Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2020»

Москва, 17–20 июня 2020 г.

ИНТЕРНЕТ-КОРПУС КАК ИНСТРУМЕНТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ: ДИФФЕРЕНЦИАЛЬНОСТЬ, АВТОРИЗАЦИЯ, ТЕМАТИЧЕСКИЕ СМЕЩЕНИЯ (ИЛИ КОРПУСЫ, КОТОРЫМ ТАК ХОЧЕТСЯ ВЕРИТЬ)

Беликов В.
ABBYY Lab, МФТИ

Селегей В. (vladimir_s@abbyy.com)
ABBYY

Селегей Д. (daniil_s@abbyy.com)
ABBYY Lab, МФТИ

Статья посвящена вопросам надежности выдачи в интернет-корпусах на примере корпуса ГИКРЯ. Несколько лет использования корпуса для лингвистических исследований дали нам пищу для размышлений и некоторых выводов. Рассматриваются проблемы, общие для любых интернет корпусов: важность учета социолингвистической вариативности, влияние ложноатрибутированных текстов, тематические смещения при нетематической классификации, перспективы и недостатки новых методов агрегации результатов поиска.

Ключевые слова: дифференциальный корпус, социолингвистическая вариативность, тематические смещения, ГИКРЯ

DOI: 10.28995/2075-7182-2020-19-62-75

WEB-CORPUS AS A TOOL FOR LINGUISTIC RESEARCH: DIFFERENTIATION, AUTHORIZATION, THEMATIC BIASES (OR CORPORA WE WANT SO MUCH TO BELIEVE)

Belikov V.
ABBYY Lab, MIPT

Selegey V. (vladimir.selegey@abbyy.com)
ABBYY

Selegey D. (daniil_s@abbyy.com)
ABBYY Lab, MIPT

The paper presents the General Internet Corpus of the Russian Language (GICR) as a tool for linguistic research. Problems are identified that are common to any WEB-corpus that affect the reliability of such research. Among the problems considered: the importance of taking into account sociolinguistic variability, the influence of falsely attributed texts, thematic biases, the prospects and disadvantages of new methods for corpora output aggregation. A distinctive feature of our approach is the emphasis on linguistic significance, reliability, and interpretability of the results obtained.

Keywords: differential corpora; sociolinguistic diversity; thematic bias; aggregated corpora output; GICR

1. Дифференциальные корпуса и социолингвистическая вариативность

Дифференциальные корпуса (ДК) предназначены для жанровых и социолингвистических исследований языка, прежде всего языка социальных медиа. Например, межсегментные сопоставления в рамках ДК позволяют объективировать стилистические пометы словарей (разг., жарг. и т. п.)

Определение «дифференциальный» означает, что любая корпусная выдача: статистика, частотные словари, скетчи, — параметризуется в ДК дифференциальными признаками, такими как пол, возраст, регион автора, сегмент интернета или жанр текста [1], [2], [12]. В рамках последовательного дифференциального подхода любые данные или цифры, не связанные с такого рода параметрами, не признаются вполне лингвистически интерпретируемыми (хотя, как известно, обобщенная статистика оказывается полезной при решении различных задач, когда объем данных оказывается важнее, чем их состав. Так делаются, например, неспециализированные частотные словари.

ГИКРЯ версии 1.0, собранный в основном в 2013–2014 гг. содержит около 20 млрд слов, распределенных по сегментам, состав которых отражен в **таблице 1**.

Таблица 1. Состав ГИКРЯ версии 1.0

Подкорпус	Слов (млн)	Док (млн)	Авторы (млн)	Gender (%)	Age (%)	Year (%)	Region (%)
Журналы	300	0,06	0,005	82	0	100	0
Новости	850	3	—	—	—		—
ЖЖ	8 100	73	1.0	0	23		70
Блоги	700	9,9	0,26	94	46		67
Соцсети	9 800	193	22	49	16		41

Такой объем исходных данных значительно превосходит объем национальных корпусов, которыми обычно пользуются лингвисты (и в которых, как правило, нет метатекстовой разметки такого типа).

Несколько лет (с 2015 г.) использования ГИКРЯ для дифференциальных исследований языка выявило много интересных явлений значимого смещения корпусной статистики, связанной с социолингвистическими параметрами и регионом, причем не только на уровне лексики и фразеологии, но также и лексикализованном синтаксисе (диатезы), морфологии и грамматике конструкций [2], [3].

Такие смещения должны предостерегать исследователя от чрезмерного доверия к любым «средним» результатам. Аккуратный анализ неагрегированных результатов поиска (сниппетов) почти всегда демонстрирует «лукавость» таких цифр, связанных с грязными данными, ошибками автоматической разметки, неправомочными обобщениями смещенных корпусных данных.

Приведем только один пример такого смещения, связанного с возрастом пишущих.

Тридцать лет назад с легкой руки Горбачева М. С. стали говорить *определимся по месту встречи*, ср. зафиксированное его *Определимся по Марии* (=Марийской АССР, Марий Эл). Ранее *определялись*, например, *по компасу*. Новое для стандартного языка значение глагола *определиться* — ‘определить свое отношение к кому/чему-л.’ мгновенно внедрилось в язык журналистов и депутатов, вкл. достойных (упомянем А. Д. Сахарова). В 1990-х оно «пошло в народ», но управление при глаголе быстро менялось: *по* → *в* → *с*.

Что происходит в повседневном узусе «простого» человека сейчас можно проверить по ГИКРЯ. Приведем некоторую статистику (**таблица 2**) по правому контексту дедублицированной выдачи ЖЖ на *определ*(*и/я*)*ться* + предлог.

При такого рода выдаче невозможно увидеть существенное смещение результатов: вероятность использования предлогов при этом глаголе значимо коррелирует с возрастом. Статистика для лиц 1950–1999 г. рожд. (учтены ошибочные написания типа *опр**и**д**и**литься*) приведена в **таблице 3**.

Таблица 2. Абсолютная статистика правого контекста для ***определ(и/я)ться*** +предлог

слово	с(о)	в(о)	по
количеством, -ве, -ву	252	8	159
с тем, в том, по тому	3 404	226	106
выбором, -ре, ру	3 807	209	5
его, ее (её)	238	23	45
материалом.	107	1	0
местом, -те, -ту	1 388	5	31
результатом…	11	60	215
темой, теме	560	1	2
формой, -е	171	2	12
цветом, -е	531	3	37
временем –и	465	17	43
датой, -те	821	1	25
величин..	19	0	17
размер..	1	34	23
своим../ей/им/му	3 311	845	16
будущ…	342	14	2
дальнейш..	339	21	3
данн…	20	33	41
основн…	225	169	10
следующ…	108	20	75

Таблица 3. Выбор управления «определиться» по возрастным когортам

	предлог			отношение	
Год рождения	*с(о)*	*в(о)*	*по*	*с(о)/в(о)*	*с(о)/по*
в целом	11 163	1 779	1 194	6,3	9,3
1950–1969	941	259	196	3,6	4,8
1970–1979	3 071	557	384	5,5	8,0
1980–1989	6 727	913	589	7,4	11,4
1990–1999	424	50	25	8,5	17,0

Как видим, «в среднем» предлог *по* при глаголе *определиться/определяться* встречается в девять раз реже, чем предлог *с*. Но соотношение предлогов в старшей и младшей возрастной когорте различается в 4,5 раза, и разница в основном обусловлена выбором предлога при новом, «горбачевском» значении.

Можно было бы привести и множество других примеров социолингвистических смещений, связанных с возрастом, гендером, регионом автора. Но мы рассмотрим далее другие источники «лукавых цифр» помимо собственно вариативности.

2. Проблема неавторского текста и динамическая дедубликация

Мысль о том, что не всякий текст, опубликованный автором, им же и написан, совершенно очевидна. Другое дело, что оценить влияние этого фактора в конкретном корпусном исследовании оказывается непростой задачей.

Работ, посвященных собственно удалению псевдоавторских текстов практически не существует, можно назвать только несколько статей [4], [9], [18], посвященных скорее дедубликации, в том числе — нечеткой дедубликации как методу очистки корпуса. И совсем нет таких работ для русского языка.

Основным механизмом, позволяющим уменьшать значимость фактора чужого текста в ГИКРЯ является механизм динамической дедубликации. Динамическая дедубликация применяется (по умолчанию!) к результатам поиска и позволяет отфильтровать сниппеты, имеющие идентичный текст в некотором окне вокруг искомого слова или фразы. В ГИКРЯ имеется целый ряд параметров, позволяющих управлять динамической дедубликацией, включая и ее нечеткий вариант, близкий работе [4]. Пользоваться ими, однако (за исключением параметра ширины контекста), стоит только очень аккуратному исследователю: в конкретных запросах влияние нечеткой дедупликации непредсказуемо.

Динамическая дедубликация действительно серьезно очищает выдачу (см. таблицы далее в тексте), но у неё имеются, некоторые недостатки:

1. Она непредсказуемым образом влияет на статистику запроса, полученного на части корпуса (поскольку мы не можем рассчитывать на «равномерность» распределения дублей на всем корпусе). Это препятствие можно обойти полным поиском, но далеко не всякий пользователь корпуса станет этим заниматься.
2. Динамическая дедубликация далеко не покрывает всех случаев инкорпорации чужого текста.
3. При динамической дедубликации из обработки исчезают и автодубли (то есть, повторы некоторого текста самим автором)[1]. То есть, вообще говоря, с задачей авторской очистки она связана все же косвенно.
4. Ее нельзя применять вне контекста конкретного запроса, когда, например решается задача классификации текстов по социолингвистическим параметрам. В результате решения по классификации, полученные на «хороших» искусственных датасетах, очень плохо себя показывают при переносе на реальный интернет-корпус в миллиарды слов. Это мы много раз видели в исследованиях по гендерной, авторской, возрастной, региональной классификации, сделанных по материалам ГИКРЯ. Например, в [15], [16]. Именно эта проблема (в сочетании с проблемами тематических смещений — см. далее) заставила нас несколько отложить задачу расширения априорной метатекстовой классификации результатами работы автоматических классификаторов.

[1] Проверка сниппетов до и после дедубликации выявила около 5% «незаконно» отфильтрованных автодублей, которые должны сохраняться в корпусе в одном экземпляре, а не удаляться совсем.

Остановимся на проблеме 2 более детально.

Общим местом является утверждение о том, что язык — это большое число относительно редких событий. На таких редких событиях хорошо видно влияние неавторского текста и способность дедубликации учесть фактор чужого слова.

Например, в **Таблице 4** представлены результаты анализа реального запроса, возникшего входе исследования одного из авторов. Искалось слово *алюторцы* (этнос на Камчатке).

Таблица 4. *Алюторцы*: отсев текстов на разных стадиях обработки

Стадия анализа	Nвхождений	N авторов
До динам. дедубликации	55	39
После динам. дедубликации	44	27
После ручной обработки	23	9

Выявилось, что даже после динамической контекстной дедубликации существенная часть оставшихся упоминаний встречается во фрагментах прямого цитирования или в автоматически сгенерированных текстах, представляющих собой локально связные фрагменты или предложения из самых разных источников.

Таким образом, ручная обработка оставляет для исследования только половину якобы «очищенных» данных. Остальное — это списки этнических групп, заголовки книг и сгенерированные тексты. Примеры такого сложно диагностируемых ложных вхождений: «*Толщина асфальта все для вас работа в ставрополе* ***алюторцы****, скорость скачивания из интернета*» или «***Алюторец*** *Серёнькина мать посмеивалась, чувствуя общий настрой на дворе и огороде*».

Каждый запрос уникален, приведем еще несколько примеров, демонстрирующих ограничения стандартных корпусных процедур очистки. Мы взяли (**таблица 5**) слово *клинить* в сравнительно новом значении ‘~вводить в ступор’ и выражение *иди ты лесом*. Интересно рассмотреть также и орфо-варианты (*клинет*). Не удержимся от комментария, что при всех издержках, связанных с чистотой данных, ГИКРЯ содержит тем не менее по сотне примеров на рассматриваемые выражения, в то время как в НКРЯ они практически не встречаются (1 раз в газетном подкорпусе для *меня клинит* и 1 — *иди ты лесом* в основном).

Таблица 5. Данные по этапам очистки

Фраза	Всего в ГИКРЯ/ ВК	Динамическая дедублик. (окно 20)	После ручной фильтрации	Доля «авторского» в отбросах
Меня клинит	141	100	76	50%
меня клинет	16	14	7	71%
меня по ночам клинит	179	21	0	—
Иди ты лесом	108	78	48	58%

Важный параметр Доля «авторского в отбросах» показывает размер «бедствия» в случае, если мы будем пытаться учесть гендерные или возрастные параметры только с помощью динамической дедубликации. Такая очистка будет оставлять от 10 до 30 % нерелевантных текстов (то есть — ложно имеющих такие параметры и участвующие в статистике).

Таким образом, динамическая дедубликация, хотя и является намного более мощным инструментом чем простая статическая дедубликация, не обеспечивает все же нужной степени чистоты данных для честного дифференциального социолингвистического исследования.

Необходимо дополнительно фильтровать неавторские тексты, то есть делать то, что сейчас аккуратный исследователь проделывает с корпусной выдачей вручную.

3. Автоматическая фильтрация

Автоматическая фильтрация широко применяется сегодня при создании интернет-корпусов и состоит обычно из двух операций:

1. прямой дедубликации, основанный на различных точных и эвристических методах;
2. сигнатурной фильтрации, основанной на методах идентификации спама (на практике, с вручную написанными шаблонами). Иногда это больше похоже на цензурные соображения (при фильтрации по ключевым словам).

Проблемой собственно фильтрации чужого слова, как уже указывалось выше, до сих пор всерьез никто не занимался, и одной из основных причин является отсутствие значимого объема обучающих данных.

При создании датасетов на основе ГИКРЯ (см. раздел «ГКИРЯ как фабрика датасетов»), динамическая контекстная дедубликация не может (и не должна) применяться. Для очистки была применена двухступенчатая технология, состоящая из:

1. квазидедубликации текстов, основанный на совпадении значений некоторых хэш-функций от начальных и конечных отрезков текста.
2. сигнатурной фильтрации в виде системы правил (регулярных выражений), настраиваемых на конкретный сегмент социальных медиа.

Отметим, что простая квазидедубликация, помимо того что отбраковывает полные дубли, неплохо справляется и с некоторыми текстами, автоматически сгенерированными по некоторому шаблону. В **таблице 6** можно увидеть результаты применения двухступенчатой очистки на примере полного сегмента ВК в составе ГИКРЯ (в **таблице 7** приведены наиболее эффективно работающие фильтры).

Мы можем сделать два важных вывода:

1. Только чуть больше 20 % исходного корпуса ВК (и всего около 17 % объема в словах) представляют собой авторские тексты.
2. Этого не видно из таблицы, но далее будет показано, что еще примерно 10 % сокращения можно было бы ожидать при гипотетической ручной разметке — с учетом ошибок автоматической фильтрации (см. далее).

Таблица 6. Этапы очистки корпуса от неавторских текстов

	Всего текстов	Всего авторов с текстами	Всего слов (кириллица)
Удаление обвязки (N0)	893 921 661 (100 %)	65 443 062 (13,7 текста на 1 автора)	34 832 362 470
После квази-дедубликации (N1)	271 511 598 (30.37 %)	39 375 322 (6,90 текста на 1 автора)	8 965 228 250 (25,74 % от N0)
После фильтрации (N2)	226 700 982 (25,36 % от N0) (83,50 % от N1)	36 175 624 (6,27 текста на 1 автора)	5 971 482 642 (17,14 % от N0)

Очевидное решение, которое могло бы быть применено, состоит в создании представительного датасета с разметкой по критерию авторский vs. чужой и машинном обучении на этом датасете, чтобы попытаться уменьшить процент ложно подтвержденных авторских текстов.

Был проведен ряд экспериментов по исследованию неавторского текста, в основном в рамках НИР студентов МФТИ и РГГУ. В рамках этих исследований были созданы несколько датасетов с последовательно увеличивавшимся объемом и представительностью на базе подкорпуса социальных сетей в ГИКРЯ/ВК за период с 2013 по 2018 год. Подробнее об этих датасетах и полученной типологии неавторских текстов см. [6])

Эта работа еще не завершена, но получены важные результаты. Поскольку разметчики получили на 1-м этапе тексты, не прошедшие автоматической обработки, стало возможным провести статистически значимую проверку работы автоматической процедуры дедубликации-фильтрации, описанной выше (та самая оценка в **таблице 6**) — просто применив ее к этому датасету. В **таблице 8** представлены основные типы и распределение текстов по категориям на разных стадиях обработки: до и после применения автопроцедур.

Таблица 7. Распределение отфильтрованных текстов по шаблонам

Типы текста, отсекаемые фильтром	Отн. эффект (%)
Тексты, содержащие ссылки на другие ресурсы/страницы (предположительно объявления и реклама)	35,1
Репосты из instagram-а (мы не можем гарантировать их авторство)	8,66
Поэзия	20,7
Другие языки (из них кириллические — 90 %)	13,1
Тексты с нехарактерными для авторских текстов символами	5,41
Объявления о продаже	3,75
Тексты с нехарактерным для авторского текста форматом	2,97
Тексты. содержащие специфические шаблоны	2,77

Типы текста, отсекаемые фильтром	Отн. эффект (%)
Тексты с копирайтом (c)	2,8
Прочее	4,74
Все отфильтрованные тексты = 45 231 439 (= 5,06% от N0)	100

Таблица 8. Результаты проверки автоматических процедур на датасете Ru

Тип	Ручн. разметка → Ддубл			Ддубл. ручная разметка → Фильтр			Осталось после фильтр.	
Все	18 848	100%	35,5%	6 689	100%	84,2%	5 634	100%
Author	6 420	34,1%	**75,9%**	4 873	72,9%	**92,6%**	4 511	80,1%
Author+	454	2,4%	58,4%	265	4,0%	70,6%	187	3,3%
Mixed	363	1,9%	38,8%	141	2,1%	74,5%	105	1,9%
nonAuthor	11 611	61,6%	12,1%	1 410	21,1%	58,9%	831	**14,7%**
— advertising	1 586	8,4%	22,4%	355	5,3%	69,0%	245	4,3%
— citation	4 087	21,7%	7,8%	319	4,8%	78,1%	249	4,4%
— poem	1 468	7,8%	11,4%	167	2,5%	10,2%	17	0,3%
— article	501	2,7%	20,2%	101	1,5%	54,5%	55	1,0%
— fiction	107	0,6%	15,0%	16	0,2%	87,5%	14	0,2%
— news	108	0,6%	39,8%	43	0,6%	69,8%	30	0,5%
— autogen	1 894	10,0%	6,8%	129	1,9%	42,6%	55	1,0%
— link_header	1 772	9,4%	11,7%	207	3,1%	95,7%	198	3,5%
— foreign	427	2,3%	36,1%	154	2,3%	6,5%	10	0,2%
— other	388	2,1%	8,0%	31	0,5%	71,0%	22	0,4%

Некоторые комментарии к **таблице 8**. Каждая расширенная колонка показывает состав датасета а) в исходном состоянии б) после применения дедубликации и в) после фильтрации. Третьи столбцы в этих колонках показывают, какой процент текстов этого типа остался после этих операций.

Некоторые числа особенно важны (выделены жирным):

1. Только 75,9% помеченных разметчиками как уверенно авторские оказались уникальными (что вполне нормально для немаркированного цитирования).
2. На этапе фильтрации было отбраковано 7,4% из оставшихся уникальных авторских текстов, что хотя и является ошибкой алгоритма фильтрации (ложноположительные результаты), но не является смещающим фактором для статистки, поскольку просто выводит эти тексты из корпуса.
3. 14,7% текстов в итогом датасете относятся к неавторским, но не были опознаны методами фильтрации (ложноотрицательные). Эти тексты являются потенциальным полем применения методов машинного обучения, и обозначают текущую границу доверия к автоматическим процедурам фильтрации (см. примечание 2 к **таблице 6**).

Отметим еще раз, что в текущей версии ГИКРЯ 1.0 используется контекстная динамическая дедубликация. Результирующее решение в готовящейся ГИКРЯ 2.0 будет гибридным, сочетающим текстовые и контекстные методы фильтрации.

В завершение раздела отметим несколько направлений, по которым будет происходить развитие методов фильтрации:

- Исследование возможности выявлять внутритекстовое цитирование (сейчас разметчики отмечают эти случаи): т. н. «авторские переходы», по аналогии с жанровыми и тематическими.
- Межсегментная дедубликация помогла бы убрать не только бродячие тексты в жанре анекдотов или кулинарных рецептов, но и цитирование публицистики, новостей и беллетристики, представленные в других сегментах Интернета
- Анализ коротких псевдодублей, являющихся по существу ритуальными общепринятыми выражениями. Их исключение заметно сдвигает статистику употребления многих частотных слов.
- Использование для фильтрации жанровых классификаторов: жанровый состав «натуральных» текстов в разных сегментах ГИКРЯ отличается и соответствующая разметка могла бы указывать на потенциально «чуждые» тексты.
- Применение методов кластеризации для поиска текстов, сходных с отбракованными (такие исследования велись нашими студентами и будут продолжены).

4. Влияние тематических смещений

Явление тематического смещения состоит в том, что имеется сильная корреляция между некоторым исследуемым параметром и тематикой текстов [10]. Грубо говоря, при тематической неоднородности очень легко спутать различия в том, «как» мы говорим с тем «о чем» мы говорим. Если не принимать специальных мер, то при попытке обучения классификаторов на априорно размеченных подкорпусах ГИКРЯ в задачах автоматической классификации (например, установление авторства, жанровые, гендерные, региональные классификации) модели обучаются на тематических признаках вместо тех, которые релевантны для исследования.

Влияние тематических факторов было обнаружено практически во всех типах проводимых исследований, например:

а) при решении задач жанровой разметки [7];

б) при автоматической региональной классификации лексики. В таких исследованиях значимыми признаками оказывается не те относительно редко используемые специфические региональные слова, которые ищет лингвист, а в лучшем случае топонимы, или просто наиболее злободневная для региона лексика [16].

в) автоматическая гендерная и возрастная классификация;

г) исследование авторских идиостилей [15]

Тематические смещения в обучающих датасетах приводят к тому, что обученные на них модели оказываются бесполезными на тематически однородных корпусах. Проблема возникает не только в задачах автоматической классификации, но и в любых дифференциальных исследованиях, когда мы рискуем связать видимые различия в частотах с интересующих нас признаком, в то время как они вызваны тематическими факторами.

В **таблице 9** приведены результаты сравнительного тестирования качества автоматического определения пола на произвольном и тематически однородных датасетах (студенческое исследование [5]).

Таблица 9. Гендерная классификация на общем и тематически однородном корпусе

Метод/accuracy	Test	Selling	Games	Beaty&Fitness
tf-idf + LogReg	70,61	54,55	57,80	53,55
BoW + Naive Bayes	70,21	53,90	56,20	54,95
CharCNN	72,00	52,60	59,00	52,30
LSTM	70,84	52,50	56,60	52,90

В эксперименте использовались разные методы классификации, которые дают близкие к SOTA результаты. Модель обучалась на случайно выбранном датасете с гендерной разметкой из сегмента социальных сетей ГИКРЯ. Затем она проверялась на тестовой части этого датасета и сравнивалась с результатами на специально подобранных тематически однородных датасетах с такой же разметкой (тексты из форумов ГКИРЯ по трем темам: «продажа», «компьютерные игры» и «красота» общим объемом ок. 30 тыс. слов).

Хорошо видно, что для однородных текстов качество предсказания близко к случайному. Это подсказывает, что попытки построить «в лоб» гендерные дифференциальные словари на всем сегменте размеченных социальных сетей (а ГИКРЯ имеет API, позволяющий это делать), дадут красивые результаты, которые при этом мало что дадут гендерной лингвистике из-за тематических смещений.

В еще большей степени тематические смещения проявляется в задачах автоматического определения авторства, в особенности для текстов non-fiction. Полученные результаты отражены в работе [15]. Основной вывод состоит в том, что тематическая однородность приводит к резкому падению точности предсказания автора — что означает, что модели, демонстрирующие очень хорошие результаты на многих датасетах (например, текстах журнального подкорпуса ГИКРЯ), обучаются на тематических признаках, а не признаках идиостиля автора.

Все это, разумеется, имеет отношение и к результатам любого дифференциального лингвистического исследования.

Тематические смещения требуют специального изучения. Даже поверхностный анализ показывает сложность их интерпретации. Приведем пример из сегмента Новости ГИКРЯ, где представлены неспециализированные новостные ленты. Их тематическая универсальность не препятствует появлению статистически значимых различий. Некоторые интерпретируются легко; так,

в политике Лента.Ру больше внимания уделяет дальнему зарубежью, а Регнум — постсоветскому пространству. Не удивительно, что упоминаний США, Великобритании, Франции и Японии в текстах Ленты в полтора раза больше, чем в Регнуме, а 11 государств исходного состава СНГ в пять раз меньше.

Но другие семантические сдвиги интерпретировать сложно. Вот как выглядит, например, 10-летняя статистика (2004–2013) этих двух лент по четырем часто упоминаемым заболеваниям:

	грипп	*пневмония*	**гр./пн.**	*инсульт*	*инфаркт*	**инс./инф.**
Лента.Ру	1 377	224	**6,1**	575	426	**1,4**
Регнум	19 884	790	**25,2**	698	860	**0,8**

Флуктуации подобного рода неизбежно и непредсказуемо влияют на результаты нетематической классфикации.

5. Достоинства и недостатки агрегированной выдачи

Под агрегацией мы понимаем представление корпусной выдачи в максимально обобщенной форме (в отличие, допустим от классической выдачи в формате QWIC). Инструментами агрегации в ГИКРЯ являются, например, скетчи, частотные словари, статистические запросы. Помимо всего прочего агрегация является хорошим способом уйти от все более острого вопроса про права на сбор текстов (scraping) в социальных сетях.

У агрегации, однако, имеются серьезные проблемы с точки зрения надежности лингвистических исследований. Все они сводятся в сущности к отходу от старой мудрости «доверяй, но проверяй».

1. При агрегации мы вынуждены доверять автоматической корпусной разметке. Например, лемматизации, которая сегодня не умеет решать проблему орфографической вариативности: для отдельных слов орфографически ошибочные записи могу составлять десятки процентов в блогах и социальных сетях, в единичных случаях ошибочные записи частотнее верных, например, «телек» вместо «телик».
2. Агрегация элиминирует важные контекстные особенности, включая и те, которые при просмотре сниппетов снижают доверие к конкретному примеру.
3. Агрегация без учета дифференциальных параметров дает результат, лингвистический смысл которого в общем случае не очевиден.

Решением этой проблемы является «дифференциальная» агрегация, при которой автоматически определяются возможные смещения по каким-либо из имеющихся в данных параметрам [8], либо прямое сравнение агрегированной выдачи (например скетчей) с разным набором параметров. Соответствующая функциональность уже до некоторой степени реализована в корпусе, но ее применение будет вполне эффективным только после решения вопроса с очисткой данных в ГИКРЯ 2.0.

6. ГИКРЯ как фабрика датасетов

ГИКРЯ играет заметную роль в русскоязычной компьютерной лингвистике, выступая источником размеченных датасетов для лингвистических задач разного типа, например, автоматической морфологической и синтаксической разметке, анализу референциальных цепочек и восстановлению эллипсиса, задачам нетематической классификации. Назовем только 4 таких датасета, использованных для проведения технологических соревнований: датасеты с морфоразметкой для тестирования систем исправления опечаток и морфологического анализа, датасет для задачи анализа явлений гэппинга и определения авторства [14], [15], [17].

Наличие априорной метатекстовой и автоматической лингвистической разметки разного типа позволяет делать датасеты с разными параметрами, но вопрос чистоты исходного корпуса оказывается первостепенно важным.

7. Заключение

Авторы осознают, что изложенный материал имеет характер отчета. Однако резюмирующего текста, описывающего уже шестилетний опыт использования ГИКРЯ, до сих пор не публиковалось.

Мы надеемся, что изложенные проблемы — а это те проблемы, с которыми мы столкнулись при реализации идеи дифференциального корпуса — будут интересны корпусным коллегам. Не менее существенно и то, что корпусом пользуются сотни лингвистов, не всегда осознавая те серьезные подводные камни, о которых шла речь в докладе.

8. Благодарности

Авторы, представляющие команду ГИКРЯ, благодарят всех коллег, пользователей корпуса, результаты работы которых стимулировали исследование их надежности. Мы выражаем благодарность Физтех-Школе Прикладной математики и Информатики за поддержку проекта ГИКРЯ, который с 2020 года развивается усилиями лаборатории компьютерного зрения и анализа социальных медиа ABBYY Lab в составе этой Физтех-школы.

Литература

1. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.,* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation, Web as Corpus Workshop (WAC-8).
2. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2014) Variational Corpus Statistics Using Author Profiles In Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.
3. *Benko V., Zakharov V. P.* (2016) Very Large Russian Corpora: New Opportunities and New Challenges. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”, Moscow.

4. *Benko V.* (2019) Dedublication in Large Web Corpora. In. Proc.of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff.
5. *Iglina A., Smurov I.* (2019). Nonthematic texts Classification: Gender of the Author. Publ. in MIPT_CoLing_Bachelors Diplomas.
6. *Ivoylova A., Raskin I., Selegey D.* A New Dataset to solve the task of non-author text filtration in social networks-based corpora. In Proc. of Student Worksop at Dialogue, Russian International Conference on Computational Linguistics, Moscow.
7. *Katinskaya A., Sharoff S.* (2015) Applying Multi-dimensional Analysis to a Russian Webcorpus: Searching for Evidence of Genres , in Proc. of the Workshop on Balto-Slavic Natural Language Processing associated with the International Conference RANLP, Hissar, Bulgaria.
8. *Lagutin M. B., Kuratov Y., Kopylov N.* (2016) Statistical analysis of the search results in a differential corpora. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow
9. *Laippala V. et al.* (2020) From Web Crawl to Clean Register-Annotated Corpora, — in Proc. of 12th WaC Workshop, Marseille.
10. *Petrenz P., Webber B.* (2011). Stable Classification of Text Genres: Computational Linguistics. Vol. 37, No. 2.
11. *Piperski A., Belikov V., Kopylov N., Selegey V., Sharoff. S.* (2013) Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In : 8th WAC Workshop.
12. *Piperski A.* (2013) The General Internet Corpus of Russian and the Notion of Representativeness in Corpus Linguistics. In Proc. of Institute of Linguistics (Russian State University for the Humanities, Moscow.
13. *Pomikálek, J., Jakubícek, M., and Rychly, P.* (2012). Building a 70 billion word corpus of English from ClueWeb. In LREC, pages 502–506.
14. *Ponomareva M., Droganova K., Smurov I, and Shavrina T.*(2019). AGRR-2019: A Corpus for Gapping Resolution in Russian. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 35–43. Florence, Italy. Association for Computational Linguistics.
15. *Solonin M.* (2019) Evaluation of Authorship Attribution Methods for Russian Texts. In. Computational Linguistics and Intellectual Technologies, Moscow, Supplementary volume pp. 240–246.
16. *Sorokin A.* (2015) Automatic Regional Classification Using a Dictionary of Regional Lexics: a Preliminary Study. In Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.
17. *Sorokin A, Baytin A., Galinskaya I., Shavrina T.* (2016) SpellRuEval: the First Competition on Automatic Spelling Correction for Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow.
18. *Wenzek G., Lachaux M. et al.* (2019) CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data (2019). arXiv:1911.00359 [cs.CL].