

Morfologi och automatisk morfologisk analys och generering

Språkteknologi DH2418
Ola Knutsson
2010

Varför är morfologi viktigt?

- Ord är grunden i alla världens språk
- Alla språkteknologiska aktiviteter kräver kunskap om ordens morfologi (nästan).
- Är och var kommer från ordet *vara*
- *vaar* är ett stavfel i svenska språket

Kan man inte lägga alla ord i en lexikonfil?

Finskt exempel ...
Täckningsexempel
“Nybildning”, t.ex. låneord, namn och sammansättningar

Vad är ett ord (idag vs. i dag)

Graford – omges av mellanslag, behöver inte tillhöra språket

Lemma – grundform, inbegriper alla ordformer av ordet inom en ordklass

Lexem – lexikonord, tillhör språket. Ett lexem flera graford

Ordformer – olika former av samma lemma

Token – en enhet med språkligt innehåll (tokenisering), ord, förkortningar,

Grundläggande begrepp

- **Morfem:** *te/de* men äv. *häst* och *ligg*
- **Stam:** *skilj*, *naturreservat*
- **Rotmorfem:** *skilj*, *natur+reservat*
- **Lemma:** flicka
- **Affix:** *-or*, *-be*, *-m-*, *ge-**, *-t*
- **Prefix:** *-be* betänka, *för-* fördriva
- **Suffix:** *-or flickor*
- **Infix:** *-m-* korrump~~era~~ (korruption)

Lemma, lexem och stam

herre

ORDLED: *herr-en*

LEXEM 1. respekterad man

LEXEM 2. överordnad (förmäm) person (av manligt kön) med oinskränkt bestämmanderätt

Böjningsmorfologi och avledningsmorfologi

Böjningsmorfologi – böjer lexem

bil --> *bilar*, *bilarna*

Avledningsmorfologi – skapar nya lexem: *godhet*, *författandet* ...

Stamförändringar

Springa – sprang –sprungit (starka verb)

OMLJUD: man:män, son:söner

AVLJUD: i:a:u springa ...

y:ö:u flyga flög flugit

REDUPLIKATION (fördubbling): (latin) currit
han springer, cu-currit –han sprang/har
sprungit

STAMBYTE: är – varit, liten-små-mindre

Morfologisk typologi

- **Analytiska**
 - Isolerande språk (vietnamesiska) –överbäggande fria morfem
- **Syntetiska språk** – kombinerar morfem:
 - Flekterande** (svenska) – sammansmälta morfem med mera variation.
 - Agglutinerande** (finska, turkiska) – tydligt urskiljbara morfem (se exempel).

Exempel från en grönländsk dialekt

iglukpisuktunga =

iglu = hus

kpi = bygga

suk = (jag) vill

tu = själv

nga = mig

10

Ordbildningsexempel

järjestelmättömyydelänsäkäänköhän

järki =förnuft, förstånd

järjestää=ordna, arrangera

järjestellä=ordna, arrangera (durativt, iterativt)

järjestelmä = **system** (det man fått då man ordnat och ordnat!)

järjestelmä+ttöm+yyde+llä+nsä+kään+kö+hän

-ttöm av –tön, -ttömän: negerande
ändelse, -lös: **systemlös**

-yyde, av -yys, -yyden:
substantivderivation, -het:

systemlöshet

-llä:kasusändelsen i ordet: med,på:

med systemlöshet

-nsä:possessivsuffix 3p:

med sin systemlöshet

järjestelmä+ttöm+yyde+llä+nsä+kään+kö+hän

-kään:negerande ändelse: inte ens:

inte ens med sin systemlöshet

-kö: frågepartikel:

inte ens med sin systemlöshet?

hän:affirmativ artuikel: väl, men ju,
nog:

inte ens med sin systemlöshet, väl?

Analytisk och syntetisk

spännande	mer spännande	mest spännande	analytisk
fin	finare	finast	syntetisk, agglutiner- ande
låg	lägre	lägst	syntetisk, flekterande

Sammansättningar i svenskan

Sammansättningar: minst två fria morfem

Lexikaliserade (finns i SAOL)

äppelträd, polisbil

Nybildningar

Språkteknologi+klubben

halk+trattar

Sammansättningar i svenskan

- båtmotor -- motorbåt (determinativa, efterledet bestämmer tolkningen)
- polismisshandel – kvinnomisshandel
- Blågul, gulblå (kopulativa, samordnade led)
- strumpbyxor?

Slutna klasser

prepositioner PP: på, i, för

determinerare DT: den, en, det

pronomen PN: den, en, det

konjunktioner KN: och, eller, men

subjunktioner SN: om, som

räkneord RG: ett, 1, 2023

possessiv PS: min, ditt, hennes

Öppna klasser

substantiv NN: bilen, hund

egennamn PM: Kalle, Volvo

adjektiv JJ: gul, bäst

adverb AB: mycket, igår

verb VB: spela, springa

particip PC: kastad, sjungande

Ordklassfördelning i en svensk korpus (SUC, 1 milj. ord)

	Förekomster	Typer
Substantiv	26.64 %	64.59 %
Verb	17.03 %	10.44 %
Pron	16.45 %	0.15 %
Prep	11.40 %	0.10 %
Adv	9.78 %	1.54 %
Adj	7.52 %	17.35 %
Konj	7.38 %	0.05 %

Automatisk morfologisk analys/parsning och generering

- Målsättningar?
- Datareducerande
- Ger böjningsinformation
- Morfologiskt komplicerade språk
- Analys och generering i samma system
- Språkvetenskaplig beskrivning

Utgångspunkter

- Text
- Tokensering, inkl. dela upp i meningar
- Morfologisk processning/analys

21

Stemming

- Regler, inget lexikon
- Mycket resurssnålt
- Främsta tillämpning: informationssökning
- Programmet tar bort morfem för morfem

Sökfrågan stemmas: cykel --> cykl Orden i texten stemmas

dok1: Jag har en fantastisk cykel.
dok2: Jag gillar att cykla.
dok3: Cykling är det bästa jag vet.

-->

dok1: Jag har en cykl.
dok2: Jag gillar att cykl.
dok3: Cykl är det bästa jag vet.

23

Regelspråk för stemming

#	Kommentar
^	Negation
	slut vänster-kontext
\$	ordslut
->	skriver om vänster till höger
;	regelslut
^vänsterkontext morfem\$->ersättning;	
Regler tillämpas i tre steg	

24

Två exempelregler

arna\$->;

ning\$->;

jämför: blåsning tidning tillagning

25

Regler steg 1	Regler steg 2	Regler steg 3

Regler tillämpas toppen-ner, endast en tillämpning i varje steg. Alla regeldelar är optionella utom -> och ;

26

Mer kontroll

^tid|ning\$->;

27

Stemming versus lemmatisering

- Behövs det ett lexikon?
- **Stemming:**
banan (stam=ban) banan (stam=ban)
- **Lemmatisering:**
banan (lemma=bana) banan(lemma=banan)
Många träffar eller relevanta träffar?

Analys och generering

Analys:

Indata: bilen

Utdata: bil NN UTR SIN DEF NOM

Generering:

Indata: bil NN UTR SIN DEF NOM

Utdata: bilen

Viktiga byggstenar för automatisk morfologisk analys och generering

Lexikon: listor med stammar och affix och inte minst särdrag.

Morfotax: regler för i vilken ordning morfemen får sättas samman. (ändliga automater)

Ortografiska regler: för allomorfer och fonologisk hänsyn, t.ex fågel - fåglar (ändliga transduktorer)

Alla former i lexikon?

bil nn.utr.sin.ind.nom
 bils nn.utr.sin.ind.gen
 bilen nn.utr.sin.def.nom
 bilens nn.utr.sin.def.gen
 bilar nn.utr.plu.ind.nom
 bilars nn.utr.plu.ind.gen
 bilarna nn.utr.plu.def.nom
 bilarnas nn.utr.plu.def.gen
 häst nn.utr.sin.ind.nom

31

Ändlig automat

- ett alfabet (lexikon)
- en uppsättning tillstånd
- en uppsättning tillståndsövergångar associerade med en symbol ur alfabetet
- används för att acceptera ord – vi får ingen analys

32

Ändlig transduktor

- kan ses som en automat med ett läsband och ett skrivband.
- Def: dess alfabet är definierat som par av symboler a:b, där a kommer från ett alfabet och b från ett annat.
- Vi kan använda den för att acceptera ord men också ge dem en analys.

33

Tillstånd	Morfem	Aktion	Utdata
0	'poj'k'	gå till 1	pojke NN UTR
0	'spik'	gå till 1	spik NN UTR
1	''	gå till 2	SIN
1	'e'	gå till 2	SIN
1	'ar'	gå till 3	PLU
2	''	gå till 4	IND
2	'en'	gå till 4	DEF
3	''	gå till 4	IND
3	'na'	gå till 4	DEF
4	''	gå till 5	NOM
4	's'	gå till 5	GEN
5	#	gå till 6	
6		stop	

34

bil --> bil r1
 hus --> hus r5

	SIN IND NOM	SIN IND GEN	SIN DEF NOM	SIN DEF GEN	PLU IND NOM	PLU IND GEN	PLU DEF NOM	PLU DEF GEN
1	''	s	en	ens	ar	ars	arna	arnas
5	''	''	et	ets	''	''	en	ens

35

Tvånivåmodellen (TWOL)

- Kimmo Koskenniemi (1983)
- Två nivåer: Lexikal (abstrakt) och Ortografisk (verklig).
- Kan implementeras med ändliga transduktorer.

Utvecklades först med finska i tankarna. Nu implementerad för många språk, t.ex. svenska – Swetwol.

Par av språkliga nivåer

Lexical		f	o	x	+N	+PL		
Intermediat e		f	o	x	^	s	#	
Surface		f	o	x	e	s		

Hur ser språket ut i en vanlig text på 1500 ord?

reda 3	omfatta 1	att 48
sida 1	om 24	är 39
hemsida 1	område 1	på 38
enda 1	omtalad 1	som 36
hända 1	opinion 5	det 33
använda 6	opinionsmätning 4	i 33
förbjuda 1	opinionsteknik 1	och 30
	opinionsundersökning 5	om 30
		av 26
		en 26
		för 22
		ett 19
		till 18
		de 16
		har 15
		vi 15
		den 12
		hur 12
		inte 11
		eurobarometer 10

39

Statistiska metoder

- Statistisk suffixanalys, kräver taggad korpus
- Man får alltid ett (hyfsat) svar – i ett regelbaserat system saknas det kanske regler, eller så är de heuristiska reglerna för grova.
- Svårare med generering
- Billigt, men risk för skräp

vara	varit är Är var vara	verb	45.0
en	En ett en	determinerare	43.0
på	på På	preposition	37.0
i	i I	preposition	29.0
och	och Och	konjunktion	29.0
att	att Att	infinitivmärke	28.0
det	det Det DET	pronomen	27.0
av	av	preposition	24.0
om	om OM Om	preposition	24.0
som	som	pronomen	21.0
att	att Att	subjunktion	20.0
för	för	preposition	20.0
den	den det Den de	determinerare	19.0
ha	har HAR ha	verb	17.0
vi	vi oss VI	pronomen	17.0
till	till	preposition	16.0
opinionsmätning			
Opinionsmätningar	Opinionsmätningarna	opinionsmätningar	
Opinionsmätningarnas	opinionsmätningarnas	opinionsmätningarna	
	substantiv		14.0

40

Marsianerna nn.utr.plu.def.nom
hade vb.prt.akt.aux
alla pn.utr/neu.plu.ind.sub/obj
ganska ab
gröna jj.pos.utr/neu.plu.ind/def.nom
långa jj.pos.utr/neu.plu.ind/def.nom
fingrar nn.neu.plu.ind.nom"
. mad

41

Man vinner kunskap om språkets användning med statistik

- ar substantiv, utr, plu, ind, nom 60 %
- ar verb, prs, akt 40 %

Webbtips

<http://www2.lingsoft.fi/cgi-bin/swetwol>

SWETWOL

Skriv ett ord på svenska (i vilken böjningsform som helst).

Ord:

Ordets analys:

```
"<E0x">
  "f0ca"  V ACT PRES
  "f0ca"  V ACT IMP
  "f0c"   N UTR INDEF SG NOM
  "f0c"   <CLE> CC
  "f0c"   ADV (RAD> RADVL)
  "f0c"   PREP
```

<http://snowball.tartarus.org/>