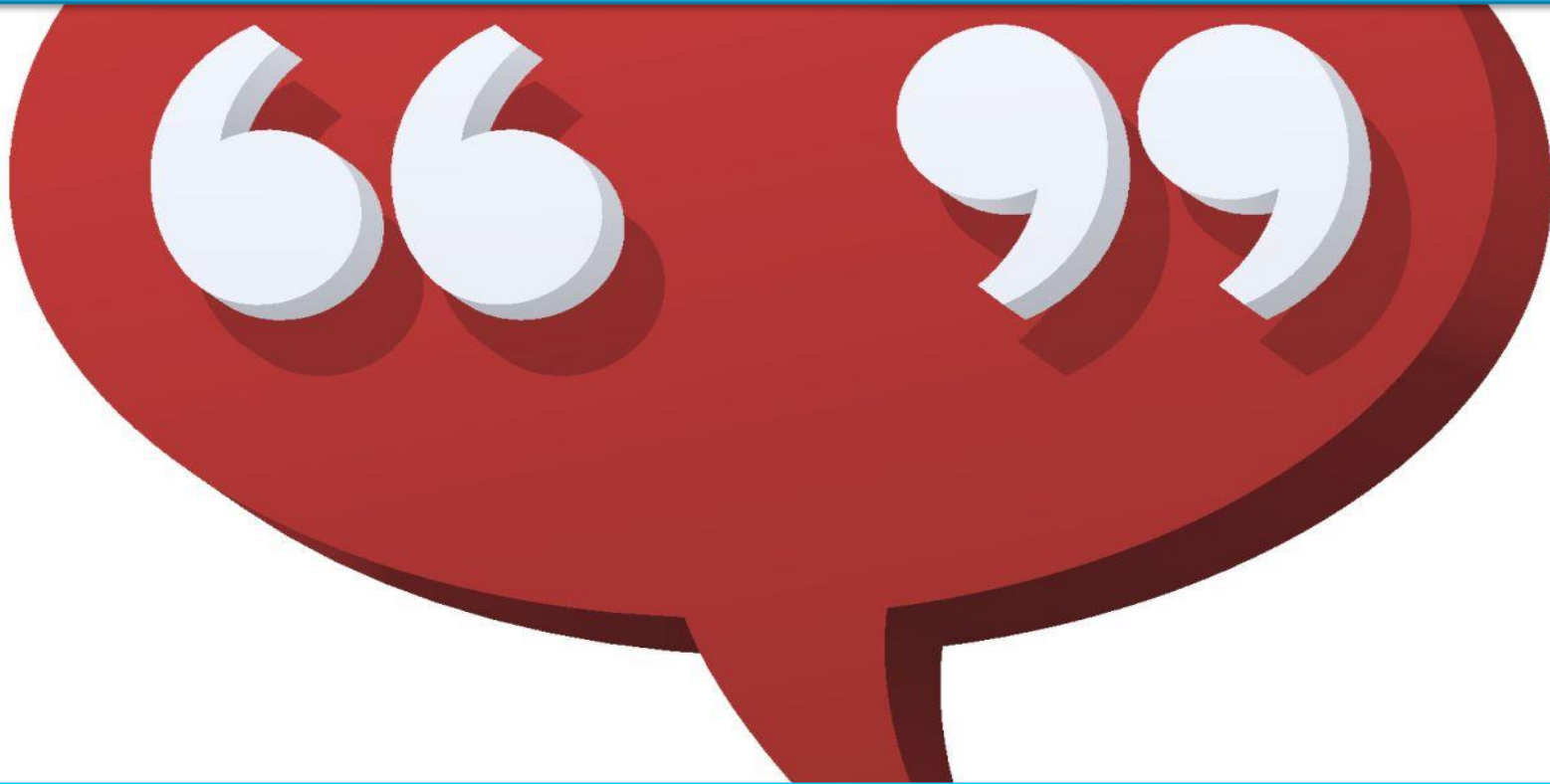


CORPUS



Dr. Riaz Hussain

Overview

- What is corpus ?
- The accurate picture of Language
- Tracking English language as a Whole
- Different Types of Corpora
- Difference b/w Dictionary & Corpus
- WHAT IS THE USE OF CORPUS?
- APPLICATIONS OF CORPUS
- CONCORDANCE & FREQUENCY COUNTS

What is Corpus ?

- A book ?
- An Article ?
- An Archive ?



What is Corpus ?

- A corpus is a collection of texts of written (or spoken) language presented in electronic form. It provides the evidence of how language is used in real situations, from which lexicographers can write accurate and meaningful dictionary entries. The plural form of corpus is corpora. Some popular corpora are [British National Corpus](#) (BNC), COBUILD/Birmingham Corpus, IBM/Lancaster Spoken English Corpus.

Recording Dynamic Aspects

- Corpus is at the heart of dictionary-making in the 21st century and ensures that lexicographers can track and record the very latest developments in language today. By analysing the corpus and using special software, they can see words in context and find out how new words and senses are emerging, as well as spotting other trends in usage, spelling, world English, and so on.

The fullest, most accurate picture of the language

- Corpus gives us the fullest, most accurate picture of the language today. It represents all types of English, from literary novels and specialist journals to everyday newspapers and magazines, and even the language of blogs, emails, and Internet message boards. And, as English is a global language, used by an estimated one third of the world's population, English Corpora contain language from all parts of the world – not only from the UK and the United States but also from Ireland, Australia, New Zealand, the Caribbean, Canada, India, Singapore, and South Africa.

The Largest lexical corpus in the world.

- For instance , the Oxford English Corpus claims to contain over 2.5 billion words of real 21st-century English. It claims that this is the largest lexical corpus in the world.

Tracking English language as a Whole

- Meanings of words and phrases change, and so do spellings, despite the existence of 'standard' or 'correct' spelling. A strength of the corpus is that it contains not only published works in which the text has been edited (and made to conform to standard spellings and grammar) but also unpublished and unedited writing like emails and blogs. Some of the most inventive uses and deliberate exploitations of language (as well as genuine mistakes) start out in this kind of informal and unselfconscious language, so tracking them is an essential part of tracking the language as a whole.

Difference b/w Dictionary & Corpus

- A **dictionary** is an abstract representation of the language, in which we express **differences** of meaning but are not engaged with specifics of **differences** of form.
The **corpus** is at the other end of the scale: the **differences** of form are immediately present but **differences** of meaning can only be inferred.

Difference b/w Dictionary & Corpus

- A corpus is an arbitrary sample of language, whereas a dictionary aims to be a systematic account of the lexicon of a language. Children learn language through encountering arbitrary samples, and using them to build systematic representations. These observations suggest a relationship between corpus and dictionary in which the former is a provisional and dispensable resource used to develop the latter.

Difference b/w Dictionary & Corpus

DICTIONARY

A dictionary presents idealized language contributed by linguists and lexicographers. These people contribute and compile language and prescriptive examples of language. A dictionary is like a cookie-cutter.

CORPUS

A real-life language is more dynamic , richer and varied than the idealized language.

A corpus contains real-life language. A corpus is like a collection of cookies if we were to use a metaphor.

Different Types of Corpora

- Monolingual corpora represent only one language while bilingual corpora represent two languages. European Corpus Initiative (ECI) corpus is multilingual having 98 million words in Turkish, Japanese, Russian, Chinese, and other languages. The corpus may be composed of written language, spoken language or both. Spoken corpus is usually in the form of audio recordings. A corpus may be open or closed. An *open corpus* is one which does not claim to contain all data from a specific area while a *closed corpus* does claim to contain all or nearly all data from a particular field. *Historical corpora*, for example, are closed as there can be no further input to an area.

WHAT IS THE USE OF CORPUS?

- Using the corpus enables lexicographers to examine a word in detail by looking at all the different contexts in which it occurs.

WHAT IS THE USE OF CORPUS?

- A corpus provides grammarians, lexicographers, and other interested parties with better descriptions of a language. Computer-based corpora allow linguists to adopt the principle of total accountability, retrieving all the occurrences of a particular word or structure for inspection or randomly selected samples. Corpus analysis provide lexical information, morpho-syntactic information, semantic information and pragmatic information.
- Linguistic information is provided by concordance and frequency counts.

WHAT IS CONCORDANCE?

- Concordances are listings of the occurrences of a particular feature or combination of features in a corpus. Each occurrence found (or hit) is displayed with a certain amount of context, the text preceding and following it. The most commonly used concordance type is KWIC which stands for Key Word In Context. It shows one hit per line of screen or print-out with principal search feature (or focus) highlighted in the centre. Concordance is used to determine the syntax in which a form is embedded. Concordances can be generated with Corpus Presenter and with Corpus Presenter Flash, programs allow one to retrieve the contexts in which a word occurs

FREQUENCY COUNTS

Frequency Counts the number of hits. Frequency counts require finding all the occurrences of a particular feature in the corpus. So it is implicit in concordance. Software is used for this purpose. Frequency counts can be explained statistically.

APPLICATIONS OF CORPUS

- Corpora are used in the development of NLP tools. Applications include spell-checking, grammar-checking, speech recognition, text-to-speech and speech-to-text synthesis, automatic abstraction and indexing, information retrieval and machine translation. Corpora also used for creation of new dictionaries and grammars for learners.

Thank you