



Introduktion till språkteknologi HT 2008

Korpusbehandling

Matias Nilsson

mattias.nilsson@lingfil.uu.se



Schemaöversikt

- 28/10: **Korpusbehandling** (kl. 12-14)
- 29/10: Labbtillfälle (korpusbeh.) (kl. 8-10)
- 29/10: Labbtillfälle (korpusbeh.) (kl. 13-15)
- 4/11: **Orduppdelning och ordklassanalys** (kl. 10-12)
- 4/11: Labbtillfälle (korpusbeh.) (kl. 13-15)
- 17/11: **Språk och kognitiv modellering** (kl. 12-14)
- 19/11: Labbtillfälle (ordklasstagning) (kl. 13-15)
- 25/11: Labbtillfälle (ordklasstagning) (kl. 10-12)
- 25/11: Labbtillfälle (ordklasstagning) kl. 13-15



Dagens föreläsning: Innehåll

- Vad är en korpus?
- Korpusar och datorer
- Korpusbaserad datorlingvistik
- Korpusurval och sammanställning
- Annotering
- Korpusexempel
- Parallellkorpusar
- Internet som korpus



Vad är en korpus?

- En större samling språkliga data hämtade från autentiskt språkbruk
- Inom språkteknologin ofta synonymt med en elektroniskt lagrad textsamling
 - sammanställd enligt fördefinierade urvalskrav
 - annoterad med lingvistisk information
 - textsamling = skriftspråk, talspråk
 - användbar resurs i både lingvistik och språkteknologi



Korpusar och datorer

- Den "empiriska revolutionen" i (dator)lingvistik:
 - Ökad användning av empiriska data
 - Uppbyggnad av stora korpusar
 - Annotering av korpusdata (syntaktiskt, semantiskt)
- Bakomliggande orsaker:
 - Teknisk utveckling:
 - Ökad tillgång på maskinläsbar text (och digitaliserat tal)
 - Förbättrad datorkapacitet
 - Lagring
 - Bearbetning
 - Vetenskaplig perspektivförskjutning:
 - Ifrågasättande av "länstolslingvistik"
 - Utveckling av statistiska språkmodeller



Korpuslingvistik

- Korpusdata som källa till kunskap om språket
 - Kvantitativa språkstudier:
 - Ordfrekvens. Testa här (för engelska): <http://www.natcorp.ox.ac.uk/>
 - Kollokation: sekvens av ord som ofta förekommer i samma kontext
 - Konkordans: lista med förekomster av ett ord eller ordsekvens med tillhörande kontext. Testa här: <http://spraakbanken.gu.se/lb/konk/>



Korpusbaserad datorlingvistik

- Korpusdata som underlag till statistiska metoder för automatisk språkanalys:
 - Statistisk maskinöversättning
 - Taligenkänning och talteknologi
 - Statistisk parsning (automatisk grammatisk analys)
 - Lexikonutveckling



Korpusbaserad datorlingvistik

- 1980-tal
 - Datorlingvistiken starkt influerad av samtida lingvistisk teori
 - Symboliska och regelbaserade metoder dominerar
- "Every time I fire a linguist the performance of the recognizer goes up" (F. Jelinek, IBM Research Group)
- Problem med regelbaserade system
 - Språkliga konstruktioner betraktas som acceptabla eller icke acceptabla. Ingen gradskillnad.
 - Inga preferensregler bland tvetydiga satser



Korpusbaserad datorlingvistik

- Datorlingvistik idag:
 - Korpusbaserad datorlingvistik dominerar fältet
 - Preferens och gradskillnad uttrycks m.h.a sannolikhet
 - Statistiska metoder är *robusta*
 - genererar alltid någon output
 - Maskininläring: Datorn lär sig automatiskt från exempel hämtade ur korpusdata



Korpustyper

- Balanserad
 - En balanserad korpus utgör ett representativt sampel av ett språk och täcker in olika texttyper och stilistiska varianter av språket
- Monitor/Open-ended
 - Ständigt växande korpus, sammansättningen växlar (t.ex. Bank of English)
- Parallellkorpus
 - Består av samma texter på olika språk (t.ex. EuroParl)
- Synkron (samtida) – diakron (historisk)



Korpusurval (*Sampling*)

- En korpus bör fungera som ett representativt "stickprov" av en större population (t.ex. ett språk)
- Genom att undersöka ett representativt sampel kan vi dra slutsatser om populationen i stort
- Därför viktigt att åstadkomma ett representativt sampel

Precis som i opinionsundersökningar!



Korpussammansättning

Att tänka på innan man påbörjar en korpusinsamling

- Vad är syftet med korpussammansättningen?
- Vilken/vilka genrer skall korpusen täcka?
- Hur väljer vi ett representativt sampel för den/de genrer vi vill täcka?
- Hur stor bör min korpus vara för att vara relevant?
- Vilken lingvistisk information skall annoteras?



Annotering

- Manuell annotering
 - görs för hand, tidskrävande metod
 - kräver expertkunskap
 - risk för inkonsekventa fel
- Automatisk annotering
 - utförs av program, snabb metod
 - kräver data att lära sig ifrån
 - konsekventa fel



Annotering

- Ordklasstagning (ord → ordklass)
- Lemmatisering (ord → grundform/lemma)
- Syntaktisk annotering (mening → syntaktisk representation)
- Semantisk annotering (mening → semantisk representation)
- Textlingvistisk annotering (stil, diskurs)
- Fonetisk annotering (grafem → fonetisk representation)



Annotering

Ordklasstagning

- bestämning av ordklass
- traditionella ordklasser: substantiv, adjektiv, verb etc.
 - dock ej självklart vilka eller hur många ordklasser som faktiskt finns
- morfosyntaktisk bestämning: genus, numerus, bestämdhet etc.

Vad avgör graden av specificitet?

- syftet med taggningen
- språkets uppbyggnad (rik morfologi – rik ordklassuppsättning)



SUC - Stockholm Umeå Corpus

- Ca 1 miljon löpord
- Balanserad, svensk 1900-talstext
- 500 texter med ca 2000 ord per text
- 9 huvudgenrer med undergenrer, t.ex.
 - K (skönlitteratur)
 - > KK allmän skönlitteratur
 - > KL deckare och science fiction
 - > KR humor
- Manuellt uppmärkt med ordklass och lemma samt morfosyntaktiska särdrag, såsom kasus genus och numerus



Exempel ur SUC

```

<s id=aa01a-007>
<w n=68>Särskilt<ana><ps>AB<b>särskilt</w>
<w n=69>smygrustningen<ana><ps>NN<m>UTR SIN
NOM<b>smygrustning</w>
<w n=70>vad<ana><ps>HA<b>vad</w>
<w n=71>gäller<ana><ps>VB<m>PRS AKT<b>gälla<
<w n=72>missiler<ana><ps>NN<m>UTR PLU IND
NOM<b>missil</w>
<w n=73>oroar<ana><ps>VB<m>PRS AKT<b>oroa</
<d n=74>.<ana><ps>MAD<b>.</d>
</s>

```



Brown Corpus

- Sammanställdes och datalagrades redan i början av 60-talet vid Brown University
- Tidig modell för många korpusar av samma typ (t.ex. SUC)
- Ren text (ej ordklassstaggad)
- Ca 1 miljon löpord skriven amerikansk engelska



BNC - British National Corpus

- Ca 100 miljoner löpord talad och skriven brittisk engelska
- Balanserad
- Automatiskt ordklasstaggad utan manuell efterredigering
- 61 olika taggar
- Delmängd på 2 miljoner ord rikare taggad och manuellt efterredigerad – 139 olika taggar



Exempel ur BNC

```
<w DT0>Each  
<w NN1>dance  
<w VVD-VVN>followed  
<w AJ0>particular  
<w NN2>rules  
<w VVD-VVN>laid  
<w AVP>down  
<w PRP>by  
<w AT0>the  
<w AJO-NN1>dancing  
<w NN2>masters
```



Lemmatiserade korpusar

- Susanne – "Surface and underlying structural analysis of natural english"
 - 130 000 löpord skriven amerikansk engelska
- SUC



Lemmatisering i Susanne

A01:0460a	-	AT	The	the	[O{S{Ns:s.
A01:0460b	-	NNlc	jury	jury	.Ns:s]
A01:0460c	-	VVDv	said	say	[Vd:Vd]
A01:0460d	-	PPH1	it	it	[Fn:o{Ni:s.Ni:s]
A01:0460e	-	VVDt	found	find	[Vd:Vd]
A01:0460f	-	AT	the	the	[Fn:o{Ns:s.
A01:0460g	-	NNJln	court	court	.Ns:s]
A01:0460h	-	YIL	<ldquo>	-	.
A01:0460i	-	VHZ	+has	have	[Vzf.
A01:0460j	-	VNVv	incorporated	incorporate	
A01:0460k	-	II	into	into	[P:t.
A01:0470a	-	APPGh1	its	its	[Np.
A01:0470b	-	VVGv	operating	operate	
A01:0470c	-	NN2	procedures	procedure	
A01:0470d	-	AT	the	the	[Np:o.
A01:0470e	-	NN2	recommendations	recommendation	.



Syntaktisk annotering

- Parsning, d.v.s. grammatisk analys av texten
- Automatisk parsning ger lägre precision än automatisk ordklasstagning
- Trädbank = syntaktiskt annoterad korpus
 - frasstruktur (S, NP, VP, AP, etc.)
 - dependensstruktur (subjekt, objekt, adverbial etc.)
- *The Penn Treebank*:
 - ca 1 miljon löpord
 - <http://www.cis.upenn.edu/~treebank/>



Frasstruktur i *The Penn Treebank*

```
( (S
  (NP-SBJ (NNP Rolls-Royce) (NNP Motor) (NNPS Cars) (NNP Inc.))
  (VP (VBD said)
    (SBAR (-NONE- 0)
      (S
        (NP-SBJ (PRP it))
        (VP (VBZ expects)
          (S
            (NP-SBJ (PRP$ its) (NNP U.S.) (NNS sales))
            (VP (TO to)
              (VP (VS remain)
                (ADJP-PRD (JJ steady))
                (PP-LOC-CLR (IN at)
                  (NP
                    (QP (IN about) (CD 1,200))
                    (NNS cars))
                  (PP-TMP (IN in)
                    (NP (CD 1990))))))))))
  (. .) ) )
```



Semantisk annotering

Två typer:

1. Uppmärkning av semantiska relationer (agent, patient etc.)

- FrameNet <http://framenet.icsi.berkeley.edu/>

2. Uppmärkning av ordbetydelse, t.ex. lexikala relationer (synonymi, antonymi, hyponymi etc.)

- WordNet <http://wordnet.princeton.edu/>



Textlingvistisk annotering

• Diskurstagg

– London-Lund Corpus of Spoken English:

- > ursäkter, *sorry*
- > hälsningar, *hello*
- > artighetsfraser, *please* m.fl.

• Anaforisk annotering (pronomenreferens)



Fonetisk annotering

• Transkribering

– MARSEC: *The Machine Readable Spoken English Corpus*

• Prosodi

– *London-Lund Corpus of Spoken English*



Parallellkorpusar

- Hansard
 - Engelsk-fransk parallellkorpus bestående av kanadensiska parlamentsprotokoll
 - ca 15 miljoner löpord
 - delvis taggad och parsad
- EuroParl
 - parallellkorpus bestående av officiella Europaparlamentstexter
 - tillgängliga på 11 olika EU-språk
 - ca 20 miljoner löpord; 740 000 meningar per språk
 - <http://people.csail.mit.edu/koehn/publications/europarl>



Parallellkorpusar - användningsområden

- Maskinöversättning
- Flerspråkiga lexikon
- Flerspråkig informationsökning
- Andraspråksinläring



Internet som korpus

- Enorm dataresurs
- Fritt tillgänglig
- Ständigt uppdaterad, blir ej omodern
- Konkordans baserad på träffar i olika sökmotorer:
 - Testa här: <http://www.webcorp.org.uk/>
- Problematik
 - annotering (endast ordformer är sökbara)
 - balans: ingen möjlighet att begränsa sökningar till en viss stil eller genre
 - Upphovsrättsliga frågor



UPPSALA
UNIVERSITET

Laboration 2: Korpusbehandling i Linux

- Mål: Att lära sig använda enklare kommandon för extraktion av information ur korpusdata
- Steg 1: välj ut en lämplig text (minst 1000 ord)
- 3 deluppgifter:
 - Räkna ord: löpord, typord
 - Skapa frekvensordlista
 - Skapa kollokationslistor
- "Handbok":
Kenneth Church, *Unix for Poets*
