

Introduktion till språkteknologi

OH-serie 1

<http://stp.lingfil.uu.se/~matsd/uv/uv08/ist/>



UPPSALA
UNIVERSITET

Mats Dahllöf
Institutionen för lingvistik och filologi
Oktober 2008

1

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Kursen

- Lärare: Mattias Nilsson och jag.
- Allmän översikt över språkteknologin: förutsättningar, metoder, tillämpningsområden, produkter och tjänster (MD).
- Grundtekniker (inkl. laborationer)
Praktiskt arbete under handledning (MN)

2

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Examination

- Tre inlämningsuppgifter
- Salstentamen: Begrepp, översiktlig förståelse

3

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Lärandemål (från kursplanen)

Efter avslutad kurs skall studenten för att förtjäna betyget Godkänd minst kunna:

- (1) översiktligt redogöra för vilka produkter, tjänster och tekniker som är typiska för området språkteknologi;
- (2) räkna upp de viktigaste delområdena inom språkteknologin;
- *Fortsätter.*

4

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Lärandemål (från kursplanen), forts.

- (3) på ett översiktligt sätt redogöra för problem och metoder inom parsning av naturligt språk, språkgranskning, dokumentsökning, informationsextraktion, maskinöversättning, taligenkänning och talsyntes, samt översiktligt redogöra för dessa teknikers prestanda och kommersiella betydelse;
- *Fortsätter.*

5

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Lärandemål (från kursplanen), forts.

- (4) implementera några elementära exempel på språkteknologisk problemlösning;
- (5) med viss självständighet finna information om ett språkteknologiskt problem eller system och muntligen och skriftligen göra en presentation av det.
- *Slut.*

6

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Kursens huvudpunkter I

- Översikt
- Formell grammatik och parsning (lab.)
- Korpuslingvistik (lab.)
Korpus = stor samling av språkliga data (oftast texter).
Hur bearbeta automatiskt?

7

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Kursens huvudpunkter II

- Tokenisering (separera ord) och taggning (etikettera ord) (lab.)
Vid korpusbearbetning.
- Talteknologi och dialogsystem.
- Maskinöversättning (+ lab.)
- Informationsökning.

8

Språkteknologi

Språkhantering i datorer med "känslighet" för språket som språk:

- Den språkvetenskapliga teorin: **datorlingvistik**
- Tillämpningsområdet: **språkteknologi**

Målet är att ge maskiner förmågan att plocka ut information ur text och tal på ett sätt som verkar förutsätta något slags förståelse, samt att använda naturligt språk för att presentera information.

9

"Tekniska" grundvalar

Primär situation: Språket manifesteras som prat här-och-nu! Akustiskt fenomen, ingen möjlighet till teknisk hantering. Sentida utveckling: lagring och överföring av "yttranden":

- Skrift (5000 år). Senare boktryckarkonsten (550 år).
- Analog elektronisk överföring av tal, telefoni: radio, etc. (120 år)
- Digitalisering av tal och skrift (50 år): media, nätverk, bearbetning. (Nu ofta allenarådande bas för tekniken.)

10

Digital representation I

- Digital = sifferbaserad.
- Datorer ytterst sett binära: Minimala komponenter (kretsar, minnesspår) i två möjliga lägen. "Ettor och nollor."
- En minimal informationsenhet, en **bit** (bi[nary digi]t), är ett eller noll.

11

Digital representation II

- Sekvenser av ettor och nollor representerar information.
2 bitar ger 4 möjligheter (00, 01, 10, 11).
3 bitar ger 8 möjligheter (000, 001, 010, 011, 100, 101, 110, 111).
4 bitar ger 16 möjligheter. Etc.
En extra bit fördubblar antalet möjliga värden.
- 8 bitar ger 256 möjligheter. 8 bitars kodning är vanligt för att representera text. Ger alltså potentiellt 256 tecken.

12

Digital text

- Sekvenser av symboler.
- Västerländska skriftsystem: små uppsättningar symboler.
- Digital representation på ett kompakt och klart sätt.
- I stort sett all professionell texthantering är digital i den industrialiserade världen idag (produktion, lagring, mångfaldigande, tryckning).
- Enorma mängder text är maskinläsbar, alltså. (Enorma mängder information.)

13

Digitalt tal I

- Överföring och lagring alltmer digital.
- Ljud är mycket informationsrikt ur mänsklig synvinkel. Man kan höra väldigt mycket i en persons tal.
- God representation av ljud måste vara informationsrik även digitalt sett.
- Exempel: En liten boksida, bara texten, motsvarar kanske 1 kB (ca 8000 bitar). Ljud i "vanlig" mp3, ca 128000 bitar per sekund (motsvarar typ 3-4 böcker per minut).

14

Digitalt tal II

- Överföring och lagring alltmer digital.
- Ljud: mycket informationsrikt mänskligt/digitalt sett. En liten boksida, kanske 1 kB (ca 8000 bitar). Ljud i "vanlig" mp3, ca 128000 bitar per sekund (motsvarar kanske 3-4 böcker per minut).

15

Digitalt tal III

- Tal — enorma variationmöjligheter: röst, "röst användning", intonation, tempo, rytm, volym, klang, etc.
- Hur tal låter beror på mänsklig anatomi, motorik, "mental dynamik". Artificiellt tal låter artificiellt.
- Svårt att urskilja den språkliga strukturen i tal automatiskt. Vi har kontinuerliga ljudskeenden. "Språkljuden" överlappar.

16

Språkteknologi: tillämpningsområden

- Informationssökning och -utvinning
- Sammandrag och sammanfattning
- Dokumentklassificering
- Språkgranskning (stavnings-, stil- och grammatikkontroll)
- Maskinöversättning
- Dialogsystem/interaktiva telefontjänster
- Språkvetenskaplig forskning, t.ex. vid lexikonbyggande (lexikografi)

17

"Problem" med språk: ord och grammatik

- Många olika ord med komplext innehåll.
- Naturliga språks grammatik är mycket rik.
- Det finns många olika typer av ord med olika grammatiskt beteende.
- Språkanvändningen trotsar ofta traditionella grammatikideal: "Lösö" fraser och ord (snarare är fullständiga meningar), (i tal) omtagningar och avbrutna enheter, etc. är vanliga.

18

"Problem" med språk: tvetydighet

Språk är rika på ord. Ord ofta tvetydiga (polysemi, homonymi). Olika språk — olika begrepp.

Det sitter ett djur i taket: ... *on the ceiling*

Det sitter ett djur på taket: ... *on the roof*

i/på → on.

tak → ceiling/roof.

Möjligheter multipliceras: "bara ben" har minst fyra möjliga översättningar till engelska.

19

"Problem" med språk: pronomensyftning

Pronomen har typiskt flera möjliga antecedenter. De fångar olika egenskaper i olika språk. Pronomina i singularis, t.ex.:

- Svenska: *den, det, han, hon*.
Saker: grammatiskt genus. Personer: kön.
- Engelska: *it, he, she*.
Saker: ett pronomen. Personer: kön.
- Franska: *il, elle*.
Saker och personer i maskulinum eller femininum.

20

"Problem" med språk: situationsanknytning

- All språkanvändning är inbäddad i sociala aktiviteter, och förstås mot denna allmänna bakgrund.
- I princip kan vi använda vi använda allt vi har av kunskaper, intelligens och sociala förmågor när vi kommunicerar.
- *Taket är liksom i kataloghallen försett med ljudisolerande plattor.* ("Roof" eller "ceiling"?)

21

"Problem" med språk: "öppenhet"

- Olika människor använder (samma) språk på olika sätt.
- Vi kan alltid vara kreativa i språket och använda "gamla" ord och uttrycksätt på nya sätt.
- Vi kan hitta på nya ord.
- Vi kan uttrycka oss indirekt, t.ex. i ironi, över- och underdrifter, metaforer, artiga frågor, etc.

22

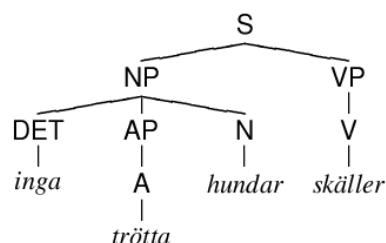
Datorlingvistik: grammatik

- Utgår från teoretisk formell syntax, där grammatik beskrivs i termer av matematiska regelsystem (Chomsky, *Syntactic Structures*, 1957).
- M.h.a. speciella beräkningsprocedurer (s.k. algoritmer) kan formella grammatiker användas för att bestämma den grammatikens strukturen hos språkliga uttryck (parsing/parsning).
- Parsning ofta utgångspunkt för annan bearbetning, t.ex. översättning.

23

Datorlingvistik: grammatik, syntaxträd

Grammatiska strukturer representeras i de flesta fall som träd:



24

Datorlingvistik: semantik

Datorlingvistisk semantik handlar om att analysera/generera naturligt språk (meningar) till/från representationer av betydelse.

- Utgår från "logisk semantik", som utvecklats ur 1900-talets logik och språkfilosofi.
- Semantik och grammatik kopplas samman.
- *Kompositionell semantik*: en sammansatt enhets betydelse kan räknas fram ur delarnas betydelse.

25

Datorlingvistik: "diskurs"

Diskurser är texter och samtal. Hur etableras ett "sammenhang" (koherens/kohesion)?

- Diskursers struktur avspeglar deras syften.
- Referens: införande av referenter, tillbakasyftande referens (ofta med pronomina).
- Referens: vilka saker står i centrum?
- "Retorisk struktur"

26

Regler eller statistisk

Datorlingvistik/språkteknologi — två huvudangreppssätt:

ca 50–80-tal	ca 80-tal–nu
kategoriska regler	probabilistiska samband
grammatikliknande system	statistiska modeller
lingvisten skriver regler	modell från automatisk analys av data
dyrt (manuellt arbete)	billigare (?)
språkspecifik metod	mindre språkspecifikt

27

Informationssökning, exempel I

www.ask.com: Who is Sahlin?

Första träff

Mona Sahlin - Wikipedia, the free encyclopedia

28

Informationssökning, exempel II

www.ask.com: How many students are enrolled at Uppsala University?

Första träff

UlfEkman.org - ULF EKMAN

Ulf Ekman enrolled at Uppsala University and studied...

29

Informationssökning, exempel III

www.ask.com: Who is the prime minister of Sweden?

Första träff:

The Government and its Offices

Angela Merkel to visit Sweden Prime Minister's Office, 21 August 2008

30

Textsammandrag

Att komprimera text, att ta ut de viktigaste bitarna

- Går ofta på ytliga "tecken": fetstil, närvaron av siffror i meningar.
- Nyckelord fastställs statistiskt: meningar med sådana bedöms som viktiga.
- Prova själv, demo/läs mer:
<http://swesum.nada.kth.se>.

31

Dokumentklassificering

För att sortera dokument i ett antal givna kategorier. T.ex.

- T.ex. skilja "spam" från "intressanta" e-brev.
- T.ex. skilja på dokument med olika innehållsteman.
- Kan baseras på samförekomst av ord i dokument.
- Sådana system baseras ofta på inläring från samlingar med redan klassificerade dokument.

32

Språkgranskning

Hjälp med stavning, grammatik och stilnivå åt författare:

- Uppslagning i lexikon enkelt: fångar vissa stavfel, kan ge stilvarningar. Typ: "bilrna" → *rödmarkering*
- Matchning mot mönster för negativt värderade konstruktioner vanligt. Typ: "was Xed" → *undvik passiva konstruktioner*
- Mer fullständig grammatikanalys svårare.
- Dessa system fångar många, men inte alla felskrivningar.

33

Exempel: maskinöversättning. Källtext

[http://www.guardian.co.uk/world/2008/sep/29/uselections2008.sarahpalin:](http://www.guardian.co.uk/world/2008/sep/29/uselections2008.sarahpalin)

However, the Obama camp believes that their man stands to do better in a poor economic climate, and that "bread and butter" issues will eventually outweigh Palin's emotional appeal to more conservative women voters.

34

Exempel: maskinöversättning. Resultat

<http://translate.google.com> (Eng.–sv. Hur bra?):

Men det Obama-lägret tror att man står bli bättre på en svaga ekonomiska klimatet, och att "bröd och smör" frågor så småningom kommer att uppväga Palin: s emotionella vädjan till mer konservativa kvinnliga väljare.

35

Exempel: maskinöversättning. Källtext II

Att härdsmältan i banksektorn långt ifrån är ett nordamerikanskt fenomen är nu uppenbar. Inte minst är det finansiell istid på Island där Glitnir, landets tredje största bankkoncern, nu hamnat under statlig kontroll för att undvika total kollaps.

36

Exempel: maskinöversättning. Resultat II

<http://translate.google.com> (Sv.–eng. Hur bra?)

That the core melt in the banking sector is far from a North American phenomenon is now evident. Not least is the financial ice age in Iceland, where Glitnir, the country's third largest banking group, now fallen under state control to avoid total collapse.

37

Maskinöversättning: användbarhet idag

- Kan ge oss hjälp att förstå dokument på språk vi inte kan.
- Bra översättningar inom begränsade domäner, särskilt i kombination med "kontrollerat" språk.
- Kan effektivisera arbetet för mänskliga översättare: Automatisk översättning i kombination med mänsklig korrigering.
- Översättning av skönlitteratur skulle man inte överlämna åt maskiner.

38

Korpuslingvistik

Språkvetenskap som undersöker stora textmängder (korpusar),

t.ex. m.h.a. s.k. *konkordanser* (KWIC, keywords in context):

av intodingarne, vilka sakna **hundar** av det ratta slaget, oc SKV:U11 [Sida](#)
 dsfall #b bebådas av tjtande **hundar**, av skata på taket. UNG:336 [Sida](#)
 aste band som finnas. Om två **hundar** behaga slåss under bord UNG:287 [Sida](#)
 är stängt överallt, och onda **hundar** bevaka entréerna. Jag f BFB:160 [Sida](#)
 kaperna däroppe. Hästar och **hundar** bilda övergången till d FAG:182 [Sida](#)
 i lätt att döda? VIKAR Många **hundar** bli harens död. MENGLÖ TRI:321 [Sida](#)
 'a av en dilettant, men många **hundar** blir harens död! #1SK ABU:147 [Sida](#)
 #b Han kunde inte tåla mina **hundar**! #1 DAMEN #b Det DAM:324 [Sida](#)
 der villt oväsen från en mängd **hundar**. De, som hade tittat in SÖ2:184 [Sida](#)
 JÄRDE KAPITLET Herrar och **Hundar**. Det hade förlutit n RÖR:039 [Sida](#)
 re hela kuren; men hade inga **hundar**. Det var en ny slucke DMP:004 [Sida](#)

39

Korpuslingvistik, forts.

Fler saker man kan göra med korpusar och datorer:

- Räkna ord
- Se samband mellan ords relativa frekvenser
- Se ordmönster linjärt
- Bas för lexikografi och grammatikstudier
- Bygga modeller och databaser för språkteknologiska tillämpningar

40

Talteknologi

Talteknologin har i hög grad blivit ett eget fält.

- Talsyntes: att artificiellt framställa tal. Eventuellt ur insamlade naturliga segment.
- Talanalys: att utifrån tal automatiskt bestämma vilka ord som uttalas. All sådan teknik utnyttjar statistiska modeller.

”Problem” med tal: olika grundton, intensitet, tempo, uttalssätt, röster, dialekter; samartikulation, reduktion, ordton, intonation, betoning, etc.

41

Talsyntes

- Naturligt tal produkten av motoriska och mentala processer av specifikt mänskligt slag. Dessa avspeglas i samartikulation, reduktion, ordton, intonation, betoning, rytm, tempo.
- Dagens tekniker för talsyntes ger inte fullt naturligt tal.
- Talsyntes kan kombineras med artificiell animation av gestik och mimik.
- Ansiktsrörelser (särskilt läpparnas rörelse) spelar roll vid perceptionen av tal.

42

Talsyntes: text till tal

- Översättning: ortografi till fonetisk representation.
- Lexikon för oregelbundna ord.
- Kräver grammatisk-semantisk analys, p.g.a. ord som *matris*, *banan*.

43

Tal till text (att identifiera ord)

- ”Problem” — tal kan låta väldigt olika: olika grundton, intensitet, tempo, uttalssätt, röster, dialekter. Bakgrundsljud.
- Möjligheter idag:
 - Få ord känns igen generellt.
 - Systemet tränas för en viss person och lärns att skilja på många ord.
- Människor är mycket bättre än maskiner i detta sammanhang.

44

Tillämpningar för talteknologi

- Dokumentuppläsning m.h.a. talsyntes för människor som ej har möjlighet att läsa.
- Dikteringssystem.
- Interaktiva telefontjänster, t.ex. biljettbokning.
- Styrning av apparater, t.ex. i hushållet eller i bilar.

45

Grundläggande steg: ”tokenisering”, etc.

- Tokenisering: Att urskilja löpord. Långt ifrån trivialt: *Som måttenhet är pixel oböjt, t.ex. 100 pixel/cm.* Viktigt steg vid i stort sett all språkteknologisk analys.
- Meningssegmentering: Att urskilja meningar. Också knepigt. (Hur långt fungerar kriteriet punkt-stor bokstav?)

46

Grundläggande steg: taggning

- Taggning: att klassificera orden i en text på något sätt, t.ex. ordklasstaggning. Svåra fall kräver djup språklig analys. För de flesta löpord ger lexikonuppslagning korrekt resultat. Statistiska metoder vanliga.
- Igenkänning av speciella uttryck: Egennamn, datum, årtal, etc. Speciell hantering typisk.
- Tvetydigheter en svårighet vid taggning.

47

Grundläggande steg: parsning

- Parsning: grammatisk analys av hur orden hänger ihop (syntax). T.ex. i termer av satsdelar (subjekt, predikat, objekt, etc.).
- Mer eller mindre fullständig.
- Svårare än taggning. Ger mer information.
- Många källor till tvetydighet.

48

Grundläggande steg: disambiguering

- Disambiguering: att välja alternativ då flera betydelser hos ett ord eller en konstruktion är möjliga.
- T.ex. Hur skilja betydelserna "roof" och "ceiling" hos *tak*? (Statistik?)
- Viktigt t.ex. vid informationssökning och översättning.

Sammanfattning I

- Mänsklig språkanvändning har en komplexitet som matchar människors intelligens och kreativitet.
- Endast människor kan förstå mänskligt språk fullt ut.
- Men maskiner kan utrustas med program som förmår behandla språk med viss "insikt".
 - Mycket av språk kan beskrivas med regler.
 - Det finns användbara statistiska samband mellan olika språkliga "variabler".

Sammanfattning II

- Enorma volymer av talad och skriven kommunikation förlitar sig på digital bearbetning, överföring och lagring.
- Informationssökningstjänster används många gånger dagligen av många svenskar. De skulle inte kunna ersättas av icke språkteknologisk teknik.
- Maskinöversättning är allmänt tillgänglig.
- Interaktiva telefonservice blir allt vanligare. (Det finns stora pengar att tjäna.)