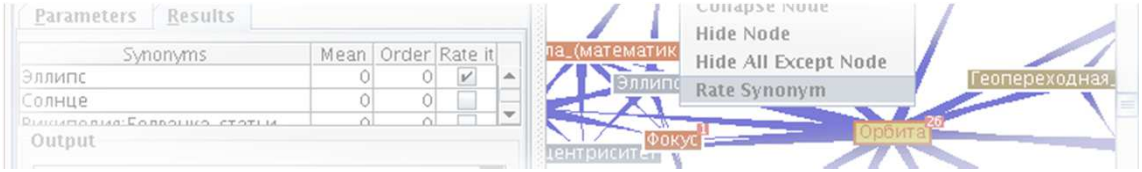


Корпусная лингвистика

Corpus linguistics

Петрозаводский государственный университет

Крижановский Андрей Анатольевич

andrew.krizhanovsky gmail.com

Основные понятия

Корпусная лингвистика – это наука о создании и использовании текстовых (лингвистических) **корпусов**, возникшая вследствие растущих потребностей лингвистики во внедрении компьютерных технологий для работы с большими массивами языковых данных.

What is Corpus Linguistics?

Corpus Linguistics is the study of language/linguistic phenomena through the analysis of data obtained from a corpus.

Лингвистический корпус (1)

- совокупность текстов,

- собранных в соответствии с определёнными принципами,
- размеченных по определённому стандарту
- и обеспеченных специализированной **поисковой системой**.

Лингвистический корпус (2)

Под текстами в этом случае понимаются не только продукты **письменного** языка (газетные статьи, романы, письма, электронные сообщения, дневники и т.п.), но и **устные** высказывания (доклады, радио- и телепередачи, телефонные разговоры и т.п.).

Why to use a *corpus*?

- Intuition alone is not enough
  - Is "*starting*" always replaceable by "*beginning*"?
  - Is it only "*time*" that is "*immemorial*"?
  - "*think of*" vs. "*think about*"

- Native speaker intuition is unreliable
  - provides no information on frequency of occurrence
  - "***head***" => body part - Is this the most used sense?

- Один раз создать корпус и многократно применять его для решения различных лингвистических задач.

Text vs. Corpus (Tognini-Bonelli 2001: 3)

TEXT	CORPUS
Read whole	Read fragmented
Read horizontally	Read vertically
Read for content	Read for formal patterning
Read as a unique event	Read for repeated events
Read as an individual act of will	Read as a sample of social practice
Coherent communicative event	Not a coherent communicative event

Text vs. Corpus

From time to time there is also the need for high quality information to support particular initiatives, such as the (successful) application for accreditation. Some progress has been made in recording data on the Polytechnic 's rooms and buildings, and on the teaching space requirements of individual courses. These data are analysed, along with the database on course details and students ' course and module registrations, using the methodology in DES Design Note 44. Ad hoc reports are an essential part of any system that aspires not merely to process data routinely but to permit management information to be creamed off the top.

N	Concordance
13	ement system. They can choose whether to enter data themselves or to use the data preparation se
14	individuals is recorded. As well as student-related data, details on courses, on modules and their
15	student on the module is entered, for subsequent data processing by the registry. In addition
16	s detailed record-keeping and effortless access to data. All of this the system provides.
17	nd passed to the Registry who, together with the data preparation service, enter and verify approxi
18	considerable use of student management system data processing. The marksheet for each module
19	ons that can assist their work. Individual student data are available to assist counselling. Registry
20	tion. Some progress has been made in recording data on the Polytechnic 's rooms and buildings, a
21	onitors to allow the whole Committee to view such data. A detailed analysis of the performance of
22	space requirements of individual courses. These data are analysed, along with the database on c
23	easy to record in a computer (unless complicated data structures are used) and are even harder to

Тексты, входящие в состав корпусов

- отобраны исходя из определенных принципов,
- специально подготовлены и размечены,
  - машиночитаемый формат + разметка
- с помощью специальных программ в них можно искать необходимые фрагменты текста по заданным параметрам.

Классификация корпусов: критерии

Характеристики анализа и сравнения корпусов:

- Тип/формат данных
- Язык текстов
- «Параллельность»
- «Литературность»
- Специфичность
- Жанр
- Доступность
- Назначение
- Динамичность
- Разметка
- Характер разметки
- Объем текстов
- Хронологический аспект
- «Общность»
- Структура

Основные понятия

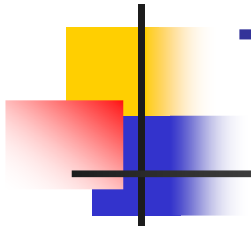
Типы корпусов:

1. Противопоставление корпусов, относящихся ко **всему языку**, корпусам, относящимся к какому-либо **подъязыку** (жанр, стиль, язык определенной возрастной или социальной группы, язык писателя или ученого и т.п.);

2. Разделение корпусов по типу лингвистической разметки.

Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам **морфологического** либо **синтаксического** типа (treebanks, «банки синтаксических структур»).

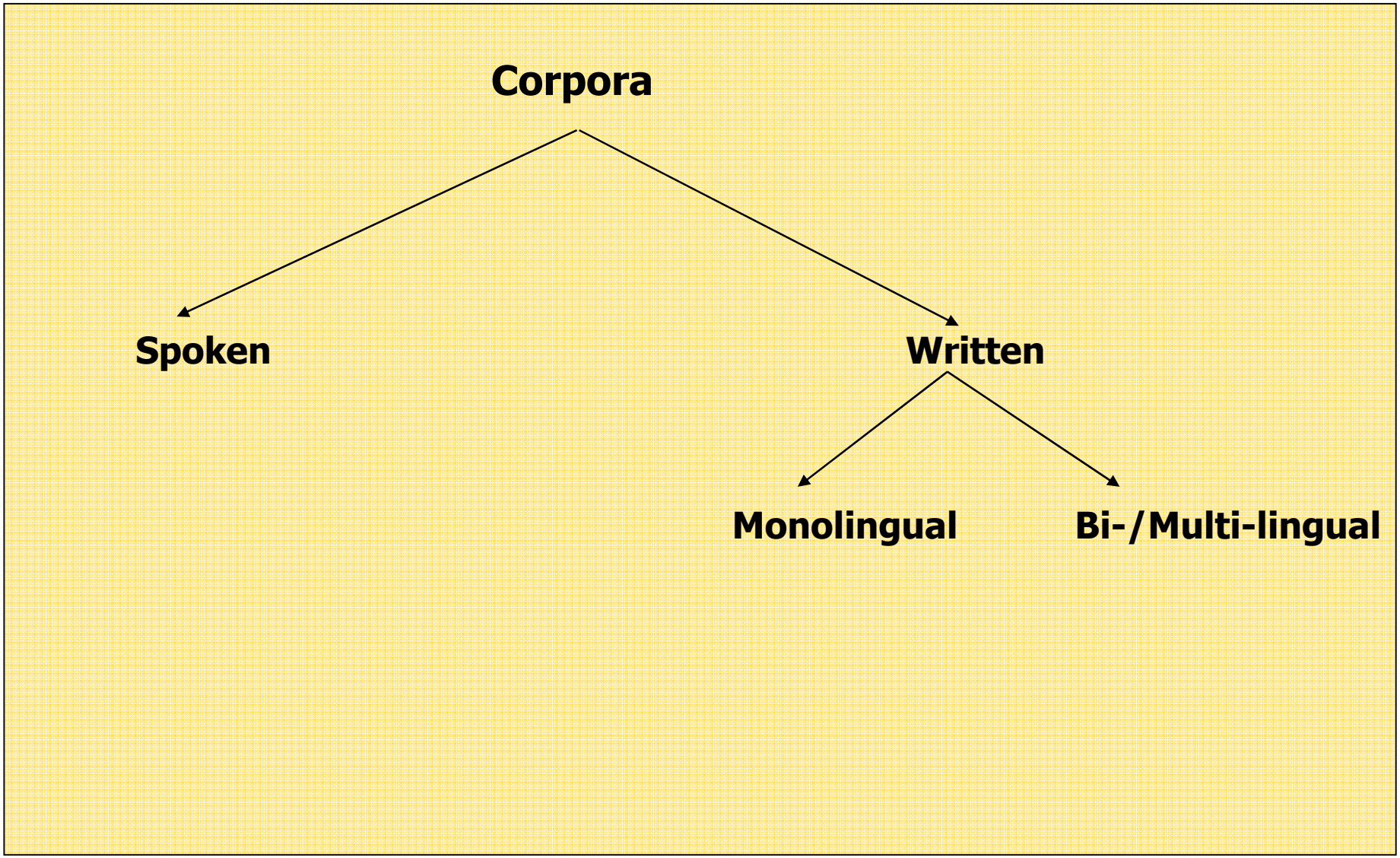
Критерий	**Типы корпусов**
Формат текста	Электронный / Неэлектронный
Полнота текстов	Полнотекстовый / Выборочный
Завершенность корпуса	Статичный / Пополняемый
Средство реализации языка	Корпус письменной речи Корпус устной речи
Языковые разновидности	Корпус стандартного языка Корпус молодежного языка Корпус экономического / компьютерного языка
Временной параметр	Корпус современного языка Корпус исторического языка
Количество языков	Одноязычный / Многоязычный



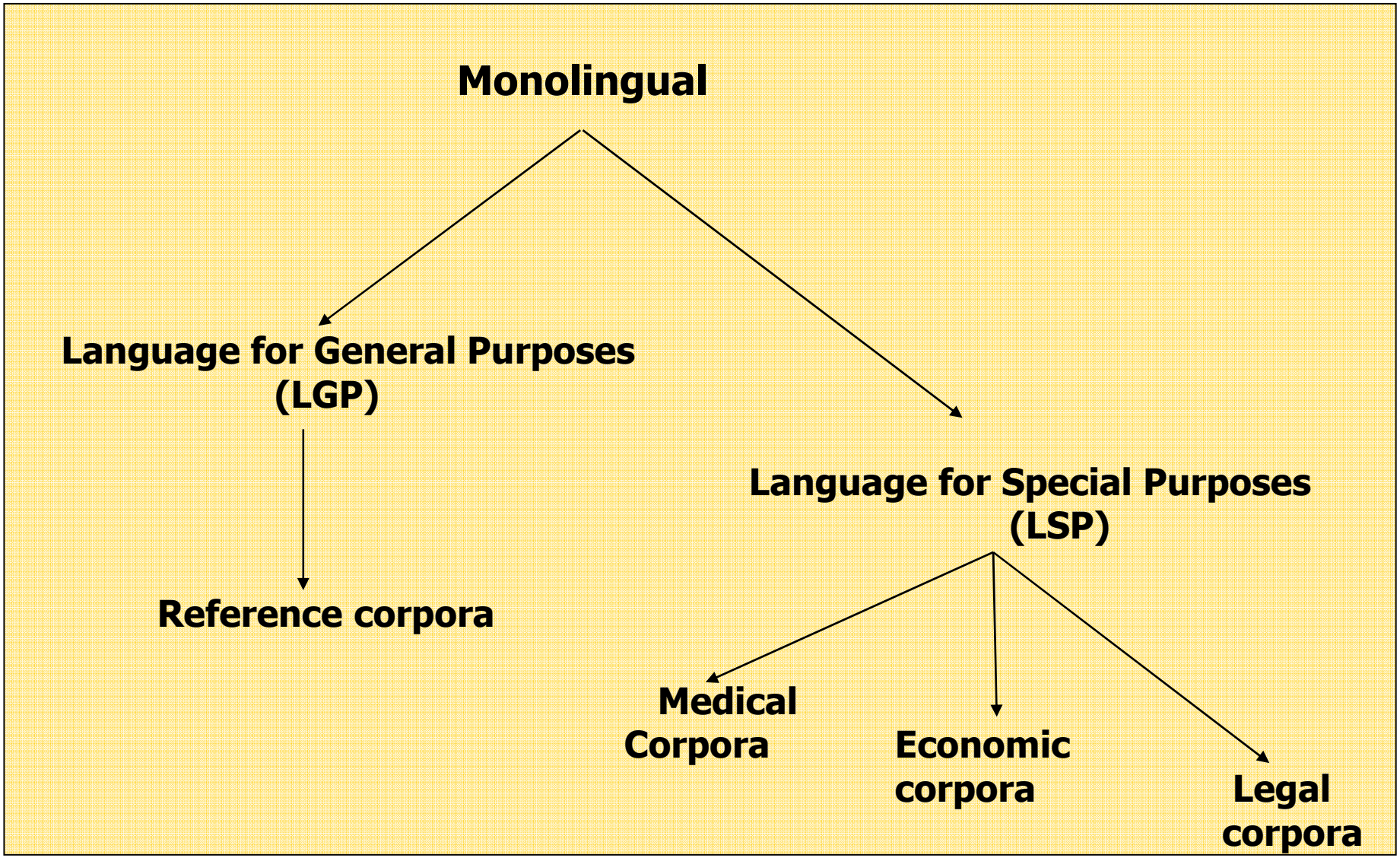
Types of corpora

- spoken vs. written
- monolingual vs. bi/multilingual
- parallel vs. comparable corpora (translation corpora)
- general language purpose vs. specialised language purpose
- diachronic vs. synchronic
- plain text vs. annotated (tagged) text

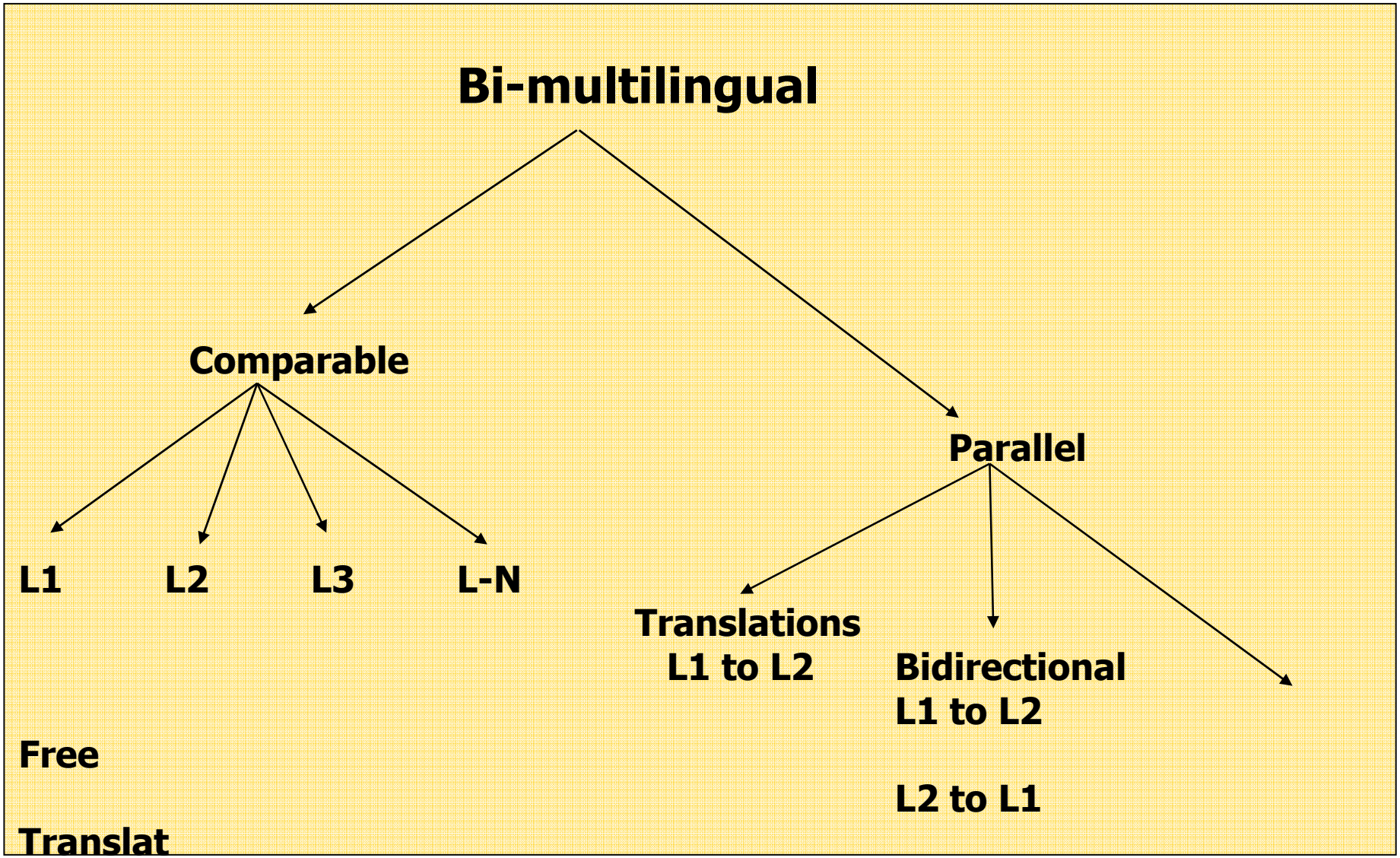
Types of corpora



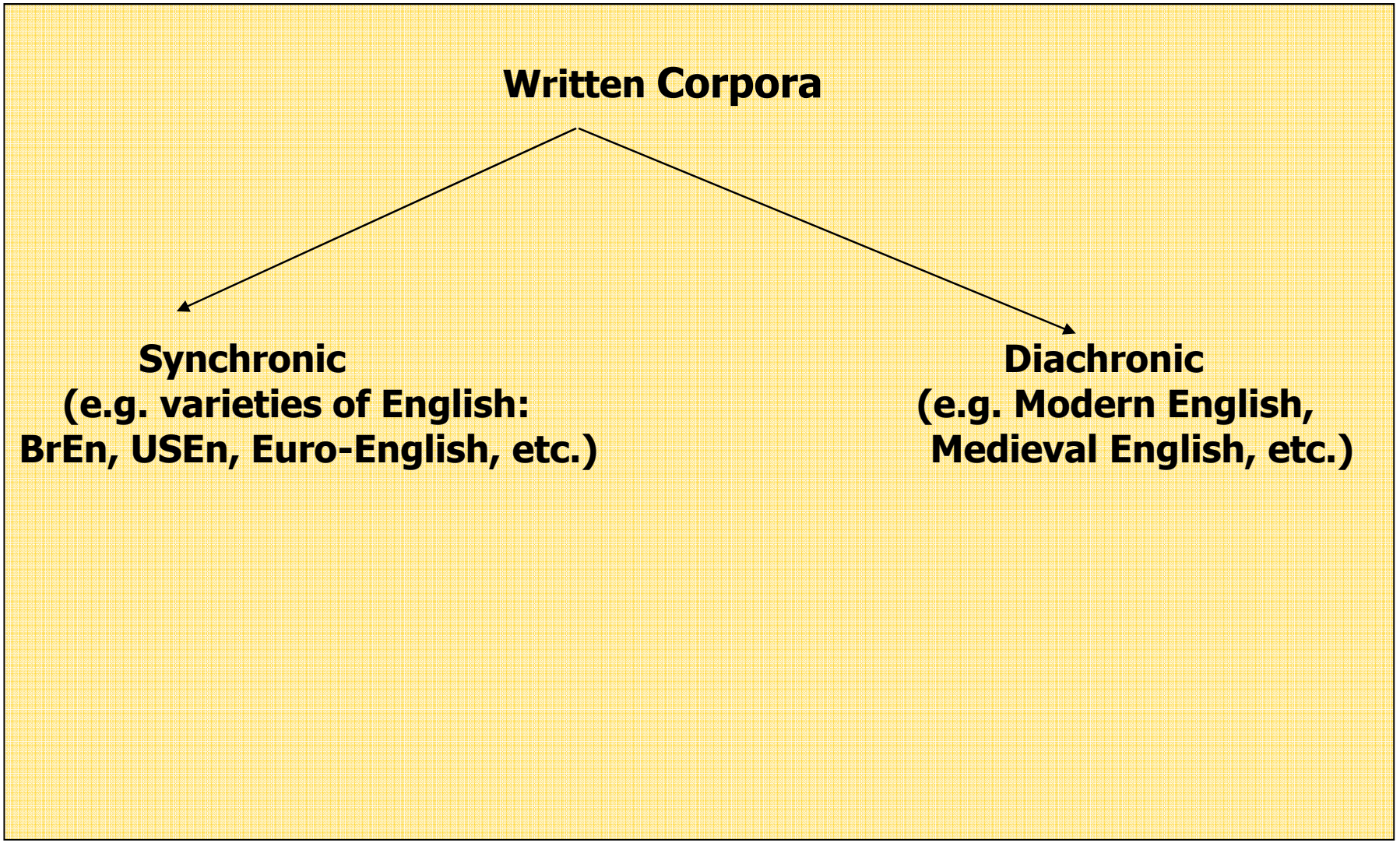
Types of corpora

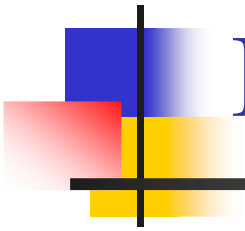


Types of corpora



Types of corpora





Корпусы(а) в Интернете

Национальный корпус русского языка **http://ruscorpora.ru**	>300 млн слов
Открытый корпус русского языка **http://opencorpora.org**	>921 тыс. слов
Компьютерный корпус текстов русских газет конца XX-го века **http://www.philol.msu.ru/~lex/corpus**	200 тыс. слов
Корпус русского языка ХАНКО (Хельсинский университет) **http://www.ling.helsinki.fi/projects/hanco/**	100 тыс. слов Ручная морфологическая разметка
Корпуса русских текстов на сайте Университета в Лидсе, Великобритания **http://corpus.leeds.ac.uk**	
Русские корпуса Тюбингенского Университета **http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html**	
Словарь-корпус языка А.С. Грибоедова **http://feb-web.ru/feb/concord/abc/**	120 тыс. слов

Уппсальский корпус русских текстов Доступен для поиска на сайте **http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html**	1 млн слов 600 текстов (публицистика 1985-1989; литературные произведения 1960-1988)
Банк английского языка (Bank of English) **http://www.collins.co.uk/books.aspx?group=153** Свободный доступ: **http://www.collins.co.uk/Corpus/CorpusSearch.aspx**	524 млн слов, 56 млн в свободном доступе (The Collins Wordbanks *Online* English corpus: 36 млн – брит. англ., 10 млн – амер. англ., 10 млн – брит. разговорн. англ.)
Британский национальный корпус **http://www.natcorp.ox.ac.uk/** или **http://sara.natcorp.ox.ac.uk**/	100 млн слов Корпусные менеджеры SARA и XAIRA (**http://www.xaira.org**)
Венгерский национальный корпус **http://corpus.nytud.hu/mnsz**/	100 млн слов

Корпус испанского языка (исторический) **http://www.corpusdelespanol.org**/	100 млн слов, тексты 13–20 вв. Создан в Иллинойском университете, США
Корпус современного датского языка **http://www.korpus2000.dk/**	50 млн слов Тексты 1998–2002 гг.
Корпус современного итальянского языка CORIS/CODIS **http://www.cilta.unibo.it/ricerca.htm**	100 млн слов
Корпус современного китайского языка (LIVAC Synchronous Corpus) **http://www.rcl.cityu.edu.hk/livac/**	720 млн слов (150 млн иероглифов)
Мангеймский корпус немецкого языка (Institut für Deutsche Sprache, Mannheim, Germany) **http://corpora.ids-mannheim.de/ccdb/**	1610 млн слов Корпусный менеджер COSMAS

Польский национальный корпус **http://korpus.ia.uni.lodz.pl/**	93 млн слов
Словацкий национальный корпус **http://korpus.juls.savba.sk**	180 млн слов Используется корпусный менеджер Manatee/Bonito
Хорватский национальный корпус **http://www.hnk.ffzg.hr/**	53 млн слов Корпусный менеджер Manatee/Bonito
Чешский национальный корпус **http://ucnk.ff.cuni.cz**	100 млн слов + 100 млн нового корпуса современной лексики Корпусный менеджер Manatee/Bonito
Эстонский корпус **http://www.cl.ut.ee/korpused/baaskorpus/**	

НКРЯ

Национальный корпус русского языка

представляет данный язык на определенном этапе его существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов.

Образовательный портал Национального корпуса русского языка:

http://studiorum.ruscorpora.ru/

НКРЯ (2)

Национальный корпус имеет две важные **особенности**:

1. Он характеризуется **представительностью**, или сбалансированным составом текстов.

2. Корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую **разметку**, или **аннотацию**).

Основные понятия

Конкорданс – это список всех употреблений заданного языкового выражения (например, слова) **в контексте**, возможно, со ссылками на источник.

Существуют специальные программы составления конкордансов по некоторому корпусу текстов, так называемые **конкордансеры**.

Основные понятия

Итак, корпуса используются, прежде всего, при

- исследовании различных языковых разновидностей; проверка лингвистических теорий;
- составлении словарей, грамматических справочников и т.п.;
- изучении и преподавании иностранных языков (составление учебных пособий, упражнений, доступ к аутентичным материалам);
- переводе (проверка сочетаемости слов, стилистических особенностей);
- социолингвистике: сравнение языков разных социальных групп;

Lexicography / Terminology (wikipedia.org)

General **lexicography** focuses on the design, compilation, use and evaluation of **general dictionaries**, i.e. dictionaries that provide a description of the language in general use.

Terminology, in its general sense, simply refers to the usage and study of terms, that is to say words and compound words generally used in **specific contexts**.

Lexicography and corpora

- Corpus-based lexicography started in England
- Corpus provides authentic uses of language
- **Extract samples (concordance) to identify different senses**
- Word Frequency information
- Help identify collocation (1), set phrase (2)

  1) Фразеологическое сочетание, коллокация
  - *ставить условия, вносить предложения*

  2) Фразеологическое выражение
  - *пословицы, афоризмы, речевые штампы*
    - всего хорошего, до новых встреч
- Most English dictionaries are now corpus-based. Oxford, Collins, Longman, Cambridge, Macmillan,

Источники литературы

1. Захаров В.П. Обзор корпусов. Презентация. – Режим доступа: download.yandex.ru/class/zakharov/CL_L9.ppt
2. Образовательный портал Национального корпуса русского языка. – Режим доступа: http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=241&Itemid=48
3. Подлесская В.И. Современные компьютерные методы в изучении и преподавании лингвистических дисциплин: корпусная лингвистика. – Режим доступа: http://zhangbyrzhan.ucoz.ru/publ/metodika/inostrannyj_jazyk/sovremennye_kompjuternye_metody_v_izuchenii_i_prepodavanii_lingvisticheskikh_disciplin/12-1-0-27
4. Портал «Национальный корпус русского языка». – Режим доступа: http://www.ruscorpora.ru/
5. Портал «Фонд знаний ЛОМОНОСОВ». Энциклопедия. Статья «Конкорданс». – Режим доступа: http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127200
6. Scherer C. Korpuslinguistik. – Universitätsverlag WINTER Heidelberg. – 2006. – 98 S.

Презентации:

1. Электронные корпуса. (безымянная презентация) http://www.slideshare.net/anna_pal/ss-13040000
2. Corpus linguistics: a general introduction (who is author?) http://www.lingue.uniba.it/dag/pagine/personale/falco/Corpus%20linguistics_introduction_Students.ppt