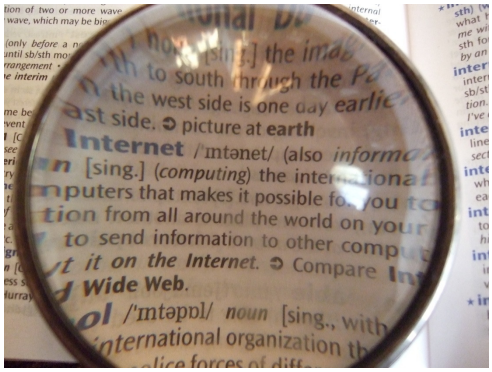
Ord och morfologi



DD2418 Språkteknologi
Johan Boye

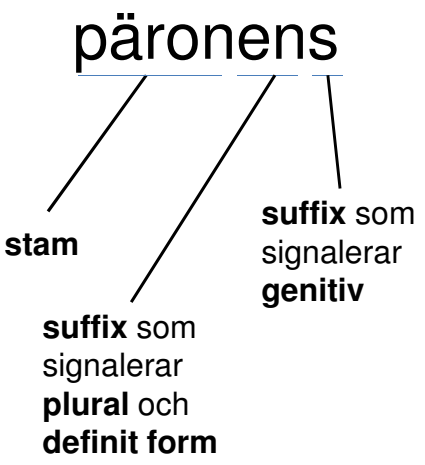
Morfologi

- 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:
  - stol, stols, stolen, stolens, stolar, stolars, stolarna, stolarnas
- 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



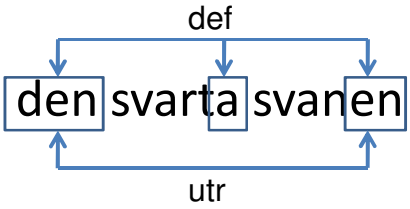
- Grammatisk info:
  - nn.neu.plu.def.gen
- **Morfologisk analys**:
  - ytform → stam + grammatisk info
- **Morfologisk generering:**
  - stam + morfologisk info → ytform

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 kluriheter:
  - foge-s: distrikt**s**sköterska
  - vokalbyte: sköterske**e**mottagning
  - vokal försvinner: läkarmottagning
- Efterledet bestämmer tolkningen (båtmotor – motorbåt)
- *polismisshandel – kvinnomisshandel*
- *strumpbyxor?*
- Lexikaliserade sammansättningar: *handboll*

Andra språk

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

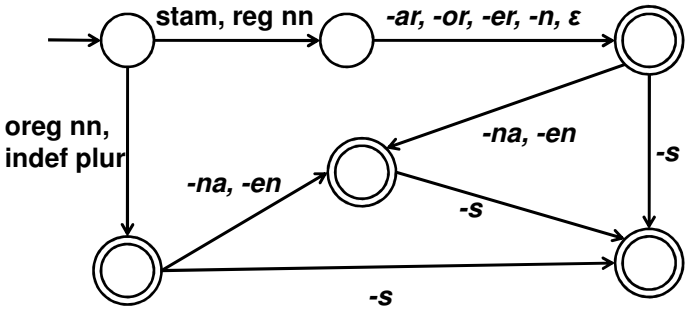
Morfologi och ändliga automater

- 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 (*sköterske**e**mottagning*)

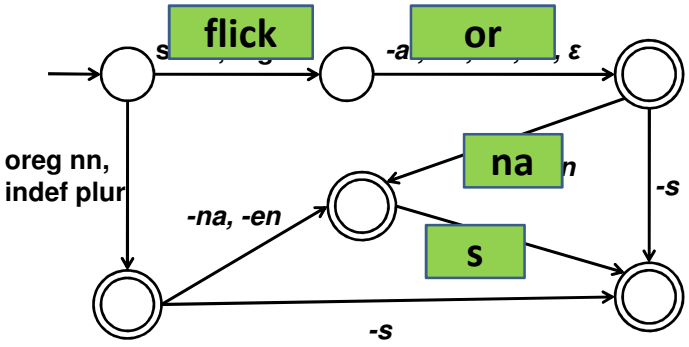
Svenskt plural

- 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

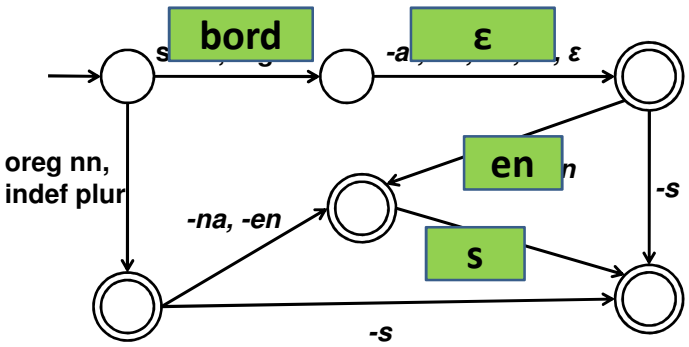


Plural: Försök 1

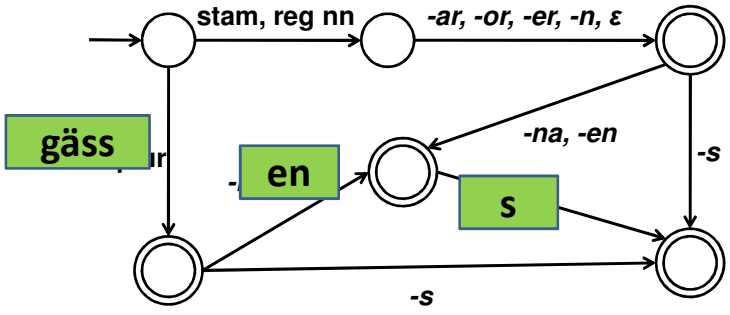


Notera att automaten förkastar ***flicknaors, *flicksorna, *flicks, *ornasflick,*** osv.

Plural: Försök 1



Plural: Försök 1



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

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

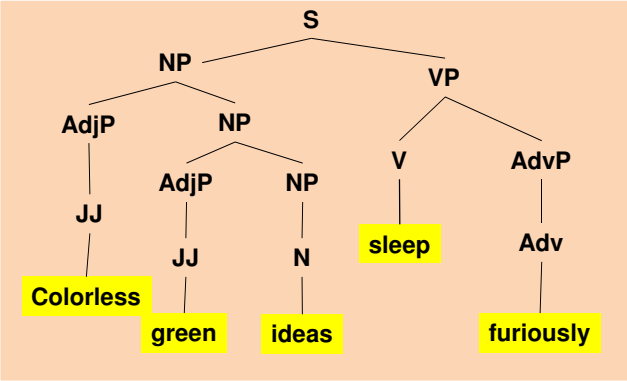
Stemming

- 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 6)

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
- I enklare tillämpningar duger det med **stemming**

Syntax



DD2418 Språkteknologi
Johan Boye

Syntax

- Frågor vi vill besvara:
  - Vilka sekvenser av ord tillhör språket?
  - Vilka relationer finns mellan orden?
- Formella beskrivningar av ett språk:
  - Ändliga automater?
  - Grammatiker, speciellt:
  - **Kontextfri grammatik**, bygger på **konstituenter** som t.ex. nominalfras, verbfras, osv.
  - **Dependensgrammatik**, bygger på **satsdelar** (funktioner)

Tillhör språket?

- Köpte Rune sin nya klocka i lördags?
- Köpte Rune i lördags sin nya klocka?
- Köpte i lördags Rune sin nya klocka?
- Hon slog han på truten.
- Jag hörde inte låten förens igår.
- Han hade en nära döden upplevelse.
- Kalle är bättre än mig.
- Stina gick ända tills skolan.

Ordklasser / kategorier

- Ord delas upp i kategorier eller **ordklasser** (**POS, parts-of-speech, taggar**) efter sina morfologiska och syntaktiska egenskaper.
- T.ex. **substantiv** kan
  - förekomma efter determinerare (*ett päron*)
  - sättas i genitiv-form (*päronets*)
  - sättas i plural-form (med vissa undantag, t.ex. *snö*)
- Finns ingen allmänt accepterad mängd kategorier
  - men flera standarder, t.ex . SUC (Stockholm Umeå Corpus)

Kategorier i SUC (öppna)

Kod	Kategori	Exempel
NN	Substantiv	päron
PM	Egennamn	Kalle
JJ	Adjektiv	grön
AB	Adverb	ofta
VB	Verb	springa
PC	Particip	slängd
IN	Interjektion	aj
UO	Utländskt ord	the

Kategorier i SUC (slutna)

DT	Determinerare	denna
HA	Frågande/relativt adverb	när
HD	Frågande/relativ determinerare	vilken
HP	Frågande/relativt pronomen	vem, som
HS	Frågande/relativt poss. pron.	vems, vars
IE	Infinitivmärke	att
KN	Konjunktion	och
PL	Partikel	på
PN	Pronomen	han
PP	Preposition	under
PS	Possessivt pronomen	din
RG	Grundtal	fyra
RO	Ordningstal	fjärde
SN	Subjunktion	att

Ordklasser / kategorier

- Kategorier är naturliga byggstenar i beskrivningar av språk
- Ord av samma kategori kan förekomma på samma plats i en sats (med vissa undantag):
  - ”Det sitter en **fluga** på väggen”
  - ”Det sitter en **rädsla** på väggen”
  - Syntaktiskt rätt även om innebörden är oklar
- Men även en **grupp av ord** kan ha samma roll:

Konstituenter

- En grupp av ord som kan sägas agera som en enhet
- T.ex. en **nominalfras**:
  - **flickan** började gråta
  - **den lilla flickan** började gråta
  - **den lilla flickan i den röda kappan** började gråta
  - **den lilla flickan i den röda kappan vars föräldrar var lärare** började gråta
- De markerade fraserna ovan kan förekomma i samma kontexter

Konstituenter

- Finns ingen allmän accepterad uppsättning konstituenter
  - precis som med kategorier
  - därför finns det många olika grammatiska teorier och alternativa formaliseringar av samma data

Kontextfria grammatiker

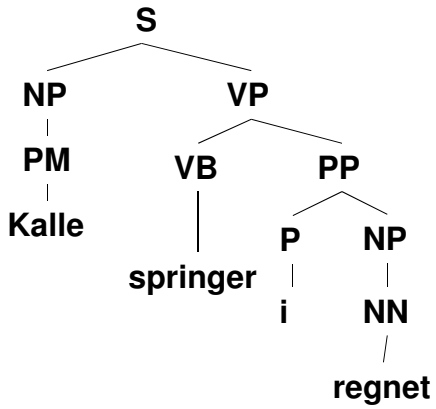
- Andra namn: **CFG**, **frasstrukturgrammatik**, **Backus-Naur Form (BNF)**
- består av:
  - terminaler (**ord**, i vårt fall)
  - icke-terminaler (**konstituenter**, i vårt fall)
  - regler på formen A → α, där α är en sekvens av terminaler och icke-terminaler

Kontextfria grammatiker: Exempel

S → NP VP	NP → PM
S → VP	NP → NN
VP → VB	PM → Kalle
VP → VB PP	NN → regnet
PP → P NP	V → springer
	P → i

Kalle springer i regnet
springer i regnet
regnet springer i Kalle

Syntaxträd



[S [NP [PM Kalle]] [VP [VB springer] [PP [P i] [NP [NN regnet]]]

Övning

- En vanlig mening från SUC-korpuset:
  - ***Viktigaste redskapen vid ympning är annars papper och penna***.
  - Med kategorier: ***Viktigaste*** **(JJ)** ***redskapen*** **(NN)** ***vid*** **(PP)** ***ympning*** **(NN)** ***är*** **(VB)** ***annars*** **(AB)** ***papper*** **(NN)** ***och*** **(KN)** ***penna*** **(NN).**
- Konstruera ett syntaxträd för denna mening! (Hitta på troliga grammatikregler)

Parsning

- **Parsning** — processen att utifrån en mening och en grammatik konstruera ett eller flera syntaxträd.
- Varje syntaxträd svarar mot en tolkning av meningen.
- Finns många olika parsnings-algoritmer.
- Mer om detta i **föreläsning 7** (Mats W)

Kongruens

- Man vill även kunna kontrollera **kongruens**:
  - *den svarta svanen* / *det svart svanen*
- Inte helt okomplicerat alla gånger:
  - *How you can claim that Bolton can beat Manchester United with all their well-payed superstars surprises me.*
- För detta krävs att man håller reda på alla ords **särdrag.**

Särdrag i SUC

Särdrag	Värde	Förklaring	Kategorier
Genus	UTR	utrum	DT HD HP JJ NN PC PN PS RG RO
	NEU	neutrum	
	MAS	maskulinum	
Numerus	SIN	singular	DT HD HP JJ NN PC PN PS RG RO
	PLU	plural	
Species	IND	indefinit	DT HD HP JJ NN PC PN PS RG RO
	DEF	definit	
Kasus	NOM	nominativ	JJ NN PC PM RG RO
	GEN	genitiv	
Tempus	PRS	presens	VB
	PRT	preteritum	
	SUP	supinum	
	INF	infinitiv	

Särdrag i SUC (2)

Särdrag	Värde	Förklaring	Kategorier
Diates	AKT	aktivum	VB
	SFO	S-form	
Modus	KON	konjunktiv	VB
Particip-form	PRS	presens	PC
	PRF	perfekt	
Grad	POS	positiv	AB JJ
	KOM	komparativ	
	SUV	superlativ	
Pronomen-form	SUB	subjektsform	PN
	OBJ	objektsform	
?	SMS	sammansättn.	Alla

Exempel

Det	DT NEU SIN DEF
är	VB PRE AKT
ingen	PN UTR SIN IND SUB/OBJ
slump	NN UTR SIN IND NOM
att	KN
den	PN UTR SIN IND SUB/OBJ
kallas	VB PRE SFO
för	PP
världens	NN UTR SIN DEF GEN
vackraste	JJ SUV UTR/NEU SIN/PLU IND/DEF NOM
sjöresa	NN UTR SIN IND NOM

Fråga: Vilka ord kongruerar med varandra?

Kongruens

- Hur kan vi koda kongruens i en kontextfri grammatik?
- *NP → DT JJ N* tillåter
  - *den svarta svanen*
  - **det svarta svanen*
  - **det svart svan*
  - osv

Verbens subkategorisering

- *Johan nös.* VP → V
- **Johan nös en nyckel.*
- *Kalle lade en nyckel på bordet.* VP → V NP PP
- **Kalle lade.*
- **Kalle lade en nyckel.*
- *Kalle gav honom ett paket.* VP → V NP NP
- *Kalle gav ett paket.* VP → V NP
- **Kalle gav honom.*
- *Eva påstod att han hade ljugit.* VP → V att S
- **Eva påstod*
- **Eva påstod ett påstående.*

Verbens subkategorisering

- Alla verb kan inte förekomma i alla VP-regler:
  - **Johan nös en nyckel.*
  - **Kalle lade en nyckel.*
  - **Eva påstod ett påstående.*
- Vi kan subkategorisera verben efter vilka VP-regler de kan förekomma i.
  - Generalisering av transitivt/intransitivt verb
  - Moderna grammatiker kan ha tiotals eller hundratals klasser av verb.
  - Subkategorisering av verb kan hanteras på samma sätt som kongruens

Alltså:

- Med kontextfria grammatiker kan vi **hantera många (alla?) syntaktiska fenomen**
- Men det finns problem:
  - t.ex. kongruens, subkategorisering
  - kan hanteras inom formalismen, men grammatikerna blir ofta stora
- Finns bättre lösningar (t.ex. unifieringsgrammatik, kap 15 i boken).
  - mer kraftfulla än kontextfria grammatiker (högre i Chomsky-hierarkin)