ЧЕГО НЕ ХВАТАЕТ В «ОЦИФРОВАННОМ МИРЕ» ЛЕКСИКОГРАФУ И СОЦИОЛИНГВИСТУ?

В. И. Беликов (vibelikov@gmail.com)

Институт русского языка им. В. В. Виноградова,
Москва, Россия

Ключевые слова: социолингвистика, лексикография, обработка информации, сегментная статистика.

WHAT ARE SOCIOLINGUISTS AND LEXICOGRAPHERS LACKING IN A DIGITIZED WORLD?

V. I. Belikov (vibelikov@gmail.com)

Vinogradov Russian Language Institute, Moscow,
Russian Federation

It is a common belief that text corpora provide the best testing ground for solving any kind of linguistic problems. As far as grammar is concerned, this may be true, but if we focus on investigating the lexicon the results often appear to be rather superfi cial. WWW contains some relatively homogeneous arrays of texts formed independently of linguists, in some cases emerging quite spontaneously. Text arrays with the most prominent social characteristics of their authors are regarded as independent Internet segments (digitized classical literature and 2010 teenager blogs are the most contrasting examples). Frequencies of the same lexical items differ greatly from one segment to another, and this statistics is very signifi cant for sociolinguistics. The main problem in applying the method of segmental statistics is the lack of a suitable instrument for automatic data processing Several case studies are presented, and the results of segmental statistics seem to be more indicative than those obtained from the Russian National Corpus

Key words: sociolinguistics, lexicography, segment statistics, data processing.

Ещё у меня есть где-то гора ссылок по корпусной лингвистике ‹…›
dimkaguarani, 29 декабря 2002 // Блогосфера

Примечание к эпиграфу: В НКРЯ словосочетание *корпусная лингвистика* не фиксируется, хотя четверть текстов с упоминанием *лингвистики* датировано там 2003 г. и позднее.

Начну с оптимистичной цитаты. «В русском языке есть глагол несовершенного вида реагировать. Его коррелятами совершенного вида могут быть несколько разных приставочных глаголов: прореагировать, отреагировать, среагировать (явление нередкое, особенно среди заимствований). Какой из этих приставочных коррелятов употребляется чаще? К каким контекстам тяготеет каждый из этих приставочных коррелятов (например, какой из них охотнее сочетается с наречием быстро)? Наконец, в какой последовательности они появляются в современном языке — одновременно или по очереди? Различается ли частота их употребления в разные периоды?»[*][1] [Плунгян 2005: 302]. На эти вопросы «лингвист может ответить с помощью Корпуса буквально за считанные минуты» [Там же].

Картина, выявляемая НКРЯ в 2011 г., представлена в Табл. 1 (число документов[2] с соответствующими глаголами в корпусе в целом, а для 1990-х и 2000-х гг. — отдельно в художественных текстах и публицистике) и в Табл. 2 (число документов, где глаголы сочетаются с наречием *быстро*, вкл. форму *быстрее*).

Комментарии начну с употребления производящего глагола. Из 22 «документов» XIX в. 14 относятся к естественным наукам. *Документ* здесь — понятие довольно условное: все тексты, взятые из трехтомника А. М. Бутлерова (1953–1958), сведены в один документ «Теоретические и экспериментальные работы по химии», датируемый «1851–1886»[3].

Таблица 1

	реагировать	**прореагировать**	**отреагировать**	**среагировать**
всего	2003	122	744	184
1800–1899	22	1	0	0
1900–1924	68	1	0	0

[1] Ограничения на объем печатаемого текста вынуждают снять некоторые частные комментарии, а сокращение аргументация делает ее декларативной. Вчетверо больший развернутый текст доклада имеется в электронной форме; в бумажном варианте наиболее значимые сокращения обозначены астериском в квадратных скобках.

[2] Произведения, представленные в Корпусе частями, считаются за один документ; ср. Прим. 3[*]. Датировки типа «1943–1999» обычно усредняются.

[3] В других случаях единый текст может достаточно искусственно делиться, так, самостоятельными документами считаются главы «Книги воспоминаний» И. М. Дьяконова; в результате документ «И. М. Дьяконов. Книга воспоминаний. Часть вторая. Последняя глава (После войны)» по объему в 25 с лишним раз уступает «единому документу» Бутлерова.

	реагировать	**прореагировать**	**отреагировать**	**среагировать**
1925–1949	120	1	2	0
1950–1989	412	31	77	20
1990–1999	249	22	152	46
2000–	1132	66	513	118
худ., 1990–99	97	4	65	26
худ., 2000–	115	12	96	28
публ., 1990–99	137	14	84	22
публ., 2000–	746	41	366	72

Вне естественнонаучной литературы примеры на *реагировать* появляются лишь с 1890 г., при этом примеры из беллетристики достаточно физиологичны: *зрачок не реагировал на свет* ‹…› (Мамин-Сибиряк), *органическая ткань* ‹…› *должна реагировать на всякое раздражение* (Чехов), *способность кожных нервов реагировать на температурные колебания* (Вересаев). В более обобщенном значении глагол фигурирует лишь у М. Горького («Мужик», 1899: *Сурков делает свои дерзости* ‹…› *на них никто не реагирует*) и текстах, отнесенных к публицистике.

Что касается глаголов совершенного вида, то *прореагировать* появляется первым (в «документе» с датой 1851–1886), и в течение длительного времени его можно считать узкопрофессиональным. В художественной прозе он фиксируется довольно поздно (1968), а до этого однажды фигурирует в дневниковой записи 1944 г. и 11 раз в химических текстах. Глагол *отреагировать* появляется в НКРЯ также как узкоспециальный — в первой половине XX в. он встретился лишь в двух текстах по психиатрии, а с 1950 г. проникает в художественную литературу. Несколько позднее (с 1960/64) в корпусе фиксируется глагол *среагировать*.

Отвлекаясь от профессиональных текстов, делаем вывод, что в повседневный узус все три глагола вошли **одновременно**, а даты появления в беллетристике — 1950–1960/64–1968 — аккуратно объясняются различиями в частотности: разрыву в 12 лет соответствует соотношение 77/31, а разрыву в 6 лет — 31/20. Новизну глаголов подтверждают лексикографы: ни одного из трех нет ни в словаре под ред. Д. Н. Ушакова, ни 17-томном БАСе, ни в МАСе.

О частоте употребления приставочных глаголов можно говорить лишь за последние 60 лет; если принять за единицу число текстов с наиболее редким глаголом *прореагировать*, то окажется, что *отреагировать* в «доперестроечном языке» появлялся в два с половиной раза чаще, а *среагировать* — на треть реже. В последующие годы употребимость двух более новых глаголов резко возросла: *отреагировать* стал абсолютным лидером, а тексты со *среагировать* стали в два раза более частотными, чем с *прореагировать*. Но сводная картина несколько маскирует реальное положение вещей. По числу текстов с глаголом несовершенного вида довольно ясно видно резкое увеличение доли публицистики: за 1950–1989 она составляла 36%, в 2000-х годах — 66%, а доля художественных текстов сократилась с 29% до 10%[*].

Попробуем выйти за пределы НКРЯ. Насколько верно, что глагол *реагировать* начал употребляться учеными-естественниками, а в общее употребление

попал не без воздействия докторов и писателей-медиков, причем лишь в самом конце XIX в.? Обращение к сегменту «Классика» Библиотеки Мошкова заставляет в этом усомниться.

Она не знала, что в этом кротком ребенке ее женственная ласка возбудит подземный огонь эротической страсти; а раз эта страсть возбудилась во мне — она реагировала на нее, и чем я больше рос, тем больше ее материнская любовь переходила в иную любовь, а моя детская привязанность в козлоногую похоть сатира, откровенничает Н. П. Огарев о своей гувернантке (1862). До первых внеестественнонаучных фиксаций в НКРЯ глагол не так уж редко встречался и в отечественной беллетристике, и в художественных переводах, и в судебных речах, и в литературной критике[*]. Видовая параллель, которую и в первой половине XX в. мы пока знаем лишь по психиатрическим текстам, столь же «древняя», ср.: *Прежде и мои «обидчики» отреагировали бы на укол в печати, а теперь все прошло без отклика* (Лесков, 1884); *Ха, ха! — отреагировал Риенцо своим странным смехом* (Э. Бульвер-Литтон, пер. 1875 г.). Об укорененности этих глаголов в общелитературной норме говорит их использование в книгах для девочек: *Эти слова встревожили Эстер. Она прижала к себе сестру и почти не реагировала на ее лепетание ‹…› Никто из подруг не отреагировал на ее слова, и после минутного возбуждения от полученных новостей девочки замолчали* (Э. Мид-Смит, пер. 1900 г.).

Глагол *прореагировать* попадает в детскую литературу также уже в начале XX в.: *Шакал не прореагировал; ему уже минуло три года, но нельзя же сердиться на оскорбление, нанесенное особой с клювом в ярд длины и сильным, как дротик* (Киплинг, пер. 1916 г.). А вот глагола *среагировать* в сегменте «Классика» нет вообще; это уже дает основание считать, что он утвердился в языке позже начала XX в.; когда именно — сказать будет можно лишь после появления достаточно представительного массива разнотипных оцифрованных русских текстов «средней половины» XX в.; пока его нет.

Рассмотрим результаты поисков на совместимость глаголов совершенного вида с наречием *быстро* (см. верхнюю часть Табл. 2). Согласно Табл. 1 глагол *отреагировать* за 1990-е — 2000-е гг. встретился в 665 документах, а *среагировать* — в 164, то есть в 4,1 раза реже.

За тот же период наречие *быстро* в непосредственном соположении с этими глаголами встретилось в 18 и 6 текстах соответственно (в три раза реже), однако при учете всех примеров, где наречие связано с глаголом (21/6) — лишь в 3,5 раза реже; десятые доли «разов» при таких цифрах вряд ли подлежат учету, так что намек на бóльшую «привязанность» наречия к глаголу *среагировать* выглядит сугубо формальным. Уровень риска каких бы то ни было рассуждений на эту тему для предшествующего периода очевиден, поэтому о динамике использования наречия с этими глаголами судить нельзя.

Таблица 2

быстро +	**прореагировать**	**отреагировать**	**среагировать**
НКРЯ, релевантные при поиске «/1» (релевантные при «/10»)			
всего	1 (1)	20 (23)	6 (6)

быстро +	**прореагировать**	**отреагировать**	**среагировать**
1950–1989	0 (0)	2 (2)	0 (0)
1990–1999	0 (0)	8 (10)	2 (2)
с 2000	1 (1)	10 (11)	4 (4)
Военная литература, тексты, опубликованные в 1950–1989 гг., релевантные при поиске «/1»			
	0	20	17
ЖЗ, релевантные при поиске «/1»			
1995–1999	0	3	2
с 2000	0	22	12

Но есть другие массивы оцифрованных текстов, где сходные задачи решаются успешнее, хотя работа с ними заметно более трудоемка, в первую очередь из-за отсутствия необходимого инструментария.

Сегмент «Военная литература» Библиотеки Мошкова (militera.lib.ru) содержит большой массив мемуарной, художественной, исторической и иной литературы (возможен поиск по соответствующим подмассивам), в основном второй половины XX — начала XXI в., но выявление датировки требует обращения к каждому тексту.

В сегменте в целом число документов с глаголом *отреагировать* троекратно превышает число документов со *среагировать*, однако имеется явное жанровое различие: если в мемуарах разница чуть больше, чем в два раза, то в текстах по военной истории она составляет 7,5; при наличии обстоятельства *быстро* различие во всех случаях уменьшается[*].

Хороший срез современного литературного языка дает сегмент «Журнальный зал» (ЖЗ); это в основном тексты XXI в., частично также 1990-х гг., но в толстых журналах изредка публикуются и заметно более ранние произведения. Здесь основная проблема в том, что подавляющее большинство текстов продублировано дважды. Несколько лет назад поисковый алгоритм Яндекса это учитывал и первоначально предлагал каждый текст по одному разу, а полную выдачу давал лишь при реализации опции «еще с сайта» (при этом на первом этапе были некоторые потери полноты, которые ликвидировались лишь при полной выдаче). Сейчас выдается сразу всё, так что близкое приближение к точному числу текстов можно получить делением брутто-результата[4] на два. При небольшой выдаче несложно выяснить подлинный результат вручную (см. нижние строки Табл. 2), но пока нас интересует соотношение, достаточно сопоставления брутто-результатов. В ЖЗ *среагировать* встречается в 5–6 раз реже, чем *отреагировать*, а в сочетании с наречием *быстро* — всего лишь в два раза. То есть сам глагол *среагировать* используется существенно реже, но его относительная сочетаемость с наречием в два, а то и в три раза выше, чем у *отреагировать*[*].

[4] Так я называю собственно выдачу по запросу, с дублетами и цитатами; переход к нетто-результату требует ручной обработки.

Общий недостаток поисковых машин — недостоверность объявленных цифр найденного, если они превышают 1000 — может быть компенсирован в этом сегменте поиском по отдельным журналам[*].

Можно ли что-то сказать о динамике процесса? Лингвистически полезным является раздел «Самиздат» Библиотеки Мошкова (zhurnal.lib.ru), где произведения размещает каждый, обнаруживший у себя творческие наклонности. Любые наметившиеся в языке изменения выражены в этом сегменте Интернета сильнее, чем в профессиональной литературе. Относительная частота *среагировать* здесь явно выше, чем у тех, кто пишет профессионально, а в сочетании с наречием *быстро* глагол *среагировать* используется даже чаще, чем *отреагировать*.

Наиболее отчетливо языковые инновации проявляются в блогосфере. Не трудно показать, что в этом сегменте популярность глагола *среагировать* несколько выше, чем в ЖЗ: число блогов, в которых он был использован, в разных регионах за разные временные отрезки составляет четверть или несколько более от числа тех блогов, где фигурировал глагол *отреагировать*[*].

Результаты поиска *быстро* /1 [глагол] по московским блогам представлены в Табл. 3, они довольно причудливы: изначально *быстро* /1 *среагировать* преобладало незначительно, в 2006–2007 гг. его доминирование существенно возросло, а затем процесс пошел вспять, да так, что к концу 2010 г. статистика уже в пользу глагола *отреагировать*.

Таблица 3

быстро /1	**2001–2005**	**2006–2007**	**2008–2009**	**2010**
реагировать	193	692	981	711
прореагировать	2	7	6	4
отреагировать	62	189	298	259
среагировать	80	334	365	250
среагировать к отреагировать, %	129	177	122	97

Разгадка проста: возрастная структура блогосферы подвержена существенным изменениям, в 2006–2007 гг. доля подростков была максимальной, позже их число снижается в связи с массовым уходом в социальные сети. А последние пару лет наметился приток новых блоггеров старших возрастов.

В отдельных регионах блогосфера оказывается в основном подростковой, и там такого рода статистика выглядит очень рельефно. В Астрахани за все годы по 2010 включительно *быстро /1 отреагировать* встретилось в 7 блогах, а *быстро /1 среагировать* — в 12, в Оренбурге, соответственно, 2 и 8; цифры мизерные, но показательные.

Не подлежит сомнению, что любые языковые инновации сильнее выражены в младших возрастах. Имея стратифицированную по возрасту синхронную статистику словоупотреблений, мы получим сведения о динамике языковых процессов. К сожалению, распределение блоггеров по возрасту,

декларированное в расширенном поиске Яндекса, с лета 2008 г. практически не работает.

В ЖЗ, отражающем более «старый» язык, *отреагировать* встречается в 5–6 раз чаще, чем *среагировать*, у блоггеров только в три-четыре; популярность второго глагола явно растет. А резкий рост сочетания *быстро среагировать* по-настоящему впечатляет: в младших возрастах оно явно обогнало синонимичное *быстро отреагировать* не только относительно, но и в абсолютных цифрах. Разница в употреблении двух глаголов может быть связана с полом автора (в блогосфере доминируют женщины, а среди авторов ЖЗ их меньше), или же с различиями в тематике. Данные за разные периоды показывают, что блоггеры, не зависимо от пола, применяют глагол *среагировать* **к себе** в 2–3 раза реже, чем *отреагировать*, а **к власти** — в 4–6 раз реже. Соответствующая статистика по всей блогосфере за последний квартал 2010 г. есть в Табл. 4, там же для сопоставления приведены данные, которые можно получить из НКРЯ и ЖЗ. Довольно очевидно, что 30 текстов ЖЗ позволяют лишь выдвигать какие-то гипотезы, а из втрое меньшего стилистически разнородного материала НКРЯ вряд ли что можно извлечь.

Разница в частотности рассматриваемых глаголов в разных контекстах особенно заметна на фоне того, что контекстно не привязанное соотношение употребимости этих глаголов находится как раз посредине[*]. Поведение наречия остается не вполне ясным; не исключено, что молодежь считает, что на любые перемены она способна *среагировать* **на порядок** *быстрее*, чем *отреагирует* власть.

Таблица 4

	Блоги, 10–12.2010	**НКРЯ, 2001–2004**	**ЖЗ, 2001–2010, нетто**
«я отреагировал»	218	2	11
«я среагировал»	100	1	6
«я отреагировала»	260	2	2
«я среагировала»	97	0	0
власть /З отреагир., муж.	253	4	9
власть /З отреагир., жен.	43	1	2
власть /З среагир., муж.	34	0	0
власть /З среагир., жен.	8	0	0

В цитате, с которой я начал, относительно трех глаголов были сформулированы четыре вопроса, ответы на которые в применении к текстам НКРЯ, действительно, можно получить «буквально за считанные минуты». Легко узнать, что чаще всего в корпусе встречается глагол *отреагировать*, он же охотнее двух других сочетается с наречием *быстро*, он же первым — в 1950 г. — появляется в беллетристике (в текстах по химии его опережает известное с XIX в. *прореагировать*). На любом относительно большом отрезке времени примеров на этот глагол больше, чем на конкурирующие. Исключение — дореволюционный период, когда глагола *отреагировать* еще не было, а *прореагировать* дважды

было употреблено химиками. Но эти ответы — про НКРЯ. Как выясняется, ответы на те же вопросы про русский язык могут заметно отличаться, если воспользоваться статистикой, полученной по разным относительно однородным массивам текстовых документов, представленных в разных сегментах Интернета. Эту методику я называю сегментно-статистической.

НКРЯ в совокупности является хорошим «макетом» русского языка: разметка и стоящая за нею идеология — это модель устройства языка, а комплект текстов корпуса — экспериментальная база для проверки модели. НКРЯ создан лингвистами и для лингвистов, а лингвиста интересует грамматика. Часть ее «зарыта» в лексиконе, это продуктивные и уникальные модели грамматического поведения лексических единиц. Слова, ведущие себя стандартно, лингвиста интересуют очень мало: важно знать объем некоторого класса однотипно ведущих себя слов (степень продуктивности модели) и иметь под рукой несколько представителей каждого класса, чтобы строить примеры типа *Seymour sliced the salami with a knife*. Прагматическая ценность такого рода примеров лингвиста не интересует[*]. Собственно лексикографу (составителю неспециализированного словаря) интересно **каждое** слово. Еще один тип специалиста по языку — социолингвист, на него НКРЯ также не рассчитан, поскольку ни в коей мере не может считаться сбалансированным с точки зрения решаемых им задач.

НКРЯ хорош для решения любых собственно лингвистических задач, часто полезен лексикографу и социолингвисту для предварительной оценки положения вещей, хотя в отдельных случаях может служить и основанием вполне серьезных выводов. Но исследователи с лексикографическими и социолингвистическими интересами иногда склонны принимать его за полигон принятия решений в последней инстанции, что досадно. Приведу два показательных примера.

Недавно была высказана претензия к «Русскому орфографическому словарю», куда включены единицы, «обнаружить наличие которых не удается даже при обращении к Национальному корпусу русского языка» [Николенкова 2011: 181]. Примеров необнаруженного нет, приведены лишь три слова, якобы встречающиеся там в единственных контекстах: *остервенить*, *окровенить* и *форсун*; эти слова традиционно включаются в толковые и орфографические словари — но зачем засорять словари тем, что даже в НКРЯ (почти) не встречается?[*].

Для исследования НКРЯ «с точки зрения отражения социальной дифференциации языка» из слов, «отражающих молодежный сленг школьников, студентов, были отобраны следующие: *училка*, *химица*, *студак*, *туса*» [Киеня 2010: 263–264]. Для слов *студак* и *химица* во всех подкорпусах нашлось лишь по одному примеру, тем не менее С. Н. Киеня делает вывод, «что национальные корпусы отражают все многообразие существования языка, генеральную совокупность языкового материала, что свидетельствует об их высокой социокультурной значимости» [Там же: 265].

Я же из этой работы делаю прямо противоположный вывод: НКРЯ с социолингвистическими задачами не справляется. Между тем,

сегментно-статистический метод выявляет, что в современном русскоязычном узусе Белоруссии *химица* используется в три раза чаще *химички*, что белорусскому (и московскому) *студаку* в Петербурге соответствует *студень*, а в Екатеринбурге *студик*, этот метод позволяет проследить диахронию замен *тусовка — туса* и *тусоваться — тусить* в молодежном жаргоне[*].

Метод этот пока довольно убог. «Оцифрованный мир» дал социолингвисту необъятный языковой материал, который «самоорганизовался» в динамично растущие относительно однородные в стилистическом отношении сегменты. Но для работы с ними специального инструмента нет, приходится пользоваться тем, что имеется: поисковым алгоритмом Яндекса, рассчитанным на принципиально иного пользователя.

Будучи оптимистом, надеюсь, что рано или поздно полнота описания языка, объявленного в России государственным, заинтересует кого-то из тех, кто в силах принять ответственное решение, ведущее к созданию специализированного инструмента для исследования больших массивов оцифрованных текстов.

А пока буду вручную доказывать необходимость такого инструмента. Gutta cavat lapidem.

References

1. *Kienia S. N.* 2010. Corpus in a Sociolinguistic Aspect [Korpusy v sotsiololingvisticheskom Aspekte]. Nauka-2010: Sbornik Nauchnykh Statei.
2. *Nikolenkova N. V.* 2011. Spelling Dictionary and Codification of the Modern Standart: the Problem of Non-corrdination [Orfograficheskii Slovar I Kodifikatsiia Sovremennoi normy: Porblemy Nesoglasovannosti]. Voprosy Kul'tury Rechi.
3. *Plungian V. A.* 2005. Why are We Making the National Corpus of Russian? [Zachem My Delaem Natsional'nyi Korpus Russkogo Iazyka?]. Otechestvennye Zapiski, (2).
4. *Russian* Spelling Dictionary [Russkii Orfograficheskii Slovar]. 2007.