УДК 811.161.1, 81’33
DOI 10.17223/18137083/56/2

В. И. Беликов

Московский государственный университет им. М. В. Ломоносова

Что и как может получить лингвист из оцифрованных текстов

Статья посвящена границам применимости онлайн-инструментов автоматической обработки цифровых текстов (поисковых машин, языковых корпусов, Google Books Ngram Viewer) в лингвостатистических исследованиях. Вопреки установившемуся мнению об объективности результатов, получаемых после автоматической обработки текстового массива, существуют ограничения и искажения данных, обусловленные множеством причин. Одной из них является частое отсутствие лингвистов в коллективах разработчиков таких автоматов. На примере анализа частотности употребления культурно значимых онимов и их орфографических вариантов, родовых форм глаголов и вариантов предложного управления по данным разных автоматических средств анализа текстов показана сложность лингвистической интерпретации результатов автоматической обработки текстовых массивов.

Ключевые слова: цифровой текст, лингвистическая статистика, онлайн-инструменты лингвистической обработки текстов, корпусная лингвистика, грамматическая вариантность.

Текст всегда был для лингвиста основным поставщиком сведений о языке: и фактов, и статистики. Традиционно, работая с бумажными текстами, языковые факты мы извлекали легко, но, прежде чем извлечь определенный факт, нужно было его найти, на что уходили часы, а то и дни. При этом поиск не всегда был результативен.

Получение статистики требовало трудоемкой ручной обработки печатных текстов (при создании частотных словарей) или столь же затратного массового анкетирования (в социолингвистике). «Объективность» результатов зависела от подборки текстов или содержания анкеты и методики ее использования.

Неимоверная сложность получения достоверной статистики вела к тому, что лингвисты прибегали к ней редко, считалось вполне естественным ограничиваться интроспективной оценкой: «говорят – не говорят», «часто – редко», «больше – меньше». На теоретическую значимость последнего противопоставления – выявление отрицательного языкового материала – впервые указал Л. В. Щерба в 1931 г.: «весьма важную составную часть языкового материала образуют имен-

Беликов Владимир Иванович – доктор филологических наук, профессор кафедры теоретической и прикладной лингвистики филологического факультета Московского государственного университета им. М. В. Ломоносова (Ленинские горы, МГУ, 1-й учебный корпус, ГСП-1, Москва, 119991, Россия; otipl@philol.msu.ru)

ISSN 1813-7083. Сибирский филологический журнал. 2016. № 3

но неудачные высказывания с отметкой “так не говорят”. <…> Роль этого отрицательного материала громадна и совершенно еще не оценена в языкознании» [Щерба, 1974, с. 32–33]. Однако в широкий лингвистический обиход примеры под звездочкой, а затем и со знаком «?» (допустимость высказывания неясна) начали входить лишь в 1960-е гг. При этом лингвист-читатель далеко не всегда разделял оценку (не)допустимости, предлагавшуюся лингвистом-автором. Возможность объективно опереться на факты узуса отсутствовала.

Информационная революция привела к появлению нового феномена – электронного (цифрового) текста. Цифровые тексты могут служить источником автоматически извлекаемых сведений о языке (и фактов, и статистики). Наиболее масштабный источник цифровых текстов – Интернет, а общедоступный инструмент работы с этими текстами – поисковая машина, для русского языка это почти всегда Яндекс [1] или Гугл [2]. «Говорят» или «не говорят» теперь выясняется «в один клик», однако совсем не просто с достоверностью уточнить, сколь часто говорят.

Но надо иметь в виду:

- что цифровые тексты и их подборки в Интернете создаются не для лингвистов, поэтому для решения лингвистических задач такие «сырые» тексты плохо приспособлены;
- для извлечения лингвистической информации нужен специальный инструментарий, применимость для этих целей «чужого» инструмента (созданного для нелингвистического манипулирования такими текстами) требует его основательного тестирования.

Поисковики создавались для поиска информации, однако при выдаче Гугла и Яндекса на первой же странице показывается некая цифра; если спрашивается что-то не особенно редкое, то это сотни тысяч и миллионы.

Ранний Яндекс по умолчанию сортировал выдачу «по релевантности», но допускал возможность упорядочить ее «по дате» (то есть ретроспективно); цифр было две: серверов (позднее сайтов) столько-то, страниц столько-то. Единицей выдачи служила одна из страниц найденного сайта, но можно было перейти и к другим его страницам. Перелистывая станицы поиска, можно было просмотреть всю выдачу до конца и удостовериться, насколько заявленные вначале цифры найденного соответствуют действительности.

Лингвиста интересует в первую очередь языковая специфика найденного поисковиком. При большом объеме выдачи для снижения трудоемкости можно было ограничиться, например, анализом содержания каждой десятой страницы выдачи [3].

Гугл изначально показывал не более тысячи единиц выдачи; позднее так стал поступать и Яндекс; единицы выдачи в обоих поисковиках именуются сейчас туманным словом «результаты». Кого-то может насторожить тот факт, что при переходе от одной страницы поисковика к другой число «результатов» меняется, иногда существенно. Но большинство радуется первой цифре и верит, что скрывающаяся за нею виртуальность – это именно та реальность, которую ищет пользователь.

---

[1] https://www.yandex.ru

[2] https://www.google.ru

[3] Иллюстрацией к методике работы с ранним Яндексом может служить выявление фактического написания сложных прилагательных [Беликов, 2003] или исследование региональных различий в управлении глагола *определиться* в распространившемся вслед за М. С. Горбачевым значении (*определиться по* / *в* / *с* / *относительно* / *по отношению* и т. п.) [Беликов, 2004].

В действительности «погуглить чего больше» – всего лишь развлекательное времяпрепровождение. Но самые разные люди серьезно относятся к такой нумерологии.

Одним из первых теоретиков гугления оказался М. Н. Эпштейн (философ, культуролог, литературовед, лингвист, эссеист, как пишет о нем Википедия), который в сентябре 2003 г. предложил термин *нумеризм* – «краткая обобщающая мысль или наблюдение, выраженное в числах, в наборе или сопоставлении цифр, статистических данных» [Эпштейн, 2003]. Первый обнародованный автором нумеризм, «Шекспир и Пушкин», сополагает два имени, снабженных числами:

Пушкин	1,911,216
Shakespeare	5,730,000

Согласно Эпштейну, предъявленные числа отражают «частоту употребления этих знаковых имен в русскоязычном и англоязычном интернете (соответственно, по Яндексу и Гуглю)».

В июне 2006 г. на конференции «Бенджамин Франклин и Россия: к 300-летию со дня рождения» Эпштейн провозглашает науку гуманетику, в рамках которой ввел «информационную единицу – один гуглик, равную одной странице, на которой встречается данное слово по статистике Гугля» [Эпштейн, 2006, с. 76]. «У имени Шекспира (Shakespeare) – огромный символический капитал, равный 105 миллионам гугликов в англоязычной сети. <…> Имя Пушкина в рунете обладает символическим капиталом в 17 млн 197 тыс. гугликов» [Там же, с. 78].

Что изменилось между сентябрем 2003 г. и июнем 2006 г.?

- Курс рубля повысился с 27,1 руб. за доллар до 30,7.
- Shakespeare приумножил свой «символический капитал» в 18,3246 раз, Пушкин – в 8,9979 раз.

Какое отношение курс доллара имеет к Пушкину? Такое же, как капитал в гугликах: н и к а к о г о. Но курс доллара легко проверяется, а про гугликовое богатство писателей мы вынуждены верить Эпштейну.

Зато мы можем выявить последующую историю «гуманетической капитализации» классиков, причем сделаем это не голословно, а со скриншотами.

Согласно грамматике Гугла, при поиске на *Shakespeare* должны выдаваться и контексты с *Shakespeare's*, однако на дизъюнкцию этих последовательностей находится на 8,6 млн больше «результатов», чем на просто *Shakespeare* (рис. 1).

В 2006 г. Гугл русскую морфологию игнорировал [4], так что Эпштейн почти справедливо писал, «чтобы вычислить частоту употребления фамилии Пушкин в Гугле, нужно провести поиск по всем шести падежам и суммировать результаты» [Эпштейн, 2006, с. 75]. Почти справедливо – поскольку находит-то Гугл документы, а не вхождения: если искать отдельно разные «падежи» [5], то любой пространный текст о Пушкине будет учтен на каждый «падеж». Так что работу по суммированию графических словоформ лучше поручить самому поисковику, что мы и сделали. Результат впечатляет (рис. 2): шекспировский капитал в начале 2011 г. составлял то ли 45 %, то ли 60 % от пушкинского.

[4] С тех пор положение с именной морфологией, а затем и глагольной улучшилось, но и Гугл, и Яндекс не для каждого слова успешно справляются со склонением или спряжением.

[5] Точнее, графические словоформы: надеемся, родительный *Пушкина* и винительный *Пушкина* Эпштейн все же не искал по отдельности с последующим сложением результатов.

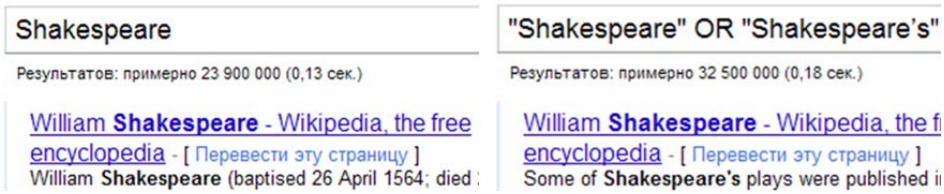

Рис. 1. Гугл о Shakespeare, два результата поиска 9 января 2011 г.

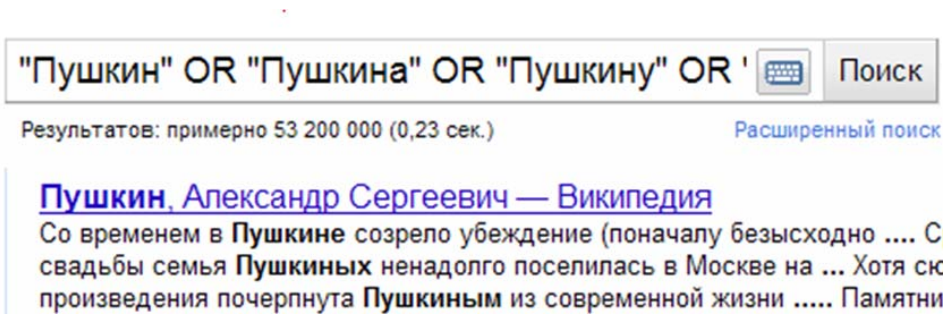

Рис. 2. Гугл о Пушкине, 9 января 2011 г.

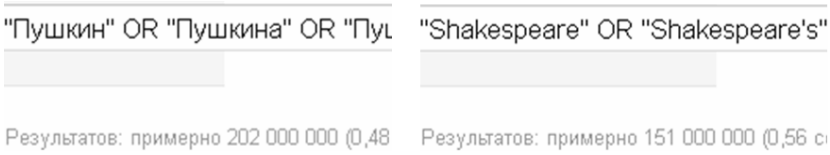

Рис. 3. Гугл о Пушкине и Shakespeare, 24 января 2012 г.

Через год (24 января 2012 г.) изменился дизайн Гугла (строка «результатов» стала гораздо бледнее), «капитализация» классиков выросла, но по-разному, *Пушкин* зарабатывал лишь на треть больше гугликов, чем *Shakespeare* – 202 млн против 151 млн (рис. 3).

Допускаем, что большинство читателей текстов Эпштейна искренне верят в гуманетические гуглики, а нашей статистике (точнее, результатам, полученным нами от того же Гугла) без скриншотов поверить невозможно: она противоречит здравому смыслу.

В чем дело? Первичная причина разницы результатов – в устройстве поисковых машин. Эпштейн искал from Atlanta, а мы из Москвы (при попытке сменить регион компьютера: на Тамбов или Владивосток Гугл позволяет, а за границу не пускает). Главное же в том, что бесплатный сыр бывает только в мышеловке: это только кажется, что мы с Эпштейном бесплатно воспользовались поиском от Гугла, за нас заплатили производители товаров и услуг. Гугл заинтересован в том, чтобы и они, и те, кто ищет информацию, получили максимум удовольствия. Производители – от результатов рекламы, а пользователи – от найденных товаров и услуг. Гуглу безразлично, что мы ищем: ондулин, ближайшую аптеку, хостел

в Стамбуле, Пушкина или сиамскую кошку. Главное – вставить в начало выдачи то, что обрадует рекламодателя и вполне удовлетворит прототипического поисковика (а он ищет товары и услуги, и не где-нибудь вдалеке, а у себя дома). Причем пользователю приятно осознавать, что то, что надо, он получил из миллиона возможностей, а не из полутора сотен.

Писатели-классики стали жертвой общего алгоритма поисков ондулина и хостелов: «вот тебе лучшие результаты, мы их отобрали *специально для тебя* из мно-о-огих миллионов (смотри число результатов)».

Лингвисты знают, что средний носитель русского языка не любит рассуждений о колебаниях языковой нормы, а всегда хочет получить однозначный ответ на вопрос: как правильно? Для большинства главным авторитетом давно стал Яндекс, у которого «Найдется всё!». Профессионалы знают, что определение правильного – дело непростое, учитывать надо разное, в том числе и статистику узуса. Многие наивно полагают, что такую статистику можно получить от того же Яндекса, который объективно отражает если не правду жизни, то фактическое положение в Интернете. «Написание *Таллинн* встречается в Интернете 6 млн раз, а *Таллин* 4 млн раз, употребление предложно-падежной словоформы *в Украине* использовано на 62 млн страниц Интернета, а словоформа *на Украине* употреблена на 60 млн страниц. Написание *Кыргызстан* использовано 6 млн, а *Киргизия* – 10 млн раз» [Кузнецов, 2009, с. 32].

Нас интересует и статистика по беспредложной форме, поэтому вместе с Яндексом будем анализировать не омонимичный дательно-предложный *Украине*, а однозначный винительный.

Из «достоверных источников» мы знаем: сочетания с предлогами *на* и *в* составляют около половины от общего числа вхождений в русские тексты винительного *Украину*. Яндекс в разное время в разных местах дает разные результаты, часто парадоксальные (рис. 4).

Решить уравнение 136 = 310 + 321 + x несложно, x = –495. Но интерпретации эта цифра не поддается.

В табл. 1 приведены несколько парадоксальных сведений про *Украину* от Яндекса; получены они были во времена, когда Яндекс нумеровал не *страницы* (как когда-то) и не *результаты* (как сейчас), а *ответы*; внизу поисковой страницы можно было поменять дефолтный поиск «по релевантности» на поиск «по дате». На основной поисковой странице можно было отметить регион, по которому происходит поиск, вплоть до райцентра (он мог не совпадать с локализацией ищущего компьютера; Угловка, где происходили поиски 6–10, находится в Окуловском районе Новгородской обл.). Цифры из любой клетки могут показаться разумными, но почти при любом сопоставлении по строкам и столбцам виден полный абсурд.

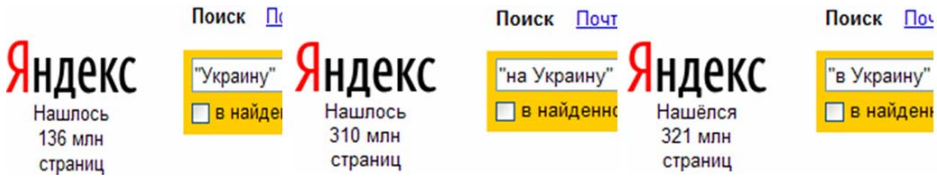

Рис. 4. Словоформа «Украину» с предлогами и без
(Угловка Новгородской обл., 12.08.2011)

Таблица 1

Результаты запросов к Яндексу в зависимости
от вариантов словоформы и поискового региона (млн страниц)

№ п/п	Запрос	А *на Украину*	Б *в Украину*	В *Украину*
Поиск от 12.08.2011 в Угловке Новгородской обл.				
1	Без ограничения региона	310	321	136
Поиск от 14.03.2013 в Петербурге				
2	Без ограничения региона	0,138	0,196	3
3	В Санкт-Петербурге	0,951	2	2
Поиск от 15.03.2013 в Москве				
4	Без ограничения региона	4	14	5
5	В Москве	3	6	69
Поиск от 23.06.2013 в Угловке Новгородской обл.				
6	В Москве	3	5	86
7	В Санкт-Петербурге	0,948	2	28
8	В Великом Новгороде	0,278	0,492	0,736
9	В Окуловке по релевантности	0,003	0,003	0,000967
10	В Окуловке по дате	0,013	0,010	0,003

Дабы не перегружать текст картинками, в подтверждение приведем лишь рис. 5, из которого следует, что в Петербурге было на порядок больше ответов, релевантных поиску *в Украину*, чем в рунете в целом (2Б и 3Б в табл. 1).

Рис. 5. Общий поиск словоформы «в Украину» и поиск по региону
(Санкт-Петербург, 14.03.2013)

Пока речь идет о поиске информации, гуглить полезно, а вникать в предъявляемые цифры – вредно. Не следует думать, что лингвист вообще не может получить полезную информацию с помощью поисковых машин. Если выдача не превышает тысячу, при аккуратной работе с их помощью можно извлечь массу интересного, особенно при поиске по конкретному сайту. Мы много раз с выгодой использовали поиск Гугла по сайтам magazines.russ.ru и feb-web.ru или их разделам [6]. До сентября 2015 г. исключительно продуктивным был расширенный

[6] Например, поиском по feb-web.ru/feb/mas автор статьи выяснял, откуда именно брались «примеры правильного употребления слова» [МАС, т. 1, с. 6]. Источниками трехсот и более цитат в этом словаре стали следующие произведения: «Порт-Артур» (671 цитата) «Воскресение» (574), «Капитанская дочка» (537), «Необыкновенное лето» (506), «Жатва»

поиск Яндекса по блогосфере, где имелась возможность поиска с учетом пола и возраста авторов как в целом, так и по отдельным временны́м отрезкам и регионам.

В декабре 2010 г. появился инструмент, который позволяет «воочию увидеть жизнь слова: когда оно появилось, когда стало модным, когда начало исчезать из текстов» [Кронгауз, 2013, с. 17]; это Google Books Ngram Viewer [7] (далее – GBN). Не удивительно, что многие лингвисты серьезно воодушевились новыми перспективами. Чрезвычайно богатые возможности корпусной лингвистики к этому времени уже все осознали, а собрание Google Books по многим приметам – это готовый многомиллиардный корпус, GBN – инструмент для работы с ним.

Однако инструмент этот создавался без участия лингвистов и не для них. Из статьи «Google Ngram Viewer» в Википедии можно узнать, что прототип этого инструмента создали J.-B. Michel и E. Lieberman Aiden, а в статье «Culturomics» они названы соавторами имени новой науки, сама же культуромика по задачам напоминает гуманетику – это «a form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitized texts». Слово «lexicology» в этом контексте не должно вводить в заблуждение: дисциплина, изучающая «human behavior» и «cultural trends», к лингвистике имеет косвенное отношение. Прежде чем использовать GBN для выявления чего-то, пока неизвестного, стоит посмотреть, как он отражает историю лексики там, где мы имеем о ней общее представление.

Мы подозревали, что не все молодые лингвисты умеют расшифровывать РККА, однако непредставительный опрос показал, что и те, кому за 50, не всегда знают, что это официальное сокращение от *Рабоче-Крестьянская Красная Армия*, которая после февраля 1946 г. стала называться *Советской Армией*. Другая аббревиатура, сопоставляемая с РККА на рис. 6, сейчас общеизвестна, хотя полвека назад в СССР она была знакома далеко не всем: «мичуринская генетика» только-только начала уступать место той, которая на два с лишним десятилетия получила в СССР статус лженауки. Графики на рис. 6 приведены со стандартным сглаживанием; отменив его, можно узнать, что частота слова ДНК в русском языке в 1964 г. составляла 19,6 ipm, за предыдущие четыре года – в среднем 11,6 ipm, в 2000 г. – 17,8 ipm, в последующие восемь лет – в среднем 12,8 ipm [8]. За последний год, о котором GBN дает сведения (2008), у ДНК ipm был ниже, чем у РККА, – 9,6 и 13,4 соответственно.

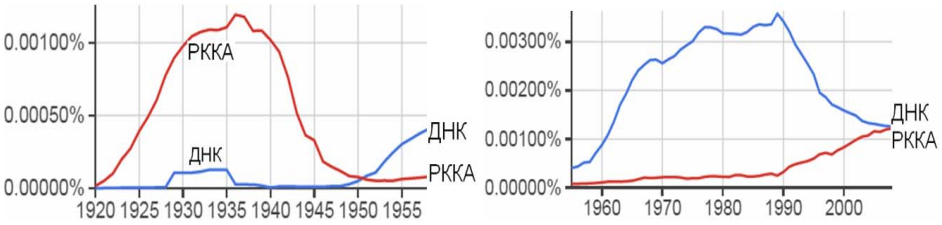

Рис. 6. «Жизнь слов» ДНК и РККА в русском языке по версии GBN

(481), «Отцы и дети» (456), «Небо и земля» (399), «Весна на Одере» (345), «Энергия» (336), «Сестры» (305), «Накануне» (309), «Братья Карамазовы» (309) [Беликов, 2006].

[7] https://books.google.com/ngrams

[8] Недолюбливая большое число нулей после запятой, мы переводим проценты GBN в общепринятый для частотных словарей показатель ipm (instances per million) – вхождений на миллион словоупотреблений.

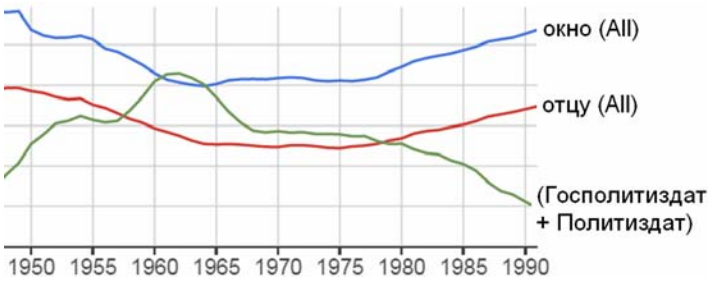

Рис. 7. Холодная война и частотность лексики

Частоты слов на любом текстовом массиве отражают не факты языка, а состав самого массива. Протестовать против неуклонного падения частотности аббревиатуры ДНК в 1990-е – 2000-е гг. при параллельной экспансии РККА бессмысленно, мы действительно воочию видим жизнь слов, только не в русском языке, а в подборке Google Books за этот период [9].

Вряд ли кто-то готов утверждать, что слова вроде *окно*, *дверь*, *дорожка*, *грызть*, *подсматривать*, *веселый*, *печально*, *утром*, *вечером*, *мать*, *отец* и им подобные в 1950-е гг. выходили из моды в обиходе и в книжно-журнальной [10] продукции, а в 1980-х мода на них вернулась. Но Google Books дают именно такую картину (рис. 7) [11].

Понятно, что в период «холодной войны» комплектация американских библиотек литературой на русском языке имела серьезный политический уклон. Разумеется, слова *Госполитиздат* и *Политиздат* обычно появлялись в книгах лишь на титуле и в выходных данных и не они потеснили обиходную лексику, а содержимое книг этих издательств. Карл Каутский (1854–1938), названный Лениным ренегатом, ко времени «холодной войны» актуальность давно утратил, но в GBN *Каутский* с 1959 по 1970 г. стабильно частотнее, чем *Лесков*, а *ренегат* с 1960 по 1978 г. частотнее, чем *племяннице*. Эта статистика отражает не «human behavior» или «cultural trends» в СССР, а старательную оцифровку брошюр Политиздата.

На работе с текстами по первую четверть XX в. включительно сказывается загадочное обращение в системе GBN с элементами дореформенной орфографии и графики (использовать дореформенные *ѣ*, *і*, *ѳ* при поиске невозможно). Существенно различается вероятность распознания ятя в разных словоформах одной лексемы, ср. частоты для *онъ ѣстъ* и *нечего ѣсть* (рис. 8) и конечного ера в разных словах, ср. *домъ* и *котъ* (рис. 9); гипотеза о том, что безъеровый *кот* – сокращение от *который* не проходит: текстовый массив с таким количеством сокращений и преобладанием *Кот.* над *Котъ* представить невозможно.

---

[9] Полезно сопоставить статистику GBN с данными подкорпуса НКРЯ за 2000–2008 гг. (56,5 млн слов); в нем ДНК имеет частоту 31,2 ipm, а РККА – 2,5 ipm.

[10] Среди Google Books представлены не только книги, но и журналы.

[11] GBN не позволяет выявлять частотность слова в целом; графики разных словоформ выше перечисленных единиц похожи, но совместить их на одной картинке не всегда удается из-за значительных расхождений по оси ординат, связанных с различиями по частоте. «All» на графиках означает учет всех вариантов капитализации (*окно*, *Окно*, *ОКНО*), «+» объединяет единицы, частоты которых складываются. На этом и ряде других рисунков ось ординат не представлена, поскольку здесь важен лишь характер изменения графиков во времени, а не их численные значения.

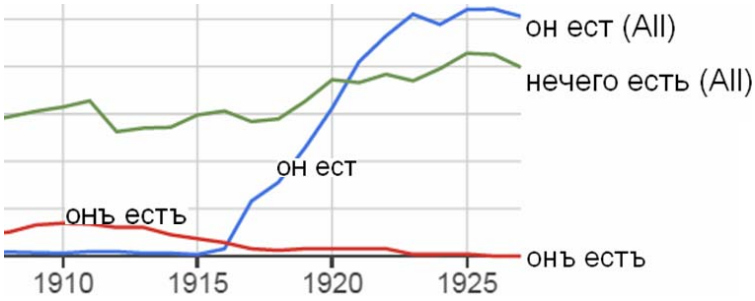

Рис. 8. Ять в разных словоформах

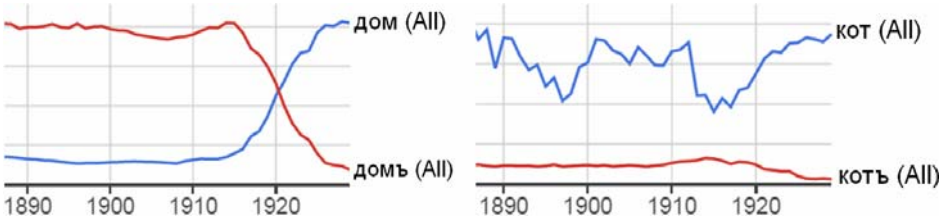

Рис. 9. Разнобой в передаче конечного ера

К каким выводам подталкивает лингвиста анализ результатов работы GBN? Видится только один, одновременно тривиальный и фундаментальный: русский язык, как и любой другой полифункциональный язык, неоднороден. Исследование с и с т е м ы русского языка «вообще» – занятие разумное, и при таком подходе любой корпус полезен для подбора материала. Для изучения н о р м ы надо сначала определить ее границы, что и прежде было непросто, а теперь еще сложнее. Корпус нормативного языка придется подбирать, все время оглядываясь на заранее постулированные границы.

А вот у з у с языка «вообще» изучать бессмысленно. Узус всегда социален, он имеет три важнейших измерения – территориальное, возрастное и гендерное; есть и другие, но они менее универсальны. Социально маркированный узус диктует и подлинные нормы, явленные в текстах, которые соответствующим социумом рассматриваются как образцовые. Жители разных регионов, молодежь и старики, мужчины и женщины не всегда будут единодушны в оценке качества текстов. Дополнительную сложность накладывает многообразие жанров и тематики.

Обращение к цифровым текстам с целью получить некую статистику, касающуюся языка, всегда будет иметь результат, выраженный числами. Иногда, как в примерах про Пушкина и Украину, разные обращения к одному и тому же материалу дают противоречивые результаты, даже если инструментарий один и тот же и время обращения к корпусу отличается совсем незначительно, например минутами (ср. строки в табл. 1). Иногда, как при обращении к GBN, разница незаметна [12]. Стабильность результатов эксперимента – положительное свойство, но оно

[12] Конечно, объем оцифрованных текстов в корпусе Google Books быстро растет, но завершающийся 2008 годом подкорпус, на котором работает Ngram Viewer, последние годы

не гарантирует его осмысленности. Выявление «общеязыковой частотности» слов, тем более словоформ, на материале черного ящика с неизвестным содержимым представляется достаточно увлекательной языковой игрой [13], но, в отличие от старых игр – Scrabble-Эрудит, «В балду», «В города», игра в GBN не носит развивающего характера.

В Интернете легко находятся работы, в той или иной степени опирающиеся на результаты GBN-поиска, среди них есть синхронические (сюда отнесем и те, которые касаются микроистории – нескольких недавних десятилетий) и диахронические, подвергающие анализу русский язык на протяжении двух столетий. На любые сомнения в результатах GBN-поиска у его сторонников есть контраргумент: в статистике, основанной на многомиллиардном корпусе, нельзя сомневаться. Это верно, если не с чем такие данные сопоставлять. Но в отношении длительной истории есть НКРЯ [14], незначительность его объема в сравнении с корпусом GBN отчасти компенсируется возможностью оперировать словом в целом, а не отдельной словоформой и – что для нас главное – прозрачностью состава корпуса, на котором получен результат. Если результаты просто разные, это мало о чем говорит. А вот если они прямо противоположны [15], то результат от НКРЯ правильный, а результат от GBN вздорный. Впрочем, это вопрос веры.

В том, что касается синхронии, мы свою аргументацию строим на материалах Генерального интернет-корпуса русского языка [16] (далее – ГИКРЯ), который разрабатывается с 2012 г. и в конце 2015 г. вышел из стадии начальной разработки, свободный доступ к нему открыт пока при условии регистрации. Поиск в ГИКРЯ ведется раздельно по сегментам, содержащим разные типы текстов: большие выборки из «Живого журнала» (ЖЖ) (2001 – начало 2014 г.), из социальной сети «ВКонтакте» (выборка за 2014–2015 гг.), из новостных лент за 1999–2013 гг. и «Журнального зала» (собственно журнальные статьи с портала magazines.russ.ru по состоянию на апрель 2014 г.). Общий объем ГИКРЯ приближается к 20 млрд словоупотреблений.

Кратко остановимся на языковых фактах, упоминавшихся выше, но начнем с комментариев к GBN.

Один автор [17] в 2015 г. опубликовал две работы (вторая в соавторстве), посвященные разным аспектам современного поведения русских глаголов в сравнении с ситуацией в XIX в. [Язык и мысль, 2015, с. 478–487; Труды…, 2015, с. 425–434]; в первой из них указаны и точные временные рамки – 1800–1860 гг. и 1993–2008 гг.

GBN не дает возможности работать с лексемами, поэтому, проанализировав результаты для словоформ *старался*, *стараясь*, *стараться*, *стараюсь* и *пытался*, *пытаясь*, *пытаться*, *пытаюсь*, автор счел достаточным «работать только с формой прошедшего времени, как обеспечивающей наибольший массив данных. Общие закономерности, найденные для этой формы, остаются справедливыми и для других форм» [Язык и мысль, 2015, с. 483] (исследовались инфинитивы, следовавшие за формами *старался* / *пытался*).

не меняется. Во всяком случае, наши поиски двухлетней давности и современные дают одинаковые результаты, то же и при повторении более давних чужих поисков.

[13] Речь идет лишь о русском варианте GBN.

[14] НКРЯ – Национальный корпус русского языка. URL: http://www.ruscorpora.ru

[15] Объем публикации не позволяет вдаваться в детали, но автор статьи проверял: так бывает.

[16] www.webcorpora.ru

[17] Нынче принято за всякое упоминание конкретной работы начислять автору разнообразные очки. Не разделяя эту точку зрения, предпочитаем цитировать некоторые работы без имен, хотя и вполне прозрачно.

Начало цитаты сформулировано несколько небрежно: форм прошедшего времени больше одной, причем само их существование заставляет усомниться в справедливости вывода про «общие закономерности». Лишь часть маркированных по полу глаголов не имеет отношения к автору высказывания (*умер* / *умерла*), но не таковы *старался* и *пытался*. Поскольку одни действия более свойственны мужчинам, другие женщинам, вероятность появления гендерно маркированных глаголов после форм *старался* и *старалась* заведомо различна. Но вторая работа [Труды…, 2015] на этом фоне преподносит сюрприз: там репрезентативными словоформами служат инфинитивы (анализируются тройки *дохнуть* [18] / *подохнуть* / *подыхать* и *грызть* / *разгрызть* / *разгрызать*).

В теории для некоего глагола может найтись полноценная словоформа-представитель, но для другого глагола таковой окажется другая словоформа. И представительность таких словоформ надо доказывать. Как – не очень ясно, но очевидно, что на фоне выявления свойств всех других словоформ этой же лексемы. Тем самым искать представительную словоформу занимательно, но бессмысленно: все равно предварительно надо все узнать про лексему. И – если рассуждать о русском языке «вообще» – для *всех* вариантов русского языка.

В применении к языковым корпусам часто используется характеристика «сбалансированный»; в теории это означает, что тексты, разнящиеся в социальном и жанровом отношении, представлены в корпусе в «правильных» долях. Но исчислением всех таких вариантов и выявлением их стандартного для русского языка соотношения никто никогда не занимался, поскольку это занятие утопическое. Фактически сбалансированным называется корпус, который его создатели считают сбалансированным.

То, что слова в различных вариантах языка используются несколько поразному, удивлять не должно. Следствием этого является различие в частоте употребления как лексем, так и отдельных словоформ. На рис. 10 отражена эволюция частот двух словоформ в 1990–2008 гг., графики не сглажены, так что известны точные погодовые частоты: ipm у *старался* в течение периода, по сути, неизменен, около – 19. Частоты формы *пытался* равномерно растут от 25,8 ipm до 34,1 ipm (в 1992 г. отмечается небольшое проседание для обеих словоформ, 15,9 и 22,6 ipm). Соотношение частот неуклонно повышается в пользу словоформы *пытался*: в начале периода она на 30 % частотнее старался, в 2000 г. – на 48 %, в 2008 г. – на 81 %.

У других пар словоформ соотношение складывается либо похожим образом (*пытается*/*старается*, *пытаясь*/*стараясь*) – частота форм от *стараться* меняется мало, формы от *пытаться* становятся частотнее, либо их развитие идет параллельно, хотя и по-разному: в первом лице частота форм множественного числа не меняется, а формы единственного числа становятся частотнее (рис. 11).

Таким образом, в корпусе GBN за 19 лет у лексемы *пытаться* частота выросла явно заметнее, чем у *стараться*. Но справедливо ли это наблюдение для русского языка? Посмотрим, как складывается судьба тех же глаголов в НКРЯ.

В табл. 2 сопоставлены данные за два периода – НКРЯ-1 (1990–1999 гг., 23,2 млн слов) и НКРЯ-2 (2000–2008 гг., 56,5 млн слов). Частоты большинства словоформ этой пары глаголов в 1990-е гг. в НКРЯ заметно выше, чем GBN (для *пытался* и *пытаясь* – в 3 раза, для *пыталась* – в 4,5). Во-вторых, у обеих лексем

[18] Исследование инфинитива *дохнуть* без обращения к контекстам грозит обернуться курьезом. Имея некоторый опыт чтения старых и современных русских текстов, мы предполагали, что соотношение *дохну́ть* и *до́хнуть* в них заметно разнится. Так и оказалось: в НКРЯ (май 2016 г.) за 1800–1899 гг. есть 133 *дохну́ть* и 8 *до́хнуть*, с 1990 г. соответственно 18 и 9.

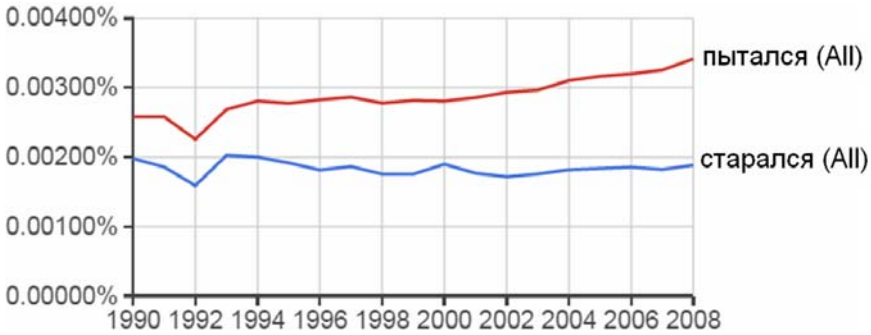

Рис. 10. Словоформы *пытался* и *старался* в GBN, 1990–2008 гг. (без сглаживания)

Рис. 11. Различные словоформы глаголов *пытаться* и *стараться* в GBN, 1990–2008 гг. (стандартное сглаживание)

Таблица 2

Динамика частотности отдельных словоформ
по данным НКРЯ (ipm)

Словоформа	НКРЯ-1	НКРЯ-2
Старался	46,0	25,4
Пытался	80,9	54,4
Старалась	15,6	11,8
Пыталась	26,9	20,9
Старались	21,9	15,3
Пытались	39,0	34,5
Стараюсь	11,4	12,5
Пытаюсь	12,1	11,2
Стараемся	4,2	7,5
Пытаемся	3,4	5,9
Старается	14,0	12,9
Пытается	30,6	34,2
Стараясь	44,1	27,4
Пытаясь	47,4	36,0
Лексема *стараться*	187,0	143,9
Лексема *пытаться*	285,6	251,8

в XXI в. они снизились: у *пытаться* на 12 %, у *стараться* на 23 %. Зато в текстах после 2008 г. (10,7 млн слов) частоты обоих глаголов в НКРЯ выросли, а разрыв частот стал уже двукратным (152,3 у *стараться*, 319,6 ipm у *пытаться*).

Итак, мы знаем как обстоит дело в двух корпусах. А что на самом деле происходит с рассматриваемыми словами в современном русском языке? По-видимому, ничего особенного. В этом отношении полезно выяснить положение в разных сегментах ГИКРЯ, язык каждого из которых относительно однороден. Данные табл. 3 (где, как и в табл. 4, в объем лексем не включены причастия; учитываются спрягаемые формы, прошедшее время, инфинитив, императив, деепричастие) показывают, что в зависимости от типа текста существенно меняется и частота лексем, и вклад отдельных словоформ: в «Журнальном зале» и *старался*, и *пытался* дают четверть вхождений лексемы, а в «Новостях» – 8 и 17 % соответственно. В «Журнальном зале» частотность словоформ *старался* и *пытался* разнится менее чем в два раза, в «Новостях» – более чем в 10 раз. Проверка по этим двум сегментам ГИКРЯ показывает, что диахронических изменений в поведении двух рассматриваемых глаголов нет: колебания частот не превышают нескольких процентов, то увеличиваясь, то уменьшаясь. Достаточно очевидно, что рост частоты *пытаться* в GBN и ее падение в НКРЯ, а также увеличение разрыва частот этих глаголов в обоих корпусах – всего лишь следствие изменений в жанрово-стилевом и тематическом наполнении самих корпусов.

Таблица 3

Частотность лексем и словоформ в зависимости от типа текста

Слово	Журнальный зал	Новости	ЖЖ	ВКонтакте
Лексема *стараться*	121,9	35,7	116,7	124.7
Лексема *пытаться*	236,2	175,5	275,0	227,8
Старался	30,3	2,9	11,2	8,6
Пытался	58,8	30,2	39,3	26,1

Таблица 4

Вариативность статистической структуры парадигм
в текстах разных сегментов ГИКРЯ

Сегмент ГИКРЯ	Журнальный зал	Новости	ЖЖ	Журнальный зал	ЖЖ
Глагол	*Стараться*	*Стараться*	*Стараться*	*Подохнуть*	*Грызть*
Всего вхождений	38 271	30 457	1 021 066	747	81 801
Инфинитив, %	4,04	12,80	10,55	14,32	**33,3**
1 л. ед. ед. ч., %	7,76	4,90	24,74	9,50	7,84
1 л. мн. ч., %	1,98	12,32	1,98	5,62	1,56
3 л. ед. ч., %	10,39	16,59	9,31	12,05	21,44
3 л. мн. ч., %	6,35	22,72	9,88	5,09	8,79
Прош. муж. р., %	24,76	8,02	9,58	20,74	7,66
Прош. жен. р., %	9,23	3,17	8,17	7,76	6,63
Прош. мн. ч., %	10,91	12,75	7,33	11,51	5,11
Деепричастие, %	21,16	4,94	7,63	0,14	1,60

Как обстоит дело в профессиональных текстах химиков и математиков, в детективах и детской литературе – знать не дано. Зато мы получаем очередное подтверждение важному, мы бы сказали, основополагающему для статистического

исследования языка тезису: «для большинства лингвистических и лексикографических задач корпусной анализ должен проводиться с точностью до четко определенных жанровых и социолингвистических границ» [Беликов и др., 2013, с. 84].

Рассмотренных фактов достаточно для констатации наивности поиска словоформ, способных адекватно представлять лексему, но в дополнение приведем табл. 4, раскрывающую вариативность статистической структуры парадигм в текстах разных сегментов ГИКРЯ.

Как видим, доля различных словоформ меняется не только в текстах разных типов, но и существенно зависит от глагола. Кроме того, априори понятно, что, хотя использование формы *старался* не обязательно характеризует пол автора, выбор для анализа форм прошедшего времени в мужском роде сдвигает материал в сторону лиц мужского пола.

Исследование гендерных характеристик узуса до недавнего времени было затруднено. Единственным источником достаточно больших спонтанных текстов, размеченных по полу авторов, был поиск Яндекса по блогам, но ограничение выдачи тысячью записей создавало серьезные трудности.

В сегменте ГИКРЯ «ВКонтакте» тексты маркированы по полу в соответствии с профилем авторов. Из общего числа словоупотреблений (9 820,5 млн) по полу размечено 47,38 %, из которых почти две трети составляет объем женского подмассива. Тем самым прямое сравнение статистики мужского и женского узуса невозможно, количество фактических словоупотреблений следует сопоставлять с показателем, нейтрализующим преобладание женщин в этом сегменте. Назовем отношение женских словоупотреблений (3 009,15 млн) к мужским (1 644,11 млн) – 1,8 – гендерным коэффициентом «ВКонтакте» ГИКРЯ (ГК). Этот показатель рассчитан по полному корпусу «ВКонтакте» ГИКРЯ, содержащему значительное число дублирующихся текстов, истинное авторство которых в общем случае неустановимо; по умолчанию выдача производится с дедубликацией.

Поскольку тематика текстов серьезно влияет на словоупотребление, не стоит видеть гендерного маркирования в том случае, если ГК находится в диапазоне 1,2–2,7 (это 50 %-е преобладание в выдаче мужского vs. женского узуса) [19]. А вот бóльшее отклонение от нейтрального ГК = 1,8, например, двойное преобладание мужского или женского узуса (0,9 и ниже либо 3,7 и выше) при достаточно больших абсолютных числах явно свидетельствует о неслучайном преобладании в анализируемом материале мужских vs. женских текстов.

Разумеется, переносить гендерную статистику, полученную на материале молодежной социальной сети, на другие варианты русского языка следует с осторожностью. Но вот поучительные данные, ставящие под серьезное сомнение репрезентативность форм мужского рода не только для лексемы, но и для использования ее в прошедшем времени (табл. 5). Говорить о гендерной маркированности всех четырех приведенных в таблице глаголов нельзя, хотя 320 тыс. примеров на глагол *пытаться* от распределенных по полу авторов – достаточно убедительное основание, чтобы считать эту лексему более свойственной мужчинам, чем женщинам, в частности, при сопоставлении со *стараться*.

[19] Это касается лексики. В том случае, если использование какой-то грамматической категории дает регулярное отклонение ГК в мужскую или женскую сторону, можно говорить о гендерном маркировании. Например, у прилагательных ГК сравнительной степени регулярно заметно отклоняется в мужскую сторону в сравнении со склоняемыми словоформами. Впервые автор статьи на это обстоятельство натолкнулся в июне 2012 г., анализируя блогосферу [Беликов, 2014, с. 126–127], но ограничения Яндекса на объем выдаваемых данных сдерживали уверенность в обнаруженном явлении.

То, как формы прошедшего времени используются мужчинами и женщинами, ясно показывает, что при поиске на глаголы в мужском роде мы получаем сведения не о русском языке «вообще», а в основном о его мужской разновидности. Не исключено, что в исследуемом явлении гендерных противопоставлений не окажется, но знать этого заранее нельзя.

ГИКРЯ дает ясный ответ и на вопрос, каков реальный узус при склонении слова *Украина*. Априори ясно, что он достаточно быстро меняется и не может быть одинаков в текстах разных регионов России и Украины. Сегмент ЖЖ НКРЯ позволяет получить надежные данные в динамике лишь по столицам [20] (табл. 6); в 2013 г., последнем представленном в таблице году наблюдения, в Москве все еще лидирует предлог *на* (3 483 против 3 338 для *в Украине*), в Петербурге в том же году впервые возобладал предлог *в* (929 против 905 для *на Украине*).

Таблица 5

Формы прошедшего времени в мужском и женском узусе

Словоформа	Примеры, где пол автора известен	ГК	Словоформа	Примеры, где пол автора известен	ГК
Всего от *стараться*	184 526	1,6	Всего от *пытаться*	320 680	1,3
Старался	16 162	0,7	*Пытался*	37 858	0,7
Старалась	12 225	5,7	*Пыталась*	24 273	4,0
Старались	12 089	1,7	*Пытались*	25 510	1,2
Всего от *подохнуть*	1 750	0,8	Всего от *грызть*	7 612	1,7
Подох	22	0,4	*Грыз*	507	1,1
Подохла	199	1,1	*Грызла*	356	2,8
Подохли	26	0,4	*Грызли*	475	1,1

Таблица 6

Варианты выбора предлога при склонении лексемы «Украина» по данным «Живого журнала» в динамике

Регион	2004–2008 гг.			2009–2013 гг.		
	на Украине	*в Украине*	*на*/*в*	*на Украине*	*в Украине*	*на*/*в*
Москва	2 565	976	2,6	9 596	7 250	1,3
Петербург	668	247	2,7	2 455	2 007	1,2
Киев	774	3 687	0,2	2 938	19 640	0,1

«Живой журнал» во многом отражает спонтанное повседневное словоупотребление блоггеров, узус «Журнального зала» более консервативен (табл. 7).

[20] Ниже мы рассматриваем предложный падеж как дающий больше результатов. Для достоверной статистики по другим городам с большим числом блоггеров ЖЖ выборка ГИКРЯ пока мала. Так, по состоянию ЖЖ ГИКРЯ на июнь 2016 г. результаты по Новосибирску показывают неожиданный скачок: в 2004–2008 гг. соотношение предлогов *на* / *в* составляет 68/37 = 1,8, в 2009–2011 гг. – 161/77 = 2,1, а в последующие два года – 140/230 = = 0,6. Такой резкий перелом в пользу нового предлога невероятен и объясняется недостатком данных.

Таблица 7

Варианты выбора предлога при склонении лексемы «Украина» по данным «Журнального зала» в динамике

Период	*На Украине*	*В Украине*	*На* / *в*
1993–2002	467	93	5,0
2003–2006	618	248	2,5
2007–2013	1 291	627	2,1

Таблица 8

Варианты выбора предлога при склонении лексемы «Украина» по данным «Новостей» в динамике

Период	*На Украине*	*В Украине*	*На*/*в*
1999–2002	396	50	7,9
2003–2007	18 351	5 590	3,3
2008	9 609	1 842	5,2
2009–2013	40 559	5 578	7,3

Наиболее интересной оказалась ситуация в новостном сегменте, где процесс перехода от предлога *на* к *в* после середины 2000-х гг. пошел вспять [21] (табл. 8).

Вернемся к заголовку статьи: «Что и как может получить лингвист…». Формулировка вроде бы подразумевает достижение положительных результатов, но речь в основном была не о них. Нам представлялось очень важным показать, что нередко лингвисты, сами того не подозревая, получают из оцифрованных текстов недостоверный результат.

Закончим цитатой, которая на первый взгляд констатирует очевидное: «Интернет… мог бы внести статистическую ясность во множество вопросов, которые сейчас обсуждаются на уровне интуитивных догадок или метафизических умозрений. Речь идет о частоте употребления тех или иных слов, имен, терминов, понятий, идиом в разных национальных сегментах интернета. Интернет уже сейчас самое большое хранилище информации, накопленной в живых языках. А главное, все данные, которые в нем хранятся, поддаются мгновенной статистической обработке» [Эпштейн, 2006, с. 71].

Как раз в *главном* философ ошибается: полезной для лингвиста (или культуролога) «мгновенной статистической обработке» данные Интернета не поддаются. Результаты обработки этих данных средствами, созданными с заведомо нелингвистическими целями, лишь множат «метафизические умозрения». Интернет – необозримый источник оцифрованных текстов, которые становятся материалом для языковых корпусов, но такие корпуса (и непременная часть корпуса – инструментарий для работы с ним) должны создаваться при активном участии профессиональных лингвистов.

Лингвист может извлечь пользу и непосредственно из Интернета, но если он хочет получить нечто большее, чем просто примеры словоупотребления, его ждет масса подводных камней, частично предсказуемых, частично неожиданных.

[21] То же и с винительным падежом, но абсолютные цифры там существенно меньше – 1999–2002 гг.: 135/7 = 19,3; 2003–2007 гг.: 4 216/1 137 = 3,7; 2008 г.: 3 194/619 = 5,2.

Список литературы

Беликов В. И. Интернет и орфография // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. «Диалог'2003». М.: Наука, 2003.

Беликов В. И. Yandex как лексикографический инструмент // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. «Диалог' 2004». М.: Наука, 2004.

Беликов В. И. Словарь «Языки русских городов»: подбор примеров и Интернет // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. «Диалог 2006». М.: ИПИ РАН, 2006.

Беликов В. И. К методике корпусного исследования лексики // Русский язык и новые технологии. М.: НЛО, 2014.

Беликов В. И., *Копылов Н. Ю.*, *Пиперски А. Ч.*, *Селегей В. П.*, *Шаров С. А.* Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегод. Междунар. конф. «Диалог». Вып. 12(19). Т. 1. М.: РГГУ, 2013.

Кронгауз М. А. Самоучитель олбанского. М.: АСТ, 2013.

Кузнецов С. А. Языковая норма и правила речевой деятельности // Комментарий к Федеральному закону «О государственном языке Российской Федерации». Ч. 1: Доктринальный и нормативно-правовой комментарий. СПб.: Изд-во С.-Петерб. ун-та, 2009.

МАС: Словарь русского языка: В 4 т. / Под ред. А. П. Евгеньевой. 2-е изд., испр. и доп. М.: Рус. яз., 1981–1984.

Труды международной конференции «Корпусная лингвистика – 2015». СПб.: СПбГУ, 2015.

Щерба Л. В. О трояком аспекте языковых явлений и об эксперименте в языкознании // Языковая система и речевая деятельность. Л.: Наука, 1974.

Эпштейн М. Н. Слово недели: нумеризм // Дар слова. Проективный лексикон Михаила Эпштейна. 2003. № 71(111). 8 сент. URL: http://www.emory.edu/INTELNET/dar71.html

Эпштейн М. Н. Мысли в числах: Америка и Россия в зеркалах интернета // Философский век: Альм. Вып. 32: Бенджамин Франклин и Россия: к 300-летию со дня рождения. Ч. 2. СПб.: Центр истории идей, 2006.

Язык и мысль: Современная когнитивная лингвистика. М.: ЯСК, 2015.

V. I. Belikov

Moscow State University, Moscow, Russian Federation; otipl@philol.msu.ru

What and how can a linguist get from digitized texts?

The article is devoted to the limits of applicability of the online tools for automatic processing of digital texts (search engines, corpora, Google Books Ngram Viewer) to a linguostatistic study. Despite the common opinion about the objectivity of the results obtained after automatic processing of the text array, there are limitations and distortions of the data due to many reasons. One of them is the frequent lack of linguists in the teams of developers of such machines. In the article, the analysis of the frequency of use of culturally significant names and their spelling variants, generic forms of the verb and the prepositional variants of the control according to the different automatic means of analysis of the texts shows the complexity of interpreting the results of automatic processing of text arrays.

Keywords: digital text, linguistic statistics, linguistic online tools, word processing, corpus linguistics, grammatical variation.

DOI 10.17223/18137083/56/2

References

Belikov V. I. Internet i orfografiya [Internet and spelling]. In: *Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunar. konf. «Dialog'2004»* [Computational linguistics and intellectual technologies: Works of Intern. Conf. «Dialog'2004»]. M., Nauka, 2004.

Belikov V. I. K metodike korpusnogo issledovaniya leksiki [To the methodology of the corpus study of vocabulary]. In: *Russkiy yazyk i novye tekhnologii* [Russian language and new technologies]. M., NLO, 2014.

Belikov V. I. Slovar' «Yazyki russkikh gorodov»: podbor primerov i Internet [The dictionary of the «Languages of Russian cities»: a selection of examples and the Internet]. In: *Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunar. konf. «Dialog'2006»* [Computational linguistics and intellectual technologies: Works of Intern. Conf. «Dialog'2006»]. M., IPI RAN, 2006.

Belikov V. I. Yandex kak leksikograficheskij instrument [Yandex as a lexicographic tool]. In: *Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunar. konf. «Dialog'2003»* [Computational linguistics and intellectual technologies: Works of Intern. Conf. «Dialog'2003»]. M., Nauka, 2003.

Belikov V. I., Kopylov N. Ju., Piperski A. Ch., Selegej V. P., Sharov S. A. Korpus kak jazyk: ot masshtabiruemosti k differencial'noj polnote [Corpus as language: from scalability to differential completeness]. In: *Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunar. konf. «Dialog»*. Vyp. 12 (19). T. 1 [Computational linguistics and intellectual technologies: Materials of Intern. Conf. «Dialog».]. Iss. 12 (19), vol. 1. M., RGGU, 2013.

Epshtein M. N. Mysli v chislah: Amerika i Rossija v zerkalah interneta [Thoughts in numbers: America and Russia in the mirror of the Internet] In: *Filosofskiy vek. Al'manakh. Vyp. 32. Bendzhamin Franklin i Rossiya: k 300-letiyu so dnya rozhdeniya. Ch.' 2* [Philosophical Age: Alm. Vol. 32: Benjamin Franklin and Russia: the 300th anniversary of his birth. Pt 2]. SPb., Tsentr istorii idey, 2006.

Evgen'eva A. P. (ed.) *Slovar' russkogo yazyka v 4 t.* 2-e izd., ispr. i dop. [Dictionary of Russian language in 4 vols. 2nd ed., rev. and ext.]. M., Rus. yaz., 1981–1984, vols. 1–4.

Jepshtejn M. N. Slovo nedeli: numerizm [Word of the week: numerism]. In: *Dar slova. Proektivnyy leksikon Mihaila Epshteina* [Projective lexicon of Mikhail Epstein]. 2003, 8 sept., no. 71 (111). Available at: http://www.emory.edu/INTELNET/dar71.html

Krongauz M. A. *Samouchitel' olbanskogo* [Tutorial of Olbany]. M., AST, 2013.

Kuznetsov S. A. Yazykovaya norma i pravila rechevoy deyatel'nosti [Language norm and speech activity rules]. In: *Kommentarij k Federal'nomu zakonu «O gosudarstvennom jazyke Rossijskoj Federacii». Ch. 1: Doktrinal'nyj i normativno-pravovoj kommentarij* [Commentary to the Federal Law «On state language of the Russian Federation». Pt 1: Doctrinal and legal commentary]. SPb, SPbSU, 2009.

Shherba L. V. O trojakom aspekte jazykovyh javlenij i ob jeksperimente v jazykoznanii [On the threefold aspect of language phenomena and about the experiment in linguistics]. In: *Jazykovaja sistema i rechevaja dejatel'nost'* [Language system and speech activity]. Leningrad, Nauka, 1974.

Trudy mezhdunarodnoy konferentsii «Korpusnaya lingvistika – 2015» [Proceedings of the international conference «Corpus linguistics – 2015»]. SPb, SPbSU, 2015.

Yazyk i mysl': sovremennaya kognitivnaya lingvistika [Language and thought: the modern cognitive linguistics]. M., 2015.