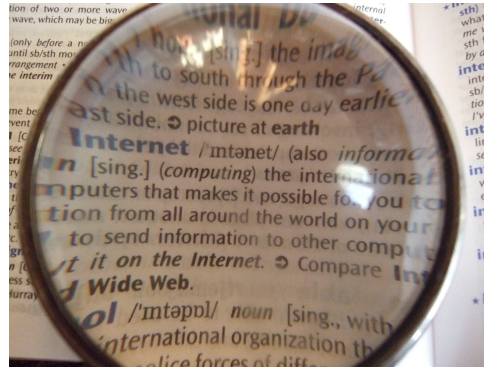


Ord och morfologi



DH2418 Språkteknologi
Johan Boye

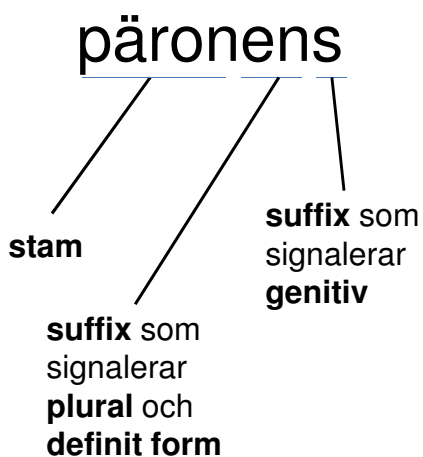
Morfologi

- Läran om hur orden är uppbyggda av mindre betydelsebärande enheter som kallas **morfem**.
- Morfem tillhör en av två klasser:
 - **stam**: den grundläggande innehållsbärande enheten
 - **affix**: små enheter som läggs till stammen för att signalera olika grammatiska funktioner
- Affix kan i sin tur delas upp i **prefix** (**be**-tänka), **suffix** (bord-**et**), **infix** (korru-**m**-pera), **cirkumfix** (**ge**-sag-**t**)

Morfologi

- Ord kan **böjas** för att signalera grammatisk information:
 - päron, pärons, päronet, päronets, päronen, päronens
- Ord kan också **avledas** och skapa nya ord, eventuellt av en annan ordklass:
 - motiv → motivera → motivering
- Ord kan **sättas samman** till nya ord:
 - distrikt + sköterska + mottagning → distriktssköterskemottagning

Böjningsmorfologi



- Grammatisk info:
 - nn.neu.plu.def.gen
- **Morfologisk analys:**
 - ytform → stam + grammatisk info
- **Morfologisk generering:**
 - stam + morfologisk info → ytform

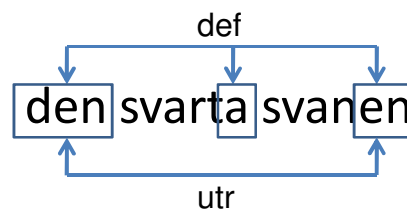
Morfologisk analys

- Att utifrån ett ord identifiera dess stam och grammatiska information.
- Kan man inte lagra alla former av ett ord i ett lexikon?
 - Jo, ibland (t.ex. taligenkänning).
 - Men oftast inte, pga **sammansatta ord, förkortningar, nybildning av ord, avledningar, låneord, namn.**
 - Vissa språk (t.ex. finska) har en mycket invecklad morfologi och en otrolig mängd olika ordformer.

Morfologisk analys

- Varför är det viktigt?

- **Stavningskontroll**
- **Grammatisk kontroll**



- **Informationssökning**
bättre resultat om man söker på alla former av ett ord

Substantiv

- Substantiv har följande parametrar:
 - genus: **utrum** (*stol*), **neutrum** (*bord*)
 - species: **indefinit** (*stol*), **definit** (*stolen*)
 - numerus: **singular** (*stol*), **plural** (*stolar*)
 - kasus: **genitiv** (*stolens*) eller **nominativ**
- Inte hela sanningen; finns olika pluralmönster
 - *apelsin, äpple, päron, häst, pojke, flicka, gås, stimulus, ...*

Verb

- Regelbundna verb:
 - *cykla, cyklar, cyklade, cyklat, cyklande, cyklandes*
 - *springa, springer, sprang, sprungit, springande, springandes*
- Oregelbundna verb:
 - *vara, är, var, varit, varande, varandes ...*
- Äldre svenska har många fler verbformer:
 - *äro, ären, voro, vorit, vore, vare, finge, ginge, ...*
 - ett fåtal av dessa lever kvar (t.ex. *vore*)
- Många språk har många fler verbformer:
 - spanska 50, klassisk grekiska 350, turkiska 2 miljoner(!)

Adjektiv

- Kan kompareras:
 - *spännande* – *mer spännande* – *mest spännande* (analytisk)
 - *fin* – *finare* – *finast* (syntetisk agglutinerande)
 - *låg* – *lägre* – *lägst* (syntetisk flekterande)

Avledningar

- Svenska (och engelska) har enkel böjningsmorfologi jämfört med andra språk.
- Hur är det med avledningar?
 - **-era**: *motiv, motivera* [nn → vb]
 - **-are**: *jaga, jagare, (jägare)* [vb → nn]
 - **-ing**: *motivera, motivering* [vb → nn]
 - **-het**: *god, godhet* [adj → nn]
 - **-o-**: *trolig, otrolig* [adj → adj]
 - **-bar**: *äta, ätbar* [vb → adj]
- Flera steg: *station* → *stationera* → *stationering*

Sammansättningar

- Svenska kan bilda långa sammansatta ord:
 - distriktssköterskemottagning
- Några klurigheter:
 - foga-s: distriktssköterska
 - vokalbyte: sköterskemottagning
 - vokal försvinner: läkarmottagning
- Efterledet bestämmer tolkningen (båtmotor – motorbåt)
- *polismisshandel – kvinnomisshandel*
- *strumpbyxor?*
- Lexikaliserade sammansättningar: *handboll*

Andra språk

- Tyska: ***Lebensversicherungsgesellschaftsangestellter***
 - “livförsäkringsbolagsanställd”
- Grönländska: ***iglu kpisuktunga***
 - *iglu = hus, kpi = bygga, suk = (jag) vill, tu = själv, nga = mig*
- Finska: ***järjestelmättömyydellänsäkääntököhän***
 - “*inte ens med sin brist på ordning*”

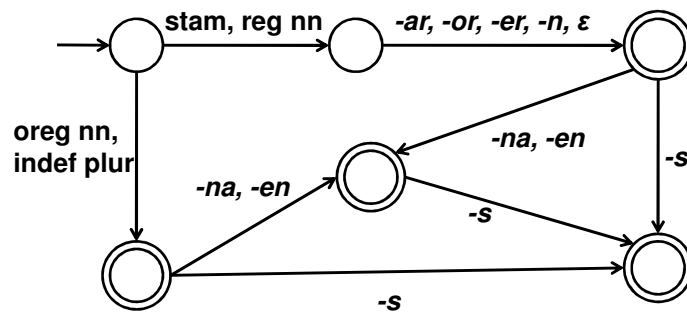
Morfologi och ändliga automater

- Vi vill koda all denna morfologiska information med ändliga automater:
 - acceptera korrekta ord
 - förkasta inkorrekta ord
 - göra detta på ett sätt som inte kräver att vi listar alla ordformer i språket
- Detta krävs:
 - **lexikon**: alla stammar och affix
 - **morfotax**: ordning på affix (*päronens*, **päronsen*)
 - **stavningsregler** för vokalförändring (*sköterskemottagning*)

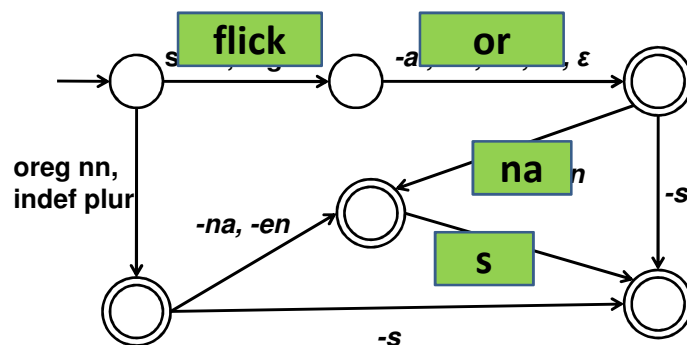
Svenskt plural

- Vi provar göra en automat som accepterar svenska substantiv i plural:
 - *stolar, stolarna, stolars, stolarnas klockor, klockorna, apelsiner, apelsinerna, bord, borden, böcker, böckerna, gäss, gässen*
- Stam + suffix funkar för *stol, klocka, apelsin, bord*
- *Bok, gås* måste specialbehandlas eftersom stammen förändras

Plural: Försök 1

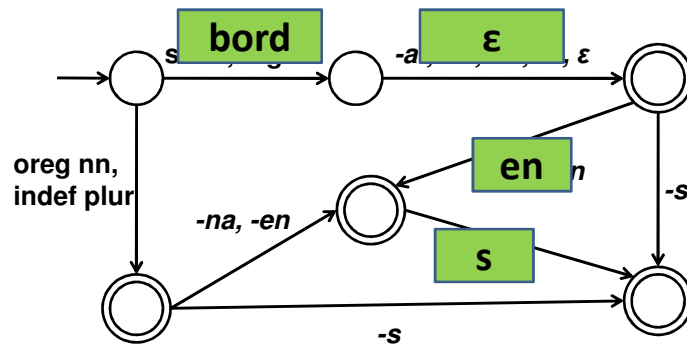


Plural: Försök 1

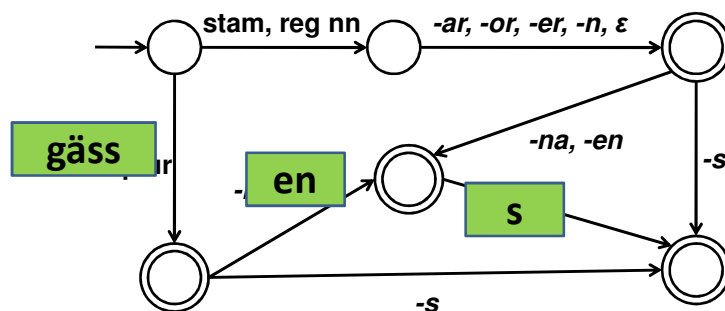


Notera att automaten förkastar **flicknaors*, **flicksorna*, **flicks*, **ornasflick*, osv.

Plural: Försök 1



Plural: Försök 1



Automaten klarar även oregelbundna substantiv. Men vad är det för fel?

Plural: Försök 2



Analys/generering - igenkänning

- Vi kan nu använda automaten för att känna igen korrekta substantiv
 - (fast automaten överaccepterar)
- Men vi vill även kunna göra **analys**
 - *päronens* → *päron* nn.neu.def.plu.gen
- ... och **generering**
 - *päron* nn.neu.def.plu.gen → *päronens*

Tvånivåmorfologi (TWOL)

- Vi tänker oss att en automat har en "remsa" som input läses ifrån
- Lägg nu till en outputremsa
- Bågarna i automaten förses med läs- och skrivinstruktioner

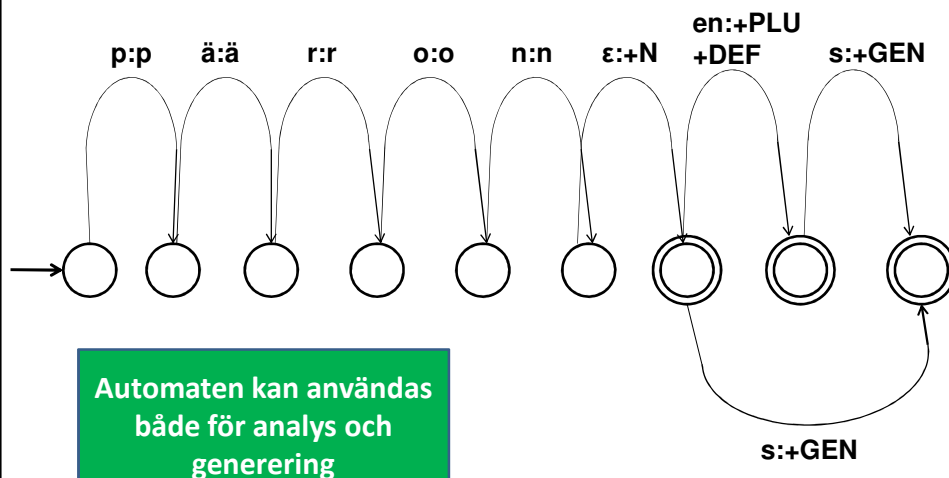
– från ena remsan läser vi
päronens

– på andra remsan skrivs
päron +nn +def +plu +gen

p ä r o n e n s

p ä r o n N D P G

Instruktioner på bågarna



Automaten kan användas
både för analys och
generering

Flertydighet

- Ett ord kan ha flera vägar genom automaten till ett sluttillstånd.
 - Olika vägar genererar olika output.
- Motsvarar flera riktiga analyser, t.ex. *talare*
 - *talare +N +SIN +DEF +NOM*
 - *talat +N +SIN +DEF +NOM*
- Hur hanterar vi det?
 - Returnerar alla svar
 - Returnerar mest troliga svaret (krävs att vi utökar lexikon med sannolikheter)

Att fundera över

- Hur skulle man utforma automaten så den kan hantera
 - *fåglar – fågel +N +IND +PLU +NOM*
 - *fåglars – fågel +N +IND +PLU +GEN*
 - *fåglarna – fågel +N +DEF +PLU +NOM*
 - *fåglarnas – fågel +N +DEF +PLU +GEN*

Lemmatisering

- Vid **informationssökning** vill man även söka på böjningsformer av sökorden
 - cykel, cykla, cykling, osv
- Vid **lemmatisering** ersätts varje ord med sitt **lemma** (grundform)
 - Pojkarnas bilar hade olika färger → pojke bil ha olik färg
 - Detta kräver morfologisk analys

Stemming

- För språk med enkel morfologi (t.ex. engelska) är det inte alltid nödvändigt med morfologisk analys
 - Onödigt att veta att "foxes" är plural av "fox"
- **Stemming** – hugg av morfem för morfem
- Resurssnålt – inget lexikon krävs
- Framför allt användbart i informationssökning (lab 5)

Porters algoritm

- Regelbaserad stemming för engelska
 - ATIONAL → ATE
 - SSES → SS
 - ING → ε
- Viss kontextkänslighet
- (W>1) EMENT → ε
 - REPLACEMENT → REPLAC
 - CEMENT → CEMENT
- Kan bli väldigt fel ibland
 - stocks → stock, stockings → stock

Regelspråk för stemming (lab 5)

#	kommentar
^	negation
\$	ordslut
	slut vänsterkontext
->	skriv om vänsterled till högerled
;	regelslut

```
# Ta bort genitiv-s
s$->;
```

Regler i 3 steg

Steg 1

s\$->;

Steg 2

eln\$ -> el;

en\$ -> ;

Steg 3

cyk|el\$ -> l;

a\$ -> ;

Tokenisering

- Indata: "**äpplen, päron och apelsiner**"
- Utdata: tokens
 - **äpplen päron och apelsiner**
- Vanligtvis avskiljs tokens av mellanslag och skiljetecken.
- Men inte alltid:
 - Birger Jarlsgatan, Johan III, et cetera, ...
 - <http://www.kth.se>, jboyen@nada.kth.se
 - :-)

Tokenisering

- Apostrofer:
 - Finland's → Finland's? Finlands? Finland?
 - don't → don't ? don t ? do not ?
- Ett token eller fler?
 - state-of-the-art → state-of-the-art?
state of the art?
 - Microsoft Word
 - Hewlett-Packard
 - ess-dur

Numeraler

- Kan innehålla mellanslag och skiljetecken
 - 123 456.7 eller 123,456.7 eller 123 456,7
- +46 (8) 790 60 00
- 3/20/91 Mar. 12, 1991 20/3/91
- B-52
- Min PGP-nyckel är 324a3df234cb23e
- 131.169.25.10

Språksspecifikt

- Franska:
 - **L'ensemble** → Le ? L ? L' ?
 - **un ensemble** bör matcha **l'ensemble**

Språksspecifikt

- Kinesiska och japanska har inga mellanslag mellan orden:
 - 莎拉波娃现在居住在美国东南部的佛罗里达。
 - Ej alltid möjligt att tokenisera unikt
- Japanska har flera alfabet
 - **Katakana** and **Hiragana** (stavelsebaserade)
 - **Kanji** (kinesiska tecken)
 - **Romaji** (latinska tecken)
 - alla dessa kan blandas i samma mening

Språkspecifikt

- Höger-till-vänster-språk
 - arabiska, hebreiska, farsi, urdu, pashtu, ...
 - några token (numeraler, årtal, ...) läses vänster till höger
- استقلت الجزائر في سنة 1962 بعد 132 عام من الاحتلال الفرنسي.
- 
- ‘Algeriet uppnådde självständighet år 1962 efter 132 år av fransk ockupation.’
 - Med Unicode är den lagrade formen lätt att tokenisera

Sammanfattning

- Ett språks **morfologi** beskriver hur ord böjs, avleds, sätts samman, och nybildas.
- Morfologisk processning viktigt steg i många applikationer
 - **igenkänning** av korrekta ord
 - **analys** – från ytform till stam + grammatisk info
 - **generering** – andra hållet
- Komplicerad morfologi kan beskrivas med ändliga automater (**tvånivåmorfologi**)
- I enklare tillämpningar duger det med **stemming**