Natuurlijke Taalverwerking
Natural Language Processing

Gosse Bouma and Begoña Villada

3e trimester 2002/2003

What is Natural Language Processing?

- Building applications which require knowledge of (the structure and meaning of) natural language.
- Examples:
  - Correcting spelling errors, Grammar checking
  - Text to Speech applications (*SMS naar vast*)
  - Spoken Dialogue Systems (*NS reisinformatie*)
  - Automatic Translation (`babel.altavista.com`)
  - Automatic reply to e-mails,
  - etc.

What does NLP use?

- Linguistics provides theories of
  - the structure of words (morphology)
  - pronunciation of words (phonology)
  - structure of sentences (syntax)
  - meaning of words and sentences (semantics)
  - use of language (pragmatics)

What does NLP use?

- Computer Science provides formalisms and algorithms:
  - finite state automata (for recognizing and translating strings),
  - parsers for context-free grammar,...
- AI and Information Science provide
  - Machine learning, Ontological Knowledge,
  - Encoding of large amounts of data (XML),
  - Applications (Dialogue systems, Summarization, Information extraction, ...)

Application: Find Spelling Errors

- Seems easy
  - Find a dictionary,
  - Flag words in a text that are not on the list
- But is hard
  - No list is exhaustive,
  - New words appear in the language every day

More Dictionary Statistics

- Could you build a dictionary by collecting the most frequent words from a large text collection (University of Twente):

Words	Corpus	OOV
20K	110M	6.6%
40K	145M	4.5%
60K	125M	3.6%

- OOV = *out of vocabulary rate*, number of word tokens missing in the dictionary

Some Statistics

- Dictionary Size
  - 125K (*Groene Boekje*) - 500K+ (*van Dale*).
- In a given text, up to 40% of the types may not occur in a dictionary.
  - Tokens: the number of words in a text,
  - Types: the number of different words in a text

Application: Question Answering

- Find sentences in text (web-pages) that contain the answer to a question.
  - Wie is de trainer van Ajax?
  - `Na wat plussen en minnen toonde Ajax-trainer Ronald Koeman zich gistermiddag uiteindelijk tevreden over de loting`
  - `De trainer van Ajax zei dit gisteravond in het televisieprogramma Barend en Van Dorp`
- Answer to a *who*-question should contain the name of person

What do we do in this course?

- Weeks 1-5: Words
  - ⋆ Word lists, hyphenation, text-to-speech, morphology
  - ⋆ Finite State Automata, regular expressions, Transducers

What do we do in this course?

- Weeks 6-11: Sentences
  - ⋆ Grammar-checking, dialogue systems, translation, information extraction, ...
  - ⋆ Grammar (word order, agreement, main and subordinate clauses, questions,...
  - ⋆ Context-free and Definite-clause grammar,
  - ⋆ Shift-reduce parsing, chart-Parsing.