Статистическая оценка функциональных свойств лексики по материалам Интернета

Владимир Иванович Беликов,
ИРЯ им. В. В. Виноградова РАН

Мария Вячеславовна Ахметова,
журнал «Живая старина», Москва

функциональные свойства лексики — любые нетривиальные грамматические и стилистические особенности слова или фразеологизма, упоминание которых оказалась бы полезным в словарной статье

Статистический анализ того типа,

который будет демонстрироваться ниже,

позволяет подтвердить, уточнить, иногда опровергнуть то, что говорится о словах и фразеологических единицах в стандартном толковом словаре,

а также выявлять совсем «новые» свойства такого рода

Задачи доклада

- Показать, что в настоящее время существуют легко-доступные и достаточно простые способы **объективного** выявления разнородных функциональных свойств лексики.
- Нетрудно убедиться, что многие частные ошибки в лексикологии и лексикографии — результат ориентации на интроспекцию, а также малые и непоказательные для языка в целом текстовые выборки. От критики таких частных решений пора переходить к созданию систематической «стратегии обработки лексики», ориентированной на современные информационные технологии, использующей всё многообразие языкового материала, которое мы получили в результате информационной революции.

Материал и инструмент

(1) Типы текстов;

(2) Некоторые релевантные свойства Яндекса.

Типы оцифрованных русскоязычных текстовых материалов

Закрытые, в первую очередь базы СМИ.

Очень полезны, но в силу ограничений на доступ не могут использоваться всегда и всеми (о них ниже речь пойдет минимально).

Интернет-материалы «общего пользования»:

- **Корпуса**, в первую очередь НКРЯ.

  Задуманы как собрание текстов языка-объекта, пополняются целенаправленно по определенной программе.

- **Текстовые массивы Интернета**, корпусами их можно называть лишь метафорически.

  Создаются с разными целями, наполняются во многом стихийно.

Интернет-материалы «общего пользования»:

- **Корпуса**, в первую очередь НКРЯ.
  - параметры задаются и контролируются;
  - поиск изначально рассчитан на лингвиста, технические затруднения носят случайный характер, поисковые возможности при развитии корпуса совершенствуются;
  - объем ограничен, некоторые задачи невыполнимы.
- **Текстовые массивы Интернета**, корпусами их можно называть лишь метафорически.
  - известны лишь самые общие характеристики (однако и их часто вполне достаточно);
  - поиск не рассчитан на лингвистические задачи, со временем поисковые возможности могут существенно ухудшаться;
  - объем неограничен.

Важнейшие текстовые массивы

- Библиотека Максима Мошкова (БМ) с подмассивами:
  - «Собрание классики»,
  - «Современная русская проза»,
    [часть текстов раздела БМ «Современная литература» (lit.lib.ru) могла бы оказаться здесь, но там есть и сетевая литература]
  - «Самиздат»,
  - • • • • • • • •
- Другие собрания литературных и стилистически близких к ним текстов,
  в первую очередь «**Журнальный зал**».
- Русскоязычная блогосфера.

В Библиотеке Мошкова граница между «Классикой» и «Современной русской прозой» достаточно условна, «современность» охватывает значительную часть советского периода.

Собранием **собственно современного** профессионального литературного творчества является «Журнальный зал» (magazines.russ.ru), где сосредоточены журнальные публикации с 1990-х гг.

«Самиздат» БМ — **очень** большое собрание самодеятельных текстов разного жанра; многие авторы имеют достаточно смутные представления о литературной норме, высока доля разговорной и просторечной лексики в авторском тексте, в целом лексикон «Самиздата» близок к разговорному узусу.

Инструмент:
Весь ли народ против Яндекса?

Сильно против только интернет-зависимые граждане в составе следующих категорий:

- *нижегородцы,*
- *новгородцы,*
- *производители КАМазов,*
- *газовики Ямала,*
- *металлурги Нижнего Тагила и Старого Оскола,*
- *ряд менее значимых обиженных Яндексом групп*.

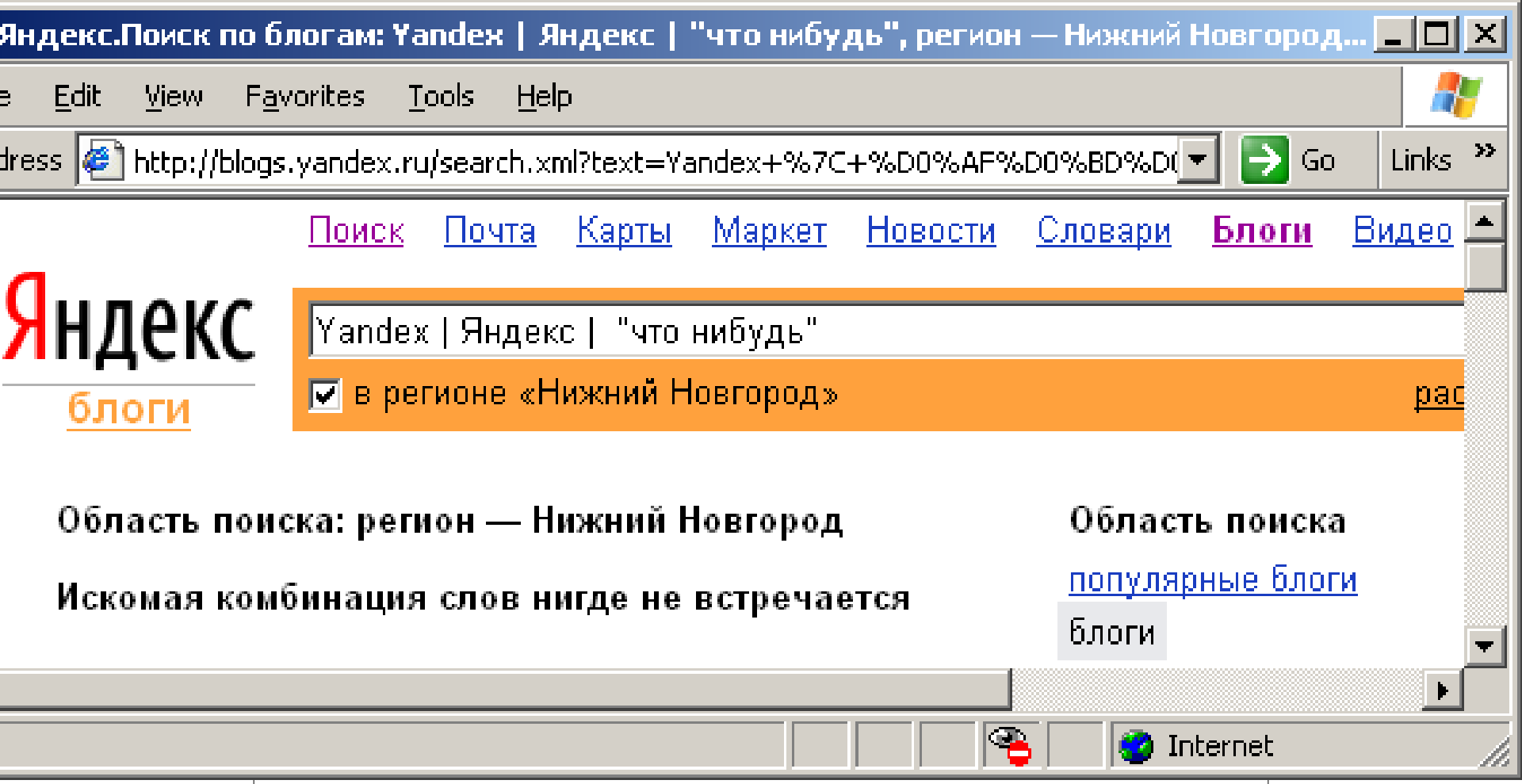
Лингвисты — «за», но с оговорками.

Больше всего оговорок у «лексикологов нетрадиционной ориентации».

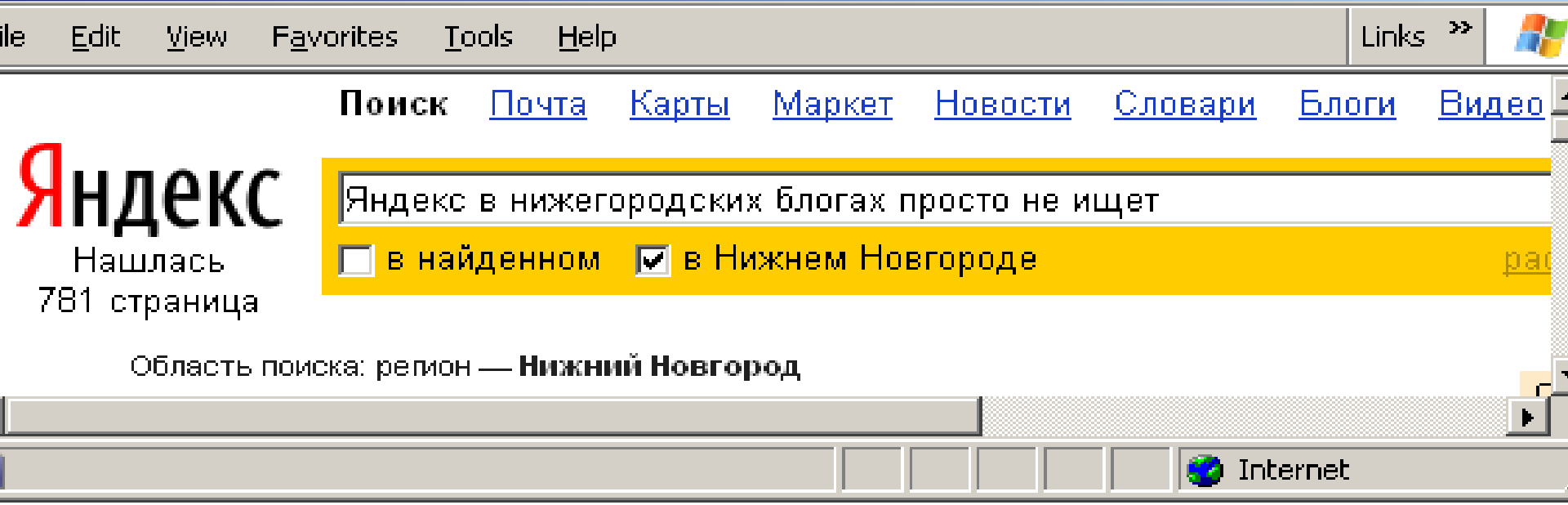
Язык блогов во многом является отражением повседневного молодежного словоупотребления.

Теоретически Яндекс допускает поиск в блогах с заданием отдельных параметров и их комбинаций: региона, пола и возраста (3 когорты) блоггеров, а также с выделением конкретного фрагмента блогосферы (livejournal.com, liveinternet.ru, diary.ru).

На практике же **Яндексу не удаются поиски в блогах** одного из крупнейших городов России, **Нижнего Новгорода** (1313689 чел. по переписи 2002, 1275 тыс. чел. на начало 2008).



Общий поиск возможен и по региону (например, *Омская область*), и по городу (например, *Омск*), **поиск по блогам — только по городам**. Потеря большая: Нижегородская область оказывается представленной только Дзержинском (248 тыс. чел.), Арзамасом (106 тыс. чел.) и более мелкими городами.



Как видим, **общий поиск в Нижнем Новгороде** возможен.

В поиске по блогам «теряются»:

- В Новгородской обл. — Великий Новгород и Старая Русса; крупнейший городом оказываются Боровичи (58 тыс. чел.).
- В ЯНАО — **крупнейший город** Новый Уренгой (118 т. ч.).
- **Вторые по величине** города Татарстана (Набережные Челны, 506 т. ч.), Свердловской обл. (Нижний Тагил, 376 т. ч.), Белгородской обл. (Старый Оскол, 220 т. ч.), Псковской обл. (Великие Луки, 100 т. ч.).
- И многое другое …

Специфика отношения Яндекса к «лексикологам нетрадиционной ориентации»

- Не особенно дружественная политика Яндекса по отношению к любителям лексической статистики постоянно ужесточается. С самого начала появления поиска по блогам Яндекс вынуждал пользователей получать информацию блоками по 10 записей*, уже года два, как он отказывается показывать 1001-ю и последующие найденные единицы, а с лета 2008 г. поиск с разделением по возрасту стал невозможен — со второй страницы результатов Яндекс сбивается с ограничения по возрасту и выдает лишь общий результат.

*Совет: если надо «скакнуть» на дальнюю страницу, можно не листать страницы выдачи, а в адресной строке заменить последнюю цифру на 99 (это сотая страница выдачи по блогам).

Несмотря на технологические недостатки, при анализе лексики Яндекс позволяет верифицировать многое известное и выявить кое-что новое, иногда неожиданное.

Продемонстрируем отдельные типологически разнородные наблюдения над лексикой и фразеологией.

корректив (м. р.) или *корректива* (ж. р.)?
«… Справочники устаревают и требуют корректив» (А. Н. Рыбаков, «Тяжелый песок», 1975—1977)

Редкий случай, где материалов НКРЯ достаточно для доказательного анализа:

	по 1940	1941—90	с 1991
Ед. число	**47**	7	7
Мн. число	31	43	**294**
Муж. род	**53**	11	16
Жен. род	1	2	**22**

Глаголы *лазить* (*лажу*, *лазишь*,…) и *лазать* (*лазаю*, *лазаешь*,…) признаются синонимичными и описываются обычно в одной статье; в московском словаре Шведовой [2007] второй снабжается пометой *разг*., что имеет естественное объяснение: «на слух» лазать в Москве говорят заметно реже, чем в Петербурге. Но в действительности положение с отдельными словоформами этих глаголов различно.

Судя по блогам (2007—2008), в петербургском узусе преобладают личные формы «от *лазать*», соотношение: *лазаешь*/*лазишь* — 52/30, *лазает*/*лазит* — 204/112; но с заметно более частотными инфинитивами положение обратное: *лазать*: 797*, лазить*: **1120**.

В московских блогах преобладание «строго нормативного» инфинитива выражено очень явственно: за IV квартал 2008 соотношение: *лазить*/*лазать* составило **696**/167, но с личными формами происходят странные вещи, за 2007—2008 гг. *лазаешь*/*лазишь* — 115/140, а *лазает*/*лазит* — 647/590.

Попытка отыскать «правильную форму» 1 лица ед. ч. глагола *лазить* почти безнадежна: Яндекс находит только *лажу* именную.

Глагольная форма *лажу* быстро устаревает и не всегда используется даже в старшем поколении; показательна реакция одного известного ученого (не русиста), чл.-корр. РАН:

«Говорю *лазию*, пишу *лазаю*».
«А *лажу*?»
«Ну, это какое-то вульгарное просторечие. *Из кичмана не вылажу*».

Не здесь ли разгадка письменной частотности *лазаешь* [устное *лазиешь*?] и *лазает* [устное *лазиет*?]

определиться$_2$:

Ушаков: Определить своё местонахождение, положение (*спец.*). *Лётчик определился с помощью компаса.*

ОШ [практически то же]: ... *с помощью приборов.*

ТСРЯ (Шведова, 2007) — без помет и дополнено:

Определить своё местонахождение, положение; вообще установить, решить что-н. для себя. *Лётчик определился с помощью приборов. О. в своих планах, целях, намерениях.*

М. С. Горбачев определялся ***по*** планам, целям, намерениям. После него определялись самым разным образом.

Государству российскому необходимо четко определиться по своим внешнеполитическим целям и задачам («Время новостей»; 2006)

Руководящий состав союза [РСПП] *должен обновиться, а сама организация — определиться в своих целях и задачах* («Ведомости»; 2005).

В Пентагоне пока не определились относительно планов долгосрочного присутствия в Центральной Азии (Sobkor.Ru; 2002).

Отвечая на вопрос корр. ИТАР-ТАСС, он [Аленичев] *сказал, что пока не определился в отношении своих ближайших планов, отложив все до окончания чемпионата Европы по футболу* (ИТАР-ТАСС; 2004).

Правительство России на своем сегодняшнем заседании намерено определиться с вопросом о целях и принципах реформирования железнодорожного транспорта в стране (РИА «Новости»; 2000).

Субъективное ощущение, что чаще всего так:

- *К этому времени правительство должно* ***определиться с*** *планами налоговой реформы на 2004 год* («Известия»; 12.04.2003).
- *И если местные фирмы* [Красноярские IT-компании] *не сумеют выработать общий подход,* ***определиться с*** *целями, эти средства уйдут другим* («Российская газета»; 28.06.2007).
- *Правозащитники, уже убедившиеся, что приговор будет сугубо обвинительным, окончательно* ***определились с*** *намерением подавать жалобу в Европейский суд по правам человека* («Новые Известия»; 31.05.2005).

Анализ базы СМИ «Интегрум» показывает, что в самом начале 1990-х на смену

определиться по чему-л.

пришла модель

определиться в чём-л.

К середине 1990-х она стала вытесняться моделью

определиться с чем-л.,

которая к настоящему времени оказалась вне конкуренции.

Оборот в разговорном узусе не частый, но используемый.
В блогосфере та же ситуация, что и в СМИ:
поиск на

определиться /+4 (планы | цели | намерения)

по 2008 г. включительно выявил

25 случаев *определиться в* и **157** *определиться с*.

«Исконное» *определиться по* с этими словами в блогах не встретилось.

С другими управляемыми словами соотношение оказывается несколько иным, но там, где возможна конкуренция*, «с-управление» заведомо преобладает.

При аналогичном поиске с наборами

понятиями | *терминами* | *определениями* (мн. ч.)
соотношение *с-* и *в*-управления — 56/23,

понятие | *термин* | *определение* (ед. ч.)
соотношение *с-* и *в*-управления — 22/2.

* *Определиться в* несовместимо, например, с временными отрезками, ср. *определиться с отпуском* (**определиться в отпуске*).

Обращение к интернет-массивам позволяет довольно точно определить время и темпы конкретных словарных изменений. Изменения эти могут иметь разный характер: лексическая единица может «просто» устареть и выйти из употребления, может, наоборот, проявить территориальную или социальную экспансию, а может замениться другой, внешне сходной.

В конце опубликованного текста доклада написано:

Приведенные выше примеры можно легко умножить.

Чтобы не быть голословными, в докладе мы будем упоминать и такие примеры, которых нет в напечатанной версии, а кое-что из напечатанного опустим.

От *пешедрала* к *пешкодралу*

	пеше-дралом	*пешко-дралом*	*соотно-шение*
«Классика»*	6	1	6,0
«Самиздат»	35	55	0,6
блоги по 2008 г. вкл.	337	2114	0,2

*XIX век: только *пешедралом*, 4 примера (Г. П. Данилевский «Беглые в Новороссии», 1862, В. В. Крестовский «Петербургские трущобы», 1867, М. Е. Салтыков-Щедрин «Господа Головлевы», 1875—1880, А. П. Чехов «После бенефиса», 1885).

От *мне это пофигу* к *мне на это пофиг*

Блоггеры Петербурга	2005, весь год		2008, январь—май	
	по_фиг(*у*)	*по_фиг*, %	*по_фиг*(*у*)	*по_фиг*, %
все блоггеры	1388	67,4	2529	73,0
указавшие пол:				
женщины	655	**69,3**	1602	**74,2**
мужчины	607	65,4	792	68,9
указавшие возраст:				
11—19 лет	34	**67,6**	640	**75,6**
20—30 лет	614	66,1	549	69,4
старше 30	95	60,0	117	59,8

Пятьдесят лет назад *шофёры* (и *шофера́*) заправлялись на *бензоколонках*, сейчас *водители* все чаще делают это на *автозаправках*.

шофер* [*водитель*] */5 машина	*шофер*	*водитель*
«Собрание классики»	34	4
«Современная русская проза»	86	53
«Самиздат»	813	962
livejournal, Москва, 04.2009	17	237
livejournal, СПб, 01-04.2009	18	302
шофер* [*водитель*] */+1 маршрутка	*шофер*	*водитель*
Все блоги , …—2008	90	[17324]
Все блоги, Москва, …—2008	14	[2851]

Переход в литературных текстах и повседневном узусе от *бензоколонок* к *автозаправкам* — результат влияния языка СМИ.

Лексикографическая справка:

БТС (1998): **автозаправка** -и; *ж*. 1. Заправка топливом, смазочными маслами и т.п. транспортных средств. ‹…› 2. *Разг*. Автозаправочная станция; бензоколонка. ‹…›

Новый БАС (т. 1, 2004. А—Бишь): Нет слова.

No comments

Соотношение текстов с *бензоколонкой/автозаправкой* в газетах

	2002	2008
Московские газеты (“Вечерняя Москва”, “Московская правда”,“Московский комсомолец”)	41/30	32/55
Петербургские газеты (“Версия в Питере”, “Вести”, “Невское время”, “Санкт-Петербургские ведомости”, “Санкт-Петербургский курьер”)	30/10	26/31

Результаты за вычетом сочетаний *королева бензоколонки* и *…японец — человек, а японка — автозаправка…*

Сегмент Интернета	*бензо-колонка*	*авто-заправка*	*соотно-шение*
«Журнальный зал», …—1999	44	5	8,8
«Журнальный зал», 2000-07	167	32	5,2
«Журнальный зал», 2008	24	7	3,4
Москва, все блоги, …—2006	977	330	3,0
Москва, все блоги, 2007—08	1227	744	1,6
СПб, все блоги, …—2006	247	81	3,0
СПб, все блоги, 2007—08	367	224	1,6

холодец vs. *студень*

МАС (2 изд.): **студень** — б/п, **холодец** — *разг*.

БТС (1998): **студень** — б/п, **холодец** — *нар.-разг*.

В Москве равно используются оба слова (Шведова-2007: оба слова б/п), но растущее предпочтение за *холодцом*.

В Петербурге в младших возрастах побеждает *холодец*.

Воронеж:

Для воспитания «интереса и уважения к родному языку» ученики составлялся диалектный словарик «примерно из ста слов», куда вошли: «*задорга* (жердь или доска по краю русской печи), ***студень*** (холодец), *жичина* (хворостина), *жужель* (мелкая картошка)» [Голубева Г. Л. Что такое диалектные слова? // Русский язык; 2003, № 35].

Считается, что в Петербурге преобладает *студень*.

Так в прессе, так в узусе старших возрастов

Любопытны региональные отличия в использовании этих слов в переносном значении:

- *Студень дизелю не по зубам* [«Автовитрина», СПб, заголовок].
- *Дизельный холодец* [«МК-мобиль», Москва, заголовок].
- *Большая часть пассажирских автобусов работает на так называемом «летнем» дизельном топливе, которое при температуре ниже 12 превращается в холодец* [«КП в Воронеже»].

Блоги СПб по окт. 2008: студень/холодец: 1330/1026, но около половины *студней* в блогах несъедобны

	«студень», всего	*студень* ‘студент’	*студень* ‘студенческий билет’
Блоги Санкт-Петербурга:			
октябрь 2005	18	1	8
октябрь 2006	27	3	18
октябрь 2006	33	4	14
октябрь 2008	40	5	16
«всего»	118	13	**56**
Блоги Москвы (те же 4 месяца в сумме):			
«всего»	165	14	16

	«студень», всего	*студень* ‘студент’	*студень* ‘студенческий билет’
Блоги Санкт-Петербурга, октябри 2005, … 2008			
«всего»	118	13	**56**
Блоги Москвы (те же 4 месяца):			
«всего»	165	14	16

Соотношение «слов» *студень*/*холодец* по окт. 2008 в Москве — 1772/4634, в Петербурге — 1330/1026.
В «октябрьских» материалах в Петербурге 55 кулинарных *студней*, в Москве 135.
Студень ‘студент’ и особенно ‘студенческий билет’ явно **петербургские** жаргонные единицы.

Интернет-блоги во многих случаях являются наиболее эффективным инструментом выявления ареалов распространения регионально маркированных единиц

- *чойс* ‘любая лапша быстрого приготовления’ и *оптарь* ‘оптовый рынок’ находятся только в **Омске**;
- *садоогород* ‘садово-огородное товарищество или участок в нем; используется и в официальных контекстах — почти исключительно в **Удмуртии** ;
- *ссобойка* (также *собойка*) ‘набор продуктов на работу, в дорогу; школьный завтрак, взятый из дома’ — практически только в **Белоруссии.**

Эффективность анализа блогов зависит от типа лексики. Один «неблоговый» полюс — канцеляризмы.

- **Простой случай**: слово редкое и легко локализуется по месту издания документа:

  На период временного отсутствия прибора учета (ремонт, поверка) по заявке потребителя жилищно-эксплуатационная организация ‹…› *устанавливает на место прибора трубный вставыш* (Пермский край).

Вставыш в словарях отсутствует, в ГОСТе Р 50193.2-92 фигурирует *трубная вставка*.

Ареал недавнего «канцелярского» фразеологизма *мокрая печать* (поставленная непосредственно на документ, не ксерокопированная) вполне успешно определяется по доменам первого уровня при общем поиске в Интернете: **преимущественно Украина**, распространяется в России. Вот статистика числа сайтов с релевантными документами за отдельные годы:

Мокрая печать	…-2005	2006-07	2008	всего
Всего	80	232	450	762
В домене *.ru*	28	76	112	216
в домене *.ru* про Украину	13	33	41	**87**
В домене *.ua*	20	84	187	**291**

Другой полюс, плохо выявляемый в блогах: детская лексика

Блоггеры редко пишут про игру в салочки, прятки, жмурки и т. п. Между тем региональные именования таких игр различны. По блогам ареал распространения выявляется очень примерно. Помогает опрос в Интернете, который выявляет и новые региональные именования. В нашей практике, например, так выявилось смоленское именование игры «в вышибалы» (устар. *круговая лапта*):

высекалы.

Временное ограничение: слабое региональное развитие блогосферы

Чебэшка с орфографическими вариантами ‘дом, не полностью обеспеченный коммунальными удобствами; квартира в таком доме’ (от сокр. *ч/б* = частично благоустроенный) встретилось в пяти газетах (30 текстов), причем только из Якутии [Ахметова, «Диалог-2008»].

Анализ блогов практически ничего не дает: за 2007—2008 гг. слово *чебэшка* (*чэбэшка*, *чебешка* и т. п.) встретилось у 17 блоггеров в значении ‘черно-белая фотография, черно-белая фотопленка’ и лишь однажды в значении жилища — про якутский поселок Багатай, но у московского блоггера.

Выводы

- В настоящее время материала для объективного описания лексики и фразеологии достаточно, будет еще больше.
- Инструментарий для работы с ним есть, будет совершенствоваться.
- Продемонстрированные способы обработки имеющегося материала существующим инструментарием достаточно просты.

  «Простые» способы пока не автоматизированы, значит трудоемки, но **во-первых**, стоит различать объективные выводы и доказательные выводы. Выше — в силу специфики жанра доклада — мы стремились достичь не только объективности, но и высокого уровня доказательности. **Во-вторых**, можно надеяться на автоматизацию того, что пока делается вручную.

Спасибо за внимание!

Всех интересующихся методикой описания современного состояния русского лексикона приглашаем заходить на форумы

«Городские диалекты»

и

«Как это будет по-русски?»