Building and annotating a corpus.
Introduction

• Building a corpus:
  – basic concepts and terminology;
  – sampling and representativeness;
  – handling and cleaning data.

• Annotation and mark-up:
  – annotation vs. mark-up;
  – adding and using annotation and mark-up.
Building a corpus

• In this session, we will consider:
  – Representativeness
  – Sampling
  – Size
What is a corpus?

[...] a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration.

(McEnery and Wilson 2001:32).
What is a corpus?

[...] a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

Sinclair (2005)
Representativeness

• “maximally representative”
• “represent, as far as possible”
• Corpora can be a ‘representative sample’ of a language or a language variety.
• Partly to do with size ...
• ... but also depends a lot on design of corpus.
Representativeness

- A corpus, no matter how big, will only ever be a sample of a particular language variety.

- The idea (in theory at least):
  
  collect enough samples of a language variety to **represent** the *entirety* of that language variety.
Representativeness

• But what does ‘represent’ mean?

**Proportionally**, there should be the same quantity of any linguistic item we care to examine (for example, a particular word or grammatical part-of-speech) in the corpus as there is in the language variety as a whole.

Any questions???
Representativeness

• Issues:
  – The only way to be absolutely sure a particular linguistic item occurs proportionally the same number of times in a corpus as it does in the entire population of the language variety is to count the number of occurrences in each.
Issues:

- It seems unlikely that a whole language variety (such as British English) could be represented by a small sample of texts.

A corpus might be representative of some common language features, but less common features might be completely absent from the corpus, so not represented at all.

Chomsky’s (1962) famous paper criticising corpus linguistics
Representativeness

• We can never know for sure whether any corpus we build is representative of the language variety it contains or not.

• We can only strive to be as representative as is practically possible.
  – This involves:
    • knowledge of the language variety under investigation
    • careful corpus building
A representative sample

- Corpus building endeavours to gather together enough samples of a language variety (i.e. texts) to adequately represent it.

- **Sampling** is crucially important in corpus building.
A representative sample

• Sampling frame
  – working out what your corpus needs to contain
A representative sample

- Sampling frame
  - What language or language variety is being studied? (e.g. English, Australian English, Yorkshire English)

What is the location of the texts you need to collect? (e.g. UK, Australia, Yorkshire, Holmfirth.)
A representative sample

- Sampling frame
  - What is the mode of the texts? (spoken or written, or both)
  - What sorts of texts (text-types) need to be included? (e.g. newspaper articles, short stories, letters, text messages, spoken conversation)
A representative sample

• Sampling frame
  – What domain will the texts come from? (e.g. academic, legal, business)
  – What is the production/publishing date of the texts you want? (This might be one day, one year, or a span of years)
A representative sample

• The answers to such questions become the sampling parameters.

• These help to identify the sorts of texts that you will eventually collect to populate your corpus.

• The number of parameters needed depends largely on the nature of your corpus and the language variety that is being represented.
A representative sample

• Typically, the questions (and hence the sampling parameters) will start off very broad and become gradually more focused.

• Consider a corpus of British English ...
Sampling – British English?

British English

2012

Written

Books

Fiction

Non-fiction

periodicals

Newspapers

Magazines

Educational

Contextual

Business

Spoken

Everyday

Male

Female
A representative sample

• We have the components – what else do we need to consider?
  – Numbers of texts for each component
  – Sizes of texts
  – Whole texts or extracts?
A representative sample

• This is the ‘how much?’ question.
• There is no simple answer.
• Various considerations:
  – Language variety
  – Practical issues to do with data collection
  – Amount of time available
A representative sample

• Consider a corpus of text messages ....
• What would you do?
  – Many different respondents
  – Age, gender, social context, recipient
  – How many texts / respondent?
  – Practical issues?
  – Ethical issues?
A representative sample

• Distribution of texts across the corpus
  – Should the number of texts collected for each component be the same, or reflect the real-life distributions?
A representative sample

• Whole texts or extracts or both?
• Sinclair (2005) suggests that whole texts should be used regardless of size difference.
• Issues?
A representative sample

- Whole texts.
  - Texts vary in size
    - novel vs. text message
  - Large texts might skew the results obtained from the corpus.
    - Make the corpus bigger?
    - Use extracts from larger texts?
A representative sample

• Extracts
  – Any part of a text cannot represent the whole text
  – Position within a text affects language choices
    ....
A representative sample

Dear John ...
In this chapter we will ...
Fire fighters battle mile-wide moorland inferno
Can I speak to the homeowner?
Once upon a time ...

Yours sincerely,
In summary, this chapter ...
The cause of the fire has still not been established
Thank you for you time
... and they all lived happily ever after.

• A corpus of beginnings or ends will not adequately represent that text-type.
  – Sample from beginnings, middles and ends of texts?
  – Use random sampling?
Sampling

• What is a text?
  – Sinclair (2005): a text is a communicative event.
    • What about a bulletin board?
    • A Twitter feed?
    • A Facebook page?
    • Or even a plain old conversation with a group of friends?

Where do some texts (or communicative events) begin and end?
Sampling

- When building a corpus we aim to collect enough samples of a language variety to represent the *entirety* of that language variety.
- Having an appreciation of what the entire population consists of will help to answer the ‘how much?’ and ‘how many?’ questions.
- Giving careful thought to a componential structure will help with this.
Corpus size

• Corpus size (usually measured in words)
  – Either ...
    • decide the size first and the size of the components is divided to fit into that fixed maximum size;
  – or ...
    • the size results from decisions made about the number and size of the components.
Corpus size

• The approach used:
  – should be decided in the design stage;
  – will have implications for the sampling.
  • A small fixed size could mean using extracts of texts
Corpus size

• Example: Brown and LOB corpora.
  – Fixed size of 1M words
  – Attempts to represent written AmE and BrE
  – Fiction sections set to 250,000 words
  – They use 2000 word extracts from 126 works of fiction.
  – Sampling of this nature requires careful management
Corpus size

• Whatever the approach, in order to adequately represent each component of the corpus, and thus the language variety being investigated, you need:
  – sufficient words ...
  – ... spread over enough different texts
When it comes to building corpora:

- there are a number of theoretical issues...
- ...and a number of practical issues.
- One (corpus) size does not fit all.
- Knowledge of the population is important when considering how to represent it... and how to sample from it.

Keep in mind GIGO (or RIRO?):
Garbage in, garbage out!