# Corpus linguistics - A general introduction

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

> 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, *Corpus linguistics at work*, 2001:1)

**Phase 1** – before 1950s
Franz Boas and the American Structuralism. He compiles small *corpora* to analyse the phonological aspects of the Inuit language, adopting an empirical approach
**Phase 2** – after 1950s
USA – Leonard Bloomfield's *verificationism*: rejects the mental approach to language in favour of an empirical one. Language studies must rely on the observation of facts.
UK – the Firthian tradition: J.R. Firth – M.A.K. Halliday – J. Sinclair
They draw back on Malinowski's *context of culture* and *context of situation*. 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*.

Reaction to Chomsky's transformational- generative grammar (mid-20th)
Dualism between *competence* and *performance*
Distinction between *deep structures* (*competence*) and *surface structures* (*performance*)
Language has to focus on *competence* rather than on *performance*

In short, the chomskyan linguistics

- Rejects *corpus linguistics* since a *corpus* is a collection of external data (*performance*)
- Is based on *introspection* and *rationalism* vs. *empiricism*.

Firth/Halliday/Sinclair reject any dualism and opt for a *monist* view of language.
Focus on *performance*

To sum up some aspects in *CL*:

- *Empiricism* and direct observation of real data
- *Performance*
- *Form* and *content* are indivisible -> *lexico-grammar approach* to language
- *Parole* is context- and time-related. *Langue* is abstract and a-temporal
- Use of computers to study *corpora* qualitatively and quantitatively.

# What is a corpus

In linguistics, corpus (plural *corpora*) is a large and structured set of texts (now usually electronically stored and processed). A corpus may contain single texts in single language (*monolingual corpus*) or text data in multiple languages (*multilingual corpus*). Multilingual corpora that have been specially

formatted for side-by-side comparison are called *aligned parallel corpora*. (*Webster's Online Dictionary*)

A corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. (Sinclair, *Corpus, Concordance, Collocation*, 1991:171)

A corpus can be defined as a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis. Usually the assumption is that the language stored in a corpus is naturally-occurring, that it is 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. […] in general there is consensus that a corpus deals with natural, authentic language. (Tognini-Bonelli, *Corpus linguistics at work*, 2001:2)

A corpus is a collection of texts, designed for some purpose, usually teaching or research. […] A corpus is not something that a speaker does or knows, but something constructed by a researcher. It is a record of performance, usually of many different users, and designed to be studied, so that we can make inferences about typical language use. Because it provides methods of observing patterns of a type which have long been sensed by literary critics, but which have not been identified empirically, the computer-assisted study of large corpora can perhaps suggest a way out of the paradoxes of dualism. (Stubbs, *Words and Phrases*, 2002:239-40)

[A corpus is] a subset of an ETL (Electronic Text Library) built according to explicit design criteria for a specific purpose (Atkins, Clear and Osler, "Corpus Design Criteria", in *Literary and Linguistic Computing*, 7.1, 1992:1-16)

A corpus is taken to be a computerised collection of authentic texts, amenable to automatic or semiautomatic processing or analysis. The texts are selected according to explicit criteria in order to capture the regularities of a language, a language variety or a sub-language. (Tognini-Bonelli, op. cit.:55)

It follows that:

- Texts must be collected according to specific criteria: content/genre/typology/register, etc.;
- Texts must be available in machine-readable form
- Texts are collected in order to analyse specific linguistic phenomena

**Criteria**
Authenticity
Size
Sampling
Representativeness
Balance
(Tognini-Bonelli, *Corpus linguistics at work*, 2001:47-64)

**English 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, text from 1961 (compiled 1970-1978)
- **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) 1990s analogue to the Brown corpus (1 million words, written American-English).
- **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, 1990 onwards, 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

**Italian Corpora**

- **CORIS/CODIS**: A corpus of written Italian - CORIS/CODIS - being developed at CILTA - Centre for Theoretical and Applied Linguistics – […] available on-line for research purposes. The project, designed and co-ordinated by R. Rossini Favretti, was started in 1998, with the purpose of creating a representative and sizeable general reference corpus of written Italian which would be easily accessible and user-friendly. CORIS contains 100 million words and will be updated every two years by means of a built-in monitor corpus. It consists of a collection of authentic and commonly occurring texts in electronic format chosen by virtue of their representativeness of modern Italian. (http://corpora.dslo.unibo.it/)
- **Corpus di italiano televisivo (CIT)**: 250,000 ~ 500,000 words. Authentic texts from advertising, entertainment, sport, news. Annotated according to TEI standards Text Encoding Initiative, http://www.sspina.it/cit/cit.htm
- **Corpora Linguistici per l'Italiano Parlato e Scritto (CLIPS)**: promoted by Federico Albano Leoni (Università "Federico II" di Napoli). The largest corpus of Spoken Italian, http://www.cirass.unina.it/

For other corpora, go to:

- http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/index2.html
- www.federicozanettin.net

**Web Corpora**

- Adam Kilgariff - http://www.kilgarriff.co.uk/
- Marco Baroni - http://www.form.unitn.it/~baroni/
- Google: www.google.com
- www.webcorp.org.uk

**Web Corpora resource**

- BootCat: http://corpora.fi.muni.cz/bootcat/
- VIEW: VARIATION IN ENGLISH WORDS AND PHRASES, Mark Davies / Brigham Young University, http://view.byu.edu/

**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

**Uses of Corpora**

- **Lexicography / terminology**
- **Linguistics / computational linguistics**
    - **Dictionaries & grammars** (*Collins Cobuild English Dictionary for Advanced Learners; Longman Grammar of Spoken and Written English*)
    - **Critical Discourse Analysis**:
        - Study texts in social context
        - Analyze texts to show underlying ideological meanings and assumptions
        - Analyze texts to show how other meanings and ways of talking could have been used, and therefore the ideological implications of the ways that things were stated
- **Literary studies**
- **Translation practice and theory**
- **Language teaching / learning**
    - ESL Teaching
    - LSP Teaching (*exemplar texts*)

**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. Such a dictionary is usually called a general dictionary or LGP dictionary. Specialized lexicography focuses on the design, compilation, use and evaluation of specialized dictionaries, i.e. dictionaries that are devoted to a (relatively restricted) set of linguistic and factual elements of one or more specialist subject fields, e.g. legal lexicography. Such a dictionary is usually called a specialized dictionary or LSP dictionary.

*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. Terminology also refers to a more formal discipline which systematically studies of the *labelling or designating o f concepts* particular to one or more subject fields or domains of human activity, through research and analysis of terms in context, for the purpose of documenting and promoting correct usage. This study can be limited to one language or can cover more than one language at the same time (*multilingual terminology*, *bilingual terminology*, and so forth).

**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, set phrase
  - Collocation : *file … patent, move on,*
  - Set phrase : *night and day, black and white*
- Most English dictionaries are now corpus-based (Oxford, Collins, Longman, Cambridge, Macmillan)

**Research on empirical linguistics**

Study language use in various aspects

- Verify linguistic theory, e.g. the explanation of definite description
- Lexical studies e.g. study near synonymous 'little' 'small'
- Sociolinguistics : compare the different of languages produced from different social groups (m/f)
- Cultural study e.g. differences found in 2 comparable corpora (British/American)

**Language Teaching / Learning and Corpora**

**Corpus-based** vs. **Corpus-driven**

> "the term *corpus-based* is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli, *Corpus linguistics at work,* 2001:65)

**Language Teaching and Corpus-based approach**

- Corpus based : use corpus as a resource
- Knowledge :
  - Know better about English answer specific questions of certain words, phrases, structures.
  - Know where the problems are error analysis on a learner corpus
  - Know what should be taught word frequency, comparing native/learner corpora
  - Language Teaching and Corpus-based approach
- References:
  - create better references dictionary, grammar book, textbooks
  - verify certain hypotheses about languages find support examples / counter examples
  - use a native corpus as a reference see whether it is possible which one is more natural
- Corpus based: use corpus as a resource
- Syllabus design:
  - Native corpora => what are actually used
  - Learner corpora => what are the problems
  - Find out which aspects should be given priority
  - Lexical syllabus = focus on frequency of occurrence
  - How many words the students should know? What are they?
  - Knowing 90% or 95% of the words?

**Language Teaching and Corpus-driven approach**

> "In a *corpus-driven* approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories

> or a probabilistic extension to an already well defined system. […] Examples are normally taken verbatim, in other words they are not adjusted in any way to fit the predefined categories of the analyst; recurrent patterns and frequency distributions are expected to form the basic evidence for linguistic categories; the absence of a pattern is considered potentially meaningful." (Tognini-Bonelli, *Corpus linguistics at work*, 2001:84)

Corpus driven

- provides new paradigm of teaching/learning
- students as a researcher
- data driven learning
- learn how to use concordance + corpora
- extract generalization from data
- Is it possible?

**Corpus-based Translation**

- Theoretical issues: Descriptive Translation Studies: Toury, Baker, Laviosa, Teubert
- Creation of *parallel corpora* or *translation corpora*
- Alignment techniques: Olivier Kraif - *Translational Compositionality* and *Maximal Resolution Alignment*

- Corpora as a resource for translation
- Parallel corpora / Translation memory
  - Provide examples of translation
  - TM software detect the most likely translation
- Native corpora
  - Help editing translation to be native-like
  - Help understanding difficult words/concepts

- Many experiments confirm that
  - Native corpora is useful for selecting the appropriate translation check whether that translation is possible; if > 1 translation choice, select the most occurrence
  - Native corpora help understanding the source text
- Translation school should teach students how to use corpora as a resource for solving translation problems.

**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?
- Help answering questions of usage easily
  - More than one character *is/are*
  - *Worth to* do / *worth* do*ing*

- Is it *sheer* a synonym of *pure, complete, utter* and *absolute*?
- How would you translate "*assolutamente*" or "*corretto*" into English?

**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.

**Corpus Linguistics : Some basic notions**

- Concordance / Concordancer
- Collocation (Lexis)
- Colligation (Grammar)
- Semantic Preference (Semantics)
- Discourse Prosody (Pragmatics)

- Paradigmatic and Syntagmatic Dimensions
- Lexico-grammar approach
- Idiom principle vs. open-choice principle
- Phraseological tendency vs. terminological tendency
- Pattern (grammar)
- Extended units of meaning
- Cultural Keywords

**Concordance / Concordancer**

*Concordance*

A term that signifies a list of a particular word or sequence of words in a context. The concordance is at the centre of corpus linguistics, because it gives access to many important language patterns in texts. Concordances of major works such as the Bible and Shakespeare have been available for many years. The computer has made concordances easy to compile.

The computer-generated concordances can be very flexible; the context of a word can be selected on various criteria (for example counting the words on either side, or finding the sentence boundaries). Also, sets of examples can be ordered in various ways. See Sinclair 1991: Ch. 2; McEnery and Wilson 1996: Ch. 1; Collier 1994; Kaye 1990; Hockey and Martin 1988.

*Concordancer*: http://www.federicozanettin.net/sslmit/cl.htm

**Collocation**

You shall know a word by the company it keeps (Firth 1957:179)

We may use the term node to refer to an item whose *collocations* we are studying, and we may define a span as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call *collocates*. (Sinclair 1966:415)

*Collocates* are the words which occur in the neighbourhood of your search word (Scott 1999 WordSmith Help File).

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: 24).

collocates …node …collocates
---------- span ----------

**Colligation**
Colligation can be defined as 'the grammatical company a word keeps and the position it prefers': in other words, a word's *colligations* describe what it typically does grammatically (Hoey 2000:234)

Knowledge of a collocation, if it is to be used appropriately, necessarily involves knowledge of the *patterns* or *colligations* in which that collocation can occur acceptably (Hargreaves 2000:214).

**Semantic Preference**
Semantic preference is the relation, not between individual words, but between a lemma of word-form and a set of semantically related words, and often it is not difficult to find semantic label for the set. […] [An] example is the word-form *large*, which often co-occurs with words for "quantities and sizes". (Stubbs 2002: 65)

**Semantic or Discourse Prosody**
A discourse prosody is a feature which extends over more than one unit in a linear string. […] Discourse prosodies express speaker attitude (Stubbs 2002: 65)

'the consistent aura of meaning with which a form is imbued by its collocates' … prosodies based on very frequent forms can bifurcate into 'good' and 'bad', using a grammatical principle like transitivity in order to do so. For example, where *build up* is used transitively, with a human subject, the form of the prosody is uniformly good … Where things or forces, such as *cholesterol, toxins,* and *armaments build up* intransitively, of their own account, they are uniformly bad. (Louw 1993:171)

**Phraseological tendency vs. Terminological tendency**
Sinclair puts phraseology at the heart of language description, arguing that the tendency of words to occur in preferred sequences has three important consequences which offer a challenge to current views about language:

1. There is no distinction between pattern and meaning;
2. Language has two principles of organization: the idiom principle and the open-choice principle;
3. There is no distinction between lexis and grammar.

**1. There is no distinction between pattern and meaning**

- Different meanings for a word tend to be used in different grammatical patterns:
  - "Maintain something"
  - "Maintain that something is true"
  - "Maintain something at a level"

- Different grammatical patterns tend to collect words with similar meanings
    - VERB one's way (in)to: *bribe, bully, cheat, fiddle, hustle, insinuate, trick, wrangle…*

**2. Language has two principles of organization: the idiom principle & the open-choice principle**

The *open-choice principle* "is a way of seeing language text as the result of a very large number of complex choices. […] This is probably the normal way of seeing and describing language. It is often called a 'slot-and-filler' model, envisaging texts as a series of slots which have to be filled from a lexicon which satisfies local restraints." (Sinclair 1991: 109)
These restraints are mainly *grammatical*.

But words "do not occur at random in a text"
"The choice of one word affects the choice of others in its vicinity. Collocation is one of the patterns of mutual choice, and idiom is another. The name given to this principle of organization [*of language*] is the *idiom principle*." (Sinclair 1991: 173)
In other words, "the language user has available to him a large number of preconstructed or semi-preconstructed phrases that constitute single choices, even though they appear to be analysable into segments". (Sinclair, quoted in Partington 1998: 19)

**The idiom principle**

- Idioms: *to get a frog in one's throat* vs. **to get an ugly frog in one's throat*
- Examples of idiomaticity: *Of course* (= *insofar as*)
- Phrases allowing internal lexical variation: *In some cases / in some instances / set x on fire / set fire to x*
- Phrases allowing internal syntactic variation: *It's not in his nature to …*
    - The verb tense can vary (*was*) or a modal may be introduced;
    - The negative *not* can be substituted with another negative (*hardly*)
    - The possessive *his* can be substituted with *my, your, 's*
- Phrases allowing some variation in word order:*to recriminate is not in his nature* vs. *it is not in the nature of an academic to*
- Words and phrases showing a tendency to co-occur with certain grammatical choices: *set about* (=inaugurate)

**There is no distinction between lexis and grammar**

- To know a word is to know how to use it
- Certain grammar attracts certain words
- Grammatical words like *a* and *the* are often used in phrases rather than being used independently
    - *A free hand* vs. *her free hand*
    - *Hurt his leg* vs. *hit someone in the leg*
    - *Turn her face* vs. *a slap in the face*