Språkteknologi (SV2122)
Föreläsning 7: Morfologi och ordklasser

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Richard Johansson

richard.johansson@svenska.gu.se

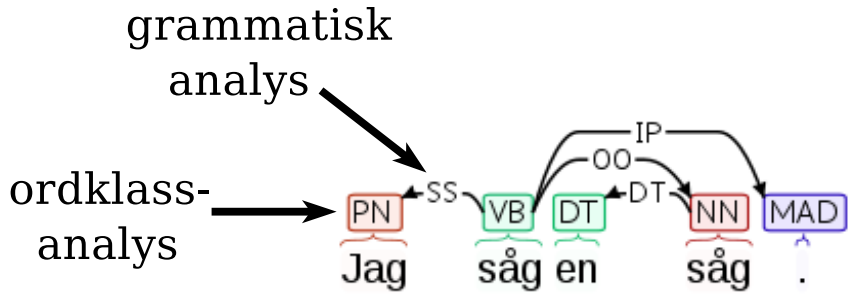
19 februari 2014

praktiska detaljer: tenta

- vilket datum föredrar ni när det gäller tentan?
  - för mig passar t.ex. 14, 17, 19, 21 mars
- jag tänker mig att tentan blir 3 timmar
- vad sägs om en extra övning inför tentan, t.ex. 7 mars?

översikt

- dagens två teman:
  - morfologisk analys – lista de möjliga anlyserna av ett ord
  - ordklassanalys – morfologisk disambiguering
- båda uppgifterna är nödvändiga för att vi ska kunna göra en mer fullständig grammatisk analys (nästa föreläsning)



översikt

morfologisk analys

ordklassanalys

fortsättningen

översikt morfologisk analys

- för en given ordform, lista vilka anlyser som är möjliga
- analys: vanligen i form av
  - grundform
  - böjning
  - ibland också analys av avledningar, sammansättningar, ...

såg $\rightarrow$ substantivet ***såg*** i obestämd form singular
verbet ***se*** i aktiv preteritum

morfologisk analys med uppslagning

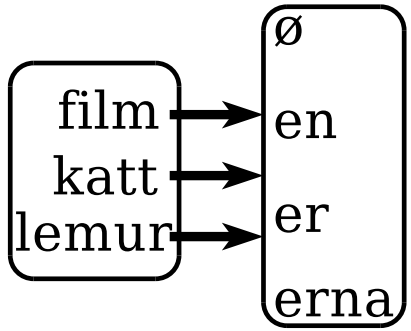
- många läroböcker i språkteknologi är kortfattade om morfologisk analys
- . . . vilket förmodligen beror på det i engelska funkar rätt så bra att bara lagra alla ordformer i en ordlista
- antal böjningsformer för regelbundna ord:
  - substantiv: *dog*, *dogs*
  - verb: *play*, *plays*, *played*, *playing*
  - adjektiv: *high*, *higher*, *highest*
- "analysen" sker därefter med uppslagning i ordlistan

när uppslagning är opraktiskt

- se `http://www.ling.helsinki.fi/~fkarlsso/genkau2.html`
- då är det bättre att verkligen beskriva (den regelbundna) morfologin!

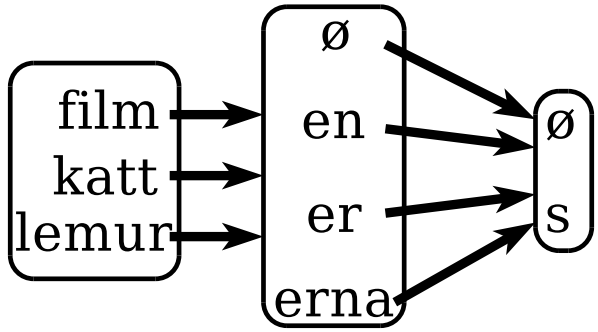
enklaste fallet: böjningstabell

katt
katten
katter
katterna



enklaste fallet: böjningstabell (2)

katt	katt**s**
katt**en**	katt**ens**
katt**er**	katt**ers**
katt**erna**	katt**ernas**

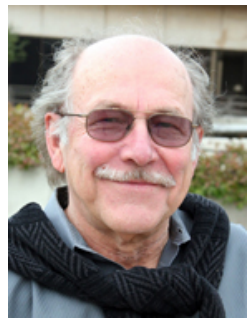


Xerox verktyg

- *Xerox Finite-state Technology Tools* är en uppsättning program för hantering av (bland annat) morfologi
- här intresserar vi oss för två av dessa program:
  - **LexC** – hanterar lexikon och böjningstabeller
  - **XFST** – hanterar omskrivningsregler
- LexC och XFST använder varsitt eget programmeringsspråk
  - till skillnad från Python så är dessa språk användbara enbart för ett syfte: hantering av text

en parentes

- en hel del av uppfinningarna inom datormorfologi har gjorts av finländare
  - ursprungliga idéer om "tvånivåmorfologi" av Kimmo Koskenniemi
  - Xerox-verktygen har utvecklats av bl.a. Lauri Karttunen
- finskans morfologi är komplex men regelbunden – därmed tilltalande att datorisera

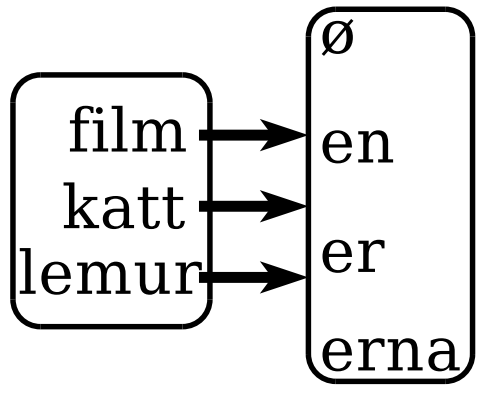


verktyget LexC

- LexC hanterar lexikon och böjningstabeller
- vi beskriver morfologin genom att säga hur deltabeller är ihopkopplade

exempel på en enkel böjningstabell i LexC

```
LEXICON Svenska
 film  Substantiv;
 katt  Substantiv;
 lemur Substantiv;

LEXICON Substantiv
 0    #;
 en   #;
 er   #;
 erna #;
```



morfologiska särdrag i LexC

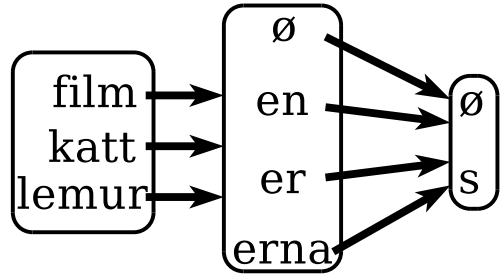
- vi vill göra mer än att bara dela upp *katten* till *katt-en*: vi vill tala om att *-en* är bestämd form singular
- i den här typen av verktyg brukar man tala om en "översida" och en "undersida"
- översidan består av grundform och särdrag; undersidan av rötter och morfem
- i LexC brukar man typiskt skriva särdrag med speciella symboler, t.ex. `+Sing` för singular

samma exempel inklusive särdrag

```
MULTICHAR_SYMBOLS +Subst +Sing +Plur +Best +Obest

LEXICON Svenska
 film  Subst;
 katt  Subst;
 lemur Subst;

LEXICON Subst
 +Subst+Sing+Obest:0   #;
 +Subst+Sing+Best:en   #;
 +Subst+Plur+Obest:er  #;
 +Subst+Plur+Best:erna #;
```

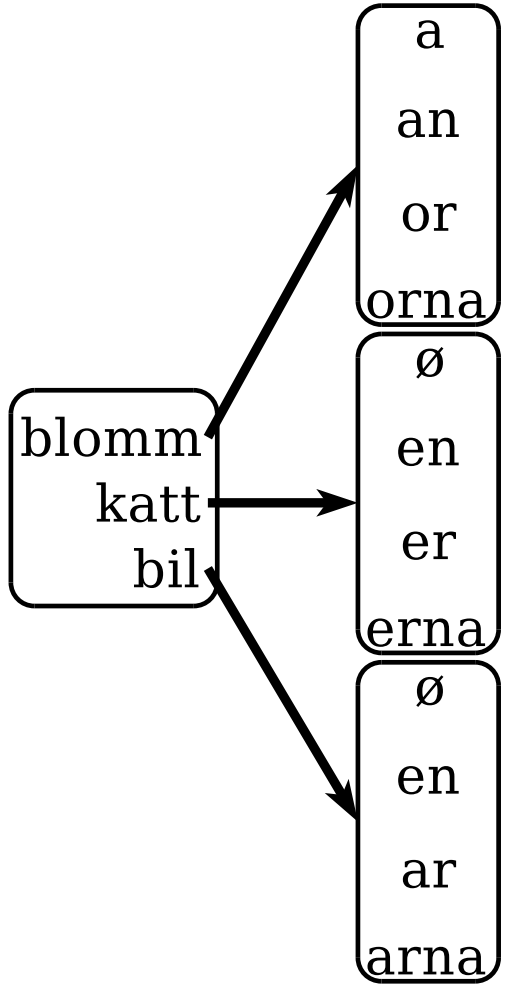
lite mer komplicerad böjningstabell

```
MULTICHAR_SYMBOLS +Subst +Sing +Plur +Best +Obest

LEXICON Svenska
 film  Subst
 katt  Subst;
 lemur Subst;

LEXICON Subst
 +Subst+Sing+Obest:0   Subst_slut;
 +Subst+Sing+Best:en   Subst_slut;
 +Subst+Plur+Obest:er  Subst_slut;
 +Subst+Plur+Best:erna Subst_slut;

LEXICON Subst_slut
 +Nom:0 #;
 +Gen:s #;
```



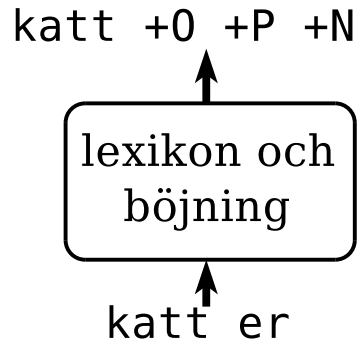
flera böjningsparadigm

```
LEXICON Svenska
 blomm Subst1;
 bil Subst2;
 katt Subst3;

LEXICON Subst1
 +Subst+Sing+Obest:a   Subst_slut;
 +Subst+Sing+Best:an   Subst_slut;
 +Subst+Plur+Obest:or  Subst_slut;
 +Subst+Plur+Best:orna Subst_slut;

LEXICON Subst2
 +Subst+Sing+Obest:0   Subst_slut;
 +Subst+Sing+Best:en   Subst_slut;
 +Subst+Plur+Obest:ar  Subst_slut;
 +Subst+Plur+Best:arna Subst_slut;
```

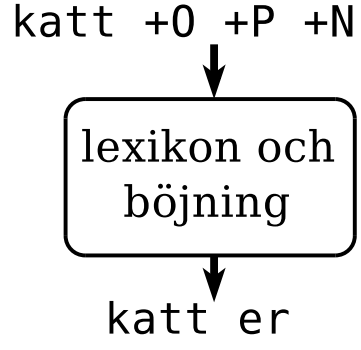


i allmänhet: finite-state morfologi

- om morfologin kan göras med hjälp av en minneslös automat så kallas den *finite-state* (ändlig-tillstånd)
  - t.ex. våra ihopkopplade böjningstabeller



- det är lika lätt för automaten att hantera "översidan" och "undersidan", så analys och generering blir två sidor av samma mynt



ett krångligare exempel

- vissa ord i svenska har ett “instabilt” e som kan falla bort när morfem tillkommer
- exempel:
  - ök**e**n, öknen, öknar
  - nyck**e**l, nyck**e**ln, nycklar
  - vack**e**r, vackra

ful böjningstabell

```
MULTICHAR_SYMBOLS +Subst +Sing +Plur +Best +Obest

LEXICON Svenska
 ök Subst2_en;

LEXICON Subst2_en
 +Subst+Sing+Obest:en   Subst_slut;
 +Subst+Sing+Best:nen   Subst_slut;
 +Subst+Plur+Obest:nar  Subst_slut;
 +Subst+Plur+Best:narna Subst_slut;
```

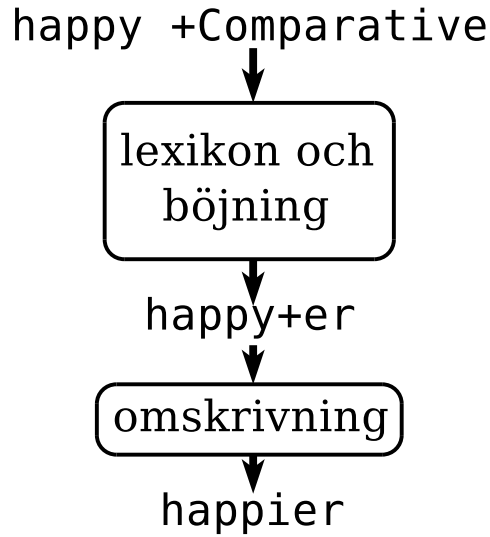
ortografiska regler

- engelska:
  - happy, happier
  - call, calls, called – try, tries, tried – men play, plays, played
- italienska:
  - compro, compri, compriamo, comprano
  - pago, paghi, paghiamo, pagano
- svenska:
  - glass + strut = glasstrut

exempel finska

- vokalharmoni:
  - talo**ssa** – kylä**ssä**; olet**ko** – syöt**kö**
  - ännu krångligare i turkiska, där harmonin har två dimensioner
- konsonantens styrka påverkas av öppen/sluten stavelse:
  - kau**pp**aa – kau**p**an

omskrivningsregler



- de beskrivna svårigheterna löses enklast med hjälp av **omskrivningsregler** som läggs "under" lexikonet
- typiska anledningar
  - ortografi, t.ex. $y \rightarrow ie$, $g \rightarrow gh$
  - fonologi, t.ex. vokalharmoni, produktivt omljud, "instabilt" *e*
- XFST är Xerox verktyg för att hantera omskrivningsregler

exempel

- här är ett XFST-program för att hantera fall som *öken/öknen*
- ett typiskt trick: vi använder en specialsymbol `E` för att representera den instabila vokalen
  - i lexikonet säger vi då att ordet är `ökEn`

```
define Vokal [ a | o | u | å | e | i | y | ä | ö ];

regex E -> 0 || _ ?* Vokal
           .o.
     E -> e;
```

- "det instabila e försvinner om det finns en vokal efter"
- . . . "och annars blir det till ett vanligt e"

exempel, finsk vokalharmoni

- vi antar att “specialvokalen” *A* realiseras som *a* eller *ä* beroende på sammanhang
  - böjningstabellen innehåller morfem i stil med `ssA`, `llA`

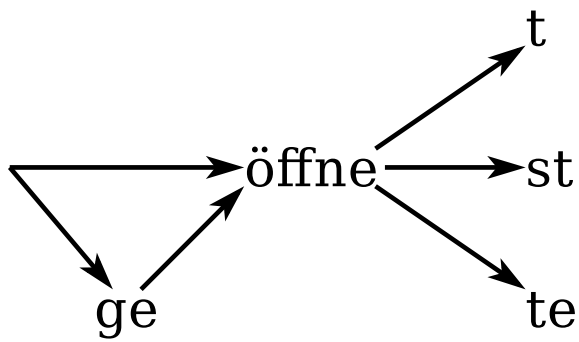
```
define BakreVokal [ a | o | u ];

regex A -> a || BakreVokal ?* _
       .o.
      A -> ä
       .o.
      U -> u || BakreVokal ?* _
       .o.
      U -> y
       .o.
      O -> o || BakreVokal ?* _
       .o.
      O -> ö;
```

designfrågor när man beskriver morfologi

- hur mycket ska gå till lexikonet och böjningstabellerna? – hur mycket ska man formulera som omskrivningsregler?
  - *nyckel*, *nycklar*?
  - *fot*, *fötter*?
- det som är lingvistiskt/historiskt korrekt behöver inte nödvändigtvis vara det mest praktiska!

exempel på fall då finite-state inte fungerar så bra

- finite-state-metoden är effektiv och lättförståelig men minneslösheten leder till vissa begränsningar
- exempel tyska:
  - finita former *öffne, öffne-st, öffne-t, öffne-te* etc
  - perfekt particip *ge-öffne-t* (**cirkumfix**)



- andra svårhanterade morfologier:
  - reduplikation, t.ex. indonesiska *burung* (fågel) – *burung-burung* (fåglelflock)
  - infix, . . .

“mallmorfologi”

- i “mallmorfologi” (*templatic morphology*) sätts böjningsmorfemen in i ett skelett
- mest känt i semitiska språk, t.ex. arabiska:
  - roten **k**-**t**-**b** betecknar skrivrelaterade saker
  - vokaler sätts in mellan konsonanterna för att böja eller avleda
  - t.ex. **k**a**t**a**b**a ‘han skrev’; u**k**tu**b** ‘skriv!’; **k**i**t**aa**b** ‘bok’
- . . . men finns även i verb med avljud i germanska språk:
  - svenska: **s**i**tt**a; **s**a**tt**; **s**u**tt**it; **s**ä**tt**a
- blandningen av lexikonrot och morfem gör livet svårt för finite-state-morfologi!

översikt

morfologisk analys

ordklassanalys

fortsättningen

ordklassanalys: översikt

jag	*såg*	*en*	*såg*
		artikel	
pronomen	**verb**	grundtal	verb
substantiv	substantiv	substantiv	**substantiv**
		pronomen	
		adverb	

- en morfologisk analys t.ex. med Xerox-verktyg ger oss alla möjliga analyser för varje ord
- **ordklassanalys** eller **morfologisk disambiguering** innebär att välja en av dem
  - "ordklasstaggning" (*part-of-speech tagging*), "ordklassmärkning"

“ordklass”?

- det finns en viss begreppsförvirring
  - om vi vill vara precisa så bör vi kanske säga “morfologisk disambiguering”
- i det enklaste fallet: skilj mellan ordklasser t.ex. verb och substantiv
- men i många fall vill man också skilja på möjliga analyser
  - *hänger*: av *hänge* eller *hänga*?
  - *koppar*: *koppar* i singular, eller *kopp* i plural?

måste man inte förstå hela meningen?

- en berättigad invändning är kanske att det finns fall då man måste läsa och förstå hela meningen för att kunna avgöra vilken böjning det handlar om
- t.ex., är *verk* singular eller plural? det kan vi inte avgöra från ett litet begränsat sammanhang!

Den nionde symfonin är Beethovens viktigaste ***verk****.*
De tidiga symfonierna är Beethovens viktigaste ***verk****.*

- . . . men i de flesta fall klarar vi oss ganska bra med ett begränsat sammanhang, och en fullständig grammatisk analys (se nästa föreläsning) är en betydligt större utmaning

annoterade korpusar

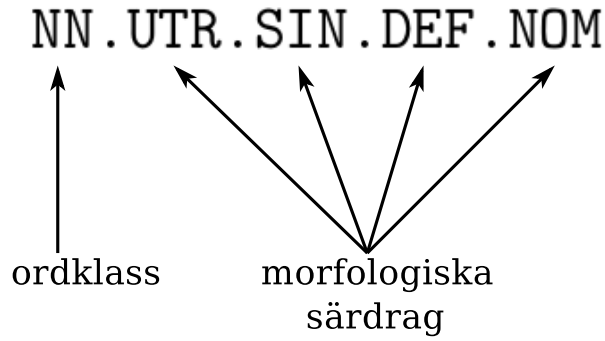
- ett ordklassanalysprogram avgör ett ords analys genom att försöka gissa vad som är **mest troligt** för ett ord **i ett sammanhang**
- vad betyder “mest troligt”? kanske “vanligast”?
- för att ha en aning om vad som är “vanligast” så behöver vi korpusar där varje ord har fått sin korrekta analys
  - dessa korpusar kan sedan användas för att samla in statistik om vad som är vanligast, för att hitta sammanhangsberoende regler, etc
  - . . . samt för att utvärdera

exempel, Stockholm–Umeå Corpus

Avspänningen	NN.UTR.SIN.DEF.NOM
mellan	PP
stormaktsblocken	NN.NEU.PLU.DEF.NOM
och	KN
nedrustningssträvanden	NN.NEU.PLU.IND.NOM
i	PP
Europa	PM.NOM
har	VB.PRS.AKT
inte	AB
mycket	PN.NEU.SIN.IND.SUB/OBJ
motsvarighet	NN.UTR.SIN.IND.NOM
i	PP
Mellanöstern	PM.NOM
.	DL.MAD

exempel på etikett i SUC

Avspänningen



- i SUC består etiketterna av ordklass, samt morfologiska särdrag som genus, numerus, bestämdhet, kasus, tempus, ...
- i det följande kommer vi att använda enbart den första delen (ordklassen) för tydlighets skull

en enkel idé

- vi ser hur långt vi kan komma med en mycket enkel metod:
  - för varje ord, välj den analysen för detta ord som är vanligast i den annoterade korpusen
  - om vi inte sett ordet förr, välj den vanligaste analysen för alla ord (för svenska blir det substantiv)

jag	*såg*	*en*	*såg*
pronomen	verb	artikel	**verb**

enkel ordklassanalys i Python

- ett Python-program blir enkelt: vi behöver bara ha en tabell där vi minns den vanligaste ordklassen för varje ord
  - . . . samt den vanligaste totalt

```
for ord in mening:
    if ord in ordklasstabell:
        print ord, ordklasstabell[ord]
    else:
        print ord, vanligaste_ordklass
```

```
jag  PN
såg  VB
en   DT
såg  VB
```

utvärdering

- för att utvärdera ett ordklassanalysverktyg kan vi göra ungefär som för textkategorisering
- korrekthet (*accuracy*):

$$A = \frac{\text{antal korrekt analyserade ord}}{\text{antal ord i korpusen}}$$

- för ett bra program brukar man typiskt ha 95–97% korrekthet
  - . . . alltså, i snitt ett fel per 20–40 ord
- hur svårt det är beror på hur detaljerad uppsättning av ordklasser man använder
  - . . . samt hur stor vår annoterade korpus var

utvärdering med SUC

- vi använder den tidigare nämnda enkla metoden och bygger tabellen utifrån 80% av SUC
- resterande 20% för att testa hur bra det funkar
- resultat: 89.7% korrekthet
- exempel på ett fall där det trasslat till sig ordentligt:

Mindes†	*de*	*att*	*de*	*hade*	*ett*	*hem*	*?*
subst	**art**	**infm**	**art**	verb	art	**part**	ip

att använda sammanhanget

- det är inte så vanligt t.ex. att artiklar följs av verb (*en såg*) eller partiklar (*ett hem*)
  - . . . så vi kanske kan rätta några av felen genom att titta på orden runtomkring?
- vi kommer att titta snabbt på två idéer för att dra nytta av sammanhanget:
  - statistiska program: räknar vilka kombinationer av ordklasser som är de vanligaste
  - transformationsbaserade program: försöker rätta de vanligaste felen

statistiska ordklassanalysprogram med sammanhang

- statistiska ordklassmärkare innehåller två slags sannolikheter:
  - **transitioner** (övergångar): om det här ordet är ett substantiv, vad är sannolikheten att nästa ord är ett verb?
    - eller ännu mer avancerat: vad är sannolikheten att en artikel och ett substantiv följs av ett verb?
  - **emissioner**: om det här ordet är ett substantiv, vad är sannolikheten att det är *såg*?
- för varje möjlig ordklass-sekvens så räknar vi ut sannolikheten genom att multiplicera ihop sannolikheterna för alla transitioner och emissioner

(start)	pronomen	verb	artikel	substantiv	(slut)
	jag	*såg*	*en*	*såg*	

exempel

- exempel: *jag såg en såg*
  - alternativ 1: pronomen verb **artikel substantiv**
  - alternativ 2: pronomen verb **artikel verb**

$$P(\text{art} \rightarrow \text{subst}) \cdot P(\text{såg}|\text{subst})$$

eller

$$P(\text{art} \rightarrow \text{verb}) \cdot P(\text{såg}|\text{verb})$$

exempel: *en såg*

- sannolikheterna får vi genom att titta i SUC:

$$P(\text{art} \rightarrow \text{subst}) \cdot P(\text{såg}|\text{subst}) = \frac{\text{antal art} \rightarrow \text{subst}}{\text{antal art}} \cdot \frac{\text{antal subst } \textit{såg}}{\text{antal subst}}$$

$$= \frac{24148}{55798} \cdot \frac{5}{234851}$$

$$= 0.43 \cdot 0.000021 = 0.0000092$$

$$P(\text{art} \rightarrow \text{verb}) \cdot P(\text{såg}|\text{verb}) = \frac{\text{antal art} \rightarrow \text{verb}}{\text{antal art}} \cdot \frac{\text{antal verb } \textit{såg}}{\text{antal verb}}$$

$$= \frac{27}{55798} \cdot \frac{616}{175255}$$

$$= 0.00048 \cdot 0.0035 = 0.0000017$$

exempel: *jag såg*

$$P(\text{pron} \rightarrow \text{subst}) \cdot P(\text{såg}|\text{subst}) = \frac{\text{antal pron} \rightarrow \text{subst}}{\text{antal pron}} \cdot \frac{\text{antal subst } \textit{såg}}{\text{antal subst}}$$

$$= \frac{1302}{57947} \cdot \frac{5}{234851}$$

$$= 0.022 \cdot 0.000021 = 0.00000048$$

$$P(\text{pron} \rightarrow \text{verb}) \cdot P(\text{såg}|\text{verb}) = \frac{\text{antal pron} \rightarrow \text{verb}}{\text{antal pron}} \cdot \frac{\text{antal verb } \textit{såg}}{\text{antal verb}}$$

$$= \frac{29073}{57947} \cdot \frac{616}{175255}$$

$$= 0.50 \cdot 0.0035 = 0.0017$$

utvärdering med SUC igen

- återigen uppdelat 80/20
- resultat: 94.6% korrekthet, dvs vi har blivit av med ungefär hälften av felen som det naiva analysprogrammet gjorde
- vi lyckas bättre i den mening som förut gick så illa:

Mindes†	*de*	*att*	*de*	*hade*	*ett*	*hem*	*?*
subst	**art**	**infm**	**art**	verb	art	**part**	ip
verb	pron	subj	pron	verb	art	subst	ip

transformationsbaserade ordklassanalysprogram

- transformationsbaserade program använder **omskrivningsregler** i stil med

  "byt verb mot substantiv om det föregås av en artikel"

- arbetsgång för att samla sådana regler utifrån en annoterad korpus:
  1. börja med den "enkla metoden" vi tidigare gått igenom
  2. konstruera den omskrivning som rättar flest fel
  3. repetera punkt 2 tills det inte längre går att hitta fler användbara regler
- brukar fungera nästan lika bra som statistiska program
  - . . . och har kanske fördelen att reglerna är lättare att tolka

exempel på ordklassanalysprogram

- Språkbanken använder ett verktyg som kallas *HunPos*
  - `http://code.google.com/p/hunpos/`
- . . . och det finns en hel del andra program
- ingår också i NLTK, vilket ni kommer att få möjlighet att prova i labbet

exempel, annoteringslabbet

`http://spraakbanken.gu.se/korp/annoteringslabb/`

översikt

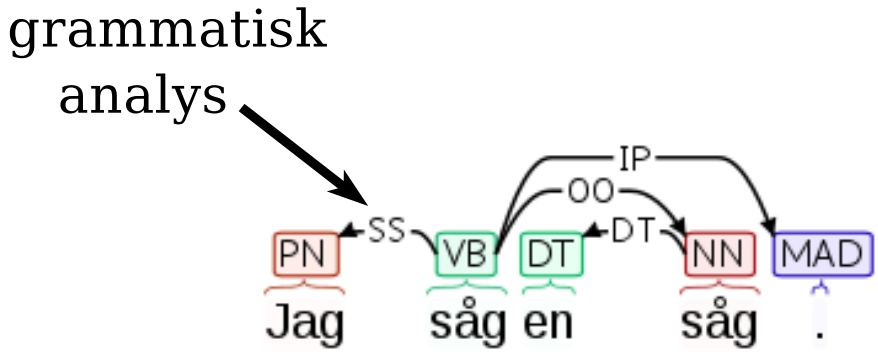
morfologisk analys

ordklassanalys

fortsättningen

nästa föreläsning

- syntaktisk (grammatisk) analys



- fredag, K235 (LT-gatan)