# Computational Linguistics Lab: Corpus Linguistics

**Elena Cabrio, FBK-Irst**
mail: cabrio@fbk.eu

# Outline:

- What is corpus linguistics?
  - Theoretical and historical background
- What is a corpus?
  - Types of corpora
  - The use of corpora
- Basic notions of Corpus Linguistics
  - Type/Token
  - Concordances
  - Collocations
  - Let's explore the BNC corpus
- Web as a corpus?

# What is Corpus Linguistics?

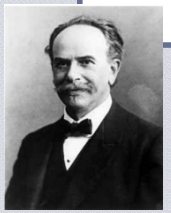
**Corpus Linguistics** is the study of the languages/linguistic phenomena through the analysis of data obtained from a corpus.

"it can be seen as a ***pre-application methodology***. […] by "pre-application" we mean that, unlike other applications that start by accepting facts as *given*, **corpus linguistics is in a position to define its own sets of rules and pieces of knowledge *before* they are applied.** […] Corpus linguistics has, therefore, a theoretical status and because of this it is in a position to contribute specifically to other applications."

(Tognini-Bonelli, 2001)

# Historical background

## Before 1950: Franz Boas and American Structuralism



Collection of small *corpora* to analyse the phonological aspects of the indigenous languages of America, adopting an empirical approach

## After 1950:

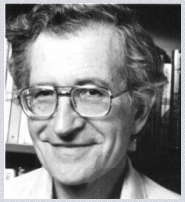
USA: Leonard Bloomfield's ***VERIFICATIONISM***: empirical approach to language: language studies must rely on the observation of facts.

- UK: J.R. Firth, M.A.K. Halliday, J. Sinclair: language is a real phenomenon, which makes sense only if it is considered in its real use, i.e. as ***PERFORMANCE*** rather than as *COMPETENCE*.

# Historical and theoretical background

- Theoretical aspects of *Corpus Linguistics*:
  - ***Empiricism*** and direct observation of real data
  - ***Performance***
  - ***Form* and *content* are indivisible**
  - ***Parole*** (context- and time-related) vs ***Langue*** (abstract and a-temporal)
  - Use of computers to study *corpora* qualitatively and quantitatively

## Mid-20: Chomsky's transformational-generative grammar



- caused a shift from empiricism to rationalism:
  - *COMPETENCE* vs *PERFORMANCE*
  - *DEEP STRUCTURES* (*competence*) vs *SURFACE STRUCTURES* (*performance*)
- the Chomskyan linguists reject ***corpus linguistics***:
  - a *corpus* is a collection of external data (*performance*)
  - regarded as 'uncreative' and passive

# Corpus Linguistics: approaches

**CORPUS-BASED**

Corpora are used mainly to expound, tests, or examplify theories and descriptions that were formulated before large corpora became available to inform language studies

**CORPUS-DRIVEN**

Strictly committed to the integrity of the data as a whole. Theorethical statement are claimed to be fully consistent with, and reflect directly the evidences provided by the corpus

- No need to achieve balanced and representative corpora
- Very large corpora

- In practise, the approaches are not so different…

# What is a corpus?

A collection of texts assumed to be representative of a given language, put together so that it can be used for linguistic analysis.

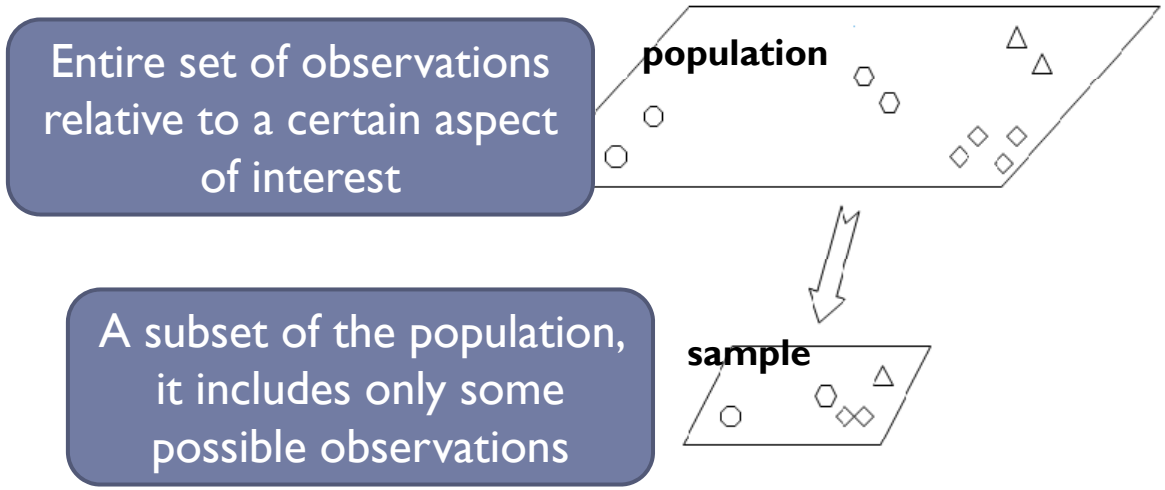
- The language stored in a corpus is assumed to be:
  - naturally-occurring
  - gathered according to explicit design criteria, with a specific purpose in mind, and with a claim to represent larger chunks of language selected according to a specific typology.

- There is consensus that a corpus deals with natural, authentic language.

(Tognini-Bonelli, 2001)

# Goals:

- Generalize the observations highlighted in the sample to the entire population



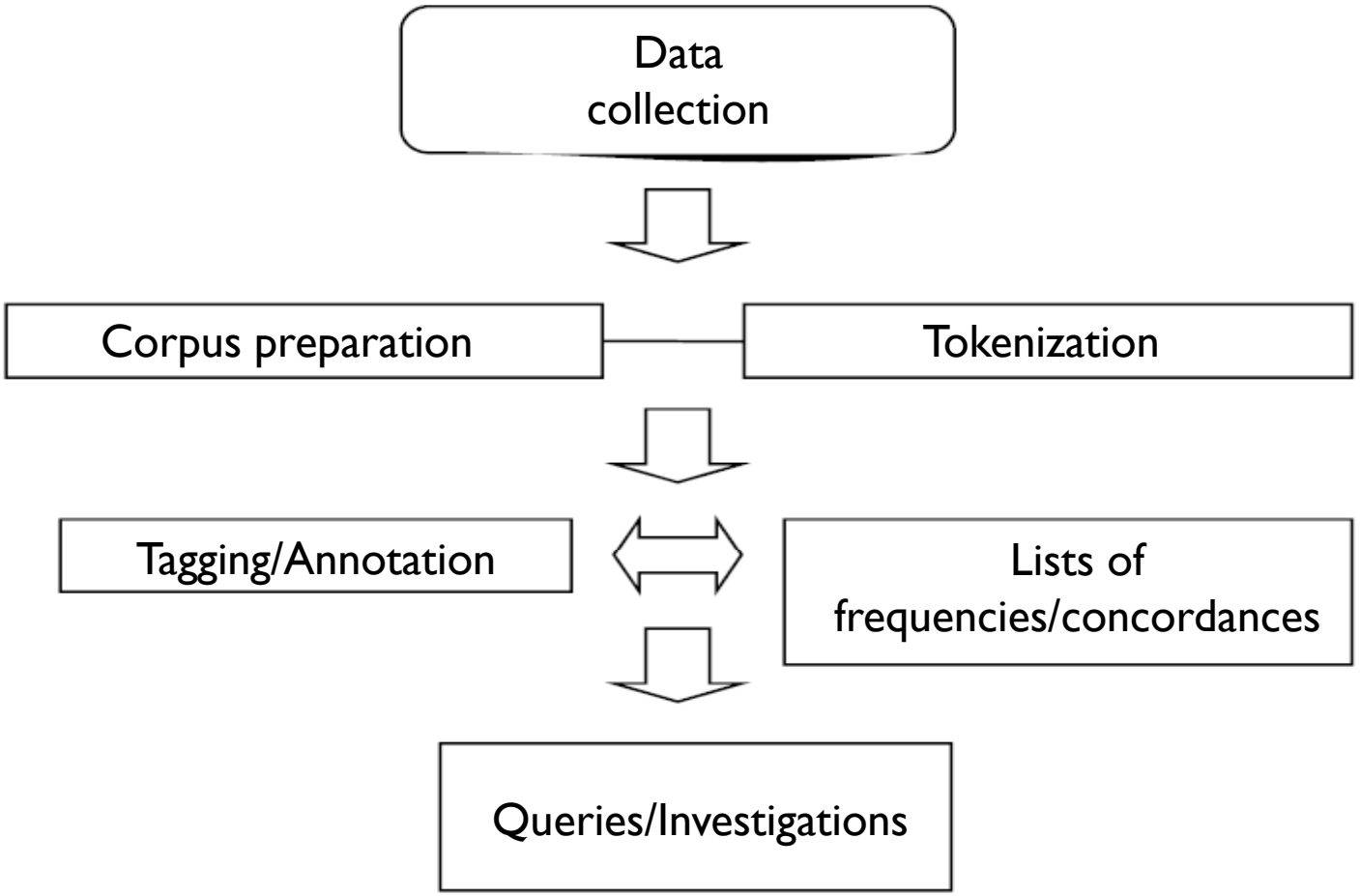

- Compare the analysis performed on different corpora, and calculate the deviation wrt a reference corpus
- Analyse specific phenomena of interest in a controlled setting

# Corpus issues:

- Authenticity
  - Need to deal with language in use
- Size
  - Standard sizes, according to the investigated phenomenon (token occurrences)
- Sampling
  - Define the target population the corpus aims to represent
- Representativeness
  - It varies according to the aspect under analysis (a corpus is not representative *per se*)
- Balance

(Tognini-Bonelli, 2001)

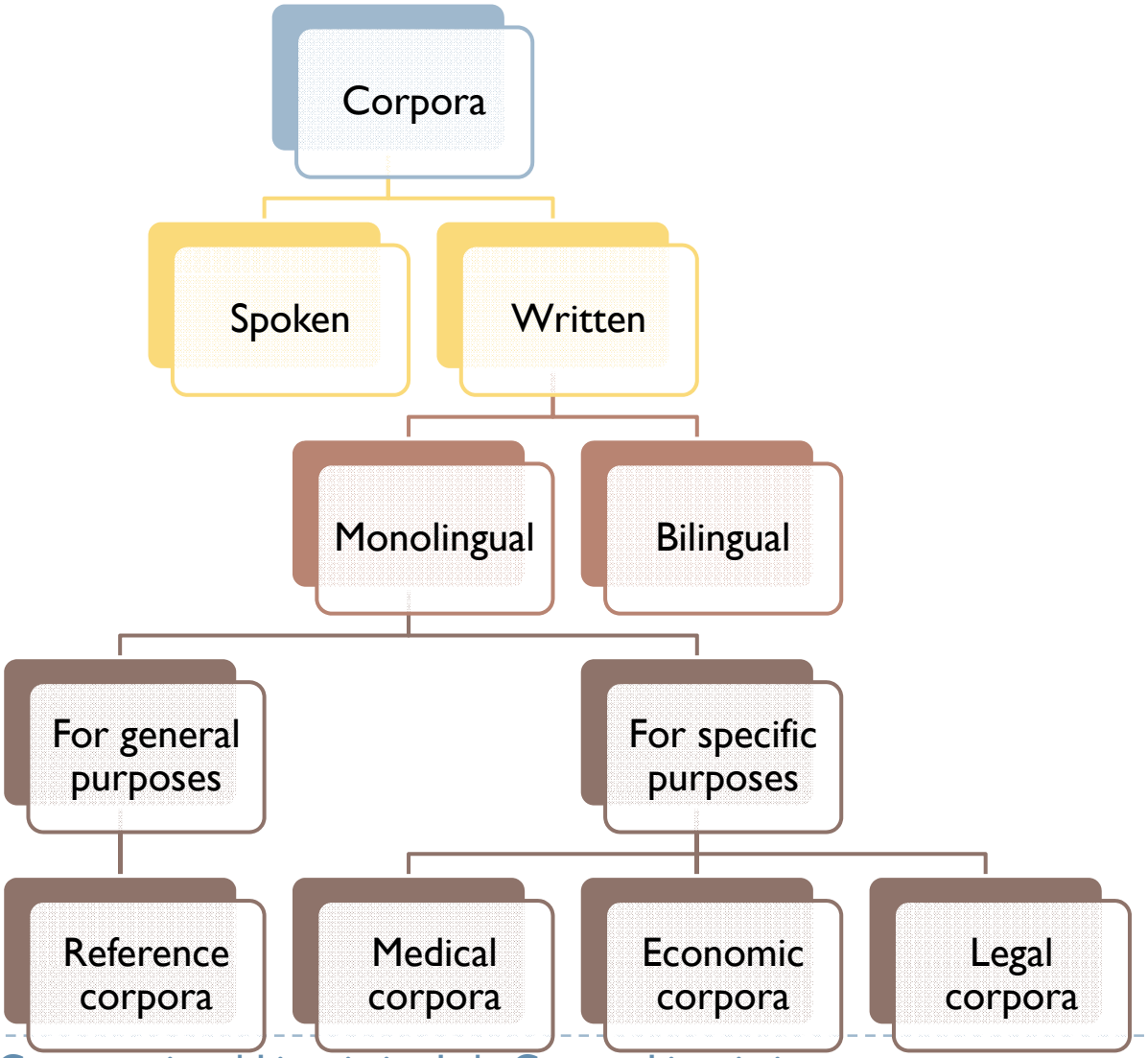
# How to build a corpus



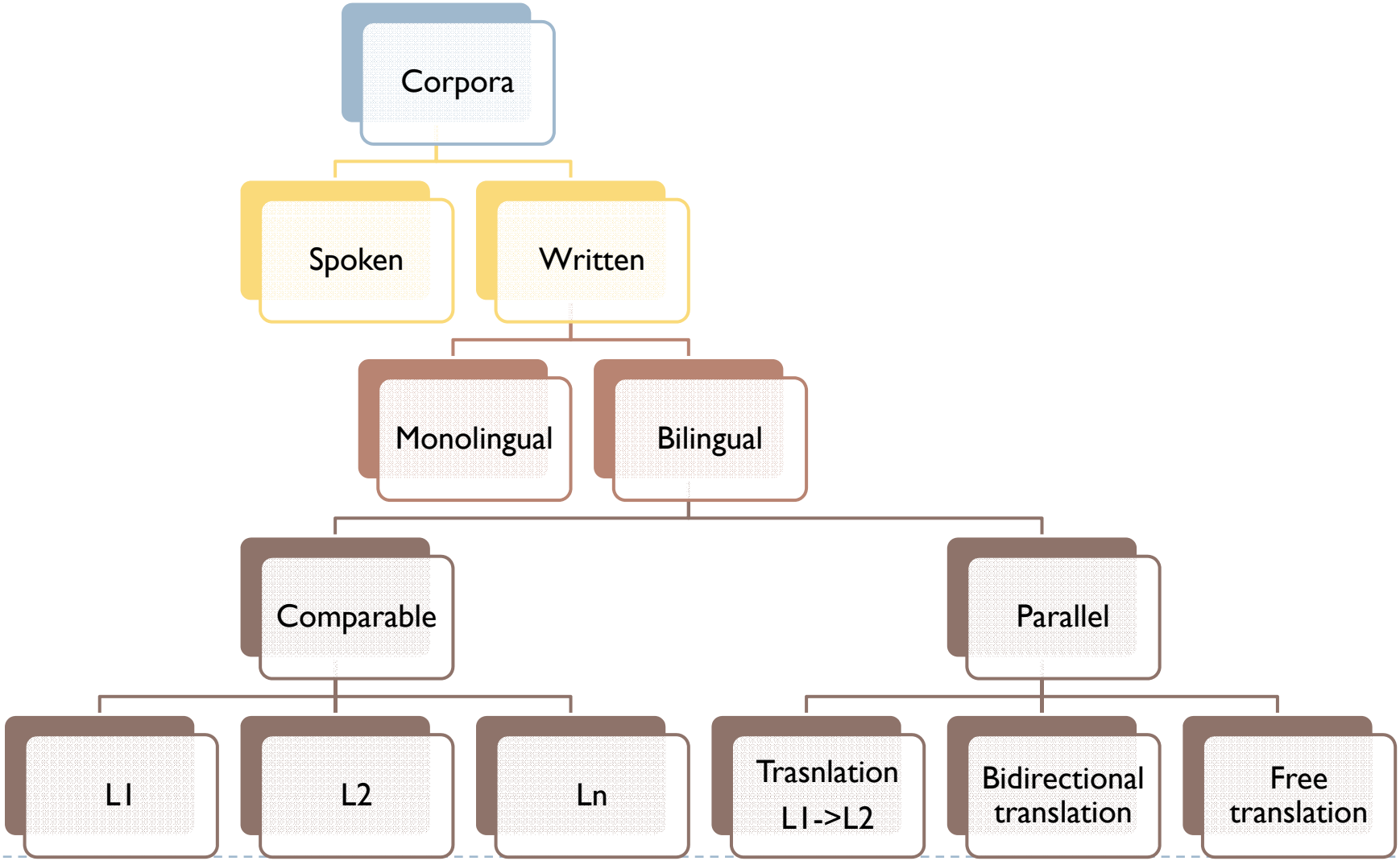
# Types of corpora:

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

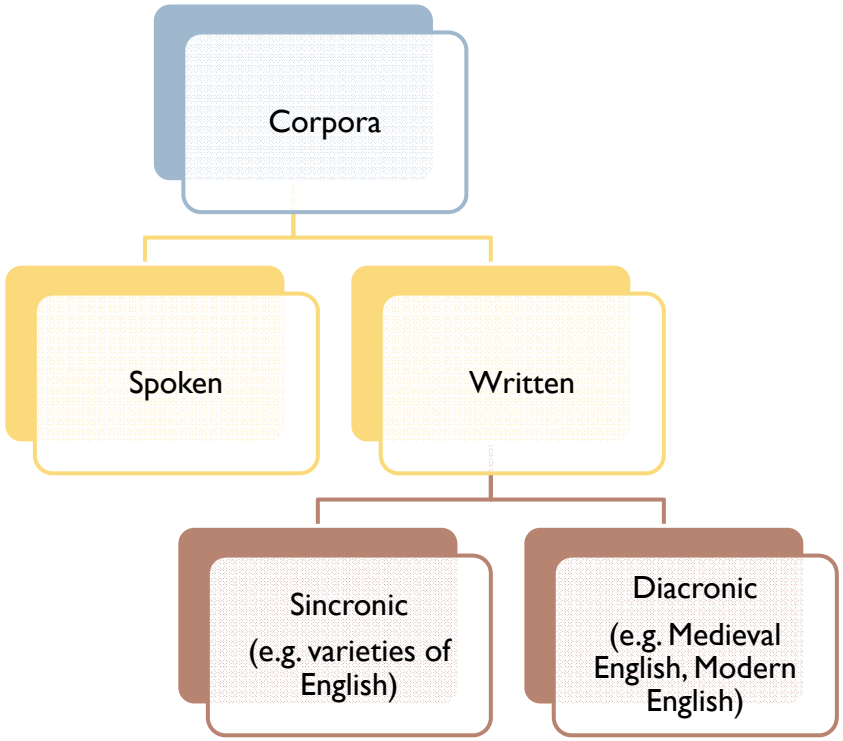
# Types of corpora:



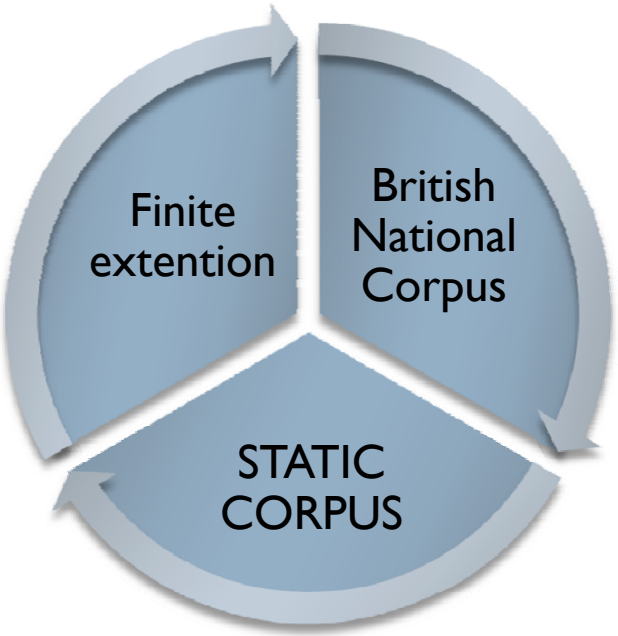
# Types of corpora:



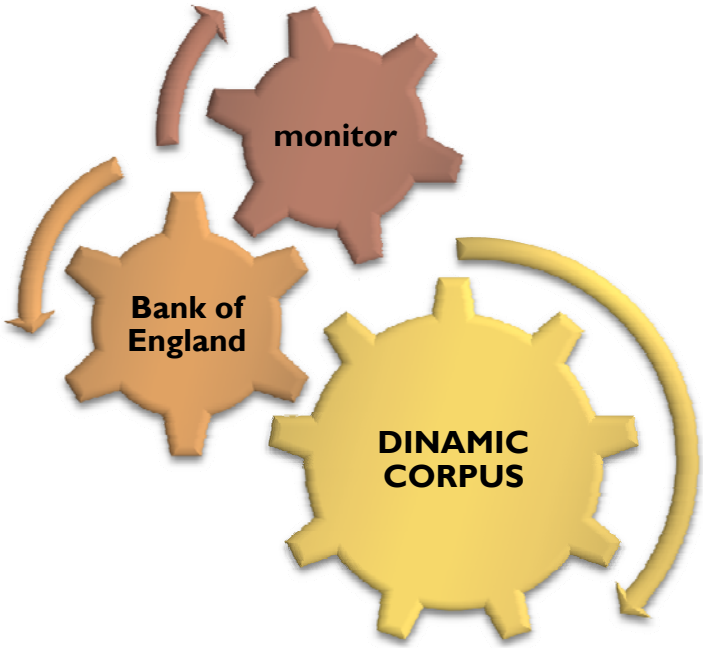
# Types of corpora:



# Static vs dinamic corpora




- Advantages:
  - Analysis and investigations can be repeated
  - Comparable




- Advantages:
  - Updates
  - Diachronical analysis

# English reference corpora:

- **The Brown Corpus (1964)**
  - 1 million words (500 samples/2,000 words, written American English, texts published in the US in 1961
- **The Lancaster-Oslo/Bergen (LOB) Corpus (1978)**
  - similar to the Brown corpus, British English
- **The London-Lund Corpus (LLC)**
  - 200 samples, ~5000 words each, 1953-1987, spoken British English, transcribed.
- **The Frown Corpus,** Freiburg-Brown Corpus of American English **(1992)**
  - analogue to the Brown corpus
  - 1 million words, written American-English.

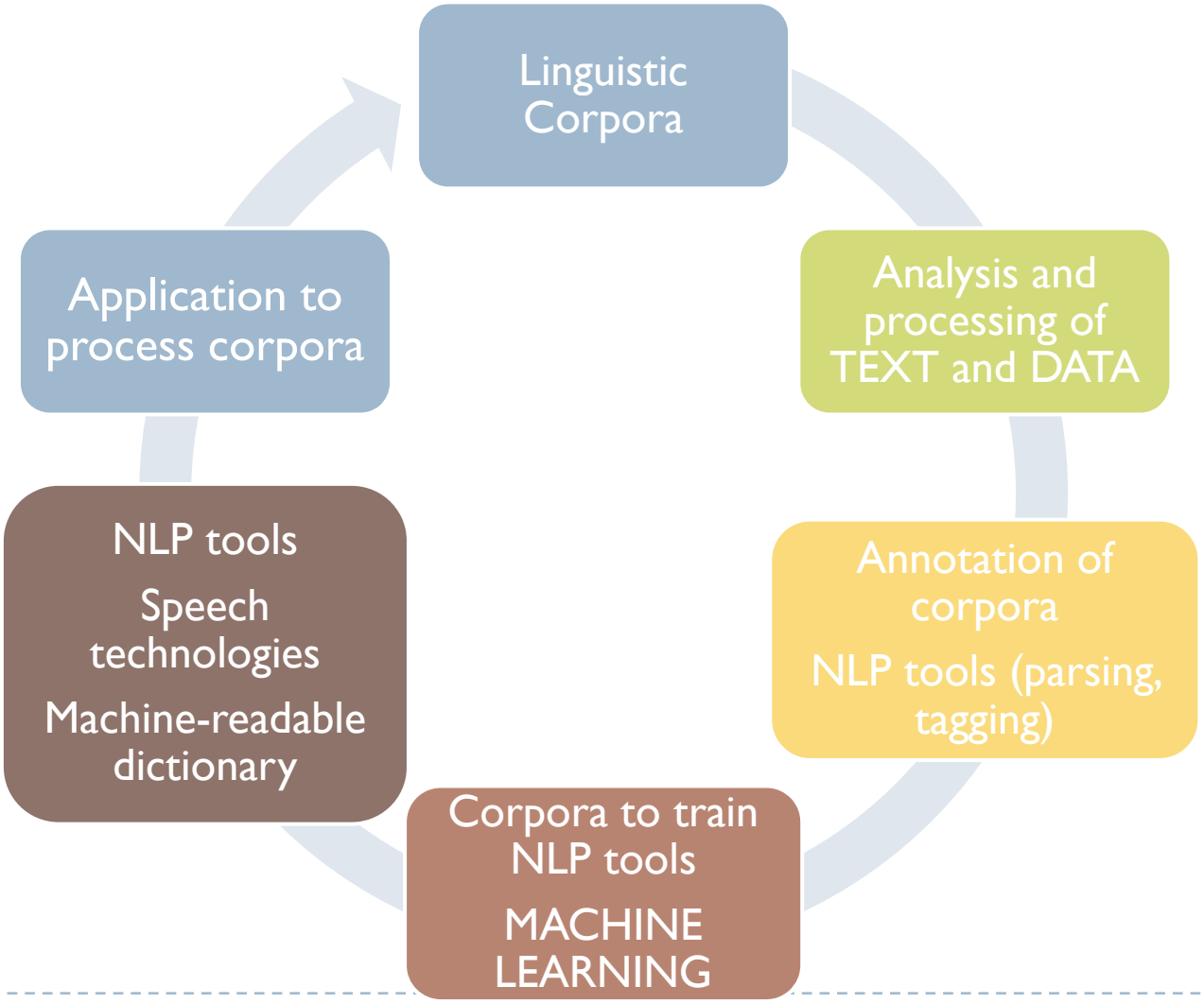
# English reference corpora (cont.):

- **The FLOB Corpus,** Freiburg-LOB Corpus of British English (1990s)
  - analogue to the LOB corpus
  - 1 million words, written British English
- **The British National Corpus (BNC)**
  - 100 million-word
  - samples of written texts (90m words) and spoken language (10m words).
- **The International Corpus of English (ICE)**
  - 500 samples (300 spoken, 200 written), ~2,000 words each
  - 20 national varieties of English (e.g. UK, India, Singapore, Australia, India, Jamaica)
- **The BoE Corpus (The Bank of English Corpus)**
  - 450M words, full texts, open, written and spoken
  - mainly US and UK

# The use of corpora:

- Corpus-based lexicography/terminology
  - Dictionaries and grammars
  - Corpus-based Machine Readable Dictionaries
- Training corpora for NLP tools
  - Tagger and parsers
- Machine Translation
  - Corpus-based MT
  - Example-based MT
- Speech technologies
  - Training for Speech recognition
  - Corpus-based Text-to-Speech
- Machine Learning
- Language teaching/learning

# The virtuous cirle of Computational Ling.



# Text vs Corpus

| 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 |

(Tognini-Bonelli, 2001)

# Critiques to Corpus Linguistics

**finiteness**

- By definition, a corpus is a **finite** collection of elements, i.e. it cannot be representative for an infinite language

**incomple-teness**

- It is **incomplete**, i.e. it excludes potential utterances

**imperfection**

- Utterances are biased by accidental factors, i.e. they are **imperfect**

**inadequacy**

- A corpus provides information on the frequency of the tokens, not on their "grammaticality"

# Basic notions of CL: type and token

Given a text…

- a TOKEN is each individual linguistic expressions (i.e. the use of word in text)
- a TYPE is the abstract class of which these tokens are members

- TYPE/TOKEN ratio gives us the richness of the vocabulary
- It is a value beween 0 and 1: the closer to 1 the richer the text is in terms of variety of the vocabulary.

# Given a sample of text:

> Miss Bingley's letter arrived, and put an end to doubt. The very first sentence conveyed the assurance of their being all settled in London for the winter, and concluded with her brother's regret at not having had time to pay his respects to his friends in Hertfordshire before he left the country.
>
> (*Pride and Prejudice*, *J. Austin*)

- Calculate the number of tokens:
- Calculate the number of types:
- Calculate type/token ratio:

# Given a sample of text:

Miss Bingley's letter arrived, and put an end to doubt. The very first sentence conveyed the assurance of their being all settled in London for the winter, and concluded with her brother's regret at not having had time to pay his respects to his friends in Hertfordshire before he left the country.

(*Pride and Prejudice*, *J. Austin*)

- Calculate the number of tokens: 53
- Calculate the number of types: 46
- Calculate type/token ratio: 0.86

# Basic notions of CL: Concordances

- A **CONCORDANCE** is a list of a particular word or sequence of words in a context.
- **Concordance programs** are basic tools that turn the electronic texts into databases which can be searched. Since most corpora are incredibly large, it is a fruitless enterprise to search a corpus without the help of a computer.

| N | Concordance | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 551 | d proper nouns. The initial construction | of the | data | structure | is of little | importance | to the | user | |
| 552 | ); the efficiency of representation | of the | data | so that its | particular | features | are | succinctl | |
| 553 | structure; the ease of alteration | of the | data | structure | ( i.e. adding | and | deleting | items); | |
| 554 | Alternative data structures Looking at possible | data | structures | for | representing | such | a | word | list |
| 555 | pointer to the next word in the list. This | data | structure | is | extremely | simple | to | implement | |
| 556 | it is rarely performed. Alternative | data | structures | Looking | at | possible | data | structu | |
| 557 | the movement of the stylus across its surface. | Data | is | collected | in the | form | of | x, y | co-ordinates |
| 558 | nd ensuring that facilities are available for these | data | to be | reported, | analysed | and | evaluated. | Ri | |
| 559 | iding managers with easy access to high- quality | data | and | ensuring | that | facilities | are | available | for |
| 560 | er can make rapid comparisons between sets of | data. | This | can | be | used | to | highlight | changes fro |

(BNC World Edition)

# Basic notions of CL: Collocations

- *"Collocates* are the words which occur in the neighbourhood of your search word" (Scott 1999).

- "This a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text. For example, PROVIDE frequently occurs with words which refer to valuable things which people need, such as *help* and *assistance*, *money*, *food* and *shelter*, and *information*. These are some of the frequent *collocates* of the verb". (Stubbs 2002)

| N | Collocations |
|---|---|
| 10289 | and output to the supply rails. The RX data input is clamped to the supply rails by diodes |
| 10290 | and a project to clear backlogs of registrations and data input for borehole logs, with the intention of pr |
| 10291 | were required to update the PMIS. The ideal data input document |
| 10292 | ndardised accounts automatically from accounting data input by the analyst. An alternative is |
| 10293 | phase in direct proportion to the value of a 4-bit data input. In the required circuit (figure, |
| 10294 | . This process is repeated for each source of data input. The randomized input map data are the |
| 10295 | input/output lines are buffered from the computer 's data input/output lines by IC5. This chip is an |
| 10296 | bit device with built-in Lithium battery. Its eight data input/output lines are buffered from the compu |
| 10297 | circuit. If a 2-bit number is set up on Data inputs D1 and D2 using switches S2 and S3, |
| 10298 | 1 and D2 to avoid possible confusion later with the data inputs D1, D2 etc. Following the |

# Some approaches to select collocations:

- **FREQUENCY**
- **MUTUAL INFORMATION**

# Some approaches to select collocations:

- **FREQUENCY**
- **MUTUAL INFORMATION**

Finding collocations by counting the number of occurrences.

PROBLEM: Usually results in a lot of function word pairs that need to be filtered out.

➡ Pass the candidate phrases through a part of-speech filter which only lets through those patterns that are likely to be "phrases". (Justesen and Katz, 1995)

| $C(w^1\ w^2)$ | $w^1$ | $w^2$ | tag pattern |
|---|---|---|---|
| 11487 | New | York | A N |
| 7261 | United | States | A N |
| 5412 | Los | Angeles | N N |
| 3301 | last | year | A N |
| 3191 | Saudi | Arabia | N N |
| 2699 | last | week | A N |
| 2514 | vice | president | A N |
| 2378 | Persian | Gulf | A N |
| 2161 | San | Francisco | N N |
| 2106 | President | Bush | N N |

# Some approaches to select collocations:

- **FREQUENCY**
- **MUTUAL INFORMATION**

Evaluate whether the co-occurrence of two words is purely by chance or statistically significant. (Church et al. 1989, 1991; Hindle 1990).

Mutual information between the occurrence of a word *x* and a word *y* is defined as follows:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)\ P(y)}$$

It compares the probability of observing *x and y together (the joint* probability) with the probabilities of observing x and *y independently (chance)*. If there is a genuine association between x and y, then the joint probability *P(x,y)* will be much larger than chance *P(x) P(y)*, and consequently *I(x,y) >> 0*.

# Let's explore the BNC corpus…

http://corpora.lancs.ac.uk/BNCweb/home.html

http://bncweb.lancs.ac.uk

# Web as a corpus?

- World Wide Web is a mine of language data of unprecedented richness and ease of access, why do not exploit the Web as a linguistic corpus? (Kilgarriff and Grefenstette, 2003)

- **Web Corpora** (e.g. www.webcorp.org.uk)

- Web Corpora resources:
  - WaCky (http://wacky.sslmit.unibo.it/doku.php?id=start)
  - BootCat (http://corpora.fi.muni.cz/bootcat/)

- **VIEW:** Variation In English Words and phrases
  **Mark Davies / Brigham Young University**
  http://view.byu.edu/

***Pay attention:***

*NO LAB SCHEDULED FOR NEXT WEDNESDAY !*

# BIBLIOGRAPHY:


- BNC World Edition. http://bncweb.lancs.ac.uk
- Chiari, I. (2007), *Introduzione alla linguistica computazionale, Laterza, Roma-Bari.*
- Church, K., Gale, W., Hanks, P., Hindle, D. (1989) *Parsing, Word Associations and Typical Predicate-Argument Relations,* Workshop on Parsing Technologies, CMU.
- *Corpus linguistics – a general introduction.* www.lingue.uniba.it/
- Hindle, D., (1990) *Noun classification from predicate argument structures*
- Justeson, J., Katz S. (1995), *Technical Terminology: Some Linguistic Properties and an Algorithm for Identification,* in Natural Language Engineering
- Kilgarriff , A. ,Grefenstette, G. (2003) *Web as Corpus.*
- Mihalcea, R., Collocations, Reading: Chap 5, Manning & Schutze
  www.cse.unt.edu/~rada/CSCE5290/Lectures/Collocations.ppt
- Scott, M. (1999), *Wordsmith Tools version 3*, Oxford: Oxford University Press.
- Stubbs, M. (2001). *Words and Phrases*. Oxford: Blackwell Publishers Ltd.
- Tognini-Bonelli E. (2001), *Corpus Linguistics at Work,* Studies in Corpus Linguistics 6, John Benjamins (publ.).