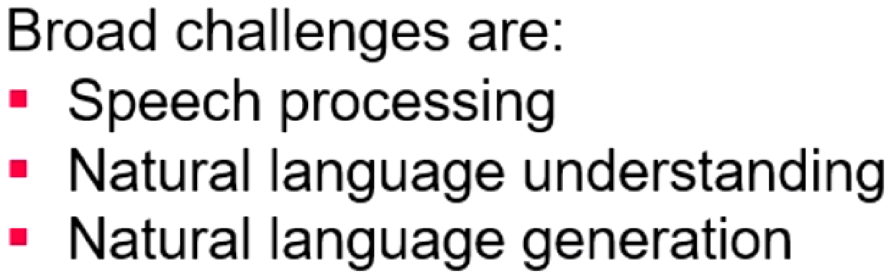
Overview

- how can computers understand natural (human) language
- intersection between
    - linguistics
    - computer science
    - electrical engineering (speech synthesis)
- 

  Broad challenges are:

  - Speech processing
  - Natural language understanding
  - Natural language generation
- NLP is language depending
    - most progess made for English

NLP in [[Information Retrieval]]

- preprocessing before creating dictionary vector
    - remove unwanted characters
        - html tags
        - punctuation
    - split up
        - e.g. on whitespace
    - detect common phrases
    - remove common/stop words
        - e.g. a, an, and, the, it, …
    - stem tokens to word roots
        - e.g. computational ==> compute
- challenges
    - semantic information
        - synonyms
        - one word wtih two different meanings
            - e.g. jaguar ==> animal, car
    - structural syntatic information

[[Information Retrieval]]