Definition of Corpus

"A corpus is a representative collection of a fairly large number of examples (or, texts) that share similar contextual or situational characteristics. These texts are then analyzed collectively in order to understand how language is used in these different contexts."

(Crawford and Csomay, 2016, 6)

Text vs. Corpus

Table 2.1 A qualitative comparison of a text versus a corpus

A text	A corpus
read whole	read fragmented
read horizontally	read vertically
read for content	read for formal patterning
read as a unique event	read for repeated events
read as as individual act of will	read as a sample of social practice
instance of *parole*	gives insights into *langue*
coherent communicative event	not a coherent communicative event

19

Tognini-Bonelli, E. (2020). Theoretical overview of the evolution of corpus linguistics. In A. O'Keeffee and M. McCarthy (Eds.): Routledge Handbook of Corpus Linguistics, p. 19.

Group Work: Corpora

- Familiarise yourself with your specific corpus type (here: BNC, COCA, Malaysian Wikipedia, Works of Dickens)

- Outline characteristics

- Find possible areas of research where it might be suitable

- Think of one specific research question that would be possible to answer with this corpus

Then partner with another group member, explain your corpus type and find similarities and differences. Compare your research fields and discuss each other's research question.

Different Corpora

- **General corpora**: consists of general texts, no single text type; e.g., BNC

- **Reference corpora**: attempts to represent nature of whole language

- **Sublanguage corpora**: from a particular variety
    - **Geographical (**e.g., variations): e.g., The International Corpus of English (ICE) on English varieties
    - **Temporal**: Sample corpora: e.g. The Brown Corpus; diachrone – synchrone

- **Parallel corpora**: texts exact translations in more than one language

- **Comparable corpora**: collection of similar texts; e.g., weather broadcasts

- **Specialised corpora**: very specific body of texts, maybe smaller

Types of Corpus Research

- Corpus-**driven**: inductive process, corpora are investigated from the bottom up (no prior hypothesis) and patterns found therein are used to explain linguistic regularities and exceptions of the language variety/genre

- Corpus-**based**: deductive process, corpora are used to test preformed hypotheses or exemplify existing linguistic theories

- Corpus-**informed**: combining both methods, consult a corpus & various text types when making a decision (e.g., translation)

Research Questions

Should be:

- **clear**: enough specifics that one's audience can easily understand its purpose without needing additional explanation;

- **focused**: narrow enough that it can be answered thoroughly in the space available;

- **concise**: expressed in the fewest possible words;

- **complex**: not answerable with a simple "yes" or "no," but requires analysis of ideas and sources prior to composition of an answer;

- **arguable**: potential answers are open to debate rather than accepted facts;

- **open**: does not ask for „yes" or „no", but is an open-ended question;

Research Questions

What is something that you are passionate about?

Please try to formulate your research question for our next meeting.