ДИФФЕРЕНЦИАЛЬНАЯ КОРПУСНАЯ СТАТИСТИКА НА ОСНОВАНИИ НЕАВТОМАТИЧЕСКОЙ МЕТАТЕКСТОВОЙ РАЗМЕТКИ

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

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

Статья основывается на исследовательских работах, проводящихся в рамках проекта создания Генерального Интернет-Корпуса Русского Языка (ГИКРЯ). В настоящее время одной из самых актуальных задач, ждущих своего решения в проекте, является автоматическая метатекстовая разметка. Тем не менее, в первую пробную версию корпуса включено большое количество материала, позволяющее проводить дифференциальный статистический анализ на большом объеме размеченных данных из разных сегментов интернета уже сейчас.

В настоящее время растет понимание того, что объем данных в ручных корпусах недостаточен для многих типов лингвистических исследований. При этом идея, что не только размер имеет значение, еще не стала настолько же популярной. В данной работе мы пытаемся показать критическую важность дифференциального анализа материала в корпусах-миллиардниках.

VARIATIONAL CORPUS STATISTICS USING AUTHOR PROFILES

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

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

This paper is based on research carried out in the framework of our project on the General Internet Corpus of Russian (Geekrya) . The need to use large-scale corpora automatically collected from the Web was first recognized in computational linguistics. Recently, the lack of data in "manually-built" corpora led to recognition of the importance of Web-derived corpora in traditional linguistic research.

The principal difference of Geekrya from the two other large web corpora of Russian (RuWac and RuTenTen) is that the latter were produced by indiscriminate crawling of the Russian Internet, resulting in no metatext markup available for their data.

GEEKRYA is different since its contents is split into "segments" which we define as a compact set of webpages sharing a general communicative purpose expressed in text-rich content. We extracted information about the authors from their profiles when this was specified.

The total size of indexed Geekrya amounts to 12 billion words, the segments with known a priori metatext parameters are listed below (size given in millions of words).

Segment	Gender	Age	Region
blogs.mail.ru	164	81	113
livejournal.com	0	1,800	5,600
vk.com	2,000	1,600	1,600
news	0	0	0
magazines.russ.ru	258	0	0
forums (adw.ru)	163	0	0
Total:	2,585	3,481	7,313

The magazines.russ.ru segment, for example, contains all the texts from this resource (mostly published fiction and literary criticism). Author's gender has been extracted for 84.3% its texts, the size of the male subcorpus is—194 MW, the female one is 64 MW.

Information about the author' profiles within the individual segments helps in variational analysis. The paper lists several studies on the gender profiles of discourse words, collocations and idioms, as well as on the regional distribution, for example, comparing word uses in Siberia against the rest of the Russian-speaking world.

Мегакорпуса русскоязычного интернета

Необходимость привлечения к языковому анализу автоматически собранных корпусов была осознана сначала в компьютерной лингвистике. Но в последнее время нехватка данных в корпусах «ручной сборки» стала заметной и для авторов многих собственно лингвистических исследований. При этом речь идет о следующих проблемах, связанных с объемом, разнообразием и способом разметки текстов в корпусе:

1. собственно недостаточный объем данных;
2. недостаточный объем лингвистически размеченных данных;
3. недостаточный объем и типологическое разнообразие текстов с метатекстовой разметкой по различным релевантным параметрам варьирования (т. н. дифференциальная неполнота [Belikov, Kopylov, 2013]).

Потенциально исчерпывающе полный источник данных — интернет, не является корпусом сам по себе, поскольку не имеет разметки и не дает сколько-нибудь точной статистики даже по поддерживаемым ограниченным видам запросов. Поэтому с середины нулевых годов бурно развивается направление WAC (Web As a Corpus), целью которого является получение автоматически размеченных мега-корпусов из интернета.

Примерами таких корпусов для русского языка являются RuWac [Sharoff, Nivre, 2011] и недавно появившийся, но уже получивший популярность у исследователей, сделанный на основе SketchEngine Адама Килгариффа корпус RuTenTen, содержащий более 10 млрд слов [Jakubíček, 2013] и собственно разрабатываемый авторами ГИКРЯ_1.0 [Belikov, Piperski, 2013].

Все три корпуса используют сейчас одну и ту же процедуру первичной морфоразметки (таггер С. Шарова [Sharoff, Nivre, 2011]). RuTenTen и ГИКРЯ_1.0 являются близкими по объему (хотя в плане развития ГИКРЯ установлен ориентир в 50 млрд слов и более).

Принципиальным отличием ГИКРЯ_1.0, о котором и идет речь в данной статье, является то, что два других интернет-корпуса (RuWac и RuTenTen) получены в результате «слепого» кроллинга русскоязычного интернета, что не позволяет решить проблему дифференциальной полноты. В частности, в них полностью отсутствует метатекстовая разметка. Кроме того, по соображениям эффективности кроллинг в таких корпусах ограничивается ресурсами в домене .ru, представленными текстами в HTML без интерфейса, что исключает многие полезные ресурсы, в том числе и такие важные, как vk.com или blogs.mail.ru.

ГИКРЯ с точки зрения закладываемой в корпус информации отличается тремя важными особенностями:

1. неслучайным способом набора корпуса с учетом сегментной структуры интернета;
2. более жесткой процедурой отбора и очистки страниц, включая их декомпозицию на отдельные объекты анализа в случае структурной неоднородности (например, пост — комментарии);
3. максимально широким использованием априорной метатекстовой разметки, которую можно получить аккуратным анализом источников корпусных данных (прежде всего — разбором профилей авторов).

Эти отличия в построении корпуса оказываются очень существенными в отношении результатов запросов к нему в сравнении с другими автоматическими интернет-корпусами.

Сегменты интернета и система метатекстовой классификации

Было показано [Беликов 2006; 2010 и др.], насколько различаются результаты лингвистического анализа (прежде всего в области лексикографии и лексикализованного синтаксиса), если перейти от усреднения данных к учету распределения авторов текстов по различным параметрам.

При этом в отсутствии адекватных инструментов подобный анализ производился с помощью требовавшего колоссальных усилий «ручного» анализа данных, полученных помощью интернет-поисковиков.

Появление мегакорпусов с метатекстовой разметкой открывает дорогу к получению подобных результатов «легко и непринужденно».

При этом остается открытым вопрос, как добиться дифференциальной полноты данных в интернет-корпусе при наличии большого числа параметров социолингвистического и жанрового варьирования [Belikov V., Kopylov N., 2013].

Если по социолингвистическим и региональным параметрам априорная разметка может быть получена из самих данных, то жанровая полнота требует как минимум ясного понимания устройства жанровых категорий, чего, увы, не наблюдается [Crowston, 2010]. Эта цель может быть достигнута не сразу, а в результате нескольких итераций создания корпуса с параллельно ведущейся работой по созданию надежной системы параметров жанровой разметки и методов автоматической жанровой классификации.

В настоящее время в проекте ГИКРЯ идет такая работа (см. статью [Сорокин, Катинская, 2014] в данном сборнике).

И пока такая система разрабатывается, эффективным способом обеспечить относительную типологическую полноту корпуса на первых этапах его сбора является опора на сегментную структуру интернета. Мы добиваемся полноты сегментной структуры корпуса, коррелирующей с жанровой полнотой.

Сегментом интернета мы называем [Belikov V., Selegey V. 2012] компактное множество страниц Интернета, объединенное некоторой общей коммуникативной целью создателей text-rich контента. При этом диапазон варьирования социолингвистических характеристик авторов внутри таких сегментов может

быть весьма значителен. К таким функционально однородным «авторизованным» сегментам интернета, которые мы используем для сбора первой версии ГИКРЯ относятся:

- Блоги
- Микроблоги
- Форумы
- Социальные сети
- Сайты, аккумулирующие авторизованные тексты художественного и публицистического характера
- Энциклопедические ресурсы в которых имеются авторы статей (а не анонимные группы авторов и редакторов, как в Википедии)
- Новостные ресурсы.

Различие в структуре и принципах отбора сравниваемых интернет-корпусов очевидно из таблиц 1 и 2. Как мы видим, типичный подход создателей интернет-корпуса состоит в достижении некоего усреднения языковой картины с помощью набора данных из максимально большого числа источников (многих тысяч!). В проекте ГИКРЯ подход принципиально отличается: сбор корпуса происходит именно в соответствии с априорной схемой сегментной структуры интернета.

Таблица 1. Доменный состав интернет-корпусов RuWac и RuTenTen

RuWac		**RuTenTen**	
Число документов	Домены	Число документов	Домены
323 300	livejournal.com	114 427	spb.ru
37 860	narod.ru	82 606	narod.ru
8 293	lib.ru	39 649	tomsk.ru
6 117	germany-rest.com.ua	38 383	org.ru
5 966	bibliotekar.ru	34 752	net.ru
5 862	sites.google.com	32 419	gov.ru
5 844	subscribe.ru	22 668	ucoz.ru
5 755	shkolazhizni.ru	19 433	karelia.ru
5 423	lenta.ru	19 227	mos.ru
5 297	russ.ru	18 598	edu.ru
4 814	hotmail.ru	18 372	com.ru
3 602	football.hiblogger.net	17 736	nnov.ru
3 528	yandex.ru	17 623	msu.ru
3 491	org.ua	16 961	perm.ru
3 487	spb.ru	15 597	rospotrebnadzor.ru
3 478	eka-mama.ru	13 986	forum2x2.ru
3 462	mail.ru	13 490	msk.ru
3 171	falppo09.ru	13 228	rfn.ru
3 160	org.ru	12 910	academic.ru
...	...	...	...
Всего: 2 млн		Всего: 35 млн	

Таблица 2. Сегментная структура ГИКРЯ_1.0

Сегмент	Слов (млн.)	Документов
Блоги мейл.ру (комментарии к топикам)	186	~6 млн
Живой журнал	7900	~ 50 млн
В контакте	2000	~100 млн
Журнальный Зал	306	56 тыс.
Новости	700	~ 2 млн
Форум awd.ru	190	~2 млн
Всего	11 282	>100 млн

Комментарии к таблице 2

1. С развитием блогосферы в ней все шире развивается новостной и иной (гороскопы, кулинарные рецепты, «мудрые притчи», анекдоты и т. п.) репостинг, который для лингвиста представляет собой шум. Дублирование записей, нередко достигающее нескольких тысяч, особенно характерно для первичных записей в блогах мейл.ру. Поэтому в версию 1.0 ГИКРЯ с этой платформы включены лишь комментарии. В дальнейшем предполагается неновостной репостинг этой и иных блоговых платформ выделить в отдельные тематические сегменты, что, в частности позволит оценить их поло-возрастную привязку.
2. В первую версию ГИКРЯ не входят в качестве сегментов тексты, представляющие т.н. языки для специальных целей (LSP/ЯСЦ). В узком понимании это языки науки, но при широком термин покрывает и любое профессиональное, и религию с эзотерикой, и самые разные «клубы по интересам», фан-движения и т.п. В той или иной, но сильно разной степени эти языки находят реализацию в интернете. Многое реализуется в «жанре» форума, но имеется и огромное число специализированных сайтов. Обеспечение дифференциальной полноты по «тематическим» параметрам является отдельной и очень сложной задачей, но она не является все же первоочередной с точки зрения исследования языковой вариативности.
3. Также пока не учитываются частично ортогональные различия по социолингвистическим характеристикам типичного адресата. Адресат и адресант в некоторых сегментах идентичны, но есть много сайтов со специализированным адресатом: детским, подростковым, женским (отдельно — *мамским*, как сейчас говорят, то есть адресованных беременным и матерям грудных младенцев).

С точки зрения пользователя корпуса отличие ГИКРЯ состоит в том, что можно задавать сегмент в качестве параметра запроса, реализуя тем самым «как бы жанровое» ограничение на исследуемый материал. В случае RuWac и RuTenTen такое ограничение задать практически невозможно (что в отсутствие и любых других параметров метатекстовой разметки делает задачу получения дифференциальной выдачи невозможной).

Первая версия ГИКРЯ с точки зрения объема дифференцированных данных

Априорная информация об авторах, связанная с отдельными сегментами, дает (в не всегда достижимом идеале) возможность проводить дифференциальный анализ по следующим социолингвистическим параметрам:

- Возраст
- Пол
- Регион
- Образовательный уровень

Объем априорной социолингвистической разметки в ГИКРЯ_1.0 (11 282 млн слов на 1.04.14) представлен в табл. 3.

Комментарии к таблице 3

1. Суммарный объем проиндексированных данных в ГИКРЯ к июню 2014 года составит около 12 млрд слов, что соответствует примерно 20% того объема, который первоначально планировалось иметь в корпусе в конечном итоге. Более точные оценки, связанные с расчетом дифференциальной полноты по релевантным параметрам, можно будет дать несколько позже.
2. Относительно невысокая скорость набора данных в корпус объясняется необходимостью проведения двух операций, требующих участия программистов и лингвистов, анализирующих очередной сегмент интернета:
   - разработки соответствующего метода очистки страниц (удаления обвязки);
   - извлечения данных об авторах
3. Даже при текущем объеме первой версии ГИКРЯ количество документов с априорной разметкой исчисляется десятками миллионов с общим объемом в несколько миллиардов словоупотреблений. На таком объеме материала уже можно основывать серьезные дифференциальные исследования.

Таблица 3

Сегмент	Пол	Возраст	Регион
Блоги мейл.ру	164	81	113
Живой журнал	0	1800	5600
Вконтакте	2000	1600	1600
Новости	0	0	0
Журнальный зал	258	0	0
Форумы adw.ru	163	0	0
Всего:	2585	3481	7313

О надежности априорной разметки

Априорная разметка в некоторых сегментах корпуса безусловно не является абсолютно надежной. Она колеблется для разных сегментов интернета в диапазоне от 85 до 95 % в зависимости от исследуемого параметра.

Исследования по автоматической социолингвистической разметке различных социальных сетей, и отдельные работы, посвященные оценке достоверности авторских данных (прежде всего, в Twitter, например [D. Nguyen, 2013]) показывают, что имеются существенные систематические факторы, сдвигающие такие данные. Кроме того, не является очевидным, на каких шкалах нужно производить такие оценки в случае выставления автоматических и смешанных признаков в корпусе. В частности, имеются аргументы в пользу небинарных гендерных шкал [J. Lorber, 1996].

В этой статье мы не будем касаться вопросов, связанных с анализом процессов «самопозиционирования» авторов интернета, а также сравнением результатов априорной и автоматической социолингвистической и региональной атрибуции авторов текстов.

Для целей этой статьи достаточно указать, что имеется достаточно высокая корреляция между реальными данными (насколько они извлекаются из самих текстов), авторской самоидентификацией и результатами автоматической классификации.

В целом это позволяет изучать как собственно объективные корреляции между языковыми и социолингвистическими параметрами, так и систематические девиации между прогнозируемыми и априорными характеристиками.

Таким образом, использование априорной разметки дает достаточно надежные результаты (с погрешностью в единицы процентов). При возрастном анализе языка блогосферы для единиц, употребление которых существенно зависит от возраста, у лиц 12–69 лет «статистические результаты обычно хорошо укладываются в „правильную“ картину, что позволяет предполагать серьезное преобладание здесь тех, кто указал фактический возраст. ‹…› При анализе возрастного распределения сниженной лексики возраст, начиная с которого получаемые данные становятся недостоверными, снижается» [Беликов 2012], в таких случаях использовать данные о лицах старше 60 нецелесообразно.

Примеры дифференциального анализа запросов в ГИКРЯ по различным параметрам

Гендерное варьирование

В последнее время гендерному анализу социальных сетей уделяется большое внимание. Можно выделить три основных направления:

- гендерная лингвистика (просто гендерная вариативность)
- гендерная психолингвистика (попытки интерпретации)
- гендерная реклама: тут важны не собственно лингвистические отличия, но любые признаки, позволяющие выявить индивидуальные предпочтения пользователя.

Последнее направление преобладает, поскольку его продвигают рекламодатели. Гендерная атрибуция основывается на побочных признаках, но зато пополняет массив атрибутированных текстов.

Журнальный сегмент ГИКРЯ (ЖС) содержит все собственно журнальные тексты Журнального зала (ЖЗ) по состоянию на апрель 2014 г. (56 тыс. текстов, 306,0 млн словоупотреблений); в этом сегменте допускается классификация по конкретным изданиям, годам публикации и создателям (создателем текста считается его автор или переводчик).

В дальнейшем предполагается пополнять ЖС новыми публикациями в ЖЗ, текстами тех изданий, которые на собственных сайтах представлены полнее, чем в ЖЗ, а также систематически включать доступные в интернете в оцифрованном виде публикации толстых журналов, не входящих в ЖЗ, как «центральных» («Москва», «Юность» и др.), так и региональных — «Вологодская литература» (Вологда), «Дарьял» (Владикавказ), «Бельские просторы» (Уфа), «Дальний Восток» (Хабаровск) и др.

В настоящее время главной задачей обработки этого сегмента является подготовка полноценных профилей создателей текстов (пол, год рождения, региональная привязка), что связано со значительным объемом ручной работы. В ходе тестовых поисков по ЖС наиболее интересными оказались результаты анализа гендерных различий в узусе.

Поиск с учетом пола проводится на 84,3 % текстов ЖС, общий объем мужских словоупотреблений — 194,2 млн, женских — 63,8 млн. Мужской подкорпус превышает женский в 3,04 раза, то есть соотношение гендерно нейтральных единиц должно быть близко к 3.

На этом подкорпусе тестировалась гендерная предпочтительность различных дискурсивных слов, коллокаций, фразеологизмов. Невысокие абсолютные цифры не позволяют делать серьезных выводов, но там, где выдача составляет тысячи вхождений, различия явно достоверны, ср. «феминизированность» выражения *каждый раз* и «маскулинизированность» по *меньшей мере* в Табл. 4:

Таблица 4

Выражение	**M**	**F**	**M/F**
стерпится слюбится	31	13	2,4
как снег на голову	105	44	2,4
каждый раз	**3657**	**1475**	**2,5**
как заведенн(ый/ая/ые)	224	84	2,7
всего лишь	5954	1998	3,0
не на шутку	648	219	3,0
наверняка	4931	1667	3,0
по крайней мере	6819	2178	3,1
как правило	5291	1711	3,1
авось (кроме *на авось*)	670	209	3,2
в несколько раз	419	131	3,2

Выражение	**M**	**F**	**M/F**
небось	1622	498	3,3
в полном разгаре	68	20	3,4
по меньшей мере	**1533**	**438**	**3,5**
почем зря	225	54	4,2

Остановимся детальнее на двух случаях гендерного противопоставления в узусе.

1. Выявились различия при описании количества: импрессионистическое описание (типа *очень много*, *так много*) почти нейтрально, хотя несколько более свойственно женщинам, а сопоставительное (типа *заметно больше*, *в … раз больше*) явно оказывается более мужским, ср. данные в Табл. 5:

Таблица 5

	муж.	**жен.**	**муж./ жен.**
очень много	3278	1090	3,01
так много	3204	1267	2,53
раз больше	536	110	4,87
раза больше	460	107	4,30
чересчур много	115	23	5,00
существенно больше	63	20	3,15
заметно больше	38	10	3,80

Абсолютные цифры в последних трех строках невелики, и их пока не стоит принимать во внимание, но гендерные различия между *очень/так …* и *в … раз* вполне очевидны и подтверждаются аналогичными конструкциями с другими наречиями, ср. отношение мужских словоупотреблений к женским в Табл. 6:

Таблица 6

		много	**мало**	**Высоко**	**низко**	**быстро**
1	*очень*	3,0	2,8	2,8	2,7	2,5
2	*так*	2,5	2,3	2,8	2,3	2,6
5	*раз(а)* + **comp**	4,6	4,7	3,3	3,3	4,4

2. Имя *Кондратий* нередко ассоциируется со смертью; нет сомнений что восходит этот факт к фразеологизму, этиология которого иногда возводится к Кондратию Булавину. Отвлекаясь от деталей, можно констатировать, что исконно во фразеологизме имя употреблялось в уничижительном варианте *кондрашка*, а семантика была привязана к параличу. Ученые выявили точное значение фразеологизма (‘скоропостижно умереть, скончаться’) и его форму — *кондрашка* сочетается с глаголом *хватить* в прошедшем времени

(порядок компонентов и род глагола не фиксированы), а Партия и Правительство утвердили такую норму при использовании русского языка в качестве государственного[1].

Между тем литераторы (как и иные носители русского языка) продолжают использовать во фразеологизме полную форму имени, а также инфинитив глагола; имя может использоваться и вне фразеологизма как персонификация явления, ср.: *Кондратий, как говорят, у всех за левым плечом* (Ирина Богатырева, «Приступ»); *От всех пережитых волнений в новогоднюю ночь у меня случился инсульт. Кондрашка. Удар* (Эдуард Русаков, «Валерик»).

Статистика употребления имени *Кондратий/Кондрашка* в таком значении отдельно и в сочетании с глаголом (*с*)*хватить* в Журнальном зале приведена в Табл. 7:.

Таблица 7

пол автора	**Имя *кондр…* как нарицательное (всего)**		**Фразеологично со (*с*)*хватить***	
	кондрашка	*кондратий*	*кондрашка*	*кондратий*
муж.	**62**	43	**54**	30
жен.	9	**15**	7	**11**

Несмотря на невысокие цифры, предпочтение мужчинами уничижительного варианта, а женщинами — полного выглядит достаточно убедительно.

Региональное варьирование

В предисловии к словарю «Языки городов» [2008] говорилось: «в Приуралье и Сибири *уколы* и *прививки* часто не *делают*, а *ставят*, то есть у слов *укол* и *прививка* есть региональная специфика»; этот вывод делался на основании газетных материалов базы СМИ «Интегрум», где на август 2007 г. в Урало-Сибирском регионе имелось 828 текста с сочетанием (*по*)*ставить укол* при 772 соотношение текстах с глаголом (*с*)*делать*. На периферии ареала — в Казахстане и но Дальнем Востоке глагол ставить употреблялся в этом контексте в пять раз реже, чем делать, а у ближайших западных соседей разница увеличивалась до 52 раз.

ГИКРЯ позволяет соотнести данные газетных текстов с повседневным словоупотреблением и узусом профессиональных литераторов.

Повседневный узус тестировался в Живом журнале по записям с этими глаголами и уколом в одном предложении[2]. В Урало-Сибирском регионе в целом соотношение глаголов (*по*)*ставить* и (*с*)*делать* в этом контексте составляет

---

[1] Подробнее см. Беликов 2010-b.

[2] Шум типа *делает уколы, ставит банки* на Урале и в Сибири в обоих типах выдачи распределяется относительно равномерно, в прочих регионах он завышает статистику на (*по*)*ставить укол*.

1:0,9[3]; лишь в характеризующихся трудовой иммиграцией автономных округах Тюменской области наблюдается двукратное превосходство глагола (*с*)*делать*. Таково же соотношение этих контекстов на Дальнем Востоке и в Казахстане. В ближайших западных субъектах федерации суммарное соотношение контекстов различается в шесть раз.

В ЖС проводился только поиск контекстов с глаголом (*по*)*ставить*[4]. *После отсеивания шума, который включает и контексты, где укол* и (*по*)*ставить* синтаксически связаны (ср. *ставить ей в счет все булавочные уколы*; *поставили на ноги каким-то уколом*), нашелся 81 релевантный текст 68 авторов. Среди них лишь у 13 не устанавливается явная связь с урало-сибирским ареалом.

У литераторов связанных с ареалом рассматриваемой конструкции, такое употребление глагола (*по*)*ставить* явно не имеет стилистических ограничений, встречается в дневниковых записях, драматургических ремарках и т. п. Для наиболее публикуемых авторов проводилось сопоставление частотности в этом контексте обоих глаголов. У Н. Горлановой (Пермь) в 12 публикациях (часть — в соавторстве с В. Букуром) каждый из них встретился по 8 раз, у Э. Русакова (Красноярск) в шести публикациях — трижды (*с*)*делать*, пять раз (*по*)*ставить*. Характерная для региональной нормы синонимия глаголов позволяет избегать нелюбимых отечественной стилистикой тавтологических повторов, ср.: ‹…› *если бы не Шура, которая насильно делает ей уколы, она бы давно уже спокойно умерла. А Шура ставит и ставит уколы* — Р. Солнцев (Красноярск), «Старица»; *Никто так и не понял, что десятиклассники делали со шприцами, — уж точно не уколы себе ставили, это было бы абсурдно, ведь никто не любит уколы* — Андрей Юрич (Якутия/Кемерово), «Ржа».

Выводы и обсуждения

Нужны или не нужны дифференциальные корпуса — ответ на этот вопрос не является самоочевидным. Возможно, для каких-то задач полезно тотальное усреднение данных на максимально больших массивах языковых данных, что хорошо укладывается в идею RuTenTen. Известны и другие стратегии «нормализации», например, проведение языкового анализа на материале статей Википедии, в которых в теории многослойное редактирование вымывает индивидуальные характеристики (в русской Википедии, впрочем, немало опечаток и явных грамматических ошибок, много неудачных переводов и калек с английского, но встречаются даже с украинского).

Но результаты наших исследований показывают, что при изучении языковых конструкций дифференциальные особенности в употреблении могут обнаруживаться не только там, где они интуитивно ожидаются.

---

[3] Нет данных по Туве.

[4] Выяснение того, как часто литераторы Урала и Сибири пользуются в этом контексте глаголом (*с*)*делать*, возможно лишь после создания полноценных профилей для всех авторов.

Проект ГИКРЯ переходит из стадии «корпусного строительства» в стадию экспериментального использования: разработчики не готовы пока открыть корпус для свободного пользования всем желающим, но все заинтересованные исследователи могут получить доступ к нему на условиях участия в тестировании.

Литература

1. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* Corpus as language: from scalability to register variation In Proc. Int. Conf. on Computational Linguistics "Dialog", 2013
2. *Belikov V., Piperski A., Selegey V., Sharoff S.* Big and diverse is beautiful: A large corpus of Russian to study linguistic variation (in co-authorship with V. Belikov, A. Piperski, S. Sharoff) — In Proc. of the 8th Web as Corpus Workshop (WAC-8) / Corpus Linguistics Conference 2013
3. *Belikov V., Selegey V., Sharoff S.* Preliminary considerations towards developing the General Internet Corpus of Russian. — In Proc. Int. Conf. on Computational Linguistics "Dialog", 2012
4. *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.
5. *Jakubíček Miloš, Kilgarriff Adam, Kovář Vojtěch, Rychlý Pavel, Suchomel Vít.* The TenTen Corpus Family // Int Conf on Corpus Linguistics, Lancaster, July, 2013.
6. *J. Lorber.* 1996. Beyond the binaries: Depolarizing the categories of sex, sexuality, and gender. Sociological Inquiry, 66(2): 143–160.
7. *D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder.* 2013. "How old do you think I am?" A study of language and age in Twitter // Proceedings of ICWSM 2013
8. *Rosenthal S. and McKeown K.* 2011. Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations // Proceedings of ACL 2011.
9. *Sharoff S., and Nivre, J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. // Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo, 2011.
10. *Беликов В. И.* 2006. Словарь «Языки русских городов»: подбор примеров и интернет // «Компьютерная лингвистика и интеллектуальные технологии». Труды Международной конференции Диалог 2006. М.: Ин-т проблем информатики РАН, 2006.
11. *Беликов В. И.* 2010-а. Методические новости в социальной лексикографии XXI века // Slavica Helsingiensia 40. Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian, Helsinki, 2010 A. Mustajoki, E. Protassova, N. Vakhtin (eds.).

12. *Беликов В. И.* 2010-b. О словарях, «содержащих нормы современного русского литературного языка при его использовании в качестве государственного языка Российской Федерации» // Грамота.Ру [http://gramota.ru/biblio/research/slovari-norm/].
13. *Беликов В. И.* 2012. К методике корпусного исследования лексики. Рукопись. (Для сборника Русский язык и новые технологии. Сост. Г. Ч. Гусейнов. М.: НЛО.)
14. Языки городов. 2008. Материалы к словарю региональной лексики. В составе электронного издания: ABBYY Lingvo X3 ME: CD. M.: ABBYY, [http://www.lingvo.ru/goroda/].

References

1. *Sharoff, S.,* 2007. Classifying Web corpora into domain and genre using automatic feature identification. // Proc. of Web as Corpus Workshop, Louvain-la-Neuve.
2. *Sharoff S.,* 2007. Creating General-Purpose Corpora Using Automated Search Engine Queries.
3. *Беликов В. И., Селегей В. П., Шаров С. А.*, 2012. Пролегомены к проекту Генерального интернет-корпуса русского языка. // Труды конференции Диалог 2012.
4. *D. Bamman, J. Eisenstein, and T. Schnoebelen*. 2012. Gender in Twitter: styles, stances, and social networks. CoRR.
5. *M. Ciot, M. Sonderegger, and D. Ruths. 2013.* Gender inference of Twitter users in non-English contexts. In Proceedings of EMNLP 2013.
6. *C. Fink, J. Kopecky, and M. Morawski.* 2012. Inferring gender from the content of tweets: A region specific example. In Proceedings of ICWSM 2012.
7. *S. Goswami, S. Sarkar, and M. Rustagi.* 2009. Stylometric analysis of bloggers' age and gender. In Proceedings of ICWSM 2009.
8. *A. Mukherjee and B. Liu.* 2010. Improving gender classification of blog authors. In Proceedings of EMNLP 2010.
9. *C. Peersman, W. Daelemans, and L. Van Vaerenbergh.* 2011. Predicting age and gender in online social networks. In Proceedings of SMUC '11.
10. *D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta.* 2010. Classifying latent user attributes in Twitter. In Proceedings of SMUC 2010.
11. *S. Rosenthal and K. McKeown.* 2011. Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations. In Proceedings of ACL 2011.