Superficial & Lexical level 1

- Superficial level
- What is a word
- Lexical level
- Lexicons
- How to acquire lexical information

Superficial level 1

- Textual pre-process
  - Getting the document(s)
    - Accessing a BD
    - Accessing the Web (wrappers)
  - Getting the textual fragments of a document
    - Multimedia documents, Web pages, ...
  - Filtering out meta-information
    - tags HTML, XML, ...

Superficial level 2

- Text segmentation into paragraphs or sentences
- Tokenization
  - orthographic vs grammatical word
  - Multiword terms
  - dates, formulas, acronyms, abbreviations, quantities (and units), idioms,
  - Named entities
    - NER, NEC, NERC
  - Unknown word
- Language identification

Beeferman et al, 1999
Ratnaparkhi, 1998

Bikel et al, 1999
Borthwick, 1999
Mikheev et al, 1999

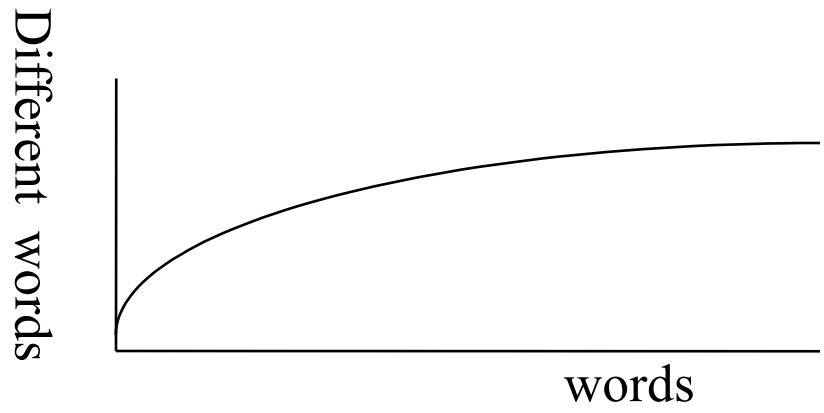
Elworthy, 1999
Adams,Resnik, 1997

Superficial level 3

- Vocabulary size (V)
  - Heap's Law
    - $V = KN^{\beta}$
    - K depends on the text $10 \leq K \leq 100$
    - N total number of words

    $\forall \beta$ depends on the language, for English $0.4 \leq \beta \leq 0.6$
  - Vocabulary grows sublinealy but does not saturate

$\forall \beta$ tends to stabilize for 1Mb of text (150.000w)



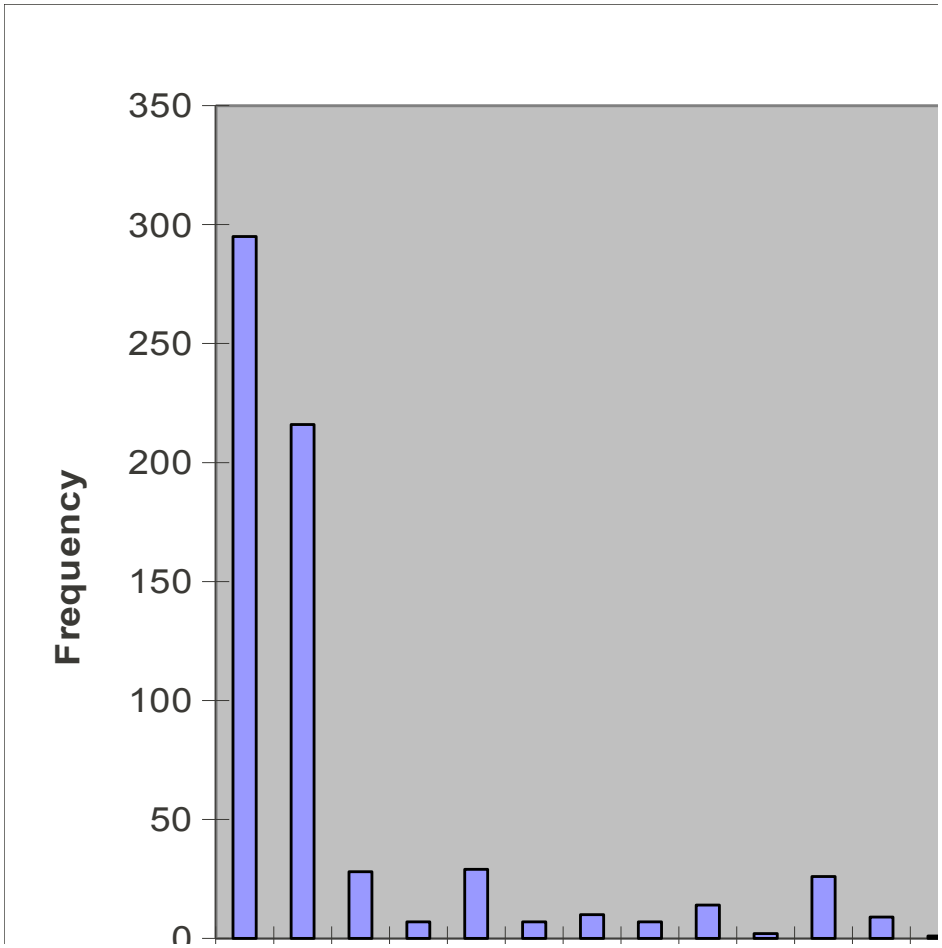
Superficial level 4

- word tokens vs word types
- Statistical distribution of words in a document
  - Obviously non uniform
  - Most common words cover more than 50% of occurrences
  - 50% of the words only occur once
  - ~12% of the document is formed by word occurring less than 4 times.

Superficial level 5

Zipf law:
We sort the words occurring in a document by their frequency. The product of the frequency of a word (f) by its position (r) is aproximatelly constant

$$f = C * 1/r$$
$$C \simeq N/10$$



Lexical level 1

- Part of Speech (POS)
  - Formal property of a word-type determining its acceptable uses in syntax.
- A POS can be seen as a class of words
- A word-type can own several POS, a word-token only one
- Plain categories
  - open, many elements, neologisms, independent and semantically rich classes
  - N, Adj, Adv, V
- Functional categories
  - closed

Lexical level 2

Lexicon

- Repository of lexical information for human or computer use
- Two aspects to consider
  - Representation of lexical information
  - Acquisition of lexical information

Lexical level 3

Lexicon content

- **Orthografic** Transcription
- **Phonetic** Transcription
- **Flexion** model
- **diathesis** alternations, **subcategorization** frames
  - LOVE VTR (OBJLIST: SN).
  - LOVE
    - CAT = VERB
    - SUBCAT = <SN, SN>

Lexical level 4

- **POS**
- **Argument structure**
- S**emantic** information
  - dictionaries => definition
  - lexicons => semantic types predefined in a hierarchy.
- **Lexical Relations**
  - derivation
- **Equivalence** with other languages

Lexical level 5

Problems

- Form
  - attribute/value pairs, binarr or n-ary relations, coded values, open domain values…
- Multiple assignments
  - One to many and many to one relations
  - Contextual dependencies …
- Facets of features
  - Mandatory or optional, cardinality, default values
- Grading
  - Exact values, preferences, probabilistic assigments.