<https://en.wikipedia.org/wiki/Text_corpus>

Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora. There are two main types of parallel corpora which contain texts in two languages. In a translation corpus, the texts in one language are translations of texts in the other language. In a comparable corpus, the texts are of the same kind and cover the same content, but they are not translations of each other.[1] To exploit a parallel text, some kind of text alignment identifying equivalent text segments (phrases or sentences) is a prerequisite for analysis. Machine translation algorithms for translating between two languages are often trained using parallel fragments comprising a first language corpus and a second language corpus which is an element-for-element translation of the first language corpus.[2]

In order to make the corpora more useful for doing linguistic research, they are often subjected to a process known as annotation. An example of annotating a corpus is part-of-speech tagging, or POS-tagging, in which information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags. Another example is indicating the lemma (base) form of each word. When the language of the corpus is not a working language of the researchers who use it, interlinear glossing is used to make the annotation bilingual.

Some corpora have further structured levels of analysis applied. In particular, a number of smaller corpora may be fully parsed. Such corpora are usually called Treebanks or Parsed Corpora. The difficulty of ensuring that the entire corpus is completely and consistently annotated means that these corpora are usually smaller, containing around one to three million words. Other levels of linguistic structured analysis are possible, including annotations for morphology, semantics and pragmatics.

Corpora are the main knowledge base in corpus linguistics. The analysis and processing of various types of corpora are also the subject of much work in computational linguistics, speech recognition and machine translation, where they are often used to create hidden Markov models for part of speech tagging and other purposes. Corpora and frequency lists derived from them are useful for language teaching. Corpora can be considered as a type of foreign language writing aid as the contextualised grammatical knowledge acquired by non-native language users through exposure to authentic texts in corpora allows learners to grasp the manner of sentence formation in the target language, enabling effective writing.