Steps involved in corpus creation

- **Documentation –** very important at every stage
- **Selection of Texts –** purpose of corpus? Availability?
  - External criteria (features of texts) or internal criteria (language features within texts)? “Corpus should be designed and constructed exclusively on external criteria” (Clear, 1992)
  - External criteria: Language(s) / variety(-ies); Mode (speech, writing, electronic?);Text type (written: book, article, etc. spoken: lecture, seminar, etc.); Genre (academic, journalistic, etc.); Domain (sport, business, politics, etc); Vintage (date of recording/publication of texts); Others: language proficiency, gender, age of speaker/writer…
- **Copyright permission –** for corpus distribution (not for personal use/research)
- **Text Acquisition –** eg download from web (+ metadata); Amount? Subcorpora?
- **Conversion - to plain text*** **+ Cleanup -** remove non-text

Corpus analysis

- **Software: plain text*, tokenization, indexing**
  - e.g. AntConc, WordSmith Tools, etc
- **Outputs:**
  - 1. Word Frequency Lists - if a language feature occurs frequently, it must be significant in some way
  - 2. Concordances – context (+ sorting, expansion)
  - 3. Collocations – lists, patterns
  - 4. N-grams (multi-word sequences)
  - 5. Keywords [NOT for this presentation]
- **Antconc software used for this presentation**