

# Introduktion till språkteknologi

Ola Knutsson  
knutsson@csc.kth.se

DH2418 Språkteknologi 2009  
sprakt09

# Vad är det bra för?

Vetenskapliga mål  
Kommersiella mål  
Samhällsbetydelse  
Tvärvetenskapligt ämne

## Viktiga tillämpningar (existerande och framtida)

- Informationssökning
- Frågebesvarande system, äv. flerspråkig
- Dialogsystem
- Språkverktyg för skrivande och läsning
- Maskinöversättning och datorstödd översättning
- Affärskommunikation, automatisk kundtjänst
- Språkutbildning
- Hjälpmedel för språkliga funktionshinder
- Text mining

## Hur började det?

- En av datorns första tillämpningar (MÖ) 1940-talet, 1963 i praktiskt bruk (MÖ sågat samma år i USA)
- Djupa rötter inom matematik och datalogi
- Chomsky (CFG) – Backus and Naur (ALGOL)

## Datorlingvistik eller datalingvistik eller vad heter det?

Talteknologi, Textprocessing,  
Språkteknologi, NLP, LT och LE

## I denna kurs fokuserar vi på text

- Relationen språk-tal-text
- Vad gör text speciellt när det skall hanteras automatiskt?
- Skriftsystemets påverkan
- Vad förutsätter en text?

## Grundläggande termer

- Tokenisering/Segmentering
- Morfologi
- Fonetik och fonologi
- Syntax
- Semantik
- Diskurs
- Pragmatik

## Kärnproblem: flertydighet

Exempel: *En man såg vatten i båtens för*

60 % av orden i svenskan har flera betydelser.

## Historiska skiljelinjer

- Formella språk (generativ grammatik) och parsningsalgoritmer (sent 1950-tal)
  - Artificiell intelligens (1960)
- 
- Bayesianiska metoder OCR (sent 1950-tal)
  - Brown corpus I milj. Ord (1960-tal)

## Metoder för språklig analys och generering

Vad är analys?

Vad är generering?

Regelbaserade: Jag såg [den lilla bilen]

Statistiska: Stockholm är vanligare än Kårböle

Hybrider

## Språkliga resurser

- Korpusar
- Lexikon
- Fria resurser?
- En teknik för alla språk?

## Banbrytande program

Eliza

MÖ av väderleksrapporter

Babelfish

Grammatikkontroll för engelska i Word

## Lösta problem? Olösta problem?

- Morfologi? Taligenkänning?
- Syntax? Yt, ja. Djup, nej.
- Semantik, nej.
- Framtida utmaningar?

## Var är vi nu?

Webben (enorma datamängder)  
Statistiska metoder tar täten (ingenjörerna tar över!)  
Resurssnåla metoder (Surprise languages)  
Kommersiellt mogen (ung. 35 företag i Sverige)

## CSC:s forskning 1992-

- Stava
- Språkliga datorstöd
- Granska
- Informationssökning, klustring, textsammanfattning
- Parsning och utvärdering
- Nya metoder för språkgranskning
- Datorstödd språkinläring: Grim och Cake

## Webbtips

<http://www.nada.kth.se/theory/humanlang/>

<http://sprakteknologi.se>

<http://www.lt-world.org/>