

Один корпус — хорошо,  
а много — лучше

Александр Пиперски

Научный семинар ИЛ РГГУ

21.11.2013

# Основные методы лингвистических исследований

1. Интроспекция
2. Эксперимент
3. Наблюдение над действительностью

# Основные объекты лингвистических исследований

- Объектом абсолютного большинства исследований является тот или иной язык (или языки):  
русский язык, английский язык, киргизский язык, язык ландума, ...
- Намного реже встречаются исследования разновидностей языка, выделенных по тем или иным принципам:  
вологодский диалект русского языка, русский молодежный жаргон, язык Пушкина, ...

# Методологические проблемы

- Проблема **применимости методов**:  
Интроспекция и эксперимент неприменимы к некоторым объектам исследования (напр., язык Пушкина)
- Проблема **масштабируемости**:  
Можно ли обобщать результаты, полученные при помощи того или иного метода на том или ином материале, на весь язык / на интересующую исследователя разновидность языка?

# Корпуса русского языка

- Национальный корпус русского языка ([www.ruscorpora.ru](http://www.ruscorpora.ru))
- А что еще?
- ⇒ остальными корпусами мало кто пользуется

# Почему другими корпусами мало кто пользуется?

- Они плохие?
- Они маленькие?
- Они неудобные?
- НКРЯ хватает для всех нужд лингвистов?
- Они плохо разрекламированы?

# НКРЯ и исследовательская программа русистики

- Практически все корпусные исследования выполняются на материале НКРЯ
- Неверно, что НКРЯ хватает для всех нужд лингвистов: на самом деле НКРЯ во многом определяет исследовательскую программу:
  - что лингвисты делают
  - что лингвисты не делают

# Пример 1: диахроническое варьирование

- В НКРЯ есть хронологическая разметка (дата написания текста выдается при поиске) ⇒ исследователи обращают внимание на диахроническое варьирование
- НКРЯ включает в основной подкорпус тексты с XVIII по XXI век, тем самым косвенно утверждая, что все это — современный русский язык
  - Британский национальный корпус (BNC): 1975–1994
  - Корпус современного американского английского языка (COCA): 1990–2012



# Пример 1: диахроническое варьирование

- Как образуется имя действия от глагола *переадресовать / переадресовывать?*
- Словари: *переадресование > переадресовка > переадресация*
  - БАС: *переадресовка* — см. *переадресование*, *переадресация* отсутствует
- НКРЯ:  
*переадресовка* 25 > *переадресация* > 10 > *переадресование* 1
- Не соответствует современному узусу

# Пример 2



- *голубика или гонобобель?*

# Пример 2:

## установка на авторитет

- НКРЯ содержит много текстов классической русской литературы и при показе примеров сразу выдает имя автора

11. В. В. Набоков. Пильграм (1930) [омонимия не снята] [Все примеры \(1\)](#)

Он посещал и север — болота Лапландии, где мох, **гонобобель** и карликовая ива, богатый мохнаты альпийские пастбища, с плоскими камнями, лежащими там и сям среди старой, скользкой, колтунн чем приподнять такой камень, под которым и муравьи, и синий скарабей, и толстенная сонная ночь же, в горах, он видел полупрозрачных, красноглазых аполлонов, которые плывут по ветру через гор отделенный широкой каменной оградой от пропасти, где бурно белеет вода. [В. В. Набоков. Пильграм (193

12. М. М. Пришвин. Дневники (1929) [омонимия не снята] [Все примеры \(1\)](#)

Сила солнечного жара сохранялась еще в вечерних лучах, на опушке бора, обращенной к западу, пр бруснику и **гонобобель**, по сухой тропинке бегали очень быстро и ровно, как заведенные, молодые не снята] ←...→

# Пример 2:

## установка на авторитет

1. *гонобобель* встречается в НКРЯ 21 раз, *голубика* — 135 раз
  2. *гонобобель* встречается в текстах Пришвина, Каверина, Астафьева и Вознесенского
- Утверждение типа 2 невозможно было бы сделать, например, на материале Британского национального корпуса (BNC), составители которого не ставили перед собой цель включать тексты известных авторов

# Выдача Британского национального корпуса (поисковый сервис Британской библиотеки) Results of your search

---

Your query was

controversial

---

Here is a random selection of 50 solutions from the 2118 found.

**A6D** 1162 Like the cross-dressing explored in the last chapter, gender inversion remains controversial because it allegedly only inverts, 1

**A9D** 803 Known for its gleaming and airy entrance lobby, bedecked with trees and tropical plants which became a night-time illuminatic  
City's recently constructed edifices.

**ACP** 31 One of New York's most controversial (and successful) artists has now teamed up with Italy's most controversial MP, Cicciorir

**AJM** 310 Ken Hargreaves, the sitting member, was probably helped last time by having the controversial Liverpool councillor Keva Cc

**AK2** 524 The assault on Labour's tax policies appeared to miss its target; the Tories' handling of Labour's controversial election broadc  
and Mr Major and his advisers seemed uncertain how he should be presented.

# Пример 3: коллокации

- НКРЯ нет инструмента для того, чтобы получить список коллокаций (частотных сочетаний с другими словами) для данного слова ⇒  
изучение коллокаций в русскоязычной традиции корпусной лингвистики намного менее развито, чем в англоязычной

# Поиск коллокаций в Corpus of Contemporary American English

**DISPLAY**

LIST  CHART  **KWIC**  COMPARE

**SEARCH STRING**

WORD(S)

COLLOCATES

POS LIST

**RANDOM**

# Поиск коллокаций в Corpus of Contemporary American English

	<input type="checkbox"/>	CONTEXT
1	<input type="checkbox"/>	MOST
2	<input type="checkbox"/>	ISSUES
3	<input type="checkbox"/>	ISSUE
4	<input type="checkbox"/>	HIGHLY
5	<input type="checkbox"/>	DECISION
6	<input type="checkbox"/>	REMAINS
7	<input type="checkbox"/>	SOMEWHAT
8	<input type="checkbox"/>	SUBJECT
9	<input type="checkbox"/>	TOPICS
10	<input type="checkbox"/>	PROPOSAL
11	<input type="checkbox"/>	TOPIC
12	<input type="checkbox"/>	DECISIONS
13	<input type="checkbox"/>	REMAIN



# Пример 3: коллокации

- Журнал «Русский язык в научном освещении»
- 19 номеров доступно онлайн
- Термин *коллокация* встречается в двух статьях, автором (или соавтором) которых является Д. О. Добровольский
- Похожее явление в русской лингвистической традиции изучается в теории лексических функций, но это другой подход: ономасиологический (от функции к форме), а не семасиологический (от формы к функции)

# Пример 4: региональная разметка

- В НКРЯ нет региональной разметки ⇒ изучение региональных вариантов русского языка считается маргинальным
- Ср. корпус GloWbE (Corpus of Global Web-Based English)

# Слова *truck* и *lorry* в GloWbE

	<input type="checkbox"/>	CONTEXT	ALL <input type="checkbox"/>	US <input type="checkbox"/>	CA <input type="checkbox"/>	GB <input type="checkbox"/>	IE <input type="checkbox"/>	AU <input type="checkbox"/>	NZ <input type="checkbox"/>	IN <input type="checkbox"/>	LK <input type="checkbox"/>	PK <input type="checkbox"/>
1	<input type="checkbox"/>	TRUCK	37606	26.16 10120	34.62 4665	11.79 4569	12.20 1233	20.67 3063	23.82 1939	15.36 1481	18.63 868	17.58 903
2	<input type="checkbox"/>	LORRY	3647	0.25 97	0.08 11	4.28 1660	3.07 310	0.44 65	0.52 42	1.36 131	3.93 183	0.31 16
		TOTAL	41253	10217	4676	6229	1543	3128	1981	1612	1051	919

BD <input type="checkbox"/>	SG <input type="checkbox"/>	MY <input type="checkbox"/>	PH <input type="checkbox"/>	HK <input type="checkbox"/>	ZA <input type="checkbox"/>	NG <input type="checkbox"/>	GH <input type="checkbox"/>	KE <input type="checkbox"/>	TZ <input type="checkbox"/>	JM <input type="checkbox"/>
17.80 703	12.59 541	15.20 633	20.12 870	18.00 728	26.54 1204	13.11 559	18.34 711	22.87 939	28.07 987	22.49 890
0.96 38	2.65 114	3.82 159	0.60 26	1.93 78	0.66 30	1.66 71	6.29 244	3.99 164	5.46 192	0.40 16
741	655	792	896	806	1234	630	955	1103	1179	906

# Другие корпуса русского языка

- Уппсальский корпус: 1 млн слов
  - 1 млн слов
  - Нет морфологической разметки
  - Транслитерация
  - Нет системы онлайн-поиска

# Уппсальский корпус: образец текста

- %%sgid1™ SGID0101™ @Ideologi,,  
obnovleni,,@@@Reweni,, \*Plenuma \*CK \*KPSS  
zovut k aktivnym dejstvi,,m@@&17-18 fevral,,  
sosto,,ls,, \*Plenum \*Central'nogo \*Komiteta \*KPSS.  
Na\*Plenume s reč'~ "Revol~cionnoj perestrojke  
\_ideologi~ obnovleni,, "vystupil \*General'nyj  
sekretar' \*CK \*KPSS \*M# \*S# \*Gorbačev. V nej  
danglubokij analiz novogo "tapa perestrojki,  
izlohena programma eeideologičeskogo  
obespečeni,,.&\*Plenum rassmotrel vopros "O xode  
perestrojki srednej i vyswej wkolyi zadačax partii po  
ee osuqestvleni~".

# OpenCorpora

- Проект группы компьютерных лингвистов из Санкт-Петербурга
- Имеет морфологическую разметку
- Разрешение омонимии при помощи краудсорсинга
- Тексты доступны для скачивания в формате xml
- Нет веб-интерфейса для поиска

# OpenCorpora: образец текста

- ```
<paragraph id="224"> <sentence id="725"> <source>Правозащитники
убеждены: обвинять Наталью нельзя.</source> <tokens> <token
id="14515" text="Правозащитники"><tfr t="Правозащитники"><v><l
id="265706" t="правозащитник"><g v="NOUN"/><g v="anim"/><g
v="masc"/><g v="plur"/><g v="nomn"/></l></v></tfr></token> <token
id="14516" text="убеждены"><tfr t="убеждены"><v><l id="352903" t="
убежден"><g v="PRTS"/><g v="perf"/><g v="past"/><g v="pssv"/><g
v="plur"/></l></v></tfr></token> <token id="14517" text=":"><tfr t=":"><v><l
id="0" t=":"><g v="PNCT"/></l></v></tfr></token> <token id="14518" text="
обвинять"><tfr t="обвинять"><v><l id="193423" t="обвинять"><g
v="INFN"/><g v="impf"/><g v="tran"/></l></v></tfr></token> <token
id="14519" text="Наталью"><tfr t="Наталью"><v><l id="176736" t="
наталья"><g v="NOUN"/><g v="anim"/><g v="femn"/><g v="Name"/><g
v="sing"/><g v="accs"/></l></v></tfr></token> <token id="14520" text="
нельзя"><tfr t="нельзя"><v><l id="183884" t="нельзя"><g v="PRED"/><g
v="pres"/></l></v></tfr></token> <token id="14521" text="."><tfr t="."><v><l
id="0" t="."><g v="PNCT"/></l></v></tfr></token> </tokens> </sentence>
</paragraph>
```

# Интернет-корпуса русского языка

- RuWac (Russian Web as Corpus), С. А. Шаров
- ruTenTen, А. Килгаррифф
- Эти корпуса состоят из текстов,  
автоматически собранных из Интернета и  
автоматически размеченных



# ruTenTen

- В составе проекта SketchEngine
- TenTen =  $10^{10}$  слововхождений
- На самом деле — уже  $\approx 16$  млрд слововхождений  $\Rightarrow$  самый большой из существующих корпусов всех языков
- Особенность SketchEngine — составление *word sketches* (списки частотных сочетаний, распределенных по синтаксическим функциям)

# капуста: word sketch (1)

капуста

ruTenTen11 freq = [374907](#) (18.6 per million)

| <u>subject_of</u> | <u>16766</u>        | <u>0.0</u> | <u>a_modifier</u> | <u>131244</u>         | <u>0.2</u> | <u>и/или</u> | <u>112172</u>         | <u>0.2</u> | <u>prec_prep</u> | <u>57608</u>          | <u>0.0</u> | <u>gen_modifies</u> | <u>54889</u>         | <u>0.1</u> |
|-------------------|---------------------|------------|-------------------|-----------------------|------------|--------------|-----------------------|------------|------------------|-----------------------|------------|---------------------|----------------------|------------|
| вариться          | <a href="#">199</a> | 7.19       | квашеный          | <a href="#">21472</a> | 12.13      | морковь      | <a href="#">10250</a> | 10.0       | из               | <a href="#">15432</a> | 3.53       | кочан               | <a href="#">5878</a> | 11.0       |
| свариться         | <a href="#">75</a>  | 6.83       | белокочанный      | <a href="#">10571</a> | 11.21      | свекла       | <a href="#">7090</a>  | 9.59       | вместо           | <a href="#">313</a>   | 3.27       | квашение            | <a href="#">795</a>  | 8.0        |
| шинкуется         | <a href="#">52</a>  | 6.65       | цветной           | <a href="#">23823</a> | 11.02      | огурец       | <a href="#">6208</a>  | 9.1        | из-под           | <a href="#">103</a>   | 3.03       | саvoyской           | <a href="#">554</a>  | 8.0        |
| тушиться          | <a href="#">58</a>  | 6.57       | брюссельский      | <a href="#">4604</a>  | 10.05      | шпинат       | <a href="#">2316</a>  | 8.81       | кроме            | <a href="#">409</a>   | 2.86       | рассада             | <a href="#">1373</a> | 8.0        |
| подешеветь        | <a href="#">149</a> | 6.43       | тушеный           | <a href="#">4398</a>  | 9.87       | картофель    | <a href="#">5998</a>  | 8.73       | про              | <a href="#">391</a>   | 2.39       | рассол              | <a href="#">470</a>  | 7.0        |
| уродиться         | <a href="#">47</a>  | 6.26       | пекинский         | <a href="#">3609</a>  | 9.56       | кабачок      | <a href="#">2258</a>  | 8.67       | с                | <a href="#">17308</a> | 2.23       | засолка             | <a href="#">267</a>  | 7.0        |
| квасится          | <a href="#">37</a>  | 6.17       | морской           | <a href="#">19147</a> | 9.12       | помидор      | <a href="#">3855</a>  | 8.64       | вроде            | <a href="#">45</a>    | 1.91       | г                   | <a href="#">3999</a> | 6.0        |
| подорожать        | <a href="#">189</a> | 6.1        | кислый            | <a href="#">3359</a>  | 9.02       | репа         | <a href="#">1893</a>  | 8.62       | насчет           | <a href="#">42</a>    | 1.76       | калорийность        | <a href="#">542</a>  | 6.0        |
| выращиваться      | <a href="#">73</a>  | 5.67       | квашенной         | <a href="#">1581</a>  | 8.6        | картошка     | <a href="#">3196</a>  | 8.48       | от               | <a href="#">2286</a>  | 0.89       | сорт                | <a href="#">2046</a> | 6.0        |
| горчить           | <a href="#">30</a>  | 5.57       | свежий            | <a href="#">8180</a>  | 8.18       | редис        | <a href="#">1582</a>  | 8.46       | подобно          | <a href="#">18</a>    | 0.88       | грамм               | <a href="#">691</a>  | 6.0        |
| остыть            | <a href="#">54</a>  | 5.43       | краснокочанной    | <a href="#">1072</a>  | 8.05       | редька       | <a href="#">1639</a>  | 8.42       | у                | <a href="#">1087</a>  | 0.85       | сок                 | <a href="#">2519</a> | 6.0        |
| Отцвела           | <a href="#">21</a>  | 5.35       | вареный           | <a href="#">1064</a>  | 7.43       | морковка     | <a href="#">1641</a>  | 8.41       | помимо           | <a href="#">56</a>    | 0.82       | кочерыжка           | <a href="#">148</a>  | 6.0        |
| размягчиться      | <a href="#">22</a>  | 5.31       | заячий            | <a href="#">694</a>   | 7.34       | брокколи     | <a href="#">1310</a>  | 8.37       | под              | <a href="#">524</a>   | 0.46       | соцветие            | <a href="#">228</a>  | 6.0        |
| возделываться     | <a href="#">25</a>  | 5.24       | кочанный          | <a href="#">540</a>   | 7.05       | томат        | <a href="#">1667</a>  | 8.12       | включая          | <a href="#">37</a>    | 0.43       | вилка               | <a href="#">455</a>  | 6.0        |
| осесть            | <a href="#">63</a>  | 5.13       | листовой          | <a href="#">675</a>   | 6.76       | лук          | <a href="#">5169</a>  | 8.07       | без              | <a href="#">433</a>   | 0.24       | шинковка            | <a href="#">122</a>  | 6.0        |
| закваситься       | <a href="#">16</a>  | 4.96       | нашинкованной     | <a href="#">435</a>   | 6.75       | тыква        | <a href="#">1747</a>  | 8.02       | вокруг           | <a href="#">44</a>    | 0.16       | лист                | <a href="#">4617</a> | 6.0        |
| закиснуть         | <a href="#">17</a>  | 4.94       | белокочанной      | <a href="#">431</a>   | 6.74       | брюква       | <a href="#">789</a>   | 7.71       |                  |                       |            | моркой              | <a href="#">110</a>  | 6.0        |
| отбеливать        | <a href="#">33</a>  | 4.93       | краснокочанная    | <a href="#">422</a>   | 6.71       | горох        | <a href="#">1140</a>  | 7.58       |                  |                       |            | выращивание         | <a href="#">746</a>  | 6.0        |
| перекисла         | <a href="#">15</a>  | 4.87       | маринованный      | <a href="#">424</a>   | 6.56       | фасоль       | <a href="#">1133</a>  | 7.51       |                  |                       |            | вилок               | <a href="#">103</a>  | 5.0        |
| подрумяниться     | <a href="#">16</a>  | 4.77       | отварной          | <a href="#">459</a>   | 6.51       | спаржа       | <a href="#">712</a>   | 7.38       |                  |                       |            | кочаном             | <a href="#">96</a>   | 5.0        |
| развариться       | <a href="#">14</a>  | 4.7        | китайский         | <a href="#">2431</a>  | 6.36       | яблоко       | <a href="#">2242</a>  | 7.3        |                  |                       |            | гр                  | <a href="#">195</a>  | 5.0        |

# капуста: word sketch (2)

| object4 of     | 23767                | 0.2  | pp obj c  | 13612                | 0.2  | pp obj из   | 12939                | 0.5  | pp c        | 7994                | 0.1  | pp в       | 6056                | 0.0 |
|----------------|----------------------|------|-----------|----------------------|------|-------------|----------------------|------|-------------|---------------------|------|------------|---------------------|-----|
| квасить        | <a href="#">604</a>  | 9.5  | пирожок   | <a href="#">1918</a> | 9.55 | салат       | <a href="#">6289</a> | 8.95 | расти-      | <a href="#">40</a>  | 7.35 | кляре      | <a href="#">191</a> | 9.5 |
| нашинковывать  | <a href="#">574</a>  | 8.94 | пирог     | <a href="#">1689</a> | 8.25 | салатик     | <a href="#">147</a>  | 7.52 | морковка    | <a href="#">204</a> | 6.96 | буртах     | <a href="#">12</a>  | 5.0 |
| рубить         | <a href="#">844</a>  | 8.14 | вареник   | <a href="#">232</a>  | 7.94 | солянку     | <a href="#">88</a>   | 7.43 | раст        | <a href="#">25</a>  | 6.26 | пароварке  | <a href="#">22</a>  | 5.6 |
| шинковать      | <a href="#">189</a>  | 7.85 | кулебяка  | <a href="#">78</a>   | 7.22 | голубец     | <a href="#">75</a>   | 6.77 | черносливом | <a href="#">17</a>  | 6.12 | опте       | <a href="#">11</a>  | 5.6 |
| солить         | <a href="#">262</a>  | 7.7  | сосиска   | <a href="#">260</a>  | 7.01 | салянку     | <a href="#">42</a>   | 6.7  | баранина    | <a href="#">154</a> | 6.11 | кляр       | <a href="#">11</a>  | 5.0 |
| шинкуем        | <a href="#">130</a>  | 7.4  | борщ      | <a href="#">157</a>  | 6.44 | шницель     | <a href="#">50</a>   | 6.55 | морковь     | <a href="#">454</a> | 6.06 | сотейник   | <a href="#">22</a>  | 5.4 |
| заквашивать    | <a href="#">95</a>   | 6.95 | рулька    | <a href="#">37</a>   | 6.34 | суп         | <a href="#">839</a>  | 6.47 | сосиска     | <a href="#">94</a>  | 5.67 | парник     | <a href="#">41</a>  | 5.3 |
| Шинкуем        | <a href="#">82</a>   | 6.81 | грядка    | <a href="#">183</a>  | 6.3  | оладья      | <a href="#">69</a>   | 6.4  | бекон       | <a href="#">48</a>  | 5.61 | пароварку  | <a href="#">12</a>  | 5.2 |
| выращивать     | <a href="#">862</a>  | 6.59 | колбаска  | <a href="#">81</a>   | 6.13 | гарнир      | <a href="#">142</a>  | 6.26 | копченость  | <a href="#">27</a>  | 5.59 | горшочек   | <a href="#">54</a>  | 5.2 |
| отваривать     | <a href="#">169</a>  | 6.55 | кадушка   | <a href="#">36</a>   | 6.05 | пелюстки    | <a href="#">29</a>   | 6.18 | клюква      | <a href="#">81</a>  | 5.44 | панировка  | <a href="#">12</a>  | 5.0 |
| тушить         | <a href="#">262</a>  | 6.47 | свинина   | <a href="#">237</a>  | 5.76 | рага        | <a href="#">43</a>   | 6.02 | Картофелем  | <a href="#">9</a>   | 5.18 | салатник   | <a href="#">13</a>  | 5.1 |
| кушать         | <a href="#">619</a>  | 6.35 | гусь      | <a href="#">162</a>  | 5.75 | запеканка   | <a href="#">75</a>   | 5.8  | тмин        | <a href="#">48</a>  | 5.16 | борщ       | <a href="#">51</a>  | 5.0 |
| потушить       | <a href="#">171</a>  | 6.33 | винегрет  | <a href="#">40</a>   | 5.59 | котлета     | <a href="#">123</a>  | 5.76 | курица      | <a href="#">215</a> | 5.09 | чехии      | <a href="#">10</a>  | 4.8 |
| стричь         | <a href="#">158</a>  | 6.32 | смешать   | <a href="#">60</a>   | 5.49 | обертывание | <a href="#">55</a>   | 5.67 | горошек     | <a href="#">55</a>  | 4.98 | кадка      | <a href="#">19</a>  | 4.8 |
| приготовливать | <a href="#">1107</a> | 6.21 | блинчик   | <a href="#">56</a>   | 5.48 | солянка     | <a href="#">24</a>   | 5.51 | чернослив   | <a href="#">41</a>  | 4.88 | диетологии | <a href="#">31</a>  | 4.7 |
| сажать         | <a href="#">338</a>  | 6.13 | борщ      | <a href="#">18</a>   | 5.42 | блюдо       | <a href="#">1051</a> | 5.33 | кочерыжка   | <a href="#">10</a>  | 4.84 | сухарь     | <a href="#">38</a>  | 4.7 |
| отварить       | <a href="#">56</a>   | 5.92 | смешивать | <a href="#">304</a>  | 5.39 | борщ        | <a href="#">68</a>   | 5.25 | картошка    | <a href="#">161</a> | 4.81 | кастрюля   | <a href="#">151</a> | 4.7 |
| есть           | <a href="#">620</a>  | 5.79 | суп       | <a href="#">397</a>  | 5.39 | Суп-пюре    | <a href="#">16</a>   | 5.2  | брусника    | <a href="#">41</a>  | 4.72 | кочан      | <a href="#">17</a>  | 4.4 |
| протыкать      | <a href="#">63</a>   | 5.77 | рульку    | <a href="#">18</a>   | 5.31 | кимчи       | <a href="#">14</a>   | 5.1  | яблоко      | <a href="#">283</a> | 4.67 | голубец    | <a href="#">10</a>  | 4.3 |
| срубать        | <a href="#">69</a>   | 5.77 | зраза     | <a href="#">20</a>   | 5.29 | начинка     | <a href="#">167</a>  | 4.91 | кальмар     | <a href="#">38</a>  | 4.65 | рассадник  | <a href="#">15</a>  | 4.3 |
| засаливать     | <a href="#">40</a>   | 5.63 | салат     | <a href="#">471</a>  | 5.21 | кочерыжка   | <a href="#">15</a>   | 4.91 | гриб        | <a href="#">326</a> | 4.56 | маринад    | <a href="#">24</a>  | 4.1 |



# капуста с раст

- 1217074 или вареной рыбы, 200гр салата из свежей **капусты** с **раст** .маслом.6-й день завтрак- черный
- 1156585 . Это намного вкуснее и приятнее. Вместо **капусты** с **раст** .маслом (я ее ненавижу) съедала
- 919963 день Завтрак (3) - кофе. Обед (0) - вареная **капуста** с **раст** . маслом; 2 вареных яйца. Ужин (
- 919963 жареный кабачок; У: пара вареных яиц, салат из **капусты** с **раст** . маслом, 200 гр вареной говядины
- 530849 яйца, 200гр. отварной говядины, салат из **капусты** с **раст** . маслом. </p><p> 26 ДЕНЬ </p><p> УТРО
- 933119 Кофе, сухарик Обед рыба несоленая, салат из **капусты** с **раст** .маслом Ужин 200 грамм вареной несоленой
- 933119 ,200 г вареной говядины, салат из свежей **капусты** с **раст** маслом </p><p> 4 день Завтрак кофе
- 107331 Столичный», грибы консервированные, морская **капуста** с **раст** . маслом, свежие овощи с раст. маслом
- 236360 кабачок в раст. масле 2 яйца вкрутую, салат из **капусты** с **раст** . маслом, 200 г. варёной говядины
- 625235 с молоком Обед. 4 картофелины, салат из **капусты** с **раст** .маслом Полдник . 2 яблока Ужин.
- 629727 Черный кофе. Обед: 2 крутых яйца, вареная **капуста** с **раст** . маслом, 1ст. томатного сока. Ужин
- 629727 вареная или жаренная, салат из моркови и **капусты** с **раст** . маслом. Ужин: 200гр. вареной говядины
- 1320 мебели салат из тертой сырой свеклы,моркови и **капусты** с **раст** .маслом если просто поробовать?
- 877470 помидора с зеленым луком и раст. маслом 16.00 **Капуста** с **раст** . маслом 18.00 100гр говядины, 100
- 771321 яйца, 200гр. Отварной говядины, салат из **капусты** с **раст** . Маслом. </p><p> 26 ДЕНЬ УТРО: кофе
- 830948 изменения?И что значит "Салат из овощей, **капусты** с **раст** .маслом",кроме капусты можно еще
- 800367 помидора с зеленым луком и раст. маслом 16.00 **Капуста** с **раст** . маслом 18.00 100гр говядины, 100
- 855523 Обед . 100г отварной говядины, салат из **капусты** с **раст** .маслом и 1 стакан фруктового сока
- 855523 с молоком Обед . 4 картофелины, салат из **капусты** с **раст** .маслом Полдник . 2 яблока Ужин
- 855523 Обед . 2 отварных картофелины и салат из **капусты** с **раст** .маслом Полдник . 1 яблоко и 1 стакан

# *капуста* в корпусах

- RuWas ( $\approx$  2 млрд слововхождений): 33589
- ruTenTen ( $\approx$  16 млрд слововхождений): 374907
- НКРЯ ( $\approx$  230 млн слововхождений): 4711
  
- НКРЯ на порядок меньше RuWas,  
а RuWas на порядок меньше ruTenTen

# Особенности НКРЯ

- Ручной отбор и добавление текстов
- Приоритет отдается текстам высокой культурной значимости
- Ручное разрешение омонимии

# Особенности НКРЯ: +

- Ручной отбор и добавление текстов
- Приоритет отдается текстам высокой культурной значимости  
⇒ высокое качество отбора материала в соответствии с общими представлениями о том, что входит в СРЛЯ
- Ручное разрешение омонимии  
⇒ высокое качество грамматического разбора в подкорпусе со снятой омонимией

# Особенности НКРЯ: –

- Ручной отбор и добавление текстов  
⇒ практическая ограниченность объема корпуса (сейчас — ок. 230 млн слов в основном подкорпусе)
- Отбор текстов по культурной значимости  
⇒ НКРЯ хорош как корпус русской классической литературы, а многие другие жанры представлены в нем недостаточно
- Ручное снятие неоднозначности  
⇒ ограниченность объема корпуса со снятой омонимией (сейчас — ок. 6 млн слов)



# Чему мешают особенности НКРЯ?

- НКРЯ слишком мал для изучения некоторых низкочастотных слов и конструкций
  - новые слова и конструкции
  - регионализмы
  - слова и конструкции за пределами художественной литературы, напр. в языке Интернета

# Ручное разрешение ОМОНИМИИ

- Небольшое количество разметчиков
- В ручной разметке тоже встречаются ошибки!

*А. А. Зализняк. Лингвистика по А. Т. Фоменко // «Вопросы языкознания», 2000*

Почему бы не предположить, например, что Венеция — это Винница, Парма — это Пермь, Лукка — это Великие **Луки**, Кельн — это Клин, Глазго — это Глазов, Верден — это Бородино...

# Ручное разрешение ОМОНИМИИ

- о + родительный падеж в подкорпусе НКРЯ со снятой омонимией:
- *журналы о кино, в память о погибших японских друзьях, вопрос о доказуемости постулата о параллельных, слухи о неких «зеленых призраках», теория Троцкого о Клемансо*

# Автоматическое разрешение ОМОНИМИИ

- Автоматическое разрешение омонимии основывается на грамматической разметке соседних слов
- Точность автоматического разрешения омонимии у разных таггеров составляет > 95%
- NB: важны не числовые показатели, а наличие/отсутствие типовых случаев, не поддающихся разбору

# Условный пример автоматической разметки

- о + прилагательное на *-ой* + слово женского рода на *-е* (в начальной форме — на *а*)
- *о яровой пшенице*  
*о русской смекалке*  
*о случайной отставке*
- Экономно ли в таких случаях использовать ручное разрешение омонимии?

# Пример задачи, неразрешимой при помощи НКРЯ (1)

- Как образуется в современном русском языке множественное число от слова *свитер*:  
*свитеры* или *свитера*?
- *свитеры* 25, *свитера* 347
- Чтобы получить точную статистику, надо либо просматривать все 347 примеров вручную, либо делать аппроксимацию
  - из 30 случайно отобранных примеров на *свитера* ко множественному числу относятся 17  $\Rightarrow$   $\approx$ 196 примеров из 347 — множественное число

# Пример задачи, неразрешимой при помощи НКРЯ (1)

- ruTenTen:

  - [word= "свитеры"]: 2175

  - [word= "свитера" & tag="N..p.\*"]: 31239

- В автоматическом разрешении омонимии есть ошибки, но они влияют на общий результат незначительно

# Пример задачи, неразрешимой при помощи НКРЯ (2)

- С какими словами употребляется собирательное числительное *двое*, а с какими — словосочетание *две пары*?
- *двое очков* или *две пары очков*,  
*двое туфель* или *две пары туфель*?

[Микаэлян, Зализняк 2013]



# НКРЯ / Яндекс.Блоги

|               |                          |                   |        |
|---------------|--------------------------|-------------------|--------|
| двое суток    | 2080/4900                | *две пары суток   | 0/4    |
| двое ворот    | 29/1937                  | две пары ворот    | 0/52   |
| двое саней    | 26/219                   | две пары саней    | 1/5    |
| двое часов    | 15/69                    | две пары часов    | 4/439  |
| двое носилок  | 6/138                    | две пары носилок  | 1/5    |
| двое очков    | 6/1038                   | две пары очков    | 11/49  |
| двое брюк     | 6 (из них 3 в 19 в.)/998 | две пары брюк     | 9/1189 |
| двое трусов   | 3/1235                   | две пары трусов   | 3/843  |
| двое весов    | 1 (19 век)/1097          | две пары весов    | 0/6    |
| двое ножниц   | 0/289                    | две пары ножниц   | 0/165  |
| двое родов    | 2/824                    | *две пары родов   | 0/0    |
| двое валенок  | 1/20                     | две пары валенок  | 4/202  |
| двое перчаток | 1/200                    | две пары перчаток | 7/1261 |
| двое сапог    | 0/322                    | две пары сапог    | 13/58  |
| двое туфель   | 0/109                    | две пары туфель   | 3/38   |

# Генеральный Интернет-корпус русского языка (ГИКРЯ)

- Разработчики:
  - кафедра компьютерной лингвистики ИЛ РГГУ
  - АВВУУ
  - Университет Лидса
- NB: ГИКРЯ не претендует на то, чтобы заменить собой все упомянутые выше корпуса — это новый корпус с новым уникальным набором достоинств и недостатков

# Генеральный Интернет-корпус

## русского языка (ГИКРЯ)

- Корпус автоматически собранных из Интернета текстов
- Автоматическая морфологическая разметка
- Автоматическое извлечение метаразметки
- Автоматическое присвоение неразмеченным текстам метаразметки (в т. ч. жанровой разметки)

# Генеральный Интернет-корпус русского языка

- Автоматическое скачивание текстов с ресурсов, список которых определяется вручную:
  - LiveJournal
  - Журнальный зал (<http://magazines.russ.ru>)
  - Новостные порталы (Lenta.ru, Regnum и т. д.)
  - Крупные форумы (Форум Винского и т. д.)
  - ...
- ⇒ **дифференциальная полнота**

# Репрезентативность, сбалансированность, дифференциальная полнота

- Все неспециализированные корпуса претендуют на репрезентативность и сбалансированность:
- *Болгарский национальный корпус постоянно развивается и пополняется новыми текстами, ставя перед собой цель достичь представительности и сбалансированности благодаря включению текстов разных способов бытования (письменных и устных), разных эпох и разнообразной тематической и жанровой принадлежности.*

# Репрезентативность, сбалансированность, дифференциальная полнота

- *Национальный корпус ... характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода* [НКРЯ]

# Репрезентативность, сбалансированность, дифференциальная полнота

- Что стоит за понятиями «репрезентативность» и «сбалансированность»?
- Более осязаемое понятие — **дифференциальная полнота**: в корпусе должны быть представлены различные типы текстов, для каждого из которых можно оценить его репрезентативность для того иного подъязыка (русский язык блогов, русский язык Татарстана и т. п.)

# Задача из Русского медвежонка (И. С. Рубанов, 2013)

- — Куда пошла мама? — спросила Маша у младшего брата .  
— В магазин, купить кочан этой... как ее... — ответил брат.  
Но Маша все равно сразу поняла, что мама пошла за ...
- (А) капустой; (Б) картошкой; (В) морковкой; (Г) редиской; (Д) колбасой.



# Пример использования ГИКРЯ: вилок vs. кочан

- Где говорят *кочан капусты*, а где — *вилок капусты*?
- Для ответа на этот вопрос нужен корпус с региональной метаразметкой
- Сравниваем количество результатов по запросам  
[word="вил.\*"] [lemma="капуста"]  
[lemma="кочан"] [lemma="капуста"]

# ВИЛОК КОЧАН

| Корпус         | ▲ | Экземпляров | IPM   | Экземпляров | IPM   |
|----------------|---|-------------|-------|-------------|-------|
| AWDRU          |   | 0           | 0.000 | 18          | 0.100 |
| AWDRUM         |   | 0           | 0.000 | 16          | 0.165 |
| AWDRUW         |   | 3           | 0.058 | 12          | 0.230 |
| BASHKIRIYA     |   | 0           | 0.000 | 2           | 0.206 |
| CHELYABINSKAYA |   | 0           | 0.000 | 4           | 0.483 |
| DONETSKAYA     |   | 5           | 0.447 | 0           | 0.000 |
| KIEV           |   | 0           | 0.000 | 7           | 0.317 |
| KRASNODARSKIY  |   | 0           | 0.000 | 0           | 0.000 |
| KRASNOYARSKIY  |   | 1           | 0.123 | 0           | 0.000 |
| MOSOBL         |   | 0           | 0.000 | 2           | 0.104 |
| NOVOSIBIRSKAYA |   | 1           | 0.076 | 2           | 0.152 |
| OMSKAYA        |   | 1           | 0.200 | 2           | 0.399 |
| PERMSKIY       |   | 0           | 0.000 | 1           | 0.101 |
| PETERSBURG     |   | 4           | 0.088 | 8           | 0.176 |
| ROSTOVSKAYA    |   | 0           | 0.000 | 1           | 0.098 |
| RUSSM          |   | 18          | 0.068 | 118         | 0.444 |
| RUWAC-PARSED   |   | 51          | 0.025 | 421         | 0.209 |
| SAMARSKAYA     |   | 2           | 0.139 | 5           | 0.348 |
| SARATOVSKAYA   |   | 0           | 0.000 | 1           | 0.186 |
| SVERDLOVSKAYA  |   | 0           | 0.000 | 4           | 0.244 |
| TATARSTAN      |   | 0           | 0.000 | 2           | 0.303 |
| ZHZHALL        |   | 32          | 0.034 | 191         | 0.202 |
| Totals         |   | 118         | 0.031 | 817         | 0.217 |

# ВИЛОК vs. КОЧАН: ВЫВОДЫ

- Числа в каждом из подкорпусов небольшие, но некоторые выводы все же можно сделать
- В Донецкой области говорят *вилок* (5:0)
- В Санкт-Петербурге распространены и *вилок*, и *кочан* (4:8)
- Общий счет по ЖЖ — 32:191  $\Rightarrow$  *вилок* — не такое уж редкое слово, но для того, чтобы понять, где еще оно распространено, нужно наращивать объемы корпусов

# Корпус с региональной разметкой как инструмент разрешения споров

- Форум «Городские диалекты», обсуждение слова *вилок*:
- *питерский товарищ мне тут тоже сказал, что знает, но сам не употребляет*
- *В Петербурге вообще такого слова (вилок) не слышал ни разу и даже не понрял бы о чем речь.... Всегда говорят "кочан" или, если маленький, то "Кочашок"*

# *поребрик vs. бордюр*

- Каково стандартное представление о распределении этих слов?

# поребрик vs. бордюр

| Корпус         | Экземпляров  | IPM          | Экземпляров  | IPM          |
|----------------|--------------|--------------|--------------|--------------|
| RUWAC-PARSED   | 877          | 0.436        | 3,899        | 1.937        |
| RUSSM          | 134          | 0.504        | 459          | 1.727        |
| AWDRU          | 49           | 0.271        | 357          | 1.974        |
| AWDRUM         | 40           | 0.413        | 208          | 2.146        |
| AWDRUW         | 14           | 0.269        | 123          | 2.363        |
| ZHZHALL        | 343          | 0.362        | 1,928        | 2.035        |
| BASHKIRIYA     | 8            | 0.823        | 24           | 2.470        |
| CHELYABINSKAYA | 1            | 0.121        | 35           | 4.226        |
| DONETSKAYA     | 0            | 0.000        | 14           | 1.253        |
| KIEV           | 2            | 0.091        | 37           | 1.674        |
| KRASNODARSKIY  | 1            | 0.117        | 21           | 2.456        |
| KRASNOYARSKIY  | 2            | 0.246        | 18           | 2.214        |
| MOSOBL         | 2            | 0.104        | 91           | 4.754        |
| NOVOSIBIRSKAYA | 11           | 0.835        | 30           | 2.276        |
| OMSKAYA        | 1            | 0.200        | 13           | 2.595        |
| PERMSKIY       | 4            | 0.403        | 15           | 1.511        |
| PETERSBURG     | 68           | 1.500        | 72           | 1.588        |
| ROSTOVSKAYA    | 0            | 0.000        | 24           | 2.359        |
| SAMARSKAYA     | 3            | 0.209        | 53           | 3.687        |
| SARATOVSKAYA   | 0            | 0.000        | 19           | 3.541        |
| SVERDLOVSKAYA  | 11           | 0.672        | 35           | 2.139        |
| TATARSTAN      | 3            | 0.455        | 8            | 1.213        |
| <b>Totals</b>  | <b>1,574</b> | <b>0.418</b> | <b>7,483</b> | <b>1.986</b> |

# *поребрик* в Башкортостане

- Этот шорт лист передается на оценку жюри, в состав которого войдут профессиональные фотографы, деятели искусств, коренные петербуржцы. <...> Победители получают ценные призы. Три основные номинации:  
"Золотой **поребрик**" - MacBook Air,  
"Серебряный **поребрик**" - iPad 4 и  
"Бронзовый **поребрик**" - iPhone 5.

# *поребрик* в Башкортостане

- Если бы в их жизни был месяц саженцев, граблей и краски для **поребриков**, все у них могло бы пойти совсем-совсем по-другому.

(из интервью Олега Гаркуши, участника петербургской группы «АукцЫон»)

- А когда я улетала на прошлой неделе в Петербург, дал мне один коллега спецзадание - узнать у местных, почему же все-таки у них **поребрик** вместо бордюра.



# *поребрик* в Башкортостане

- Город швырнул меня из парадной, размазал об **поребрик**. Нет, вина во мне, не стоило поддаваться чарам гордского безделья. Прошел год и я снова купил билет на randevу со столицей холода. И еще не приехав, я не хочу уезжать из Питера. У меня петербургомания, я сижу на Петербурге, глотаю его, нюхаю и колю.

# *поребрик* в Башкортостане

- Вывод: собственно башкирских примеров на *поребрик* нет

# *поребрик* в Новосибирской области

- Единственной мерой, направленной для облегчения жизни инвалидов являются пандусы, прорезанные в **поребриках** для спуска с тротуаров на проезжую часть.
- порвал пыльник при неудачном съезде с **поребрика**
- зы у нас в Нске " бордюры "редко говорят, в основном как раз многострадальный "поребрик" :))))))
- Надя обняла все столбы, посидела на всех **поребриках**, побегала заскейтами туда и обратно по параллельной трассе.

# *поребрик*

- В Свердловской области примеры на *поребрик* тоже в основном «свои»
- Вывод: в Новосибирской и Свердловской области край тротуара тоже называется *поребриком*, как и в Санкт-Петербурге

# Гендерная разметка

- ГИКРЯ снабжен гендерной разметкой
  - извлекается из профилей пользователей
  - может быть приписана автоматически
- На данный момент:
  - гендерно размеченные записи мужчин и женщин с Форума Винского (<http://forum.awd.ru/>, Форум самостоятельных путешественников)



**Alexander Piperski** shared Polit.ru's album: ProScience Театр. Максим Кронгауз.

November 11 at 11:36pm ·



Ну не мимими ли мы?  
(или это уже устаревший мем?)



**ProScience Театр. Максим Кронгауз**

By: Polit.ru  
Photos: 29

Like · Comment · Share · Unfollow Post · Promote



Irina Kolosova, Svetlana Bochaver, Наталья Брагина and 11 others like this.



**Anna Popova** розовая рубашка!

November 11 at 11:44pm · Like



**Валя Люсина** мимимиливы!

November 11 at 11:50pm · Like



**Alexander Piperski** Я протестую, она красная в мелкую крапинку

November 11 at 11:50pm · Like



**Anna Popova** эх, а раньше ты бы наоборот протестовал, говоря, что не красная, а розовая! все течет, все меняется

November 11 at 11:54pm · Like



**Alexander Piperski** Все познается в сравнении: у меня же есть настоящая розовая!

November 11 at 11:55pm · Like · 1



**Ksenia Gilyarova** Прекрасные фотографии. А мимими - это ведь женщина только может говорить...

November 12 at 3:13am · Like

# *мимими* на Форуме Винского

| Корпус        | [word="мимими"] |              |
|---------------|-----------------|--------------|
|               | Экземпляров     | IPM          |
| AWDRUM        | 307             | 3.168        |
| AWDRUW        | 155             | 2.977        |
| <b>Totals</b> | <b>462</b>      | <b>3.101</b> |

- Разница в частотности слова *мимими* у мужчин и женщин представляется незначимой
- NB: в НКРЯ 2 вхождения слова *мимими*, автор оба раза обозначен как «КОЛЛЕКТИВНЫЙ»

# Выводы (1)

- Используемые корпуса во многом определяют направления работы лингвистов
- Многие лингвистические вопросы могут быть разрешены только на очень больших корпусах (несколько миллиардов слов), которые неизбежно основываются на автоматическом сборе текстов и автоматической разметке



# Выводы (2)

- Для разных задач нужны разные корпуса
- Чтобы работать с разными корпусами, надо понимать принципиальные особенности их устройства, их достоинства и недостатки

# Список использованных ресурсов (английский язык)

- British National Corpus:  
<http://www.natcorp.ox.ac.uk/>
- Corpus of Contemporary American English:  
<http://corpus.byu.edu/coca/>
- GloWbE: Corpus of Global Web-based English:  
<http://corpus2.byu.edu/glowbe/>

# Список использованных ресурсов (русский язык)

- ruTenTen: <https://the.sketchengine.co.uk/>
- RuWac: <http://corpus.leeds.ac.uk/ruscorpora.html>
- The Uppsala Russian Corpus:  
<http://www.moderna.uu.se/slaviska/ryska/corpus/>
- Национальный корпус русского языка:  
<http://www.ruscorpora.ru>
- Открытый корпус: <http://www.opencorpora.org>

# Список использованных ресурсов (русский язык)

- Генеральный Интернет-корпус русского языка: станет доступен в 2014 году
  - Беликов В. И., Селегей В. П., Шаров С. А. 2012. Прологомены к проекту Генерального интернет-корпуса русского языка. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). Вып. 11 (18). М.: Издательство РГГУ, 2012. С. 37–50.
  - 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 ежегодной Международной конференции «Диалог» (Бекасово, 29 мая – 2 июня 2013 г.). Вып. 12 (19). — М.: Изд-во РГГУ, 2013. С. 84–95.