Дифференциальная корпусная статистика на основании неавтоматической метатекстовой разметки

Variational corpus statistics using author profiles

V. Belikov, N. Kopylov, V. Selegey, S. Sharoff

Using human classification in a computer-processed corpus

- The presentation is based on research carried out in the framework of the project on General Internet Corpus of Russian (GICR or Geekrya).
- Geekrya is a large Web-derived corpus which contains morphosyntactic annotation (with automatic disambiguation) and text-type annotation (based on partly automatic classification).
- Why and how we use human classification (from the profiles of the authors):
- For supervised machine learning
- For linguistic analysis and for making important data checks

диалог
международная конференция

Big is not always beautiful (Просто много – этого мало)

- Recent trend: linguists tasted large-scale corpora and liked them.
- For example, at this conference we have presentations based on RuTenTen (Sketch Engine, Adam Kilgariff: 10 Gigawords, 10^{10})
- We would like to show that size alone is not enough.

Anatomy of large corpora

- Method for corpus collection
- Method for morphosyntactic annotation
- Method for text-level annotation
- Using human annotations
- Two kinds of Web-derived corpora:
  - Large corpora without text-level annotation (RuWac, RuTenTen)
- Variational corpora with text-level annotation (GICR).

Segments of the Internet

- *WWW contains some relatively homogeneous arrays of texts formed independently of linguists, in some cases emerging quite spontaneously…Frequencies of the same lexical items differ greatly from one segment to another, and this statistics is very significant for sociolinguistics.* ***The main problem in applying the method of segmental statistics is the lack of a suitable instrument for automatic data processing***
- GICR should represent the major social, generic and topical segments available on the Internet

диалог
международная конференция

Two ways to compose Internet corpora: segments vs. blind crawling

- Advantages of taking the segmentation of the Internet into account:
- Regions of functional homogeneity.
- Segments can be used as generalised genres in the absence of reliable genre classification methods.
- Close analysis of segments helps in developing methods for automatic extraction of information about the authors, which gives the possibility to study variation on the basis of natural human labelling

Main segments in GICR

- Blogs
- Microblogs
- Forums
- Social networks
- Sites collecting texts of known authors, primarily fiction and feature articles
- Encyclopedic resources with information about their authors (unlike anonymous entries in Wikipedia)
- News.

Segments with known a priori metatext parameters in Geekrya

- The total size of indexed Geekrya amounts to 12 billion words. Segments with metatext parameters (size in millions of words):

Segment	Gender	Age	Region
blogs.mail.ru	164	81	113
livejournal.com	0	1800	5600
vk.com	2000	1600	1600
news	0	0	0
magazines.russ.ru	258	0	0
forums (adw.ru)	163	0	0
Total:	2585	3481	7313

Important dimensions of variation

regions,

gender,

age.

We hope that the majority of segments can be studied within variational parameters.

We are still at the early stage of experiments.

диалог
международная конференция

Территориальное варьирование

Соотношение глаголов (*с*)*делать* и (*по*)*ставить* в соположении мо словом *укол* (по данным сегмента livejournal).

Регион	делать	ставить	д/с
Урал (Удмуртия, Пермский кр., Свердл., Челяб. и Курганск. области	519	547	0,9
Сибирь (с Якутией, без Тюменских АО)	636	728	0,9
ЯНАО и ХМАО	42	19	2,2
Дальний Восток без Якутии	244	101	2,4
Казахстан	67	30	2,2
Соседи на западе и юго-западе (Коми, Кировск. обл., Татарстан, Башкирия, Оренбур. обл.)	464	73	6,4

диалог
международная конференция

Территориальное варьирование

Коллокация (*с*)*делать укол*
в Журнальном сегменте ГИКРЯ

- 81 релевантный текст 68 авторов. Среди них лишь у 13 не устанавливается явная связь с урало-сибирским ареалом.
- У литераторов связанных с ареалом рассматриваемой конструкции, такое употребление глагола (*по*)*ставить* явно не имеет стилистических ограничений, встречается в дневниковых записях, драматургических ремарках и т. п.

диалог
международная конференция

Возрастное варьирование

Для блоговых сегментов была поставлена задача предварительной очистки от дублетов, которые иногда насчитываются тысячами и даже десятками тысяч. Из блогов активно задействованы пока лишь комментарии на платформе mail.ru.

Создание полноценных профилей для авторов журнального сегмента еще только предстоит.

Однако исследование возрастных различий уже возможно.

Возрастное варьирование

Вот некоторые противопоставления в блогах mail.ru.

годы рождения	*тусоваться*	*тусить*
1960-е	73	11
1970-е	54	40
1980-е	64	53
1990-е	23	24

год рождения	*по_фигу*	*по_фиг*	доля архаизма
1945—1974	117	203	36,6%
1980—1999	112	290	27,9%

диалог
международная конференция

Гендерное варьирование

По сути гендерная лингвистика только начинается. Знаем мы слишком мало, но именно большие массивы оцифрованных текстов должны помочь.

Следует иметь в виду, что узус в текстах неиндивидуализированного содержания (беллетристике, различных жанрах нон-фикшн) может сильно отличаться от того, как человек говорит о собственных проблемах.

В гендерном отношении интересны и блоги, и неиндивидуализированные тексты.

диалог
международная конференция

Гендерное варьирование: блоги

О сложностях работы с блогами уже сказано.
До очистки от дублетов нет смысла выяснять статистику по единицам, с большой вероятностью частотным в анекдотах, притчах и т. п. Поэтому не стоит отказываться от иных, отличных от ГИКРЯ, средств анализа оцифрованных текстов. В частности, от поиска в Яндекс-блогах.

Но там есть два серьезных недостатка:

- Нет статистики по соотношению полов.
- тысячный предел достоверной выдачи.

Гендерное варьирование: блоги

Вот статистика «импрессионистических» и сопоставительных оценок

(Яндекс-блоги, 2007 г., Россия).

	жен	муж	ж/м
так мало			
1—25.01	970	615	**1,58**
12—31.12	973	568	**1,71**
гораздо меньше			
01—03	641	904	**0,71**
5.10—12	695	970	**0,72**
раз(а) меньше			
01—03	381	699	**0,55**
10—12	503	886	**0,57**

диалог
международная конференция

Гендерное варьирование: блоги

Вот сходная статистика по Журнальному сегменту ГИКРЯ, где известно соотношение словоупотреблений мужчин (194,2 млн) и женщин (63,8 млн) —3,04

	муж.	жен.	муж./ жен.
так много	3204	1267	2,5
очень много	3278	1090	3,0
раз(а) больше	996	217	4,6
чересчур много	115	23	5,0

диалог
международная конференция

Гендерное варьирование: блоги

Есть и занятные различия во фразеологии, приведем одно, где среди авторов толстых журналов вырисовывается пристрастие женщин к ненормативному варианту.

Абсолютные цифры, впрочем, невелики.

пол автора	(*с*)*хватить*	
	кондрашка	*кондратий*
муж.	**54**	30
жен.	7	**11**

диалог
международная конференция

От литературного языка к письменному стандарту

Еще четверть века назад было известно, и что такое литературный язык, и даже кто является его носителем.

Но литературный язык в традиционном понимании остался в прошлом.
На замену ему пришел стандартный общерусский узус (письменный и устный).

Представляется, что в отношении этого варианта языка достаточно репрезентативны тексты толстых журналов — Журнальный сегмент ГИКРЯ.

диалог
международная конференция

ГИКРЯ как источник знаний о новой норме

Лексикографы, создающие общие толковые словари, где фиксируется литературный (а теперь стандартный) узус, не очень стремятся мониторить современный язык.
А если мониторят, то достаточно субъективно.

диалог
международная конференция

ГИКРЯ как источник знаний о новой норме

Апрельский пример — дополнение к электронной версии «Большого толкового словаря».

Среди инноваций — *шелкография*, но по-прежнему нет многих заведомо более частотных слов. Например, *гендер* и *гендерный*.
Нет их и в 4 томе нового БАСа (2006), и в словаре Н. Ю Шведовой (2009).

ГИКРЯ как источник знаний о новой норме

Как часто в «интеллигентном обиходе» упоминаются *гендер* и *шелкография* ?
Это легко выяснить в Журнальном сегменте ГИКРЯ:

	всего	муж.	жен.
шелкография	40	20	15
гендер	591	135	279
гендерный	2477	773	1054

Любопытно, что при троекратном численном превосходстве авторов-мужчин о гендере заметно чаще пишут женщины. Казалось бы, и так ясно. Но без статистики субъективно.

диалог
международная конференция

ГИКРЯ об инновациях в словообразовании

Какие бывают «-гейты»?

Общеизвестен *Фаренгейт*; после *Уотергейта* появилось много нового. На ум сразу приходит *Ирангейт* и… что еще?

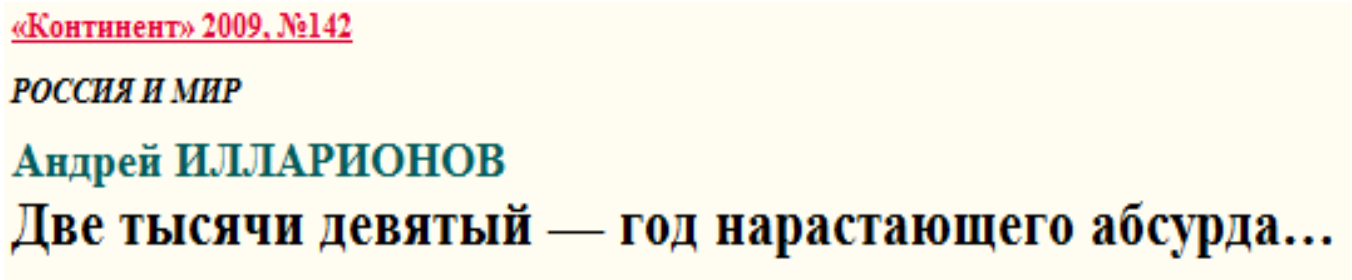
Запросив в Журнальном сегменте лемму **гейт* за вычетом лемм *фарен** и *уотер**, получаем массу интересного, например, *климатгейт*.

диалог
международная конференция

Фрагмент выдачи ГИКРЯ на запрос о **гейт*'ах

	Результат	Справа
: Большой	Климатгеит	(серия малых климатгеито
ия малых	климатгейтов	, принятие российскими вла
рии малых	климатгейтов	. Если можно , остановитес
в получила название	Климатгейта	. Надо сказать , что Климат
ыгались ликвидирова...	Климатгейт	, конечно , добавил , я бы с
и о малых	климатгейтах	, о Климатической доктрине
ыми	климатгейта...	? Этот сюжет развивался с.
ными , чем первонач...	Климатгейт	. Все-таки обнародованная
сточной Англии . На ...	Климатгейт	они отреагировали раздра
, что	Климатгейт	это одно из крупнейших соб

Это, конечно, из одной статьи:



«Континент» 2009, №142

РОССИЯ И МИР

Андрей ИЛЛАРИОНОВ

Две тысячи девятый — год нарастающего абсурда…

Нужны ли дифференциальные корпуса?

Ответ на этот вопрос неочевиден. Возможно, для каких-то задач полезно тотальное усреднение всех вариантов узуса на максимально больших массивах языковых данных.

Но как определить для тотального корпусе «правильное» соотношение объемов текстов частушек, заданий ЕГЭ по русскому языку, кулинарных рецептов, заговоров от грыжи и вузовских учебников по микробиологии?

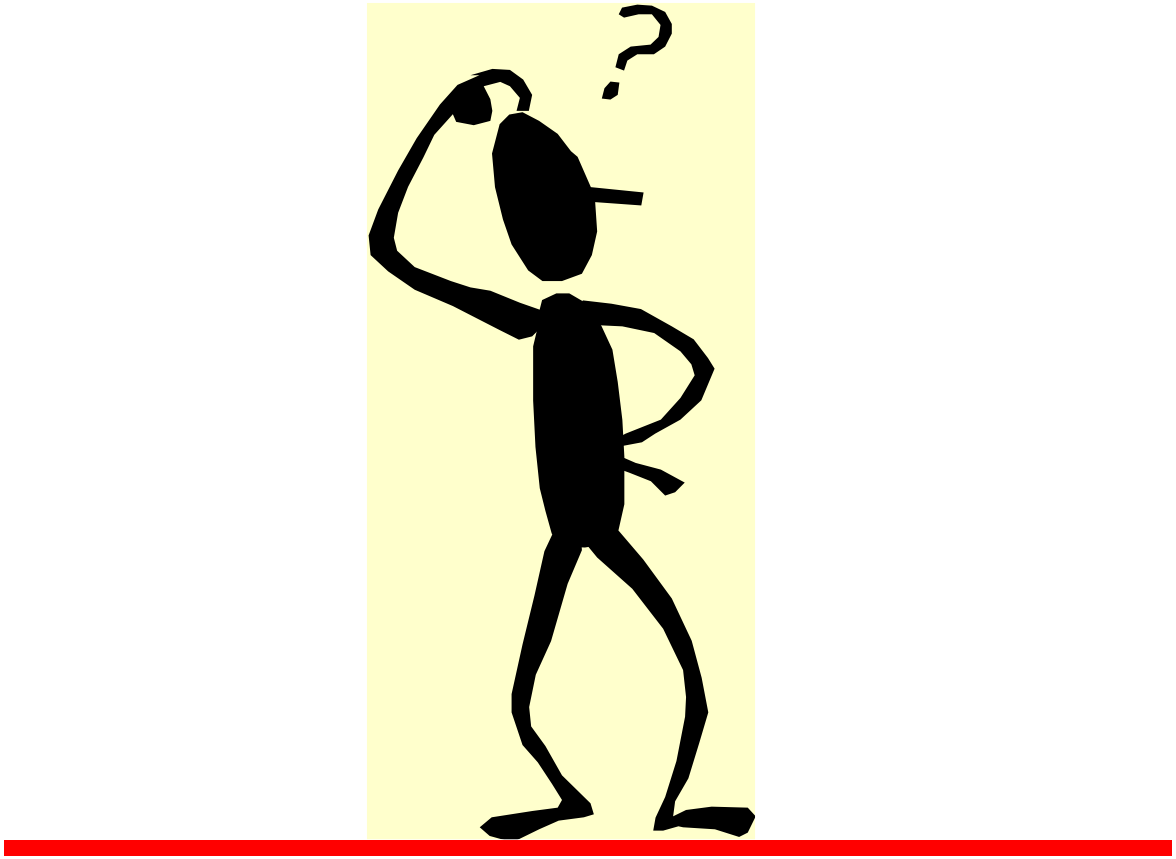
Результаты наших исследований показывают, что при изучении языковых конструкций дифференциальные особенности в употреблении могут обнаруживаться не только там, где они интуитивно ожидаются.

диалог
международная конференция

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

диалог
международная конференция

Any questions!



диалог
международная конференция