

Korpuslingvistik (SV2119)

Föreläsning 3: Annotering



Richard Johansson

`richard.johansson@svenska.gu.se`

4 oktober 2013

1. introduktion



kort info föreläsning 7

- ▶ ändring föreläsning 7



dagens föreläsning

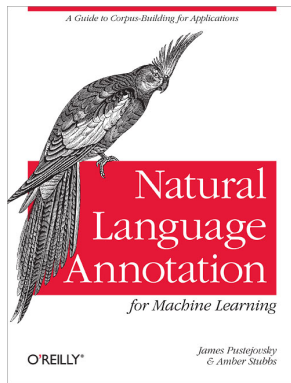
- ▶ du har fått 10 miljoner av VR för att annotera en stor korpus
- ▶ hur lägger du upp ditt projekt?



översikt

- ▶ annoteringsmodell: hur beskriver vi språkliga fenomen systematiskt? (och annat också)
- ▶ format: hur lagrar vi annoteringen i filer?
- ▶ några fallstudier
- ▶ annoteringsprocessen; kvalitetskontroll
- ▶ exempel på verktyg

dagens boktips



2. annoteringsmodeller



annoteringsmodell

- ▶ för att göra intressantare analyser behöver vi mer information än bara texten i sig
- ▶ **annoteringsmodell**: hur beskriver vi vad som finns i korpusen?
- ▶ med en **formaliserad** vokabulär och struktur
- ▶ vilka enheter finns?
 - ▶ texter, meningar, ord, ...?
- ▶ hur är enheterna relaterade?
- ▶ vilka attribut har respektive enhet?
 - ▶ textens författare, publiceringsdatum, ...?
 - ▶ ordets böjningsform, ...?
- ▶ bygger kanske på någon lingvistisk teori, men konkretiserat

att standardisera sin lingvistiska vokabulär

- ▶ definiera i förväg vilka kategorier som ska beskrivas
- ▶ exempel:
 - ▶ **ordklass**: verb, substantiv, preposition, ...
 - ▶ **fraser**: verbfras, nominalfras, ...
- ▶ i många fall kan det vara möjligt att använda kategorier som redan är definierade i en standard som ISOcat
 - ▶ då kan man slippa återuppfinna hjulet, och det blir lättare att återanvända i andra sammanhang



<http://spraakbanken.gu.se/swe/forskning/saldo/taggmangd>

<http://www.isocat.org/rest/dc/1333>

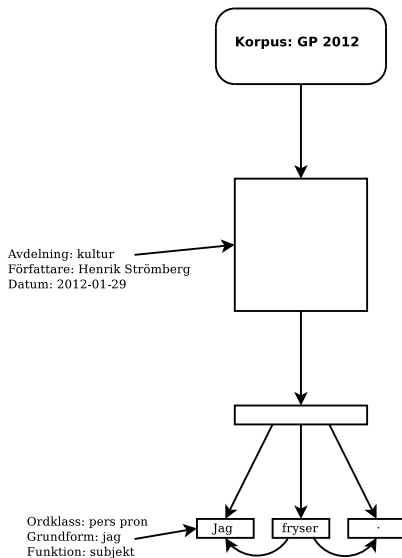
exempel: syntaktiska relationer i Talbanken/MAMBA

	MAMBA Categories
Tag	Meaning
++	Coordinating conjunction
+A	Conjunctive adverbial
+F	Coordination at main clause level
AA	Other adverbial
AG	Agent
AN	Apposition
AT	Nominal (adjectival) pre-modifier
CA	Contrastive adverbial
DB	Doubled function
DT	Determiner
EF	Relative clause in cleft
EO	Logical object
ES	Logical subject
ET	Other nominal post-modifier
FO	Dummy object
FP	Free subjective predicative complement
FS	Dummy subject
FV	Finite predicate verb

annoteringsmodell i Språkbanken

- ▶ vilka enheter finns?
 - ▶ **korpusar**, **dokument**, **meningar**, **ord**
- ▶ hur hänger enheterna ihop?
 - ▶ en korpus består av dokument
 - ▶ ett dokument består av meningar
 - ▶ en mening består av ord
 - ▶ ett ord består av fritext
 - ▶ ett ord kan vara länkat till ett annat ord genom **dependensrelation** (stöds dock inte av Korp)
- ▶ vilken information finns om enheterna?
 - ▶ ett dokument har **textattribut** (korpusberoende)
 - ▶ ett ord har **ordattribut**

exempel



3. exempel

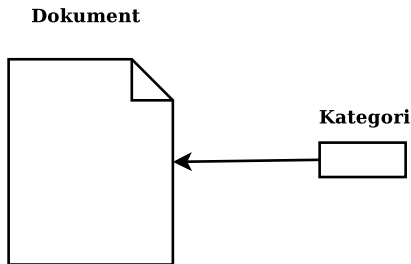


några exempel

- ▶ vi gör nu några fallstudier för att se hur man har utformat annoteringsmodellerna
- ▶ från enklast möjliga till gradvis mer komplicerat

enklast möjliga? kategorisering av text

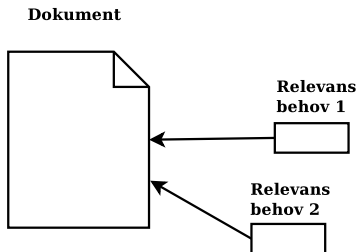
- ▶ vi talar om ett dokument som helhet
- ▶ vi tilldelar varje dokument en kategori från en **given uppsättning**
 - ▶ i enklaste fallet bara två kategorier (t.ex. författarens kön, spam/ickesпам)
 - ▶ ämneskoder ...



relevansbedömning för informationssökning



- ▶ givet ett sökbehov, t.ex. *“finns det ett samband mellan konsumtion av rött kött och tjocktarmscancer?”*
- ▶ är dokumentet är **relevant** eller inte?



exempel på annotering i text: namnannotering

KARLSTADSRESAN.

I.

*Vad man än kan ha emot **Thea Sundler**, så får man erkänna, att hon bättre än någon annan förstod sig på att handskas med **Karl-Artur Ekenstedt**.*

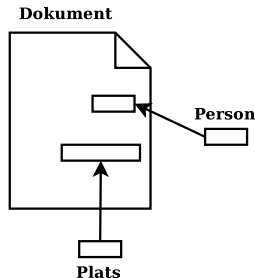
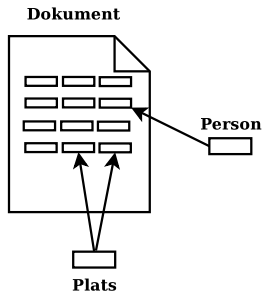
*Om man till exempel tänker på **Charlotte Löwensköld**, så hade hon också velat förmå honom att resa till **Karlstad** och försona sig med sin mor. Men för att beveka honom härtill hade hon påmint honom om allt det, som modern hade varit för honom, och till sist hade hon verkligen försökt skrämma honom med att han inte skulle kunna predika så bra som hittilldags, om han visade sig otacksam emot sin mor.*

Platsnamn: KARLSTAD, Karlstad

Personnamn: Thea Sundler, Karl-Artur Ekenstedt, Charlotte Löwensköld

representation av namnannotering

- ▶ ett par tänkbara modeller (av många tänkbara):
 - ▶ alternativ 1: ett dokument består av ord; ett namn är en grupp av ord
 - ▶ alternativ 2: ett dokument består av ren text; namnen är definierade med hjälp av positioner i texten
- ▶ vi specificerar också de giltiga typerna av namn: person, plats, organisation, ...



ordklassannotering, t.ex. SUC

- ▶ en **text** består av **meningar**, som i sin tur består av ordnade **ord**
- ▶ varje ord består av tre attribut:
 - ▶ ordform
 - ▶ ordklass – från en i förhand definierad lista
 - ▶ grundform

annotering av syntaktisk struktur



- ▶ de vanligaste typerna av syntaktisk annotation: frasstruktur (konstituentstruktur) och dependens
- ▶ en **mening** består av ordnade **ord**
- ▶ frasstruktur:
 - ▶ en **fras** består av ett antal underfraser eller ord
 - ▶ en fras har en **frasetikett** t.ex. NP, PP, ...
- ▶ dependensstruktur:
 - ▶ en **dependensrelation** sammanbinder två ord
 - ▶ **grammatisk funktion** t.ex. subjekt, objekt, ...
- ▶ det finns också hybrider
- ▶ mer i föreläsning 5!

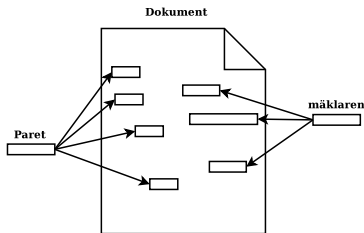
ett mer exotiskt exempel: anaforiska uttryck

Paret trodde att de köpt en vanlig bostadsrätt, men så var det inte. Nu stämmer de mäklaren. Det var under förra hösten som lägenheten annonserades och visades av en mäklare i Norrköping. I sin stämningsansökan till tingsrätten skriver parets advokater att mäklaren gjort fel ända från början. I såväl annonser som i den beskrivning som delades ut vid visningen ska mäklaren ha använt ord som fått objektet att låta som en bostadsrätt.

**Paret – de – de – parets
mäklaren – en mäklare i Norrköping – mäklaren – mäklaren
lägenheten – objektet
visades – visningen
annonserades – annonser ?
stämmer – stämningsansökan till tingsrätten ?
en vanlig bostadsrätt – en bostadsrätt ??**

anaforiska uttryck: en tänkbar lösning

- ▶ här är ett sätt att representera:
 - ▶ ett **omnämmande** är en bit text
 - ▶ en **kedja** är en samling omnämmanden
- ▶ kanske vi också behöver ett attribut som säger vilken typ av relation vi har?
 - ▶ enbart identitet?
 - ▶ eller även delmängd?
- ▶ vår specifikation får bestämma hur vi hanterar svåra fall
 - ▶ ofta hanteras enbart nominalfraser och enbart identitet

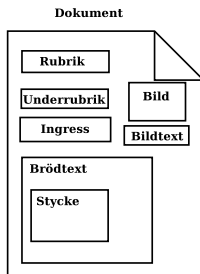


Par lurades - köpte del av hyreshus

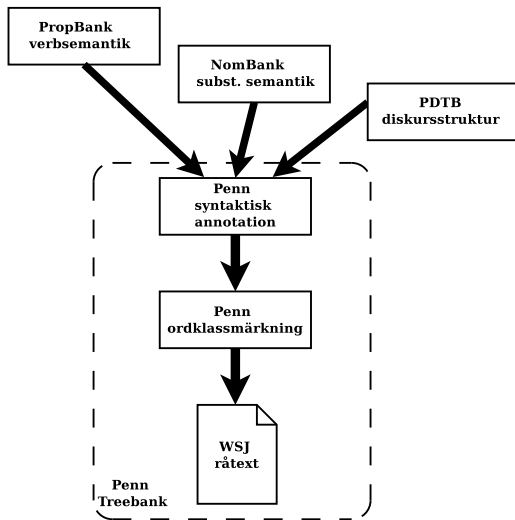
Norrköping

Paret trodde att de köpt en vanligt bostadsrätt, men så var det inte. Nu stämmer de mäklaren.

Det var under förra hösten som lägenheten annonserades och visades av en mäklare i Norrköping. I sin stämningsansökan till tingsrätten skriver parets advokater att mäklaren gjort fel ända från början. I såväl annonser som i den beskrivning som delades ut vid visningen ska mäklaren ha använt ord som fått objektet att låta som en bostadsrätt. Vid något tillfälle ska även ordet bostadsrätt ha använts.



lager på lager



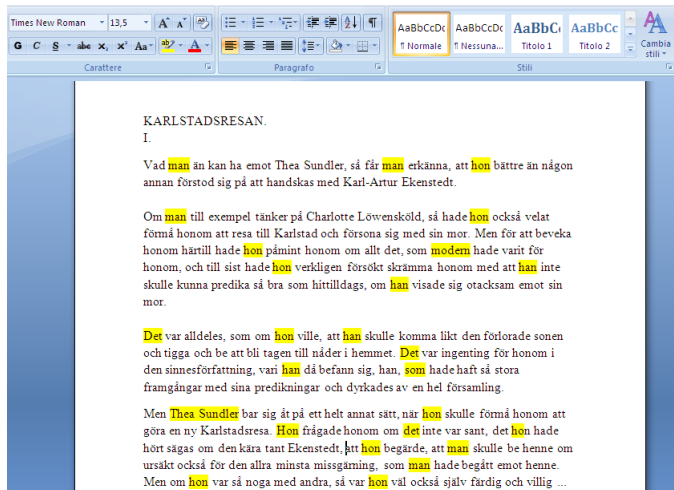
4. serialisering av annotering



lågnivåkodning (serialisering)

- ▶ när vi har en annoteringsmodell måste vi bestämma hur annoteringen ska **lagras**: vi behöver ett **format**
- ▶ återanvändbarhet är att föredra om möjligt
 - ▶ så att vårt arbete kan utnyttjas av andra
- ▶ läsbarhet för maskiner är nödvändigt: formatet måste vara **entydigt**
- ▶ mänsklig läsbarhet är inte en nackdel men inte en hög prioritet
 - ▶ med ett bra verktyg behöver vi aldrig se formatet

exempel på vad man inte vill göra!



KARLSTADSRESAN.
I.

Vad **man** än kan ha emot Thea Sundler, så får **man** erkänna, att **hon** bättre än någon annan förstod sig på att handskas med Karl-Artur Ekenstedt.

Om **man** till exempel tänker på Charlotte Löwensköld, så hade **hon** också velat förmå honom att resa till Karlstad och försona sig med sin mor. Men för att beveka honom härtill hade **hon** påmint honom om allt det, som **modern** hade varit för honom, och till sist hade **hon** verkligen försökt skrämma honom med att **han** inte skulle kunna predika så bra som hittilldags, om **han** visade sig otacksam emot sin mor.

Det var alldeles, som om **hon** ville, att **han** skulle komma likt den förlorade sonen och tigga och be att bli tagen till nåder i hemmet. **Det** var ingenting för honom i den sinnesförfattning, vari **han** då befann sig, han, **som** hade haft så stora framgångar med sina predikningar och dyrkades av en hel församling.

Men **Thea Sundler** bar sig åt på ett helt annat sätt, när **hon** skulle förmå honom att göra en ny Karlstadsresa. **Hon** frågade honom om **det** inte var sant, det **hon** hade hört sägas om den kära tant Ekenstedt, att **hon** begärde, att **man** skulle be henne om ursäkt också för den allra minsta missgärning, som **man** hade begått emot henne. Men om **hon** var så noga med andra, så var **hon** väl också själv färdig och villig ...

exempel på format, sent 1970-tal

0 100 99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81 80 79 78 77 76 75 74 73 72 71

1 2 3 4 5 6

TR³ ↑↑ 3 ↓↓

Me 2-16

18

Tunc anima accipiens puerum, et amplexibus stringens, cordi suo impressit ita, ut audiret et sentiret pulsum Cordis eius.

PF PF KC

1	1	5	2	1	8
1A	SS	1A	1A	FV	1A
			1A	FV	SS
			1	1	1
					5
					6
					10
					+C
					KC

2

Tha togh hon barneth til sik j sin fampn Ok thrykte thz jntil sith hiarta Swa ath hon badhe kände ok hördhe barnsens hiarta bostan

1-2 3 4 5 6 7

Sorto 315 - Esselte System 740023

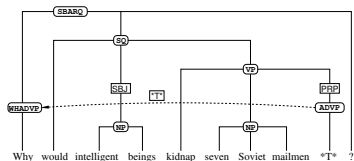
EFERTENYCK SORTO FÖRBJUDS

25

12 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

exempel på primitiva lågnivåkodningar

- ▶ för kategori: med filnamn eller kataloger
- ▶ Excel-ark
- ▶ hemmagjorda textformat
- ▶ Penn Treebank: struktur markeras med parenteser



```
( (SBARQ
  (WHADVP-2 (WRB Why) )
  (S (MD would)
    (NP-SBJ (JJ intelligent) (NNS beings) )
    (VP (VB kidnap)
      (NP (CD seven) (JJ Soviet) (NNS mailmen) )
      (ADVP-PRP (-NONE- *T*-2) )))
  (. ?) )
```

kodning av struktur med XML

- ▶ **XML** (*extensible markup language*) är en standard för att beskriva strukturerade data
- ▶ XML består av fritext blandad med strukturmärkning
- ▶ start- och slutmärken används för att visa inneslutning
- ▶ man kan använda attribut

```
<document author="Selma Lagerlöf" title="Anna Svärd">
  <chapter title="KARLSTADSRESAN">
    <paragraph>
      <sentence>
        <word pos="CONJ">Men</word>
        <word pos="NAME">Thea</word>
        <word pos="NAME">Sundler</word>
        <word pos="VERB">bar</word>
        ...
      </sentence>
      ...
    </paragraph>
    ...
  </chapter>
</document>
```

XML är också ett halvfabrikat

- ▶ XML är ett allmänt strukturbeskrivningsspråk
 - ▶ och det finns många datorverktyg som läser XML
- ▶ det vet ingenting om vad vi försöker göra
- ▶ vi måste därför själva bestämma hur våra strukturer ska kodas

exempel på kodning av namn med XML (alternativ 1)

- ▶ ett alternativ: namn märks upp direkt i texten (“inline”)

<DOCUMENT>

Om man till exempel tänker på <PERSON>Charlotte Löwensköld</PERSON>, så hade hon också velat förmå honom att resa till <LOCATION>Karlstad</LOCATION> och försona sig med sin mor.

</DOCUMENT>

exempel på kodning av namn med XML (alternativ 2)

► uppdelat i ord:

```
<DOCUMENT>
  <w>Om</w>
  <w>man</w>
  <w>till</w>
  <w>exempel</w>
  <w>tänker</w>
  <w>på</w>
  <PERSON>
    <w>Charlotte</w>
    <w>Lövensköld</w>
  </PERSON>
  <w>,</w>
  . . .
</DOCUMENT>
```

exempel på kodning av namn med XML (alternativ 3)

- ▶ “stand-off”:
 - ▶ själva texten hålls separat och oförändrad
 - ▶ den lingvistiska informationen sitter vid sidan av (kanske i en annan fil), och refererar till texten

```
<DOCUMENT>
```

```
<TEXT>
```

```
Om man till exempel tänker på Charlotte Löwensköld,  
så hade hon också velat förmå honom att resa till  
Karlstad och försona sig med sin mor.
```

```
</TEXT>
```

```
<NAMES>
```

```
  <PERSON start="31" end="51"/>
```

```
  <LOCATION start="140" end="148"/>
```

```
</NAMES>
```

```
</DOCUMENT>
```

standardiserade format

- ▶ det mest kända formatet för lingvistisk annotering är **TEI** (Text Encoding Initiative)
 - ▶ första versionen av standarden kom 1990
 - ▶ senaste versionen (version 5) kom 2007
- ▶ ett annat format som börjar sprida sig: **LAF** (Linguistic Annotation Format), en ISO-standard

7. exempel på annoteringsverktyg



exempel på ett verktyg för uppmärkning: Callisto

Callisto - O1_1.txt.aif.xml

File Edit Format Tools Help

For example , biogas can be made from sewage and organic waste that exists in huge quantities in every city , and is currently thrown away in most places . Sewage and garbage could become a part of the solution . Public heating systems are another way of delivering heat in a highly efficient way , and we should look at how Sweden does this .

These solutions are as relevant to Russia , Ukraine and the former Soviet republics as they are to EU member states . These countries are increasingly vulnerable because of their over dependence on gas , and they need to invest in local and renewable sources of energy . They should become part of the diversification of energy supplies solution , rather than being seen as part of the problem .

The case for renewables has been proven in the EU and most government 's are in favour . But it is very much a case of something `` should " be done , not something that urgently `` must " be done .

Central European EU member states are trying to block the EU 's 20/20/20 commitments as they want to protect their old and dirty energy sector . But this is a false economy ; it is in their own economic interest to diversify supplies and take the renewable road to sustainability .

The author is an independent consultant based in Romania

subject_actor article issue claim object_actor

Text	actor	name	type
------	-------	------	------

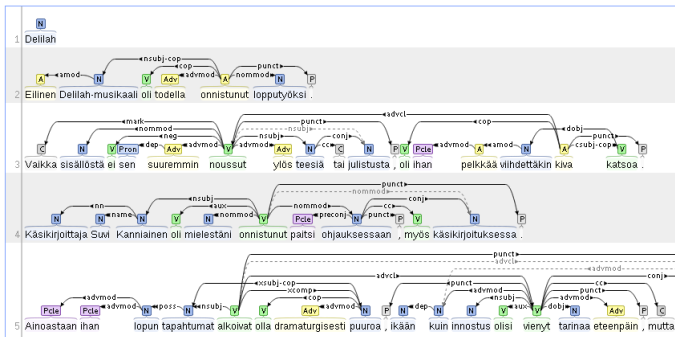
Font: 12pt. Default | Charset: UTF-8 | Task: Codebook Annotation

ett web-baserat verktyg: Brat (här namn)

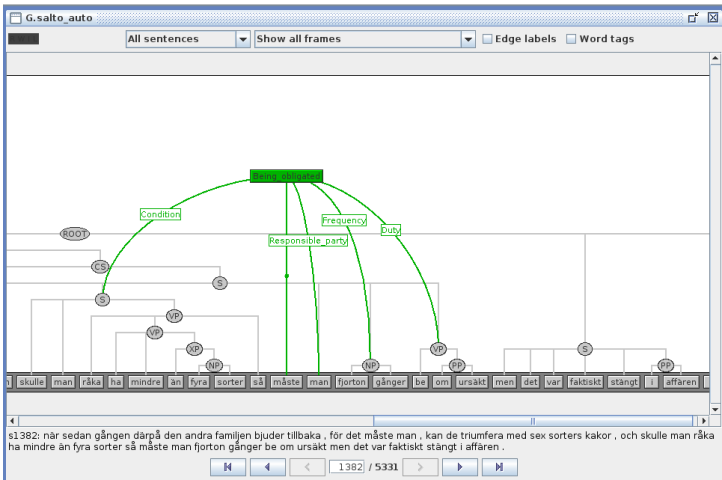


1	LOC Freetown, ORG 25 may (EFE).
2	.
3	El asesinato en LOC Sierra Leona de cuatro soldados de las fuerzas pro gubernamentales y de dos periodistas a manos de los rebeldes de las ORG Fuerzas Revolucionaria Unidas (FRU) coincidió con el recrudecimiento de la ola de violencia que desde hace casi un mes vive este país africano.
4	El periodista español Miguel Gil Moreno, que trabajaba como camarógrafo para la agencia estadounidense ORG Associated Press TV y su colega norteamericano PER Kurt Schork, corresponsal de la agencia de noticias británica ORG Reuters, fueron asesinados ayer cuando la columna de tropas gubernamentales junto a la que PER viajaban fue asaltada por los rebeldes del ORG FRU.
5	En el suceso, que tuvo lugar en el cruce de caminos de LOC Rogbeni, a unos 80 kilómetros al noreste de LOC Freetown, también resultaron heridos leves otros dos PER periodistas, el griego PER Yannis Behrakis y el sudafricano Mark Chisholm, ambos corresponsales de ORG Reuters.

dependenssyntax med Brat (Turku Dependency Treebank)



annotering av ramsemantik med SALTO

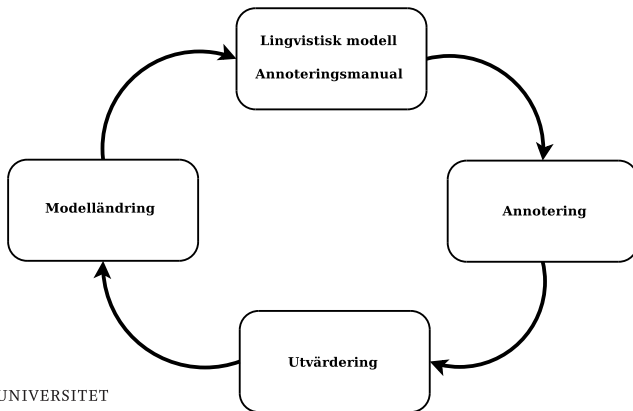


5. annoteringsprocessen



exempel på en process

- ▶ för att systematisera sitt arbete kan det vara bra att fundera ordentligt på hur processen ska fungera
- ▶ MAMA (Pustejovsky och Stubbs, 2012)
- ▶ man kan börja med snabba cykler (pilotstudier)
 - ▶ tills specifikationerna är någorlunda stabila



- ▶ när vi har vår modell är det dags att skriva ned den i en praktisk manual
- ▶ manualens kvalitet påverkar resultatets kvalitet
- ▶ några lämpliga saker att ta med
 - ▶ anoteringsprojektets syfte
 - ▶ en definition av de begrepp som modellen innehåller
 - ▶ ...och praktiska förklaringar av hur de ska användas
 - ▶ en lagom mängd exempel
 - ▶ beskriv svåra fall, "gråzoner"
 - ▶ beskrivning av praktiska frågor t.ex. arbetsgång och verktyg

exempel: Karins manual



exempel: Penn Discourse Treebank



att skaffa annoterare

- ▶ efter de inledande pilotstudierna är det kanske dags att kalla in mer arbetskraft
- ▶ hur får vi tag på annoterare?
 - ▶ hyra studenter?
 - ▶ Academic Work?
- ▶ annoterarna måste instrueras ordentligt
- ▶ låt dem gärna göra en träningsrunda och ge feedback

crowdsourcing

- ▶ **crowdsourcing**: att använda många otränade annoterare i stället för några stycken tränade
 - ▶ hur får vi tag på dem och hur får vi dem att jobba?
- ▶ det mest kända är nog Amazon Mechanical Turk
 - ▶ kräver en hel del jobb att sätta upp
 - ▶ funkar bäst om annoteringen kan delas upp i mycket små och enkla delsteg
 - ▶ kan man göra något meningsfullt om uppgiften kräver lingvistisk träning?
 - ▶ kanske svårt att göra annotering om språket inte är engelska?
 - ▶ risk för fusk! använd kontroller
- ▶ “games with a purpose”?
- ▶ risk för låg kvalitet: kompensera genom att låta många göra samma sak

exempel: folkets synonymlexikon Synlex

Folkets synonymlexikon Synlex

Folkets synonymlexikon är byggt av tiotusentals Internetanvändare och är inte en del av Lexin eller Skoldatanätet. Hjälp till att bygga vidare på synonymlexikonet genom att svara på följande fråga (eftersom vissa synonymförslag är inskickade av användare förekommer oseriösa frågor; svara 0 på sådana):

Är **pinsam** och **förarglig** synonymer?

Välj på en skala från 0 till 5.
0 = instämmer inte, 5 = instämmer helt

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ vet inte

Rösta!

Klicka [här](#) för att komma till Folkets lexikon.



6. kvalitetssäkring



“adjudication”

- ▶ slutlig genomgång innan korpusen anses färdig och publiceras
- ▶ viktigast och enklast: gå igenom de fall där annoterarna motsäger varandra
- ▶ ...men tänk på att flera annoterare kan göra fel samtidigt!
- ▶ en specialist kan detaljstudera en delmängd

meta-annotering: annotering om annotering

- ▶ det kan vara användbart att lägga till information om själva annoteringarna
- ▶ kvalitetskontroll och spårbarhet
- ▶ exempel:
 - ▶ “Lisa märkte denna ordklass den 18 juni 2013 med hjälp av verktyget Brat”
 - ▶ “jag är osäker på vilken ordklass detta är”
 - ▶ “denna ordklassmärkning är dubbelkollad”
 - ▶ “denna ordklassmärkning är automatiskt gjord med verktyget TreeTagger”

att mäta överensstämmelse

- ▶ naivt: hur stor andel överensstämmelse?
- ▶ problem: även om vi annoterar på måfå kommer vi att få en hel del överensstämmelse
- ▶ bättre idé: jämför överensstämmelse med vad vi skulle få på måfå
- ▶ exempel på sådana mått:
 - ▶ Cohens κ -mått om antalet annoterare är 2
 - ▶ Fleiss' κ om antalet är större
 - ▶ ännu mer generellt: Krippendorffs α
 - ▶ ...

- ▶ jämför sannolikheten för överensstämmelse $P(a)$ med sannolikheten för slumpöverensstämmelse $P(e)$

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

- ▶ exempel (relevansbedömning):
 - ▶ Jan tyckte att 15% var relevanta och Lisa tyckte 20%
 - ▶ Alltså, sannolikheten för slumpöverensstämmelse är $0.15 \cdot 0.2 + 0.85 \cdot 0.8$, dvs 71%
 - ▶ De gjorde samma bedömning för 90% av dokumenten
 - ▶ $\kappa = (0.9 - 0.71)/(1 - 0.71) = 0.66$
- ▶ tumregel: under 0.4 “inget vidare”, 0.4–0.6 “medel”, 0.6–0.8 “bra”; 0.8–1 “jättebra”

om κ var lågt

- ▶ problem med den lingvistiska modellen?
- ▶ oprecis annoteringsmanual? har vi beskrivit vanliga svåra fall?
- ▶ har annoterarna fått ordentliga instruktioner och blivit tränade?
- ▶ är uppgiften svår i sig?



nästa föreläsning: att räkna ord

- ▶ metodologiska frågor om undersökningar
- ▶ lite statistik:
 - ▶ ordfrekvenser
 - ▶ associationsmått
 - ▶ vokabulärspridningmått
 - ▶ jämförelser, t.ex. mellan korpusar