Лекция 5. ПРИКЛАДНЫЕ РАЗДЕЛЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

План

1.Корпусная лингвистика как раздел прикладной лингвистики.

2. Понятие корпуса, разметки. Виды корпусов.

3. Требования к корпусам.

4. Понятие компьютерной лексикографии.

5.Электронный словарь. Состав словарной статьи. Виды электронных словарей. Преимущества электронных словарей.

6.Перспективы компьютерной лексикографии.

1. Корпусная лингвистика как раздел прикладной лингвистики.

Корпусная лингвистика - раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов при помощи компьютеров.

2. Понятие корпуса, разметки. Виды корпусов.

Центральное понятие корпусной лингвистики – и лингвистический корпус – определяется как совокупность специально отобранных текстов, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска. Таким образом, корпус можно кратко охарактеризовать следующим образом:

Корпус = тексты + их разметка.

Важным этапом создания корпуса является его разметка. Разметка (англ. tagging, annotation) - это приписывание текстам и их компонентам специальных меток (англ. tag). Эти метки могут быть внешними (экстралингвистическими), включающими сведения об авторе и о тексте, или внутренними: структурными или собственно лингвистическими.

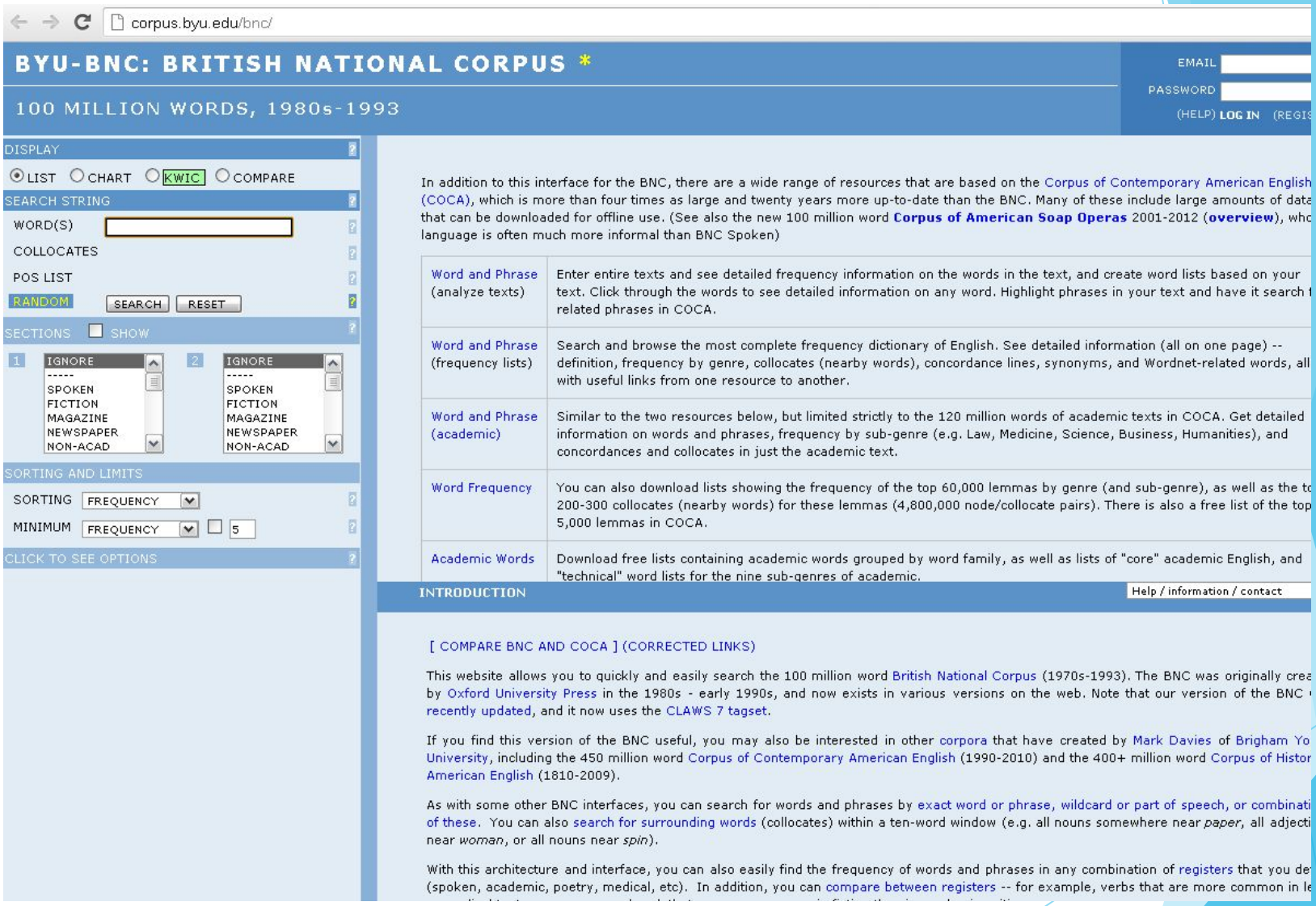
В зависимости от характера собранных в корпусе текстов, от их разметки и некоторых других факторов различают следующие виды корпусов

№	Признак	Виды корпусов
1	Форма хранения	звуковые письменные смешанные
2	Язык текстов	русский английский и т.д.
3	«Параллельность»	одноязычные двуязычные многоязычные
4	Стиль	литературные диалектные разговорные публицистические терминологические смешанные
5	Способ доступа	свободно доступные коммерческие закрытые
6	Разметка	размеченные неразмеченные
7	Характер разметки	морфологические синтаксические семантические просодические и т.д.

Наиболее важным видом корпусов является универсальный национальный корпус, создаваемый для разных национальных языков.

Универсальный национальный корпус - это собрание текстов конкретного есте-ственного языка, представительное по отношению ко всему языку, которое может слу-жить для исследования самых разнообразных явлений этого языка.

Общепризнанный образец универсального национального корпуса Британский национальный корпус (BNC) (www.natcorp.ox.ac.uk). Для русского языка таким представительным корпусом является Национальный корпус русского языка (НКРЯ) (www.ruscorpora.ru). Среди корпусов славянских языков выделяется Чешский нацио-нальный корпус (http://ucnk.ff.cuni.cz), созданный в Карловом университете Праги. Национальные корпусы существуют также для немецкого, китайского, финского и других языков.


corpus.byu.edu/bnc/
BYU-BNC: BRITISH NATIONAL CORPUS *
100 MILLION WORDS, 1980s-1993
EMAIL
PASSWORD
(HELP) LOG IN (REGIS
DISPLAY
LIST CHART KWIC COMPARE
SEARCH STRING
WORD(S)
COLLOCATES
POS LIST
RANDOM
SEARCH
RESET
SECTIONS SHOW
1 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER NON-ACAD
2 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER NON-ACAD
SORTING AND LIMITS
SORTING FREQUENCY
MINIMUM FREQUENCY 5
CLICK TO SEE OPTIONS
In addition to this interface for the BNC, there are a wide range of resources that are based on the Corpus of Contemporary American English (COCA), which is more than four times as large and twenty years more up-to-date than the BNC. Many of these include large amounts of date that can be downloaded for offline use. (See also the new 100 million word Corpus of American Soap Operas 2001-2012 (overview), whc language is often much more informal than BNC Spoken)
Word and Phrase (analyze texts)
Enter entire texts and see detailed frequency information on the words in the text, and create word lists based on your text. Click through the words to see detailed information on any word. Highlight phrases in your text and have it search t related phrases in COCA.
Word and Phrase (frequency lists)
Search and browse the most complete frequency dictionary of English. See detailed information (all on one page) -- definition, frequency by genre, collocates (nearby words), concordance lines, synonyms, and Wordnet-related words, all with useful links from one resource to another.
Word and Phrase (academic)
Similar to the two resources below, but limited strictly to the 120 million words of academic texts in COCA. Get detailed information on words and phrases, frequency by sub-genre (e.g. Law, Medicine, Science, Business, Humanities), and concordances and collocates in just the academic text.
Word Frequency
You can also download lists showing the frequency of the top 60,000 lemmas by genre (and sub-genre), as well as the to 200-300 collocates (nearby words) for these lemmas (4,800,000 node/collocate pairs). There is also a free list of the top 5,000 lemmas in COCA.
Academic Words
Download free lists containing academic words grouped by word family, as well as lists of "core" academic English, and "technical" word lists for the nine sub-genres of academic.
INTRODUCTION
Help / information / contact
[COMPARE BNC AND COCA] (CORRECTED LINKS)
This website allows you to quickly and easily search the 100 million word British National Corpus (1970s-1993). The BNC was originally crea by Oxford University Press in the 1980s - early 1990s, and now exists in various versions on the web. Note that our version of the BNC recently updated, and it now uses the CLAWS 7 tagset.
If you find this version of the BNC useful, you may also be interested in other corpora that have created by Mark Davies of Brigham Yo University, including the 450 million word Corpus of Contemporary American English (1990-2010) and the 400+ million word Corpus of Histor American English (1810-2009).
As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinati of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjecti near woman, or all nouns near spin).
With this architecture and interface, you can also easily find the frequency of words and phrases in any combination of registers that you de (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in le

3. Требования к корпусам.

При отборе текстов в корпус следует ориентироваться на следующие требования к созданию корпусов:

1) репрезентативность

2) полнота

3) достаточный объем

4) экономичность

5) структуризация материала

6) компьютерная поддержка

4. Понятие компьютерной лексикографии.

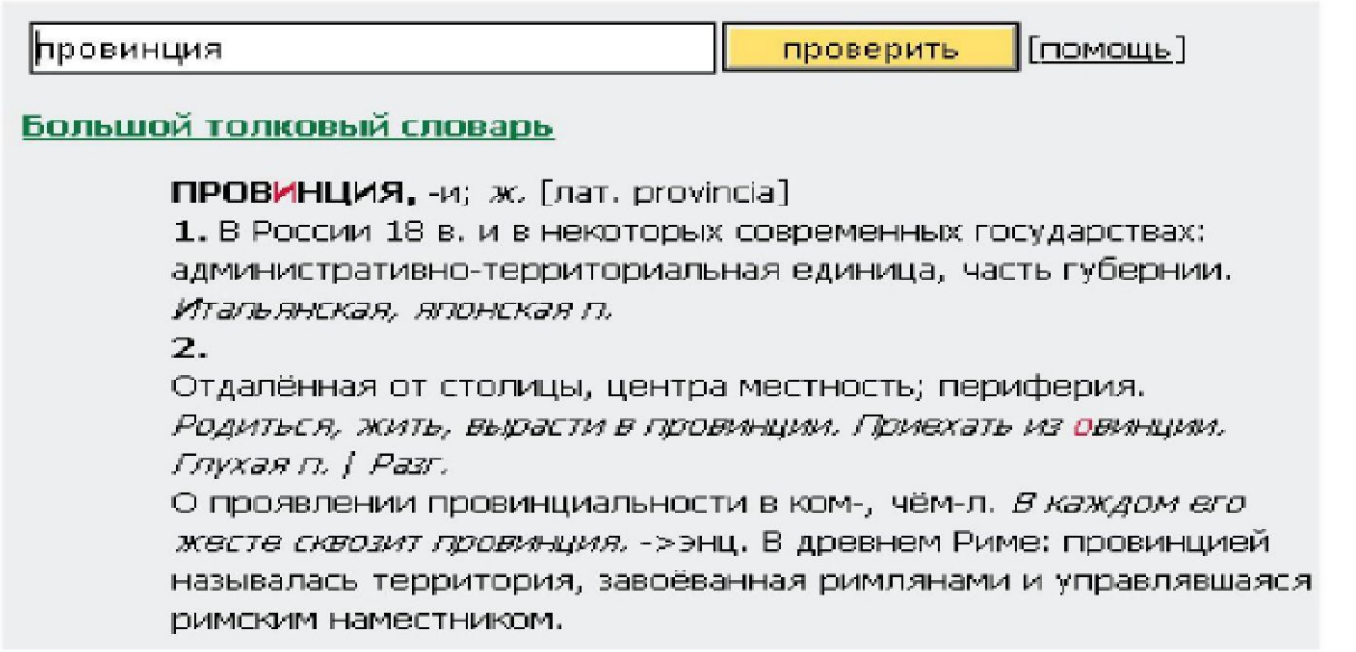
Компьютерная лексикография представляет собой раздел прикладной лингви-стики, нацеленный на создание компьютерных словарей, лингвистических баз данных и разработку программ поддержки лексикографических работ.

5. Электронный словарь. Состав словарной статьи. Виды электронных словарей. Преимущества электронных словарей.

Электронный (автоматический, компьютерный) словарь - это собрание слов в спе-циальном компьютерном формате, предназначенное для использования человеком или являющееся составной частью более сложных компьютерных программ (например, систем машинного перевода).

Соответственно, различаются автоматические словари конечного пользователя-человека (АСКП) и автоматические словари для программ обработки текста (АСПОТ).

Автоматические словари такого типа практически повторяют структуру словарной статьи обычных словарей, однако они обладают функциями, недоступными своим прототипам, например, осуществляют сортировку данных по полям словарной статьи (ср. отбор всех прилагательных), проводят автоматический поиск всех вокабул, имеющих в толковании определенный семантический компонент, и т.д..


провинция проверить [помощь]

Большой толковый словарь

ПРОВИНЦИЯ, -и; ж. [лат. provincia]

1. В России 18 в. и в некоторых современных государствах: административно-территориальная единица, часть губернии. *Итальянская, японская п.*

2.

Отдалённая от столицы, центра местность; периферия. *Родиться, жить, вырасти в провинции. Приехать из провинции. Глухая п. | Разг.*

О проявлении провинциальности в ком-, чём-л. *В каждом его жесте сквозит провинция.* ->энц. В древнем Риме: провинцией называлась территория, завоёванная римлянами и управлявшаяся римским наместником.

Структура традиционного словаря обычно включает следующие компоненты:

-введение, объясняющее принципы пользования словарем и дающее информацию о структуре словарной статьи;

-словник, включающий единицы словаря: морфемы, лексемы, словоформы или словосочетания; каждая такая единица с соответствующим комментарием представляет собой словарную статью;

-указатели (индексы);

-список источников;

-список условных сокращений и алфавит.

В электронных словарях из названных компонентов обязательным является, пожалуй, лишь словник, в онлайн-словарях нередко имеется также алфавит с заложенными за каждой буквой гиперссылками, ведущими к тексту словарной статьи.

Примерами переводных электронных словарей выступают ABBYY Lingvo (www.lingvo.ru), Translatelt! (www.translateit.ru) и Multitran (www.multitran.ru). Электронные толковые словари это, в частности, словарь Merriam Webster (www.merriam-webster.com) и словарь французского языка «Tresor de la langue francaise» (http://atilf.atilf.fr). Формальными электронными словарями являются орфо-графические словари русского (http://slovari.yandex.ru) и английского (www .spellcheckonline .com) языков.

Электронные словари имеют положительные стороны не только в процессе их создания, но и в процессе использования. В частности, выделяются следующие преимущества в использовании электронных словарей:

1) электронные словари позволяют по-разному представить содержание словарной статьи (различные «проекции» словаря), в том числе с помощью разнообразных графических и мультимедийных средств, которые не используются в обычных словарях;

2) в выдаваемой информации находят отражение различные технологии компьютерной лингвистики, например морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и т.п.;

3) становится возможным быстро получить информацию, которая содержится где-то в недрах словаря и непосредственно отвечает тому запросу, который сформулирован пользователем в удобной для него форме;

4) электронный словарь позволяет быстро реагировать на изменения в языке и мире, и выпуск каждой последующей его версии или внесение изменений в онлайн-версию не занимает много времени и труда.

6. Перспективы компьютерной лексикографии.

Специализированных программных оболочек для лексикографических целей на рынке практически нет. Для этих целей вполне подходят современные базы данных типа ACCESS или PARADOX. Для поиска примеров создатели словарей могут использовать компьютерные программы построения конкордансов, например, DIALEX.

Для создания оригинал-макета (верстки) словарей привлекаются издательские си-стемы типа Page-Maker или WinWord, которые позволяют приписывать стили зонам словарных статей, алфавитизацию, создание указателей и т.д..

Компьютерная лексикография, направленная на создание электронных словарей, представляет собой весьма перспективное и нужное направление компьютерной лингвистики, поскольку создаваемые ею продукты - электронные словари - отличаются многогранностью, мультимедийностью, интеграцией новейших технологических решений, актуальностью материала и отвечают потребностям пользователя в организации доступа к необходимой информации.

Спасибо за внимание