sketchengine.eu

Corpus types: monolingual, parallel, multilingual

5-6 minutos

What is a corpus?

A text corpus is a very large collection of text (often many billion words) produced by real users of the language and used to analyse how words, phrases and language in general are used. It is used by linguists, lexicographers, social scientists, humanities, experts in natural language processing and in many other fields. A corpus is also be used for generating various language databases used in software development such as predictive keyboards, spell check, grammar correction, text/speech understanding systems, text-to-speech modules and many others.

Types of text corpora

A text corpus can be classified into various categories by the source of the content, metadata, the presence of multimedia or its relation to other corpora. The same corpus can fall into more than one category if it fulfils the criteria for more categories.

Monolingual corpus is the most frequent type of corpus. It contains texts in one language only. The corpus is usually tagged for parts of speech and is used by a wide range of users for various tasks from highly practical ones, e.g. checking the correct usage of a word or looking up the most natural word combinations, to scientific use, e.g. identifying frequent patterns or new trends in language. Sketch Engine contains hundreds of monolingual corpora in dozens of languages.

see also What can Sketch Engine do? and Build your own corpus

A parallel corpus consists of two monolingual corpora. One corpus is the translation of the other. For example, a novel and its translation or a translation memory of a CAT tool could be used to build a parallel corpus. Both languages need to be aligned, i.e. corresponding segments, usually sentences or paragraphs, need to be matched. The user can then search for all examples of a word or phrase in one language and the results will be displayed together with the corresponding sentences in the other language. The user can then observe how the search word or phrase is translated.

see also Parallel / Bilingual Concordance and Build a parallel corpus

A multilingual corpus is very similar to a parallel corpus. The two terms are often used interchangeably. A multilingual corpus contains texts in several languages which are all translations of the same text and are aligned in the same way as parallel corpora. Sketch Engine allows the user to select more than two aligned corpora and the search will display the translation into all the languages simultaneously. When only two languages are selected, a multilingual corpus behaves as a parallel corpus. The user can also decide to work with one language to use it as a monolingual corpus.

The terms *parallel* and *multilingual* are sometimes used

interchangeably.

see also Parallel / Bilingual Concordance

A comparable corpus is a set of two or more monolingual corpora whose texts relate to the same topic., however, they are not translations of each other, and therefore, there are not aligned. When users search these corpora they can use the fact, that the corpora also have the same metadata. An example of comparable corpora in Sketch Engine is CHILDES corpora or various corpora made from Wikipedia.

see comparable corpora CHILDES corpora and corpora from Wikipedia

A learner corpus is a corpus of texts produced by learners of a language. The corpus is used to study the mistakes and problems learners have when learning a foreign language. Sketch Engine allows for learner corpora to be annotated for the type of error and provides a special interface to search either for the error itself, for the error correction, for the error type or for a combination of the three options.

see also Setting up a learner corpus

A diachronic corpus is a corpus containing texts from different periods and is used to study the development or change in language. Sketch Engine allows searching the corpus as a whole or only include selected time intervals into the search. In addition, there is a specialized diachronic feature called Trends, which identifies words whose usage changes the most of the selected period of time.

see also Trends – diachronic analysis

In addition, any of the above types of corpora can be:

A specialized corpus contains texts limited to one or more subject areas, domains, topics etc. Such corpus is used to study how the specialized language is used. The user can create specialized subcorpora from the general corpora in Sketch Engine.

see Build a subcorpus

A multimedia corpus contains texts which are enhanced with audio or visual materials or other type of multimedia content. For example, the spoken part of British National Corpus in Sketch Engine has links to the corresponding recordings which can be played from the Sketch Engine interface.

see BNC audio