Linguistic Resources

Corpus

A corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting. Its plural is corpora. They can be derived in different ways like text that was originally electronic, transcripts of spoken language and optical character recognition, etc.

Elements of Corpus Design

Language is infinite but a corpus has to be finite in size. For the corpus to be finite in size, we need to sample and proportionally include a wide range of text types to ensure a good corpus design.

Let us now learn about some important elements for corpus design −

Corpus Representativeness

Representativeness is a defining feature of corpus design. The following definitions from two great researchers − Leech and Biber, will help us understand corpus representativeness −

- **According to Leech (1991),** "A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety".
- **According to Biber (1993),** "Representativeness refers to the extent to which a sample includes the full range of variability in a population".

In this way, we can conclude that representativeness of a corpus are determined by the following two factors −

- **Balance** − The range of genre include in a corpus
- **Sampling** − How the chunks for each genre are selected.

Corpus Balance

Another very important element of corpus design is corpus balance – the range of genre included in a corpus. We have already studied that representativeness of a general corpus depends upon how balanced the corpus is. A balanced corpus covers a wide range of text categories, which are supposed to be representatives of the language. We do not have any reliable scientific measure for balance but the best estimation and intuition works in this concern. In other words, we can say that the accepted balance is determined by its intended uses only.

Sampling

Another important element of corpus design is sampling. Corpus representativeness and balance is very closely associated with sampling. That is why we can say that sampling is inescapable in corpus building.

- According to **Biber(1993)**, “Some of the first considerations in constructing a corpus concern the overall design: for example, the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples from within texts, and the length of text samples. Each of these involves a sampling decision, either conscious or not.”

While obtaining a representative sample, we need to consider the following −

- **Sampling unit** − It refers to the unit which requires a sample. For example, for written text, a sampling unit may be a newspaper, journal or a book.
- **Sampling frame** − The list of all sampling units is called a sampling frame.
- **Population** − It may be referred as the assembly of all sampling units. It is defined in terms of language production, language reception or language as a product.

Corpus Size

Another important element of corpus design is its size. How large the corpus should be? There is no specific answer to this question. The size of the corpus depends upon the purpose for which it is intended as well as on some practical considerations as follows −

- Kind of query anticipated from the user.
- The methodology used by the users to study the data.
- Availability of the source of data.

With the advancement in technology, the corpus size also increases. The following table of comparison will help you understand how the corpus size works −

Year	Name of the Corpus	Size (in words)
1960s - 70s	Brown and LOB	1 Million words
1980s	The Birmingham corpora	20 Million words
1990s	The British National corpus	100 Million words

Early 21st century	The Bank of English corpus	650 Million words