Lexicographic evidence

Ferenczi Zsanett

2019. 02. 22.

Reliable dictionary

Reliable dictionary

- ➔ generalizations about word behaviour
- ➔ subjective evidence
  - introspection
  - informant-testing
- ➔ objective evidence
  - observing language in use

Reliable dictionary

- ➔ probable, not possible
- ➔ if something is typical,

  it is **frequent** and **well-dispersed**

  - ◆ ‘core’ of the language

Citations

Citations

- ➔ until 1980: main form of empirical language data
- ➔ reading programme:
  - organized data-gathering exercise
  - automatic methods are not that precise
  - human readers tend to notice the atypical
  - labour-intensive

Reading programme

➔ data to require

- keyword
- citation
- source of the citation
- comments

Rationalism vs empiricism

Rationalism	**Empiricism**
describe competence	describe performance
Chomsky,	lexicographers,
chomskyites	corpus linguists
introspection	observation

Corpora

Corpora

- John Sinclair
- lexicographic corpus
- sample (subset of all communicative events)
- does not favour high quality language
- compromises

Design issues

➔ size, categories of text, proportion, individual texts

Brown Corpus	1 million words	1960s
Birmingham Collection of English Text	20 million words	1980s
British National Corpus	100 million words	1990s
Oxford English Corpus	1 billion words	2000s

Zipf's Law

➔ word frequency

word form	ranking in BNC	actual frequency in BNC	frequency predicted by Zipf's Law
was	10th	923,957	–
at	20th	478,177	461,978
made	100th	91,659	92,396
advice	1000th	10,316	9,240
quiet	2000th	5,295	4,619

Content

- ➔ ‘representativeness’ is unattainable
- ➔ ‘balanced’ corpus:
  - ◆ diversity of the target language
  - ◆ every text should be carefully described

Selecting texts

Internal criteria:

- ➔ linguistic features
- ➔ stylistic features

External criteria:

- ➔ situational attributes
- ➔ functional attributes

Spoken data

- ➔ demographic approach
  - features: age, gender, social class, region, etc
- ➔ context-governed component
  - educational & informative events
  - business events
  - public events
  - leisure events

‘Skewing’

- ➔ form of bias
- ➔ a particular feature is over- or under-represented
- ➔ words can occur more frequently than other more usual words
- ➔ larger corpora are more ‘forgiving’ and less likely to be affected by skewing

Attributes

- language
- time
- mode
- medium
- domain
- sublanguages
  - *deuce* in tennis
  - *serve*, *set*, *game*?

Collecting corpus data

Collecting data

- written data
  - have to be in digital form
- spoken
  - contemporary language
  - speech-recognition
- from the web
  - separate text from all other data-types
  - hard to tell the exact provenance

Processing the data

Processing and annotating

- clean-up, standardization, text encoding
- XCES
- removing parts of the content
- transcribing
- encoding:
  - tokenization
  - marking textual structure
  - lemmatization

Processing and annotating

- ➔ textual annotation: document header
  - ◆ information including feature-values which would be used in corpus queries
- ➔ linguistic annotation
  - ◆ POS-tagging
  - ◆ parsing

Corpus creation

Corpus creation

- ➔ noise
  - irrelevant data can be rapidly discounted
- ➔ lexicographers prefer size to granularity
- ➔ no such thing as ‘perfect’ corpus
- ➔ the access it gives us to the ‘regularities’ of the language