambridge, 2011.
7. *Rayson, P. and Garside, R.* 2000. Comparing corpora using frequency profiling. // Proceedings of the workshop on Comparing Corpora at ACL 2000, Hong Kong.

8. *Sharoff, S.* 2007. Classifying Web corpora into domain and genre using automatic feature identification. // Proc. of Web as Corpus Workshop, Louvain-la-Neuve.
9. *Sharoff, S.* 2007. Creating General-Purpose Corpora Using Automated Search Engine Queries.
10. *Sharoff, S.* 2010. In the garden and in the jungle: Comparing genres in the BNC and Internet. In Mehler, A., Sharoff, S., Santini, M., editors, Genres on the Web: Computational Models and Empirical Studies, Springer, Berlin/New York.
11. *Беликов* 2010. Методические новости в социальной лексикографии XXI века // Slavica Helsingiensia 40. Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian. Helsinki, 2010 / Ed. by A. Mustajoki, E. Protassova, N. Vakhtin.
12. *Беликов В. И., Селегей В. П., Шаров С. А.* 2012. Пролегомены к проекту Генерального интернет-корпуса русского языка. // Труды конференции Диалог 2012.
13. *НКРЯ. Что такое Корпус?* Сайт Национального корпуса русского языка. http://www.ruscorpora.ru/corpora-intro.html. 2012 г.
14. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка Издательство Азбуковник, 2009. Вступительная статья http://www.dialog-21.ru/digests/dialog2008/materials/html/53.htm.
15. *Плахов А.* Системы поиска в Интернете: как обрабатывается запрос пользователя. Лекция в политехническом музее 16.10.12. http://pmlectures.ru/video/Cistemy_poiska_v_Internete_kak_obrabatyvaetsya_zapros_polzovatelya-38