Корпус как язык: от масштабируемости к дифференциальной полноте

В.И. Беликов, Н.Ю. Копылов, А.Ч. Пиперски,
В.П. Селегей, С.А. Шаров

31 мая 2013

Проект Генерального Интернет-корпуса Русского Языка (ГИКРЯ)

- ГИКРЯ – проект создания корпуса для целей дифференциальной лингвистики и лексикографии, объемом ок. 50 миллиардов словоупотреблений (посредине между НКРЯ и Рунетом).

Проект Генерального Интернет-корпуса Русского Языка (ГИКРЯ)

- ГИКРЯ – проект создания корпуса для целей дифференциальной лингвистики и лексикографии, объемом ок. 50 миллиардов словоупотреблений (посредине между НКРЯ и Рунетом).
- В ГИКРЯ будут представлены все существенные социальные, жанровые, тематические сегменты Интернета. Одна из основных целей проекта – разработка соответствующих методов классификации

Проект Генерального Интернет-корпуса Русского Языка (ГИКРЯ)

- ГИКРЯ – проект создания корпуса для целей дифференциальной лингвистики и лексикографии, объемом ок. 50 миллиардов словоупотреблений (посредине между НКРЯ и Рунетом).
- В ГИКРЯ будут представлены все существенные социальные, жанровые, тематические сегменты Интернета. Одна из основных целей проекта – разработка соответствующих методов классификации
- Совместный проект Института Лингвистики РГГУ, МФТИ, ABBYY, университета Лидса.

Проект Генерального Интернет-корпуса Русского Языка (ГИКРЯ)

- ГИКРЯ – проект создания корпуса для целей дифференциальной лингвистики и лексикографии, объемом ок. 50 миллиардов словоупотреблений (посредине между НКРЯ и Рунетом).
- В ГИКРЯ будут представлены все существенные социальные, жанровые, тематические сегменты Интернета. Одна из основных целей проекта – разработка соответствующих методов классификации
- Совместный проект Института Лингвистики РГГУ, МФТИ, ABBYY, университета Лидса.
- Корпус будет размещен на сайте `www.webcorpora.ru`

Проект Генерального Интернет-корпуса Русского Языка (ГИКРЯ)

- ГИКРЯ – проект создания корпуса для целей дифференциальной лингвистики и лексикографии, объемом ок. 50 миллиардов словоупотреблений (посредине между НКРЯ и Рунетом).
- В ГИКРЯ будут представлены все существенные социальные, жанровые, тематические сегменты Интернета. Одна из основных целей проекта – разработка соответствующих методов классификации
- Совместный проект Института Лингвистики РГГУ, МФТИ, ABBYY, университета Лидса.
- Корпус будет размещен на сайте `www.webcorpora.ru`
- **Пролегомены к проекту Генерального интернет-корпуса русского языка** (ГИКРЯ) В.И. Беликов, В.П. Селегей, С.А. Шаров (РГГУ) Диалог 2012

Корпуса, корпусометрия и методики корпусных исследований

- Современная лингвистика является преимущественно корпусной: большинство лингвистов, независимо от области своих исследований, рассматривают корпуса устной или письменной речи одновременно и как объект, и как инструмент анализа.

Корпуса, корпусометрия и методики корпусных исследований

- Современная лингвистика является преимущественно корпусной: большинство лингвистов, независимо от области своих исследований, рассматривают корпуса устной или письменной речи одновременно и как объект, и как инструмент анализа.
- Ключевая роль корпуса находится сегодня в разительном контрасте с экспериментальной наивностью большинства исследователей: абсолютным доверием к цифрам, полученным неизвестным способом на неизвестном по структуре корпусе текстов (особенно заметно при использовании в качестве корпуса Интернета).

Корпуса, корпусометрия и методики корпусных исследований

- Современная лингвистика является преимущественно корпусной: большинство лингвистов, независимо от области своих исследований, рассматривают корпуса устной или письменной речи одновременно и как объект, и как инструмент анализа.
- Ключевая роль корпуса находится сегодня в разительном контрасте с экспериментальной наивностью большинства исследователей: абсолютным доверием к цифрам, полученным неизвестным способом на неизвестном по структуре корпусе текстов (особенно заметно при использовании в качестве корпуса Интернета).
- Объективные причины методических просчетов: отсутствие адекватных корпусных инструментов и/или изъяны в работе этих инструментов (что, впрочем исследователю хорошо бы осознавать).

Об объекте наблюдения в корпусном эксперименте

- Основным вопросом всякого корпусного исследования, будь то эксперименты с Интернетом, работа с НКРЯ или иным корпусом, должен быть вопрос об объекте наблюдения: изучается конкретный корпус, поисковая машина или собственно язык?

Об объекте наблюдения в корпусном эксперименте

- Основным вопросом всякого корпусного исследования, будь то эксперименты с Интернетом, работа с НКРЯ или иным корпусом, должен быть вопрос об объекте наблюдения: изучается конкретный корпус, поисковая машина или собственно язык?
- К сожалению, почти всегда исследователь принимает в качестве не требующего доказательства предположения «масштабируемость» результатов частного корпусного исследования на весь язык.

О языковых данных, средних по корпусу

- Резюме исследования (типичный пример из портфеля «Диалога») :

О языковых данных, средних по корпусу

- Резюме исследования (типичный пример из портфеля «Диалога») :
  - Материалом для исследования стали данные Национального корпуса русского языка / (вар.) данные, полученные из Интернета с помощью Яндекса. На момент написания работы количество вхождений для каждой конструкции составило: *конструкция A* – 150 вхождений; **конструкция B** – 650 вхождений; *конструкция C* – 350 вхождения. Таким образом, можно говорить о. . .

О языковых данных, средних по корпусу

- Резюме исследования (типичный пример из портфеля «Диалога») :
  - Материалом для исследования стали данные Национального корпуса русского языка / (вар.) данные, полученные из Интернета с помощью Яндекса. На момент написания работы количество вхождений для каждой конструкции составило: *конструкция A* – 150 вхождений; **конструкция B** – 650 вхождений; *конструкция C* – 350 вхождения. Таким образом, можно говорить о. . .
- Вопросы, оставшиеся без внимания:

  сколько вхождений, сколько документов, сколько авторов; с какой временной динамикой,

  с каким распределением по параметрам метатекстовой разметки (социолингвистической, региональной, жанровой)?

  нет ли в выдаче дублетов или результатов действия иных систематических факторов, "накручивающих" счетчик.

  Насколько объем данных в корпусе достаточен для сделанных выводов?

Основные методические проблемы (доверяй, но проверяй!)

- Некритичное использование для лингвистического анализа инструментов, для этого не предназначенных, прежде всего – поисковиков интернета (Google, Яндекс). Использование недокументированных возможностей таких систем (с неясными принципами работы и непредсказуемой надежностью/устойчивостью).

Основные методические проблемы (доверяй, но проверяй!)

- Некритичное использование для лингвистического анализа инструментов, для этого не предназначенных, прежде всего – поисковиков интернета (Google, Яндекс). Использование недокументированных возможностей таких систем (с неясными принципами работы и непредсказуемой надежностью/устойчивостью).
- Отождествление корпуса с языком. Некорректная интерпретация отрицательных результатов.

Основные методические проблемы (доверяй, но проверяй!)

- Некритичное использование для лингвистического анализа инструментов, для этого не предназначенных, прежде всего – поисковиков интернета (Google, Яндекс). Использование недокументированных возможностей таких систем (с неясными принципами работы и непредсказуемой надежностью/устойчивостью).
- Отождествление корпуса с языком. Некорректная интерпретация отрицательных результатов.
- Отсутствие анализа полученных результатов с точки зрения их достоверности (“область допустимых значений”).

Основные методические проблемы (доверяй, но проверяй!)

- Некритичное использование для лингвистического анализа инструментов, для этого не предназначенных, прежде всего – поисковиков интернета (Google, Яндекс). Использование недокументированных возможностей таких систем (с неясными принципами работы и непредсказуемой надежностью/устойчивостью).
- Отождествление корпуса с языком. Некорректная интерпретация отрицательных результатов.
- Отсутствие анализа полученных результатов с точки зрения их достоверности (“область допустимых значений”).
- Игнорирование дифференциальных свойств языка.

Об устойчивости результатов поиска в Интернете

	"на украину"	"в украину"	"украину"
Поиск от 12.08.2011 в Угловке Новгородской обл.			
без ограничения региона	**310 млн**	**321 млн**	136 млн
Поиск от 14.03.2013 в Петербурге			
без ограничения региона	**138 тыс.**	196 тыс.	3 млн
в Санкт-Петербурге	951 тыс.	2 млн	2 млн
Поиск от 15.03.2013 в Москве			
без ограничения региона	4 млн	**14 млн**	5 млн
в Москве	3 млн	6 млн	69 млн

Что содержит корпус X?

Корпусом X языка Y может называться собрание текстов с явно указанными принципами отбора объектов X (позволяющими в идеале оценить соответствие замысла и исполнения по некоторым критериям).

Например:

Что содержит корпус X?

Корпусом X языка Y может называться собрание текстов с явно указанными принципами отбора объектов X (позволяющими в идеале оценить соответствие замысла и исполнения по некоторым критериям).

Например:

- корпус детских рассказов о сновидениях;

Что содержит корпус X?

Корпусом X языка Y может называться собрание текстов с явно указанными принципами отбора объектов X (позволяющими в идеале оценить соответствие замысла и исполнения по некоторым критериям).

Например:

- корпус детских рассказов о сновидениях;
- звуковой корпус «Один речевой день».

Что содержит корпус X?

Корпусом X языка Y может называться собрание текстов с явно указанными принципами отбора объектов X (позволяющими в идеале оценить соответствие замысла и исполнения по некоторым критериям).

Например:

- корпус детских рассказов о сновидениях;
- звуковой корпус «Один речевой день».
- параллельный русско-немецкий корпус текстов переводов романа «Идиот»;

Что содержит корпус X?

Корпусом X языка Y может называться собрание текстов с явно указанными принципами отбора объектов X (позволяющими в идеале оценить соответствие замысла и исполнения по некоторым критериям).

Например:

- корпус детских рассказов о сновидениях;
- звуковой корпус «Один речевой день».
- параллельный русско-немецкий корпус текстов переводов романа «Идиот»;
- корпус региональных СМИ России (Интегрум, Медиалогия);

Что содержит корпус X?

Корпусом X языка Y может называться собрание текстов с явно указанными принципами отбора объектов X (позволяющими в идеале оценить соответствие замысла и исполнения по некоторым критериям).

Например:

- корпус детских рассказов о сновидениях;
- звуковой корпус «Один речевой день».
- параллельный русско-немецкий корпус текстов переводов романа «Идиот»;
- корпус региональных СМИ России (Интегрум, Медиалогия);
- Параллельный корпус документов Европарламента Europarl (на всех языках Евросоюза);

Что содержит корпус X?

Корпусом X языка Y может называться собрание текстов с явно указанными принципами отбора объектов X (позволяющими в идеале оценить соответствие замысла и исполнения по некоторым критериям).

Например:

- корпус детских рассказов о сновидениях;
- звуковой корпус «Один речевой день».
- параллельный русско-немецкий корпус текстов переводов романа «Идиот»;
- корпус региональных СМИ России (Интегрум, Медиалогия);
- Параллельный корпус документов Европарламента Europarl (на всех языках Евросоюза);
- корпус текстовых расшифровок переговоров шоферов-дальнобойщиков на трассе Москва-Ростов летом 2011 г.

Что содержит корпус X?

Корпусом X языка Y может называться собрание текстов с явно указанными принципами отбора объектов X (позволяющими в идеале оценить соответствие замысла и исполнения по некоторым критериям).

Например:

- корпус детских рассказов о сновидениях;
- звуковой корпус «Один речевой день».
- параллельный русско-немецкий корпус текстов переводов романа «Идиот»;
- корпус региональных СМИ России (Интегрум, Медиалогия);
- Параллельный корпус документов Европарламента Europarl (на всех языках Евросоюза);
- корпус текстовых расшифровок переговоров шоферов-дальнобойщиков на трассе Москва-Ростов летом 2011 г.
- Русский Национальный Корпус (что это значит?)

Корпус как язык

- Предполагается возможность создания универсальных корпусов языка L, которые содержали бы языковой материал, адекватный (по замыслу создателей) для любых исследовательских задач.

Корпус как язык

- Предполагается возможность создания универсальных корпусов языка L, которые содержали бы языковой материал, адекватный (по замыслу создателей) для любых исследовательских задач.
- «Национальный корпус Русского Языка представляет данный язык на определенном этапе (или этапах) его существования *и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т.п.*» (из авторского описания НКРЯ).

Корпус как язык

- Предполагается возможность создания универсальных корпусов языка L, которые содержали бы языковой материал, адекватный (по замыслу создателей) для любых исследовательских задач.
- «Национальный корпус Русского Языка представляет данный язык на определенном этапе (или этапах) его существования *и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т.п.*» (из авторского описания НКРЯ).
- Гипотеза: Национальный корпус языка L и есть универсальный корпус языка L.

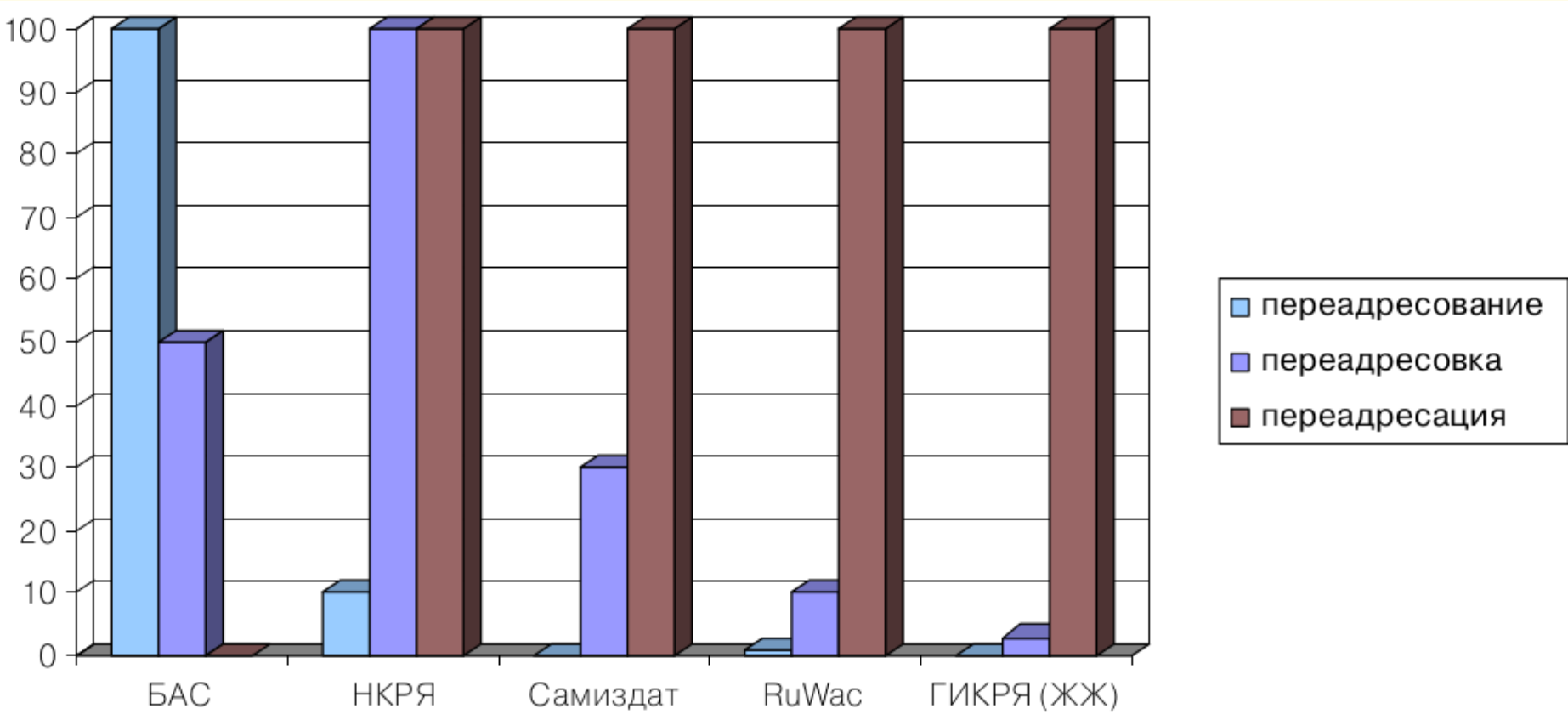
Корпус как язык

- Предполагается возможность создания универсальных корпусов языка L, которые содержали бы языковой материал, адекватный (по замыслу создателей) для любых исследовательских задач.
- «Национальный корпус Русского Языка представляет данный язык на определенном этапе (или этапах) его существования *и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т.п.*» (из авторского описания НКРЯ).
- Гипотеза: Национальный корпус языка L и есть универсальный корпус языка L.
- Аналогичные гипотезы выдвигаются и относительно «национальных» лексикографических ресурсов. По идее корпусные и словарные результаты должны соответствовать друг другу.

Корпус как язык

- Предполагается возможность создания универсальных корпусов языка L, которые содержали бы языковой материал, адекватный (по замыслу создателей) для любых исследовательских задач.
- «Национальный корпус Русского Языка представляет данный язык на определенном этапе (или этапах) его существования *и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т.п.*» (из авторского описания НКРЯ).
- Гипотеза: Национальный корпус языка L и есть универсальный корпус языка L.
- Аналогичные гипотезы выдвигаются и относительно «национальных» лексикографических ресурсов. По идее корпусные и словарные результаты должны соответствовать друг другу.
- Контрпримеры

Сравнительные частоты употребления



- «Склонение топонимов на *-ово (-ево), -ыно (-ино)*» (одна из предлагаемых на сайте НКРЯ исследовательских тем по изучению стилистической вариативности). В предлагаемом case-study подсчитываются вхождения, а не документы, используемый подкорпус (50 млн слов) составлен так, что редакционная политика отдельных изданий заметно влияет на результаты, которые оказываются неустойчивыми к изменению базы исследования. То есть, решаются не проблемы *русской стилистики*, а проблемы стилистики *конкретного собрания текстов*.

- «Склонение топонимов на *-ово (-ево), -ыно (-ино)*» (одна из предлагаемых на сайте НКРЯ исследовательских тем по изучению стилистической вариативности). В предлагаемом case-study подсчитываются вхождения, а не документы, используемый подкорпус (50 млн слов) составлен так, что редакционная политика отдельных изданий заметно влияет на результаты, которые оказываются неустойчивыми к изменению базы исследования. То есть, решаются не проблемы *русской стилистики*, а проблемы стилистики *конкретного собрания текстов*.
- Даже беглый анализ вариативности в склонении топонимов по данным **блогосферы** показывает, что на нее влияют самые разные факторы. Прежде всего – региональные. Для усмотрения каких-то тенденций и параметров, которые на них влияют, нужен существенно больший материал, чем имеется в НКРЯ.

Корпус и его значимые характеристики

- Сбалансированность и репрезентативность (представительность). Иногда эти понятия рассматриваются как эквивалентные.

Корпус и его значимые характеристики

- Сбалансированность и репрезентативность (представительность). Иногда эти понятия рассматриваются как эквивалентные.
- НКРЯ: «*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***»

Корпус и его значимые характеристики

- Сбалансированность и репрезентативность (представительность). Иногда эти понятия рассматриваются как эквивалентные.
- НКРЯ: «*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***»
- Как доказать, что корпус обладает указанными характеристиками?

Дифференциальная полнота

- В отличие от понятия сбалансированности-репрезентативности, дифференциальная полнота означает не просто предположительную типологическую полноту корпуса в некоей правильной пропорции, но полную метатекстовую разметку и наличие в корпусе статистически значимого объема текстов каждого типа.

Дифференциальная полнота

- В отличие от понятия сбалансированности-репрезентативности, дифференциальная полнота означает не просто предположительную типологическую полноту корпуса в некоей правильной пропорции, но полную метатекстовую разметку и наличие в корпусе статистически значимого объема текстов каждого типа.
- В дифференциально полном корпусе результат обработки любого запроса может быть разложен по типологическим координатам. Вопрос о составе «реального» языка (в процентах для каждого типа признается бессмысленным).

Сколько жанров можно разместить на кончике иглы?

- Знаменитый Брауновский (Brown) корпус – 15 жанров: A) Press: reportage, B) Press: editorial, C)Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure . . .

Сколько жанров можно разместить на кончике иглы?

- Знаменитый Брауновский (Brown) корпус – 15 жанров: A) Press: reportage, B) Press: editorial, C)Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure . . .
- Британский национальный корпус BNC – 70 жанров: (ac.med, ac.tech, non-ac.tech, news. . .), medium (book,periodical, ephemeral, . . .), audience, . . .

Сколько жанров можно разместить на кончике иглы?

- Знаменитый Брауновский (Brown) корпус – 15 жанров: A) Press: reportage, B) Press: editorial, C)Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure . . .
- Британский национальный корпус BNC – 70 жанров: (ac.med, ac.tech, non-ac.tech, news. . .), medium (book,periodical, ephemeral, . . .), audience, . . .
- Каталог британской библиотеки (несколько сотен жанров для художественной литературы); Adventure stories, Detective stories, Picaresque literature, Robinsonades, Sea stories, Spy stories, Thrillers, Allegories, Didactic fiction, Fables, Parables, Alternative histories, Dystopias, Bildungsromane, Arthurian romances, yec. . .

Сколько жанров можно разместить на кончике иглы?

- Знаменитый Брауновский (Brown) корпус – 15 жанров: A) Press: reportage, B) Press: editorial, C)Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure . . .
- Британский национальный корпус BNC – 70 жанров: (ac.med, ac.tech, non-ac.tech, news. . .), medium (book,periodical, ephemeral, . . .), audience, . . .
- Каталог британской библиотеки (несколько сотен жанров для художественной литературы); Adventure stories, Detective stories, Picaresque literature, Robinsonades, Sea stories, Spy stories, Thrillers, Allegories, Didactic fiction, Fables, Parables, Alternative histories, Dystopias, Bildungsromane, Arthurian romances, yec. . .
- 349 жанров, отобранных в исследовании предпочтений пользователей [Crowston, Kwasnik, Rubleske, 2010]

Сколько жанров можно разместить на кончике иглы?

- Знаменитый Брауновский (Brown) корпус – 15 жанров: A) Press: reportage, B) Press: editorial, C)Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure . . .
- Британский национальный корпус BNC – 70 жанров: (ac.med, ac.tech, non-ac.tech, news. . .), medium (book,periodical, ephemeral, . . .), audience, . . .
- Каталог британской библиотеки (несколько сотен жанров для художественной литературы); Adventure stories, Detective stories, Picaresque literature, Robinsonades, Sea stories, Spy stories, Thrillers, Allegories, Didactic fiction, Fables, Parables, Alternative histories, Dystopias, Bildungsromane, Arthurian romances, yec. . .
- 349 жанров, отобранных в исследовании предпочтений пользователей [Crowston, Kwasnik, Rubleske, 2010]
- Классификатор Адамчик [Adamzik (1995)] - около 4 тысяч жанров, являющихся сокращением еще большего

Список из Adamzik (1995)

Abänderungsantrag
Abbestellung
Abbruchgenehmigung
Abdankungserklärung
Abecedarium
Abendblatt
Abendgebet
Abendgespräch
Abendnachrichten
Abendprogramm
Abendzeitung
Abenteuerroman
Aberkennung
Abfahrtsplan
Abfindungserklärung
Abgabebewilligung
Abgabeordnung
Abgangsmeldung
Abgangszeugnis
Abgeordnetenrede
Abgesang [im Meistersang]
Abhandlung

Abrüstungsverhandlungen
Absage
Absatz
Absatzgarantie
Abschiedsbrief
Abschiedsgespräch
Abschiedsrede
Abschilderung
Abschlußarbeit
Abschlußbesprechung
Abschlußbilanz
Abschlußgespräch
Abschlußrechnung
Abschlußzeugnis
Abschnitt
Abschrift
Abschußliste
Abschußplan
Abschwörungsformel
Absichtserklärung
Absolutorium [Reifezeugnis; österr.: Bestätigung einer Hochschule über erbrachte Leistungen]

Adversaria [vor Augen liegende Kladde mit ungeordneten Konzepten, Notizen]
Agenda [Notizbuch]
Agende [Kirch]
Agentenroman
Agenturbericht
Agenturmeldung
Agitpropstück
Ahnenprobe
Ahnentafel
Akkordzettel
Akkreditiv [Beglaubigungsschreiben eines Diplomaten]
Akquisitionsliste [Anschaffungsliste]
Akte
Aktenband
Aktenfaszikel
Aktenheft
Aktennotiz
Aktenstück
Aktenvermerk

публицистические

- анкета
- беседа
- доклад
- заявление
- интервью
- комментарий
- лекция
- листовка
- монография
- некролог
- обращение
- отзыв
- очерк
- памфлет
- письмо
- письмо открытое
- поздравление
- предисловие, послесловие
- рецензия
- речь
- совет
- статья
- фельетон
- эссе

мемуарно-биографические

- автобиография
- биография
- дневник, записные книжки
- мемуары

информационные

- аннотация
- анонс
- заметка
- информационное сообщение
- календарь
- обзор
- объявление
- отчет
- репортаж
- хроника

прочие

- гороскоп
- задача
- игра
- инструкция
- миниатюра
- путеводитель
- рецепт
- тест
- характеристика
- цикл

ПРОИЗВОДСТВЕННО-ТЕХНИЧЕСКИЕ

- инструкция
- описание
- паспорт технический
- правила

научные

- аннотация
- диссертация
- дневник
- доклад
- конспект
- лекция
- монография
- обзор
- отзыв
- отчет
- рецензия
- статья
- тезисы
- трактат

учебные

- диплом
- задача
- игра
- инструкция
- лабораторная работа
- методические материалы
- памятка
- правила
- словарь
- справочник
- сочинение
- учебник
- учебное пособие
- хрестоматия

деловые документы

- автобиография
- акт
- анкета
- аттестат
- диплом
- доверенность
- договор
- доклад
- донесение
- записка докладная
- записка служебная
- заявка
- заявление
- инструкция
- контракт
- письмо деловое
- поручение
- предложение
- протокол
- прошение
- расписка
- резюме
- рекомендация
- соглашение
- справка
- телеграмма
- уведомление
- удостоверение
- характеристика

законодательные

- закон
- кодекс
- манифест
- положение
- постановление
- резолюция
- указ
- устав

правовые

- правила
- предписание
- приказ
- присяга
- распоряжение

судебные

- жалоба
- заключение
- определение
- постановление
- решение

нотариальные

- завещание
- свидетельство

дипломатические

- декларация
- документ итоговый
- коммюнике
- нота
- пакт
- протокол дипломатический
- хартия

- билет
- дневник, записные кни
- записка
- объявлен
- открытка
- письмо л

РЕКЛАМА

- анонс
- буклет
- надпись
- объявлен

ЦЕРКОВНО-БОГОСЛОВ

- беседа
- житие
- катехизис
- молитва
- отечник
- описание паломничест
- послание
- поучение
- проповед
- Священно
- слово

ЭЛЕКТРОН КОММУНИК

- блог
- конферен
- смс-сооб

Метатекстовая разметка

- принципальная неоднородность языковых данных

Метатекстовая разметка

- принципальная неоднородность языковых данных
- явное указание типа данных:

Метатекстовая разметка

- принципальная неоднородность языковых данных
- явное указание типа данных:

a ручная (например, возраст и регионы в ЖЖ)

Метатекстовая разметка

- принципальная неоднородность языковых данных
- явное указание типа данных:

а ручная (например, возраст и регионы в ЖЖ)

б автоматическая (например, непроставленные регионы и жанры)

Метатекстовая разметка

- принципальная неоднородность языковых данных
- явное указание типа данных:

а ручная (например, возраст и регионы в ЖЖ)

б автоматическая (например, непроставленные регионы и жанры)

- Надежность ручной классификации

Метатекстовая разметка

- принципальная неоднородность языковых данных
- явное указание типа данных:

а ручная (например, возраст и регионы в ЖЖ)

б автоматическая (например, непроставленные регионы и жанры)

- Надежность ручной классификации
- Точность автоматической классификации

Метатекстовая разметка

- принципальная неоднородность языковых данных
- явное указание типа данных:

а ручная (например, возраст и регионы в ЖЖ)

б автоматическая (например, непроставленные регионы и жанры)

- Надежность ручной классификации
- Точность автоматической классификации
- полнота представлена как относительное понятие: один корпус полнее другого в рамках того или иного языкового явления.

Параметры оценки

- До какой степени текст стремится побудить читателя поддержать какую-либо точку зрения (или отказаться от существующей)?
- До какой степени текст отражает точку зрения организации?
- До какой степени текст выражает чувства или эмоции автора?
- До какой степени текст является говорит о вымышленных персонажах/реалиях?
- До какой степени текст предназначен для развлечения читателя?
- До какой степени текст написан в неформальном стиле, например, с использованием просторечий или сленга?
- До какой степени текст стремится обучить пользователя делать что-либо?
- До какой степени текст похож на информационное сообщение в виде, который может появиться в газете? (например, перепост)
- До какой степени текст носит юридический характер?

Как считать значимость

- признаки и их доверительные интервалы

Как считать значимость

- признаки и их доверительные интервалы

  1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7

  $m_1 = 3.125, m_2 = 4.5$

Как считать значимость

- признаки и их доверительные интервалы

  1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7

  $m_1 = 3.125, m_2 = 4.5$

нет $p = 0.1315$ (Т-критерий)

Как считать значимость

- признаки и их доверительные интервалы

  1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7

  $m_1 = 3.125, m_2 = 4.5$

нет $p = 0.1315$ (Т-критерий)

  1, 2, 2, 3, 3, 4, 5, 5, 1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7, 2, 3, 3, 4, 4, 6, 7, 7

Как считать значимость

- признаки и их доверительные интервалы

  1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7

  $m_1 = 3.125, m_2 = 4.5$

нет $p = 0.1315$ (Т-критерий)

  1, 2, 2, 3, 3, 4, 5, 5, 1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7, 2, 3, 3, 4, 4, 6, 7, 7

да 16 наблюдений$\rightarrow p = 0.02573$

Как считать значимость

- признаки и их доверительные интервалы

  1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7

  $m_1 = 3.125, m_2 = 4.5$

нет $p = 0.1315$ (Т-критерий)

  1, 2, 2, 3, 3, 4, 5, 5, 1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7, 2, 3, 3, 4, 4, 6, 7, 7

да 16 наблюдений$\rightarrow p = 0.02573$

  1, 2, 2, 3, 3, 4, 5, 5, 1, 2, 2, 3, 3, 4, 5, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7, 2, 3, 3, 4, 4, 6, 7, 7, 1000

Как считать значимость

- признаки и их доверительные интервалы

  1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7

  $m_1 = 3.125, m_2 = 4.5$

нет $p = 0.1315$ (Т-критерий)

  1, 2, 2, 3, 3, 4, 5, 5, 1, 2, 2, 3, 3, 4, 5, 5

  2, 3, 3, 4, 4, 6, 7, 7, 2, 3, 3, 4, 4, 6, 7, 7

да 16 наблюдений$\rightarrow p = 0.02573$

  1, 2, 2, 3, 3, 4, 5, 5, 1, 2, 2, 3, 3, 4, 5, 5, 5

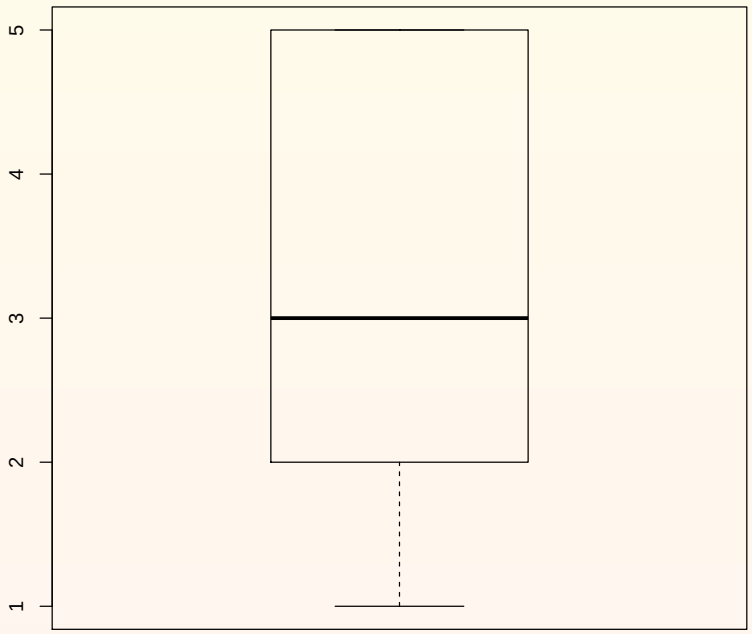
  2, 3, 3, 4, 4, 6, 7, 7, 2, 3, 3, 4, 4, 6, 7, 7, 1000

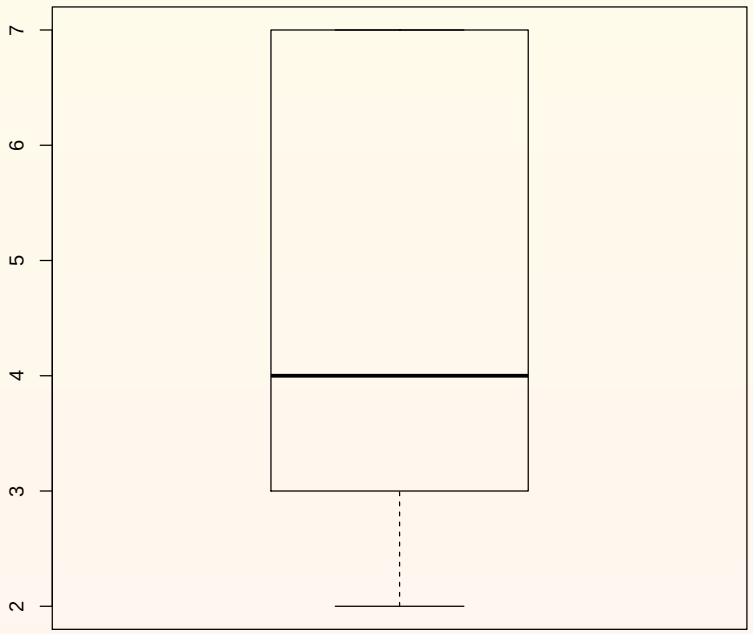
нет $p = 0.3222(\sigma = 241.45)$

Сравнение двух данных

1, 2, 2, 3, 3, 4, 5, 5, 1, 2, 2, 3, 3, 4, 5, 5

2, 3, 3, 4, 4, 6, 7, 7, 2, 3, 3, 4, 4, 6, 7, 7





Значимые различия между сегментами

1. выбор языкового явления, которое измеряется в числовом выражении;

Значимые различия между сегментами

1. выбор языкового явления, которое измеряется в числовом выражении;
2. значение этого явления в рамках отдельных текстов каждого из сегментов (например, частота или частотный ранг);

Значимые различия между сегментами

1. выбор языкового явления, которое измеряется в числовом выражении;
2. значение этого явления в рамках отдельных текстов каждого из сегментов (например, частота или частотный ранг);
3. значение этого явления в сегменте в целом.

Значимые различия между сегментами

1. выбор языкового явления, которое измеряется в числовом выражении;
2. значение этого явления в рамках отдельных текстов каждого из сегментов (например, частота или частотный ранг);
3. значение этого явления в сегменте в целом.

	Сегмент	Корпус	IPM
Частоста	a	b	$r = \frac{b}{d}$
Размер (слов)	c	d	

Значимые различия между сегментами

1. выбор языкового явления, которое измеряется в числовом выражении;
2. значение этого явления в рамках отдельных текстов каждого из сегментов (например, частота или частотный ранг);
3. значение этого явления в сегменте в целом.

	Сегмент	Корпус	IPM
Частоста	a	b	$r = \frac{b}{d}$
Размер (слов)	c	d	

4. $G^2 = a * ln(\frac{a*r}{c}) + b * ln(\frac{b*r}{d})$

Значимые различия между сегментами

1. выбор языкового явления, которое измеряется в числовом выражении;
2. значение этого явления в рамках отдельных текстов каждого из сегментов (например, частота или частотный ранг);
3. значение этого явления в сегменте в целом.

	Сегмент	Корпус	IPM
Частоста	a	b	$r = \frac{b}{d}$
Размер (слов)	c	d	

4. $G^2 = a * ln(\frac{a*r}{c}) + b * ln(\frac{b*r}{d})$

   Kilgarriff, 2005: Language is never, ever, ever, random (*CLLT*)

   `http://kilgarriff.co.uk/Publications/2005-K-lineer.pdf`

Хабарики города Питера

- Нью Йорк похож на китайский провинциальный город. Может быть, вне Манхеттена и по-другому все, но тут ужас. Грязюка, повсюду хабарики, толпы людей, временами воняет.

Хабарики города Питера

- Нью Йорк похож на китайский провинциальный город. Может быть, вне Манхеттена и по-другому все, но тут ужас. Грязюка, повсюду хабарики, толпы людей, временами воняет.
- Как же меня бесит истеричка с первого этажа. Хабарики кидают видите ли. Показать ей чтоли мою банку с коллекцией фильтров от парламента

Хабарики города Питера

- Нью Йорк похож на китайский провинциальный город. Может быть, вне Манхеттена и по-другому все, но тут ужас. Грязюка, повсюду хабарики, толпы людей, временами воняет.
- Как же меня бесит истеричка с первого этажа. Хабарики кидают видите ли. Показать ей чтоли мою банку с коллекцией фильтров от парламента

	Питер	Корпус	G^2	
хабарик	8	10	24.035	$p \leq 0.001$

Хабарики города Питера

- Нью Йорк похож на китайский провинциальный город. Может быть, вне Манхеттена и по-другому все, но тут ужас. Грязюка, повсюду хабарики, толпы людей, временами воняет.
- Как же меня бесит истеричка с первого этажа. Хабарики кидают видите ли. Показать ей чтоли мою банку с коллекцией фильтров от парламента

	Питер	Корпус	G^2	
хабарик	8	10	24.035	$p \leq 0.001$

	ГИКРЯ	Yandex	Google
хабарик	10	883	240
чибон	2	13	2

Хабарики города Питера

- Нью Йорк похож на китайский провинциальный город. Может быть, вне Манхеттена и по-другому все, но тут ужас. Грязюка, повсюду хабарики, толпы людей, временами воняет.
- Как же меня бесит истеричка с первого этажа. Хабарики кидают видите ли. Показать ей чтоли мою банку с коллекцией фильтров от парламента

	Питер	Корпус	G^2	
хабарик	8	10	24.035	$p \leq 0.001$

	ГИКРЯ	Yandex	Google
хабарик	10	883	240
чибон	2	13	2

- Текущий ГИКРЯ: 950 млрд слов, 10 млн текстов

Название	Описание	Размер (Количество слов)	Ссылка
☐ RUWAC-PARSED	RUWAC Русский интернет корпус	2 миллиарда	
☐ RUSSM	Журнальный Зал	250 миллионов	
☐ ZHZHALL	Живой Журнал	950 миллионов	
☑ bashkiriya	ЖЖ-Башкирия	9,7 миллиона	
☑ chelyabinskaya	ЖЖ-Челябинская обл.	8,3 миллиона	
☑ donetskaya	ЖЖ-Донецкая обл.	11 миллионов	
☑ kiev	ЖЖ-Киевская	22 миллионов	
☑ krasnodarskiy	ЖЖ-Краснодарский край	8,5 миллиона	
☑ krasnoyarskiy	ЖЖ-Красноярский край	8,1 миллиона	
☑ mosobl	ЖЖ-Московская обл.	19 миллионов	
☑ novosibirskaya	ЖЖ-Новосибирская	13 миллионов	
☑ omskaya	ЖЖ-Омская обл.	5 миллионов	
☑ permskiy	ЖЖ-Пермский край	9,9 миллиона	
☑ petersburg	ЖЖ-Петербург	45 миллионов	
☑ rostovskaya	ЖЖ-Ростовская	10 миллионов	
☑ samarskaya	ЖЖ-Самарская обл.	14 миллионов	
☑ saratovskaya	ЖЖ-Саратовская обл.	5,3 миллиона	
☑ sverdlovskaya	ЖЖ-Свердловская обл.	16 миллионов	
☑ tatarstan	ЖЖ-Татарстан	6,5 миллиона	

Далее >

Мы рекомендуем браузер Chrome(12.0+), Firefox(3.6.8+) или Internet Explorer(8.0.6+)

Работы выполняются при финансовой поддержке Министерства образования РФ

В/на Украину

	"на украину"	"в украину"	"украину"
Поиск от 12.08.2011 в Угловке Новгородской обл.			
без ограничения региона	**310 млн**	**321 млн**	136 млн
Поиск от 14.03.2013 в Петербурге			
без ограничения региона	**138 тыс.**	196 тыс.	3 млн
в Санкт-Петербурге	951 тыс.	2 млн	2 млн
Поиск от 15.03.2013 в Москве			
без ограничения региона	4 млн	**14 млн**	5 млн
в Москве	3 млн	6 млн	69 млн

Корпус	в Украину		на Украину	
	Экземпляров	IPM	Экземпляров	IPM
BASHKIRIYA	10	1.029	13	1.338
PETERSBURG	29	0.639	61	1.345
SVERDLOVSKAYA	11	0.672	26	1.589
KRASNODARSKIY	12	1.403	15	1.754
KRASNOYARSKIY	10	1.230	23	2.829
NOVOSIBIRSKAYA	5	0.379	22	1.669
OMSKAYA	4	0.799	4	0.799
PERMSKIY	5	0.504	21	2.115
ROSTOVSKAYA	9	0.885	16	1.573
SAMARSKAYA	7	0.487	26	1.809
SARATOVSKAYA	1	0.186	6	1.118
TATARSTAN	3	0.455	6	0.910
KIEV	356	16.111	155	7.015
DONETSKAYA	54	4.832	98	8.769
CHELYABINSKAYA	9	1.087	8	0.966
Totals	525	2.702	500	2.573

Корпус	ставить% .. укол		сделать% .. укол		поставить% .. укол		делать% .. укол	
	Экземпляров	IPM	Экземпляров	IPM	Экземпляров	IPM	Экземпляров	IPM
RUSSM	10	0.038	303	1.140	20	0.075	97	0.365
BASHKIRIYA	0	0.000	5	0.515	1	0.103	2	0.206
CHELYABINSKAYA	2	0.242	1	0.121	2	0.242	0	0.000
DONETSKAYA	0	0.000	4	0.358	0	0.000	2	0.179
KIEV	0	0.000	5	0.226	0	0.000	6	0.272
KRASNODARSKIY	0	0.000	4	0.468	1	0.117	2	0.234
KRASNOYARSKIY	1	0.123	4	0.492	2	0.246	1	0.123
NOVOSIBIRSKAYA	1	0.076	10	0.759	6	0.455	1	0.076
OMSKAYA	1	0.200	0	0.000	1	0.200	1	0.200
PERMSKIY	0	0.000	3	0.302	0	0.000	2	0.201
PETERSBURG	0	0.000	15	0.331	5	0.110	9	0.198
ROSTOVSKAYA	0	0.000	2	0.197	0	0.000	1	0.098
SAMARSKAYA	0	0.000	12	0.835	0	0.000	0	0.000
SARATOVSKAYA	0	0.000	1	0.186	0	0.000	0	0.000
SVERDLOVSKAYA	1	0.061	7	0.428	6	0.367	1	0.061
TATARSTAN	0	0.000	4	0.607	1	0.152	0	0.000
Totals	16	0.035	380	0.826	45	0.098	125	0.272

RUSSM
BASHKIRIYA
CHELYABINSKAYA
DONETSKAYA
KIEV
KRASNODARSKIY
KRASNOYARSKIY
NOVOSIBIRSKAYA
OMSKAYA
PERMSKIY
PETERSBURG
ROSTOVSKAYA
SAMARSKAYA
SARATOVSKAYA
SVERDLOVSKAYA
TATARSTAN
Totals

rison of frequency of c

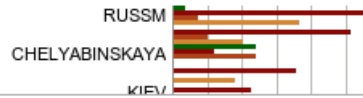
Deutsch-russische Idiome | Sketch Engine: Corpora | yarn_org | Google Groups | ASSIST: Automated Semar | https://www.aclweb.org/a | IntelliText Corpus Queries

int.webcorpora.ru/itweb/htdocs/Query.html#

Главная страница | Выбор языка | Выбор корпуса | Настройка поиска | Результаты | Построить новый корпус

Frequency Comparison Table
Collocation Table
Empty
Empty
Affix Main Table
Affix Forms Table
Lists Most Common Table
Multivariate Data
Genres Data
Parallel Table

текст	слева	совпадение	справа
3018	словно	поставил укол или взял	пробу



13766

id="13766" url=http://magazines.russ.ru/zz/2009/18/bu7.html
, дружили десятилетими – как они там сейчас, когда российские самолеты бомбят Гори и другие города … Лицо у Сусанны в это время изменилось, как у алеутского шамана, – без всякой косметики. Только что цвело – и вдруг стало будто из глубины океана всплыло что-то не наше … Я вспомнила, что в детстве она играла с сестрой в больницу и хотела циркулем из готовальни отца **поставить укол младшей сестре** … благо та решила " спросить у мамочки " … – Слушайте, зачем вы жалеете грузин! Это ужасные люди! Помните: они торговали фруктами на рынке – на нас наживались?! Потом шили подпольно джинсы и этим развалили Союз! – Сусанна, неужели за фрукты и джинсы нужно бомбить детей и женщин, стариков и больницы? Ты сама только что говорила: украинцы должны

magazines.russ.ru/zz/2009/18/bu7.html

karlson77 - Profile | Журнальный зал | Заруб

magazines.russ.ru/zz/2009/18/bu7.html

ЖЗ ЖУРНАЛЬНЫЙ ЗАЛ — Русский толстый журнал как эстетический феномен

Русский Журнал — десятые годы

Последнее обновление: 25.05.2013 / 13:49

Все проекты ЖЗ: | Новые поступления | Афиша | Авторы | Обозрения | О проек

Google™ Пользовательский п | Поиск

Опубликовано в журнале:
Зарубежные записки 2009, 18

+1 | Твитнуть 0 | Нравится 0

Нина Горланова, Вячеслав Букур

Урал-Кавказ
Тобаго (гонки крабов)

Два рассказа

версия для печати

УРАЛ-КАВКАЗ

Через неделю после окончания Пятидневной войны пришла Сусанна.

Мы не виделись лет тридцать √ с тех пор, как со своим вторым мужем она уехала в Норильск. Но после перевала жизни ведь все сползаются. Правда, оказалос дворяне, а другие √ в монастырь. Однако в гости к нам все приходят и вино полусладкое приносят, а мы дарим свои книжки.

Из-за большого слоя воли некогда пластичное лицо Сусанны теперь казалось почти мужским. Прорубая воздух прекрасной скалой носа, она подошла к столу и середину два пирога: с брусникой и сёмгой.

√ Тетя навалилась с кулинарным обучением, когда я вернулась в пермское гнездо. √ В груди у нее словно разговаривала посуда из толстого цветного стекла. √ встречу в кассах?

√ Такие незабвенные встречи меняют всю жизнь

√ Слава, больше не пей, а то опять будут белые столбы в глазах.

√ Жена не понимает, что белые столбы √ они потому, что не каждый день выпиваю

Сусанну было не сбить:

Deutsch-russische Idiome | Sketch Engine: Corpora | yarn_org | Google Groups | ASSIST: Automated Semar | https://www.aclweb.org/ | IntelliText Corpus Querie

int.webcorpora.ru/itweb/htdocs/Query.html#

Главная страница | Выбор языка | Выбор корпуса | Настройка поиска | Результаты | Построить новый корпус

Frequency Comparison Table
Collocation Table
Frequency Comparison Chart
Empty
Affix Main Table
Affix Forms Table
Lists Most Common Table
Multivariate Data
Genres Data
Parallel Table

текст	слева	совпадение	справа
>>	выбрасывающего	чибон	из

3539007

id="3539007" url=http://karlson77.livejournal.com/18047.html
да по-больше, по-больше.: -)
Проживание рядом с пожарной частью откладывает, в основном, негативный отпечаток на психику здорового человека, особливо по выходным ..., когда приперся домой в 6 утра, а в 8 пара-тройка лихих парней с молодецким гиканьем(читай сиренами) летят на вызов.
Но седня парни с утра порадовали: походу мусоровоз вывалил в себя тлеющий **чибон**, который возродился в пламя, аккурат в районе ПЧ, водила не дурак - тормознул прям перед выездом - " нате парни, я вам шабашку привез ". Парни бодренько выкатили одну свою огнеборщецкую колесницу и начали выполнять свои прямые обязанности. Однако остальным сотрудникам видать с утреца силдеть было скучно, посему они подтянулись к месту действа ... В итоге картина напоминала организацию труда в нормальной современной

karlson77 - Profile

karlson77.livejournal.com/profile

add friend | add note | view friends feed | track user | send message | send v-gift

karlson77

karlson77's Journal

Plus Account, Created on 10 November 2004 (#5099585), Last updated on 5 January 2013 [Gift]

209,696 place in User ratings

Social capital: less than 10

63 journal entries	4 memories
873 comments posted	31 photos
218 comments received	0 V-Gifts
0 tags	2 userpics

Name: karlson77

Location: Perm, Пермский край, Russian Federation

Website: смежная специализация - фото :)

External Services: karlson77@livejournal.com

Interests: (4) авто, бильярд, **путешествия**, **фото**

Bio

Имя: Евгений
Город: Пермь
Работа: телеком
ну и для начала хваит :) остальное в процессе общения если получиться

Deutsch-russische Idiome | Sketch Engine: Corpora | yarn_org | Google Groups | ASSIST: Automated Semar | https://www.aclweb.org/a | IntelliText Corpus Querie

int.webcorpora.ru/itweb/htdocs/Query.html#

Главная страница | Выбор языка | Выбор корпуса | Настройка поиска | Результаты | Построить новый корпус

Frequency Comparison Table
Collocation Table
Frequency Comparison Chart
Empty
Affix Main Table
Affix Forms Table
Lists Most Common Table
Multivariate Data
Genres Data
Parallel Table

текст	слева	совпадение	справа
>>	с	чибонами	дымит

25431

id="25431" url=http://magazines.russ.ru/urnov/2003/16/rakov.html
, моя родная, не горюй И после водки на воду не дуй. В ударе мы и не боимся спиться, Раз заглянув за раздвижные лица. Пусть этот город высосан до дна, Пусть грош ему, поганому, цена, - Я за одно твоё живое тело Его прощу у лёгкого предела. * * * (1997) Пермь-Первая, Пермь-та-ещё-подруга, Рождаешься и, как **чибон**, - по кругу, По кругу первому, Итаке, ИТК. Не успевает загореться спичка - Проскакиваешь лимб на электричке И начинаешь без черновика. Вокруг тебя родные психопаты Шагают патетически до хаты И за столом решают, кто кого, И вон душа с кишками вперемешку, Вергилий шепчет, чтобы ты не мешкал, А то одолевает естество. Вот Пермь-Вторая подаёт вагоны, Ты вытираешь

magazines.russ.ru/urnov/2003/16/rakov.html

Мужчины и женщины

	жен	муж	жен/муж
"так быстро", с 17.05.2012	942	542	1,73
"так мало", с 15.05.2012	937	623	1,50
"такой маленький", с 6.02.2012	989	677	1,46
"так много", с 26.05.2012	977	746	1,31
"такой большой", с 1.04.2012	874	766	1,14
....			
"заметно меньше", с 1.04.2011	380	940	0,40
"существенно меньше", с 1.08.2011	334	941	0,35
"существенно больше", с 1.07.2011	325	992	0,33
"заметно больше", с 1.03.2011	289	951	0,30