3. Lexicographic evidence

1. What makes a dictionary reliable?

- generalizations about word behaviour approximate to the ways in which people normally use language
- subjective evidence:
    - introspection (own mental lexicon)
        - one individual's store of linguistic knowledge is incomplete, idiosyncratic
    - informant-testing
- objective evidence: what we learn by observing language in use

Scope of the dictionary:
- 'our focus must be the probable, not the possible'
- how do we know something is typical and not idiosyncratic?
    - frequent
    - well-dispersed

2. Citations

until 1980 it is the main form of empirical language data
'reading programme': organized data-gathering exercise

- pros:
    - new vocabulary does not necessarily mean new words (compounds, multiword expressions, novel use of existing words)
    - specific subject field or particular dialect
- cons:
    - labour-intensive
    - human readers tend to notice what is remarkable and ignore what is typical (bias)

3. Corpora

John Sinclair: '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'

lexicographic corpus: a collection of language data designed specifically for use in the creation of dictionaries

there is no such thing as a perfect corpus for lexicography

- the corpus is a sample (just a subset of all the communicative events of the language)
- the corpus does not favour 'high quality' language
- pragmatism and compromise (need for compromise: design, data-collection, encoding)
    - various non-linguistic factors may force us to change our minds

4. Corpora: design issues

Decisions about:

- how large it will be,
- which broad categories of text it will include,
- what proportions of each category it will include,
- which individual texts it will include.

Zipf's Law:

- word frequency
- 'a few words occur with very high frequency while many words occur but rarely'
- the frequency with which a word appears in a collection of texts is inversely proportional to its ranking in a frequency table
- strong correlation between the word's frequency and its complexity

Content:

- goal: a corpus whose constituent texts are drawn from a wide range of sources
- there is no obvious way of creating a 'representative' corpus of a widely used living language, because:
    - impossible to define the population that the corpus should be representative of
    - since the population is unlimited, it is impossible to establish 'correct' proportions of each component
- 'balanced' corpus:
    - reflects the diversity of the target language, by including texts which collectively cover the full repertoire of ways in which people use the language

    - if every text is carefully described in terms of its key features, corpus-users will have the information they need to assess the significance of any given instance of a word, phrase
- internal properties of texts (linguistic or stylistic features)
    - noun + preposition sequences are more common in technical writing than in fiction
- external properties of texts (situational or functional attributes - such as newspaper, novel, instruction manual, conversation)
- spoken data:
    - ‘demographic’ approach to collect samples of ordinary conversations
    - context-governed component of the corpus
- ‘skewing’: form of bias in data where a particular feature is over- or under-represented
- attributes a text can have:
    - language: mono/bi/multilingual?
        - parallel corpus: a set of corpora in which the texts in language A corresponds in some way to those in language B
        - two types of parallel corpus:
            - translation corpus: translated version of the same text (EU documents)
            - comparable corpus: identical sampling frame
    - time: synchronic/diachronic?
    - mode: written / spoken / both
    - medium: channel in which the text appears
        - print media
        - spoken media
        - web: blogs, social networking, and newspapers/conference proceedings published online
    - domain: subject matter of the text
    - sublanguages: ‘core’ usages vs. ‘sublanguages’

5. Collecting corpus data

- written data:
    - electronic form (for synchronic corpora) is rarely a problem
    - scanning, keyboarding
- spoken data:
    - contemporary language
    - difficult and expensive

    - recording, transcription
    - speech-recognition technology
- from the web:
    - a source of texts from which a lexicographic corpus can be assembled
    - Oxford English Corpus - the first lexicographic corpus sourced entirely from the web
- copyright and permissions:
    - knowing who owns the copyright of each text to be included
    - short explanation about what a corpus is, how and why people use it

6. Processing and annotating the data

- clean-up, standardization, text encoding
    - wide range of sources, input texts can differ
    - generally accepted standard: XCES (XML Corpus Encoding Standard)
    - removing parts of the content? (acknowledgements, copyright information, tables, etc)
    - encoding: tokenization, marking textual structure, lemmatization
        - mark-up: enriching raw data by adding information of various kinds
        - tokenization: identifying all the tokens, hyphenation and apostrophes can be ambiguous
        - sentences: end with . (or: ?, !, ’), full stops don’t always signal sentence boundaries
        - lemmatization: headwords are generally lemmas (like *permit*, not *permitted*)
- documentation (whatever information the user might need about a text)
- linguistic annotation
    - POS-tagging (automatically assigning every word in the corpus to a wordclass)
    - parsing (not necessary)

7. Corpus creation

- lexicographers prefer size to granularity
- size/granularity trade-off in 3 areas:
    - text-selection parameters
    - level of detail in document headers
    - linguistic annotation
- no such thing as a ‘perfect’ corpus: natural language is too diverse and too dynamic
- the biggest benefit: the access it gives us to the ‘regularities’ of the language