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)
- 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, колтунно
чем приподнять такой камень, под которым и муравьи, и синий скарабей, и толстенькая сонная ночн
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, r

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, Cicciolir

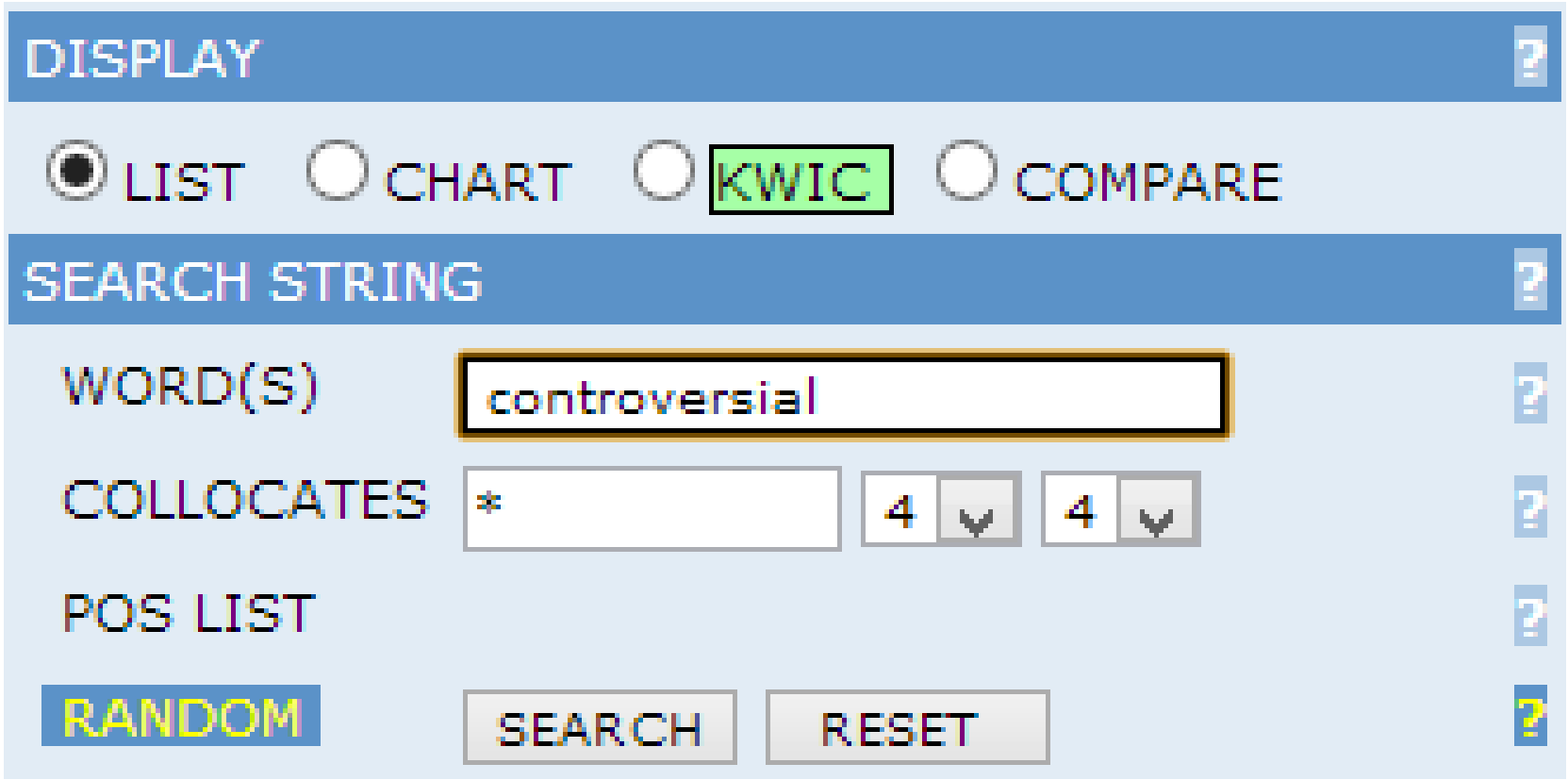
AJM **310** Ken Hargreaves, the sitting member, was probably helped last time by having the controversial Liverpool councillor Keva Cc

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



Поиск коллокаций в Corpus of Contemporary American English

		CONTEXT
1	☐	MOST
2	☐	ISSUES
3	☐	ISSUE
4	☐	HIGHLY
5	☐	DECISION
6	☐	REMAINS
7	☐	SOMEWHAT
8	☐	SUBJECT
9	☐	TOPICS
10	☐	PROPOSAL
11	☐	TOPIC
12	☐	DECISIONS
13	☐	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

		CONTEXT	ALL		US	CA	GB	IE	AU	NZ	IN	LK	PK
1		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		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	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
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 re†'~ "Revol~cionnoj perestrojke _ideologi~ obnovleni„"vystupil *General'nyj sekretar' *CK *KPSS *M# *S# *Gorba†ev. V nej danglubokij analiz novogo ”tapa perestrojki, izlohena programma eeideologi†eskogo obespe†eni„.&*Plenum rassmotrel vopros "O xode perestrojki srednej i vyswej wkolyi zada†ax 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 ≈ 16 млрд слововхождений ⇒ 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)

subject_of	16766	0.0
вариться	199	7.19
свариться	75	6.83
шинкуется	52	6.65
тушиться	58	6.57
подешеветь	149	6.43
уродиться	47	6.26
квасится	37	6.17
подорожать	189	6.1
выращиваться	73	5.67
горчить	30	5.57
остыть	54	5.43
Отцвела	21	5.35
размягчиться	22	5.31
возделываться	25	5.24
осесть	63	5.13
закваситься	16	4.96
закиснуть	17	4.94
отбеливать	33	4.93
перекисла	15	4.87
подрумяниться	16	4.77
развариться	14	4.7

a_modifier	131244	0.2
квашеный	21472	12.13
белокочанный	10571	11.21
цветной	23823	11.02
брюссельский	4604	10.05
тушеный	4398	9.87
пекинский	3609	9.56
морской	19147	9.12
кислый	3359	9.02
квашенной	1581	8.6
свежий	8180	8.18
краснокочанной	1072	8.05
вареный	1064	7.43
заячий	694	7.34
кочанный	540	7.05
листовой	675	6.76
нашинкованной	435	6.75
белокачанной	431	6.74
краснокочанная	422	6.71
маринованный	424	6.56
отварной	459	6.51
китайский	2431	6.36

и/или	112172	0.2
морковь	10250	10.0
свекла	7090	9.59
огурец	6208	9.1
шпинат	2316	8.81
картофель	5998	8.73
кабачок	2258	8.67
помидор	3855	8.64
репа	1893	8.62
картошка	3196	8.48
редис	1582	8.46
редька	1639	8.42
морковка	1641	8.41
брокколи	1310	8.37
томат	1667	8.12
лук	5169	8.07
тыква	1747	8.02
брюква	789	7.71
горох	1140	7.58
фасоль	1133	7.51
спаржа	712	7.38
яблоко	2242	7.3

prec_prep	57608	0.0
из	15432	3.53
вместо	313	3.27
из-под	103	3.03
кроме	409	2.86
про	391	2.39
с	17308	2.23
вроде	45	1.91
насчет	42	1.76
от	2286	0.89
подобно	18	0.88
у	1087	0.85
помимо	56	0.82
под	524	0.46
включая	37	0.43
без	433	0.24
вокруг	44	0.16

gen_modifies	54889	0.1
кочан	5878	11.
квашение	795	8.
савойской	554	8.
рассада	1373	8
рассол	470	7.
засолка	267	7.
г	3999	6.
калорийность	542	6.
сорт	2046	6.
грамм	691	6.
сок	2519	6.
кочерыжка	148	6.
соцветие	228	6.
вилка	455	6.
шинковка	122	6.
лист	4617	6.
моркой	110	6.
выращивание	746	6.
вилок	103	5
кочаном	96	5.
гр	195	5.

капуста: word sketch (2)

object4_of	23767	0.2
квасить	604	9.5
нашинковывать	574	8.94
рубить	844	8.14
шинковать	189	7.85
солить	262	7.7
шинкуем	130	7.4
заквашивать	95	6.95
Шинкуем	82	6.81
выращивать	862	6.59
отваривать	169	6.55
тушить	262	6.47
кушать	619	6.35
потушить	171	6.33
стричь	158	6.32
приготавливать	1107	6.21
сажать	338	6.13
отварить	56	5.92
есть	620	5.79
протыкать	63	5.77
срубать	69	5.77
засаливать	40	5.63

pp_obj_с	13612	0.2
пирожок	1918	9.55
пирог	1689	8.25
вареник	232	7.94
кулебяка	78	7.22
сосиска	260	7.01
борщ	157	6.44
рулька	37	6.34
грядка	183	6.3
колбаска	81	6.13
кадушка	36	6.05
свинина	237	5.76
гусь	162	5.75
винегрет	40	5.59
смешать	60	5.49
блинчик	56	5.48
борщь	18	5.42
смешивать	304	5.39
суп	397	5.39
рульку	18	5.31
зраза	20	5.29
салат	471	5.21

pp_obj_из	12939	0.5
салат	6289	8.95
салатик	147	7.52
солянку	88	7.43
голубец	75	6.77
салянку	42	6.7
шницель	50	6.55
суп	839	6.47
оладья	69	6.4
гарнир	142	6.26
пелюстки	29	6.18
рага	43	6.02
запеканка	75	5.8
котлета	123	5.76
обертывание	55	5.67
солянка	24	5.51
блюдо	1051	5.33
борщ	68	5.25
Суп-пюре	16	5.2
кимчхи	14	5.1
начинка	167	4.91
кочерыжка	15	4.91

pp_с	7994	0.1
расти-	40	7.35
морковка	204	6.96
раст	25	6.26
черносливом	17	6.12
баранина	154	6.11
морковь	454	6.06
сосиска	94	5.67
бекон	48	5.61
копченость	27	5.59
клюква	81	5.44
Картофелем	9	5.18
тмин	48	5.16
курица	215	5.09
горошек	55	4.98
чернослив	41	4.88
кочерыжка	10	4.84
картошка	161	4.81
брусника	41	4.72
яблоко	283	4.67
кальмар	38	4.65
гриб	326	4.56

pp_в	6056	0.0
кляре	191	9.
буртах	12	5.
пароварке	22	5.6
опте	11	5.6
кляр	11	5.
сотейник	22	5.4
парник	41	5.3
пароварку	12	5.2
горшочек	54	5.2
панировка	12	5.
салатник	13	5.1
борщ	51	5.0
чехии	10	4.8
кадка	19	4.
диетологии	31	4.7
сухарь	38	4.7
кастрюля	151	4.
кочан	17	4.4
голубец	10	4.3
рассадник	15	4.3
маринад	24	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 (З) - кофе. Обед (О) - 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. 26 ДЕНЬ 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 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. 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

- RuWac (≈ 2 млрд слововхождений): 33589
- ruTenTen (≈ 16 млрд слововхождений): 374907
- НКРЯ (≈ 230 млн слововхождений): 4711

- НКРЯ на порядок меньше RuWac,
  а RuWac 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 слововхождений 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 слововхождений)

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

- Автоматические 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 ⇒ ≈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
  - кафедра компьютерной лингвистики МФТИ
  - ABBYY
  - 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 ⇒ *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
Totals	**1,574**	**0.418**	**7,483**	**1.986**

поребрик в Башкортостане

- Этот шорт лист передается на оценку жюри, в состав которого войдут профессиональные фотографы, деятели искусств, коренные петербуржцы. <…> Победители получат 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 рандеву 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 принеудачном съезде с **поребрика**
- зы у нас в Нске " бордюр "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
Totals	462	3.101

- 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. Пролегомены 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.