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 databases
    - 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

Beeferman et al, 1999
Ratnaparkhi, 1998

- Tokenization
  - Orthographic vs grammatical word
  - Multiword terms
  - Dates, formulas, acronyms, abbreviations, quantities (and units), idioms,
  - Named entities
    - NER, NEC, NERC
  - Unknown word

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

- Language identification

Elworthy, 1999
Adams,Resnik, 1997

Superficial level 4

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.

- word tokens vs word types

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**
- **Semantic** 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.

Lexical level 6

Representation

- General purpose databases
- Textual databases
- Lexical databases
- Object oriented formalisms
- Object oriented databases
- Frames
- Unification-based formalisms

Lexical level 7

Lexical Information acquisition

- Dictionaries
  - Machine readable dictionaries (MRD)
  - Predefined internal structure
  - Some degree of coding in some contents
  - Internal relations (synonimy, hyponymy, ...)
  - (sometimes) restricted vocabulary
  - Systematics on building definitions

Lexical level 8

Information present in corpora

- Colocations
- Argument structure.
- Frecuency information
- Context
- Grammatical Induction
- Probabilistic Analysis.
- Lexical relations
- Examples of use.
- Selectional Restrictions
- Nominal compounds
- Idioms, ...

Lexical level 9

Corpus typology

- Raw corpus
- Tagged corpora
- Parenthized corpora
- Treebanks