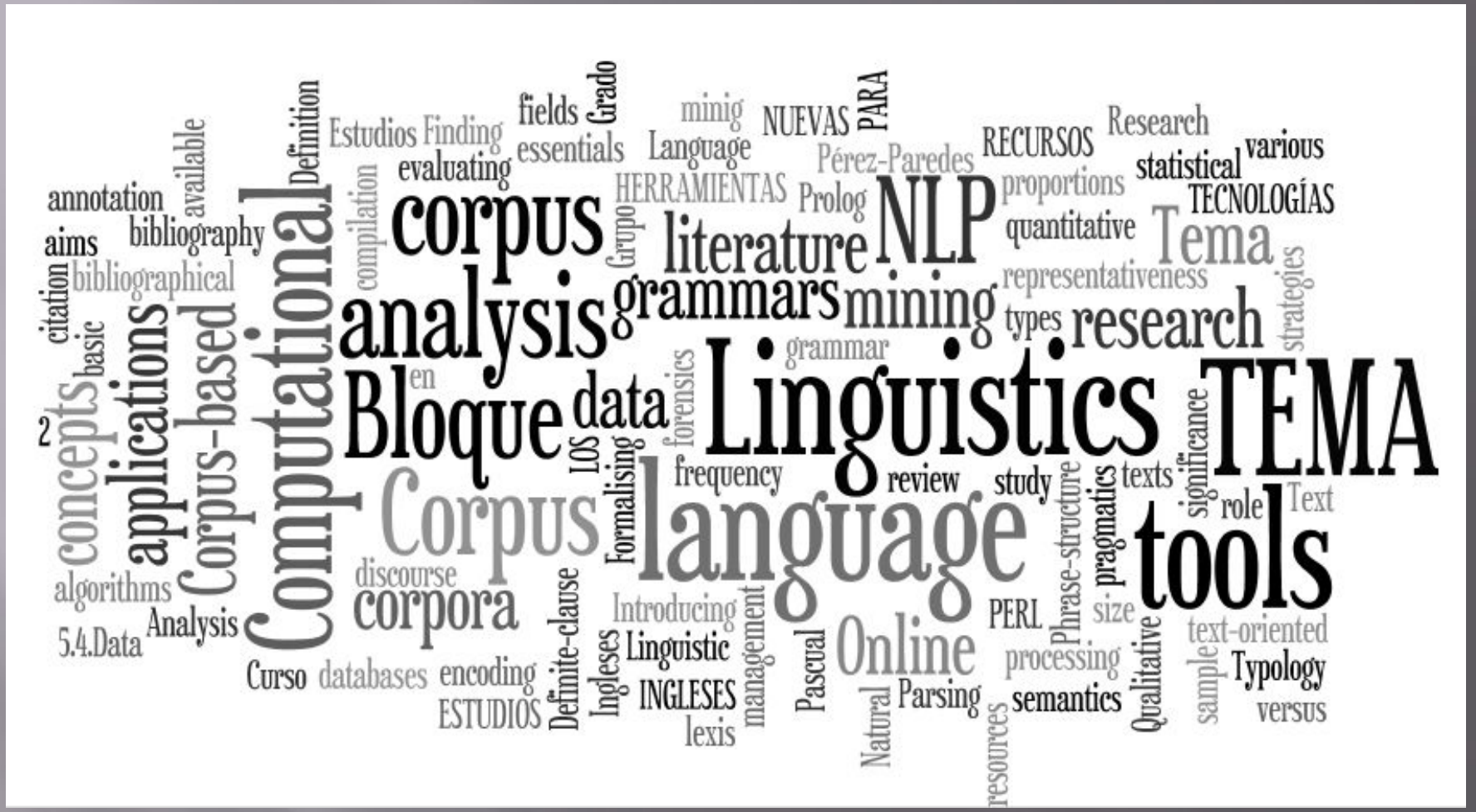


КОРПУСНАЯ ЛИНГВИСТИКА

Е.А. Долуденко

2012 г.

Введение: корпусы и корпусная лингвистика

- ***Корпусная лингвистика*** – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий.

Плунгян Владимир Александрович - член-корреспондент РАН, заведующй отделом корпусной лингвистики Института русского языка им. В.В. Виноградова РАН, профессор МГУ им. М.В.Ломоносова.

Введение: корпусы и корпусная лингвистика

- ***Лингвистический (языковой) корпус текстов*** - большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач.
- ***Корпус-менеджер*** - специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.

Введение: корпусы и корпусная лингвистика

- *Конкорданс* – результат поиска в корпусе - список всех употреблений данного слова в контексте со ссылками на источник

046.	hands. I was afraid to look… but I was afraid not to	LOOK. After a time I took my hands away from my eyes. The tr
047.	hould of said oh I just come up for a few days Blair had to	LOOK after him Joy cos he were bad when he come home like Ah
048.	ht? 23:77 GITA Good. 23:78 SANJAY Gita. 23:79 I need you to	LOOK after the stall for me for a bit. 23:80 [takes off his
049.	idower had, as the judge found, reasonably given up work to	LOOK after his children. It was decided that the damages for
050.	it in a group where we have our own particular prisoners to	LOOK after. So as far as this meeting's concerned, erm, most
051.	ized how tight Mike was going to be with you know having to	LOOK after Paul and all the rest of it I promised I'd take h
052.	licewoman but not a Wren! Yo yo got all to look af Wrens to	LOOK after, if you're a Wren. Why have you not got ? Oh I do
053.	lly with you know, maybe a mother who has a few children to	LOOK after, and she's on her own twenty four hours a day wit
054.	me one of the animals. It was part of Little Jack's work to	LOOK after the dogs. One White House dog was immortalized in
055.	ms and three miles of red carpet. Two men work full-time to	LOOK after the 300 clocks. About 700 people work in the Pala
056.	n, I can see a policewoman but not a Wren! Yo yo got all to	LOOK af Wrens to look after, if you're a Wren. Why have you
057.	o back and to for treatment to Yeah. Denbigh. And I used to	LOOK after her husband while he was, while she was away gett
058.	o worry that they're er gonna be a problem for mum Yeah. to	LOOK after you see, so I think if that's the price we gotta
059.	of having three to look after I'd probably end up having to	LOOK after a seven or eight. Yeah. But you see I, I don't kn
060.	orld, and we are the only species that can choose either to	LOOK after our world or, with the push of a button, destroy
061.	r Council is here for. Your care and our representatives to	LOOK after the interests of the population within your area.
062.	ry privilege that members of parliament have got to have to	LOOK after their own constituencies in addition to being spo
063.	s. WHAT ELSE DO YOU DO TO KEEP IT LOOKING HEALTHY? I try to	LOOK after myself although I'm a bit of a part-timer when it
064.	Salon Collection 11 Gentle Shampoo Fun to wear and great to	LOOK after. This urchin style comes from the Edmonds Salon i
065.	see why I should. Mum's at home all day, and its her job to	LOOK after the house, not mine. What do you think? ? Q: WEIG
066.	t, p793. In that case the husband, who had given up work to	LOOK after his children, sought to recover the cost of a hou
067.	ther way, I expect I shall manage," she said, determined to	LOOK on the bright side. "After all, until Edward marries an

Введение: корпусы и корпусная лингвистика

Целесообразность создания и смысл использования:

- 1) достаточно большой (репрезентативный) объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений;
- 2) данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;
- 3) однажды созданный и подготовленный массив данных может использоваться многократно, многими исследователями и в различных целях.

Введение: корпусы и корпусная лингвистика

Первый лингвистический корпус:

- Год создания: 1963 г.
- Название: Brown Corpus
- Авторы : У. Френсис и Г. Кучера
- Состав: 500 двухтысячесловных прозаических печатных текстов американского варианта английского языка;15 жанров
- Дополнительно: частотный и алфавитно-частотный словарь, разнообразные статистические распределения.

Введение: корпусы и корпусная лингвистика

Самые известные корпусы:

- Ланкастерский корпус английского языка (Lancaster-Oslo-Bergen Corpus, LOB)
- Уппсальский корпус русского языка
- Британский национальный корпус (British National Corpus)
- Международный корпус английского языка (International Corpus of English)
- Лингвистический Банк английского языка (Bank of English) и др.

Введение: корпусы и корпусная лингвистика

Функции корпуса:

- Построение конкордансов (списков всех употреблений данного слова в контексте со ссылками на источник).
- Получение разнообразных справок и статистических данных о языковых и речевых единицах: о частоте словоформ, лексем, грамматических категорий,
- Отслеживание изменений частот и контекстов в различные периоды времени,
- Получение данных о совместной встречаемости лексических единиц .
- Изучение динамики процессов изменения лексического состава языка.
- Анализ лексико-грамматических характеристик в разных жанрах и у разных авторов.
- Подготовка разнообразных исторических и современных словарей . Построение и уточнение грамматик .
- Обучение языку.

Свойства корпуса

- ***Репрезентативность -*** необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п.
- Объем: не менее100 млн словоупотреблений.
- *Разметка* (tagging, annotation) - приписывание текстам и их компонентам определенных сведений (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика),
- *Метаразметка* - приписывание структурных (глава, абзац, предложение, словоформа) и собственно лингвистических сведений, описывающих лексические, грамматические и прочие характеристики элементов текста.

Типы разметки

- *Морфологическая* (part-of-speech tagging или POS-tagging), дословно – частеречная разметка.
- *Синтаксическая* или *парсинг* (англ. parsing), описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции.
- *Семантическая* - по семантическим категориям, к которым относится данное слово или словосочетание, и более узким подкатегориям, специфицирующим его значение.

Типы разметки

- *Анафорическая* - фиксирует референтные связи, например, местоименные;
- *Просодическая*– использует метки, описывающие ударение и интонацию.
- *Дискурсная -* в корпусах устной разговорной речи для обозначения пауз, повторов, оговорок, и т.д.

Технология создания корпусов

1) Определение перечня источников
2) Оцифровка текстов
3) Предобработка текста (филологическая выверка и корректировка; подготовка библиографического и экстралингвистического описания текста)
4) Конвертирование и графематический анализ
5) Разметка текста
6) Корректировка результатов автоматической разметки
7) Конвертирование размеченных текстов в структуру ИПС
8) Обеспечение доступа к корпусу

Корпусные менеджеры

- поиск конкретных словоформ;
- поиск словоформ по леммам;
- поиск группы словоформ в виде разрывной или неразрывной синтагмы;
- поиск словоформ по набору морфологических признаков;
- отображение информации о происхождении, типе текста и т.п.;
- вывод результатов поиска с указанием контекста заданной длины;
- получение различных лексико-грамматических статистических данных;
- сохранение отобранных строк конкорданса в отдельном файле на компьютере пользователя и др.

Пользователи корпусов

- Лингвисты-теоретики используют корпусы в качестве экспериментальной базы для проверки гипотез и доказательства своих теорий.
- Прикладные лингвисты (преподаватели, переводчики и т.п.) используют компьютерные корпусы при обучении языкам и для решения своих профессиональных задач.
- Компьютерные лингвисты пытаются выявить и использовать статистические и лингвистические закономерности для создания компьютерных моделей языка.
- Специалисты по общественным наукам (историки, социологи) - для изучения своих объектов через язык, используя такие параметры текстов, как период, автор или жанр.
- Литературоведы используют корпусы для стилеметрических исследований.
- Корпусы также используются для разработки и настройки различных автоматизированных систем (машинный перевод, распознавание речи, информационный поиск).

Классификация корпусов №1

1. ***по форме хранения:***
   - в звуковой форме;
   - письменные;
   - смешанные;
2. ***по языку представления текстов:***
   - одноязычные;
   - многоязычные;
3. ***по жанровой принадлежности:***
   - литературные;
   - диалектные;
   - разговорные;
   - публицистические;
   - смешанные;

Классификация корпусов №1

4. по способам доступа:

- свободно доступные;
- коммерческие;
- закрытые;

5. по назначению:

- исследовательские;
- иллюстративные;

6. по динамичности:

- динамические (мониторные);
- статические;

7. по наличию дополнительной информации:

- аннотированные (размеченные);
- неразмеченные.

Классификация корпусов №2

1. ***по степени организации и структурированности:***
   – электронный архив – это тексты на электронном носителе, но их форма, представленная на машинном носителе, не стандартизирована и не унифицирована;
   – электронная библиотека – тексты здесь представлены однородным и стандартизированным образом;
   – корпус текстов – форма стандартизирована и унифицирована, тексты предназначены для отражения части лингвистической реальности;
   – субкорпус – это некоторая автономная часть корпуса.

2. ***по хронологическому признаку:***
   – синхронический;
   – мониторный (отслеживает текущее состояние языка);
   – диахронический.

Классификация корпусов №2

3. по индексации:

- простой;
- аннотированный.

4. по языку:

- одноязычный;
- двуязычный;
- многоязычный.

5. по способу применения и использования корпуса:

- исследовательский;
- иллюстративный;
- параллельный.

6. по способу существования корпуса:

- динамический;
- статический.

Пример использования корпуса

Как по-английски правильно сказать

«принять решение» ?

- **to take a decision** или **to make a decision**?

make a decision VS take a decision

Home> Concordancers> English Input [«Back] (Back keeps original settings) Extract-Link to this data >> here

Concordance for *equals* MAKE in Corpus **mega_corpus.txt** with assoc. decision on RIGHT side; sorted 1 wd **left** of key

Dictionary Eng_Eng

Extract >> All
Ø|any|10|20|30|50

>> equal make in All of above (>3m) sort 1 v left +assoc decision on right Hear it in Eng-US

Click any KEYWORD for more context

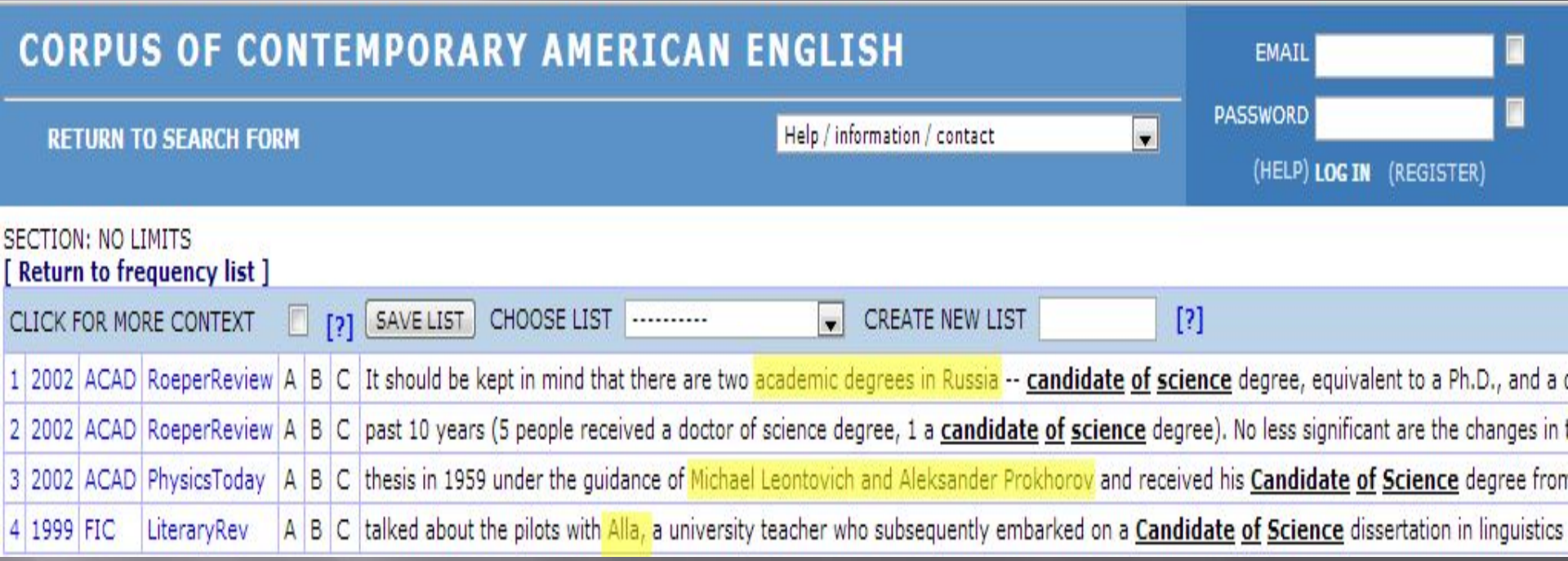
001. ur own good. Are we? Would do that? I, I s I can't, I **can't** MAKE that decision. Oh. I'd have to, I'd have to erm I'd hav
002. d program. The first thirty to sixty days after **individuals** MAKE their decision will determine their interest and partic
003. gold circuit from their homes. All could help the **President** MAKE his decision. The talk would not be in code, but neithe
004. ut about it they could still all three prove. So you **should** MAKE the decision in the first place. If they're abroad I su
005. when I go to the site again, is to look at it and er **simply** MAKE a decision as to whether or not in my opinion the land
006. application. So they'll move from one window to another to MAKE a decision on that information in the other window and
007. charges since this lot came to power, is forcing people to MAKE a decision as to whether they should get a prescription
008. problem arises, if it does arise, when the educator has to MAKE a choice or a decision within the area of his professio
009. 'd also need some kind of refreshments and er we'll have to MAKE a decision about lunch. Can I put in excellent restaura
010. d barley it contained. Arthur Guinness had the foresight to MAKE a business decision which determined the future of the
011. en his mouth, and it took him no more than three seconds to MAKE his decision. For over the years he had received many s
012. how soon after that decision will this committee be able to MAKE a decision on the implementation? Thank you Chairman. W
013. ne piece of information in the application and they need to MAKE a decision based on an information which is elsewhere i
014. perseded by the purposive. On the contrary it is my duty to MAKE my own decision as between the two". Schweitzer seems,
015. bearable and can not brook as leader is those who, after **we** MAKE that decision, will deviate and not comply with our dec
016. See what date see the availability of rooms and then **we'll** MAKE a firm decision. The remit with regard to the er one da
017. AROLE JUDGE #1 Well we have what we need here. 53:505 **We'll** MAKE our decision without the inmate present. 53:506 TUNE: H
018. w really er it's something that we're not the only ones **who** MAKE the decision. While we actually do the a the removal, w
019. part of the debate with the delegates. They, not he, **would** MAKE the decision. "What I find unbearable and can not brook
020. captain. The captain will then inform the umpires who **would** MAKE the decision whether the game should be suspended or wh
021. ant than the er, erm, the Royal Crown and Minton. Would **you** MAKE any conscious decision to move slightly downmarket or l

3 hits Standardized to **1 per million** (hits/corpus size x 1,000,000)

Click any KEYWORD for more context

001. rk noted proceed to proceed by something sinister. It **might** TAKE just a decision for St Albans oldest of the country by
002. would be fair to say we should not, at this moment in **time**, TAKE any decision on on the matter . . Taking votes on the f
003. term action before he will authorise the County Council to TAKE the decision on it. Erm if we having considered all the

candidate of science



http://corpus.byu.edu/coca/

Rector

was bringing his associates here for a specific purpose and RECTOR would not be able to fathom it until they arrived
ovember 1894, Carew H. St. John Mildmay, the Anglo-Catholic RECTOR of St. Mary's church, wrote in a letter to the Me
, and after studying with Fletcher for a time he approached RECTOR, announced that he wanted to be baptized and that
o dampen his spirits, but he was still a little wary around RECTOR for they had not yet discussed the incident. "I t
you think you can manage it"? Mavis smiled. "I'll try". As RECTOR was walking back toward the residential hall, Joh
ell. But I was deeply moved by his letter of resignation as RECTOR of St. Luke's Church in Atlanta. It was the cry o
pirit of a child going on a holiday, nodding attentively as RECTOR gave him his final instructions. He was to get in
mission as if he belonged here and had become a Christian. RECTOR was often curious; often tempted to ask questions
ce while the party was here. When everything had been done, RECTOR went back to his desk to occupy himself with his
e for a position in which he could make the decisions. Even RECTOR himself was prey to this spirit of competition an
, using the stiffness in his arm as an excuse. In any event RECTOR sent him to the local hospital to have it checked
indirection. In all of his experience in the mission field RECTOR had never seen a convert quite like Hino. From th
t the pump cuts out before the water gets too low". "Fine", RECTOR said. He looked out over the expanse of the compo
ith relief and then the telephone clicked in Rector's hand. RECTOR had no idea why Kayabashi wanted to visit the mis
, was accompanied by his personal chaplain, Michael Hudson, RECTOR of Uffington near Stamford, and Mr. Ashburnham. O

Примеры использования корпуса

- **to make a decision** или **to take a decision**

http://www.lextutor.ca/

British National Corpus

http://www.natcorp.ox.ac.uk/

Corpus of Contemporary American English
http://corpus.byu.edu/coca/

- **класть** или **ложить**

Национальный корпус русского языка
http://www.ruscorpora.ru/

Dirty Corpus

http://www.google.com

Использование корпуса в обучении ИЯ

UK: Conservation and Environment

Going for a walk is the most popular leisure activity in Britain. Despite its high _____ density and widespread urbanization, the UK has many unspoilt rural and coastal areas. **POPULATE**

Twelve National Parks are freely accessible to the public and were created to conserve the _____ beauty, wildlife and cultural heritage they contain. **NATURE**

In 1997, the UK subscribed to the Kyoto Protocol binding developed countries to reduce emissions of the six main greenhouse gases. The Protocol declares environmental _____ **PROTECT**

http://www.lextutor.ca/

http://corpora.iling.spb.ru/

Факультет филологии и искусств
Санкт-Петербургского государственного университета

Институт лингвистических исследований РАН

Корпусная лингвистика

- главная
- архив семинара
- теория
- глоссарий
- библиография
- организаторы
- ссылки
- контакты
- гостевая

ГЛАВНАЯ

Уважаемые коллеги!

Этот сайт посвящен корпусной лингвистике. Здесь вы найдете информацию об этом научном направлении, познакомитесь с основными понятиями, использующимися в корпусной лингвистике, ее историей, научными школами, узнаете, для чего применяются корпуса текстов, и какие технологии используют ученые-корпусники.

Обращаем ваше внимание на то, что в Институте лингвистических исследований РАН продолжает работу семинар по корпусной и компьютерной лингвистике. Этот семинар проводится уже более 4-х лет и посвящен всему спектру проблем корпусной лингвистики.

Адрес института: 199053, Санкт-Петербург, Тучков пер., д. 9
Телефон / факс: (812) 328-46-11
По организационным вопросам, а также по вопросам принятия участия в семинаре обращаться по адресу: spbcorpora@yandex.ru (Мария)

Участие в семинаре бесплатное! Ждем Вас!

наверх

Новости семинара

Состоялись:

▸19.05.08
А.В. Андреев.
Конструктивная логика как синтаксический формализм
(ИЛИ РАН)

▸26.05.08
Е. Чухарев.
Цель, принципы и методы формирования и разметки корпуса спонтанной компьютерно-опосредованной коммуникации
(РГПУ им. А.И. Герцена)

Архив новостей

Руководитель проекта: Виктор Захаров