Корпусная лингвистика

Corpus Linguistics

Corpus Linguistics

Corpus Linguistics is a branch of Linguistics (Computer Linguistics) that studies language/linguistic phenomena through the analysis of data obtained from a corpus using IT based tools.

Corpus Linguistics vs. Traditional Linguistics

Corpus Linguistics	Traditional Linguistics
The subject of study is *speech*	The subject of study is *language*
Aimed at *describing* a living language	Aimed at *studying and explaining* language phenomena
Goes *from speech to theory*	Goes *from theory to its reflection in language*
Applies *objective methods*	Applies *deductive methods*
Analyses a large collection of texts	Analyses a definite phenomenon

Linguistic Corpus (pl. corpora)

- **Linguistic Corpus** can be defined as a systematic collection of naturally occurring texts. To be worth linguistic analyses it must be

- ✔ representative
- ✔ consistent
- ✔ structured
- ✔ tagged

Representative

Large and broad enough to include all types of texts

- all genres: from fiction to publicistic
- all language varieties: from colloquial to scientific
- all time periods: from old to modern
- ……

Systematic (consistent)

* the structure and contents of the corpus follows certain extralinguistic principles
* “sampling principles” are principles on the basis of which the texts included were chosen for the corpus
* information on the exact composition of the corpus is available to the researcher

Tagged

Англ.: tagging, annotation.

- the practice of adding interpretative linguistic information to a corpus

- Types of tagging:
- extralinguistic (*metatags*)
- structural
- linguistic

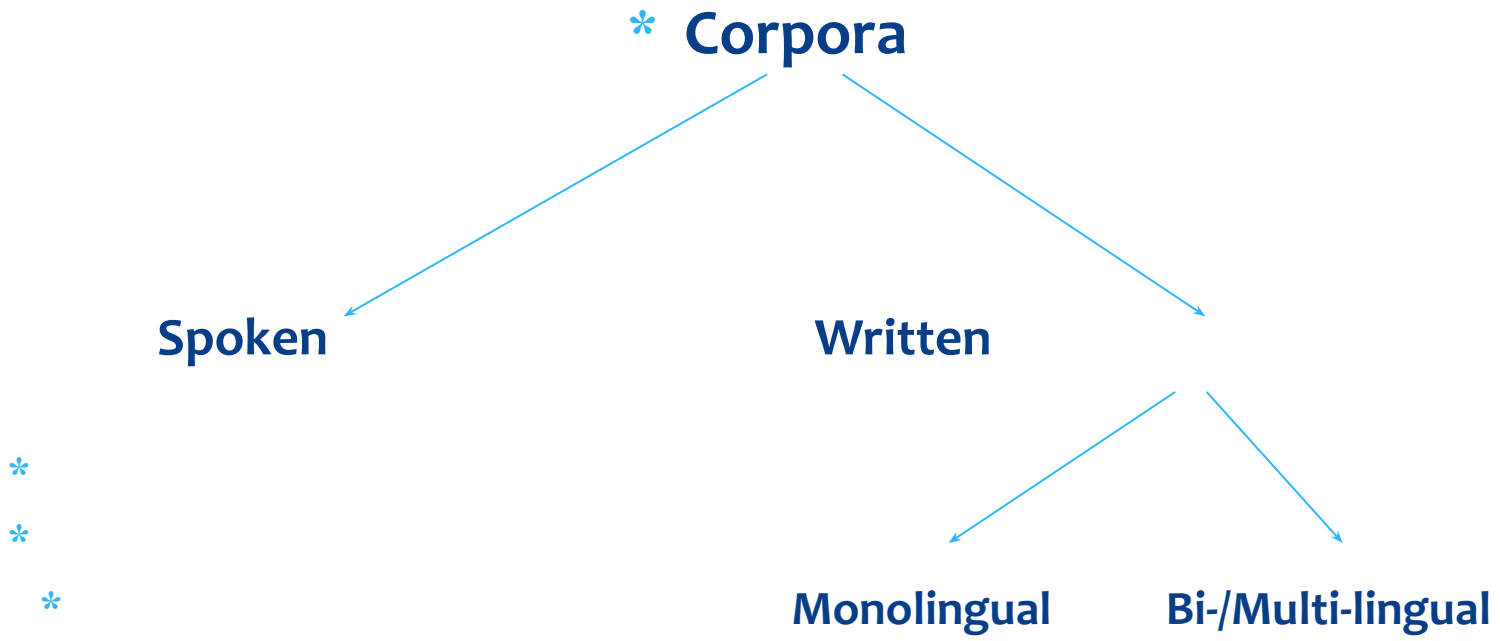
Linguistic Tagging/Annotation

1. part-of-speech tagging (POS-tagging)
2. syntactic
3. semantic
4. phonetic (prosodic)
5. …..

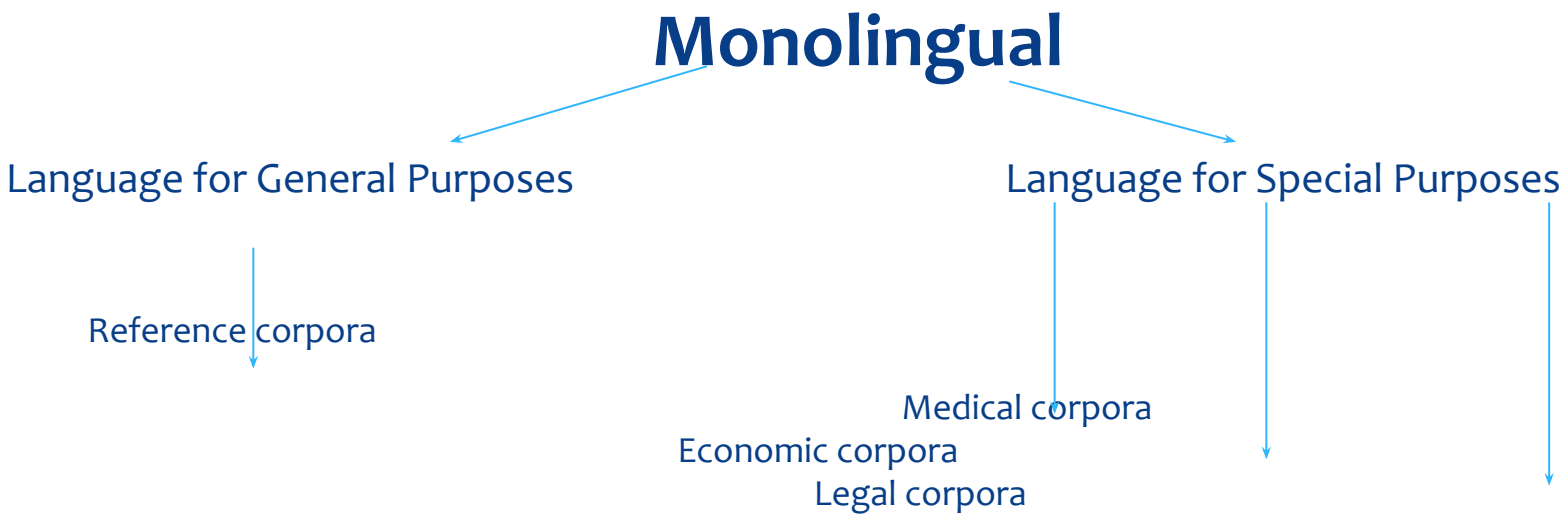
Types of Corpora

* spoken vs. written
* monolingual vs. bi/multilingual
* parallel vs. comparable corpora (translation corpora)
* general language purpose vs. specialised language purpose
* diachronic vs. synchronic

Types of Corpora



Types of Corpora



Bi-multilingual

- Comparable
- Parallel

Предпосылки создания и использования корпусов

Назначение языкового корпуса – показать функционирование лингвистических единиц в их естественной контекстной среде.

<u>На основе корпуса можно получить данные</u>:

- ✔ о частоте словоформ, лексем, грамматических категорий,
- ✔ об изменениях частот
- ✔ об изменениях контекстов в различные периоды времени
- ✔ о поведении языковых единиц разных авторов
- ✔ о совместной встречаемости лексических единиц
- ✔ об особенностях их сочетаемости, управления

Linguistic corpora

- British National Corpus
- International Corpus of English.
- Bank of English
- Национальный корпус русского языка.

British National Corpus

* http://www.natcorp.ox.ac.uk/
* http://corpus.byu.edu/bnc/

* The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English, both spoken and written, from the late twentieth century.

International Corpus of English

* http://ice-corpora.net/ice/index.htm
* The International Corpus of English (ICE) began in 1990 with the primary aim of collecting material for comparative studies of English worldwide.
* Twenty-six corpora of national or regional varieties of English.
* Each ICE corpus consists of one million words of spoken and written English produced after 1989.
*

Национальный корпус русского языка

* http://www.ruscorpora.ru/
* includes texts representing standard Russian
* modern written texts (from the 1950s to the present day)
* a subcorpus of real-life Russian speech (recordings of oral speech from the same period)
* early texts (from the middle of the 18th to the middle of the 20th centuries).

Corpus Approach

Linguistic corpus

(*data*)

+

Corpus manager

(*indexing and search tool*)

Concordance

- Concordance is used to analyse different use of a single word, word frequency and phrases or idioms.

Corpus Managers

- AntConc
- dtSearch
- TeleportPro

TeleportPro / dtSearch

TeleportPro

- Программа для скачивания сайтов
- Создает корпус текстов с различной глубиной копирования сайта

dtSearch

- Программа индексации корпусов
- Работает с корпусами любых форматов

AntConc

* Does not require installing
* Compatible with most operation systems
* Broad array of tools
* Limited to certain document types (htm, html, xml,txt – на входе и txt – на выходе)

Good luck!

* Practice the use of

* AntConc tools: KWIC-конкорданс, Word List, Key Word List, Concordance Plot, etc.
* TeleportPro + dtSearch