

**NOEN PROBLEMER VED OPPBYGGING AV ET DATAMASKINELT MORFEM-LEKSIKON.**

-----

Oppbyggingen av ei morfemordbok kan kreve en betydelig arbeidsinnsats. Arbeidsmengden vil i stor grad avhenge av hvor stort materialet for ordboka er, og hvor mye informasjon en vil legge inn i ordboka, f.eks. av statistisk art.

Hittil er det utgitt forholdsvis få morfemordbøker. Ei av de første som baserte seg på datateknikk, var "Russian Derivational Dictionary", utgitt i New York 1970. I forordet til denne ordboka forteller de tre utgiverne (Worth, Cozak og Johnson) at arbeidet med ordboka har tatt 7 år. Det står ikke noe om hvor mange som har arbeidet med dette prosjektet, men det er uten videre klart at det må ha gått med mange årsverk.

Kunne så et tilsvarende norsk prosjekt regne med bevilgninger til f.eks. 20 årsverk (4 personer i 5 år)? Det er neppe realistisk.

Den russiske morfemordboka bygger på ordbokmateriale (ca. 110 000 oppslagsord) og har således ingen informasjon om frekvenser i løpende tekst. Når det gjelder leksikalsk tekst, er det ikke gitt opplysninger om statistiske forhold. En leser som er interessert i frekvensopplysninger på leksikalsk nivå, må altså sjøl gjøre telle-arbeidet. Morfemordboka har bare én sorteringsversjon - forlengs alfabetisk sortering i KWIC-format.

Et atskillig mer ambisiøst prosjekt er den svenske morfemordboka. Arbeidet med denne ordboka har foregått i flere år under ledelse av Sture Allén i Språkdata og er nå så å si avsluttet. Rapportene som er utgitt om denne ordboka, viser at det er lagt ned et betydelig arbeid for å gi mest mulig informasjon om morfologiske forhold i svensk (jfr. f.eks. "Morf-o-log-i och poly-sem-i" av Sture Allén).

Det kan diskuteres hvor mye arbeid en skal legge ned i slike ordbøker. Jo større arbeidsinnsats, desto større informasjon for leseren. Et mål for den norske morfemordboka bør nok være at den skal gi minst like mye informasjon som den russiske morfemordboka. Men det er neppe realistisk å gå i gang med et så omfattende prosjekt som den svenske morfemordboka.

Ved denne konferansen deltar det flere av de som har arbeidet med den svenske morfemordboka. De vil sikkert kunne gi opplysende svar på en rekke av de problemene som knytter seg til oppbyggingen av ei morfemordbok. I resten av dette foredraget vil jeg begrense meg til et spesielt problem-område - nemlig den morfologiske behandlingen av innlånte ord.

Først noen opplysninger om materialet:

1. Avistekster - ca. 700 000 løpende ord og ca. 45 000 lemma
2. Bokmålsdelen av "Norsk ordregistrant" - ca. 60 000 oppsl.ord
3. Nynorskdelen " " " " 60 000 "

-----

En stor del av de innlånte orda er av romansk eller gresk opphav. Med kjennskap til morfemstrukturen i det långivende språket kan en lett skille ut de enkelte morfemene i et ord. Eksempelvis har verbet *assistere* i latin morfemstrukturen *as-sist-ere*. Men skal dette fremmedordet også få tilsvarende deling i norsk? I så fall ville det være vanskelig å angi betydningen for de utskilte elementene. Hvilken betydning skulle f.eks. ligge i *as* og *isist* *assistere*?

For noen brukere av ordboka vil det sannsynligvis være nyttig at de forskjellige morfemer i det långivende språk er skilt ut. For andre brukere kan dette føre til at morfemordboka blir unødig vanskelig å finne fram i. Spørsmålet er så om en kan tilgodese begge brukergrupper - med sorteringsversjoner tilpasset hver gruppe?

Dersom morfemene i det långivende språket ble behandlet på linje med regulære morfemer i norsk, ville en få sorteringer som denne i KWIC-format (forlengs alfabetisering av "nøkkel-morfem" med sorteringsprioritering av venstre kolonne før høyre kolonne):

```
Sorteringsversjon_A      film pro duc er
                          turtel due miljø
                          dug nød s time
                          av duk ing
                          intro duk sjon
                          lin duk
                          pro duk t
                          pro duk t iv
                          re duk sjon
                          styrke re duk sjon
                          dukk ert
                          edder dun s dyn e
                          fly dur
                          intro dus er e
                          re dus er ing
                          pro dus ent
                          sal at pro dus ent
```

Denne sorteringsversjonen har interesse for dem som vil ha en oversikt over innlånte ord med et bestemt element, f.eks. -dus-.

De som har kjennskap til latin, vil her finne rota av verbet ducere (= "føre") i forskjellige kontekster. P.g.a. ortografiske regler i norsk kommer imidlertid ikke disse beleggene direkte etter hverandre, men det har trolig mindre for seg å gruppere sammen rotvarianter som -duk-, -dus- og duc-.

De som er interessert i en slik sorteringsversjon, er sannsynligvis så godt inne i morfemsystemet i andre språk at de sjøl finner fram til slike rotvarianter. Dessuten vil det i en del tilfelle by på problemer om enheter med samme etymologiske rot skal sammenføres eller ikke. Skal f.eks. loi- i lojal (lånt fra fransk) og leg- i legal (fra latin) betraktes som rotvarianter og i så fall presenteres under hverandre? Begge disse rotvariantene skriver seg fra lat. leg. Sannsynligvis vil det være vanskelig å formulere klare regler for når slike rotvarianter skal sammenføres eller ikke.

Dersom en innfører spesialtegn som markører av morfemgrenser i det långivende språket, vil en ha mulighet for sorteringsversjoner som er bedre egnet for de fleste brukere av morfemordboka. Nedafor har vi presentert det samme ordmaterialet som på forrige side, men markert grensa mellom prefiks og rot i det långivende språket med tegnet \*. Vi kommer senere tilbake til andre grensemarkører, f.eks. mellom rot og suffiks. Vi får da denne sorteringsversjonen:

#### Sorteringsversjon\_B

```
turtel due miljø
      dug nad s time
      av duk ing
      lin duk
      dukk ert
edder dun s dyn e
      fly dur
```

---

```
intro*duk sjon
intro*dus er e
```

---

```
film pro*duc er
      pro*duk t
      pro*duk t iv
      pro*dus ent
søl at pro*dus ent
```

---

```
re*duk sjon
styrke re*duk sjon
      re*dus er ing
```

Her kommer altså ordstammer med felles betydningsinnhold samlet. Dette gjelder f.eks. ordstammen produkt (og varianter av denne) som er belagt i substantiv (f.eks. produksjon) og i verb (produsere). Prefiksene intro-, pro- og re- er ikke skilt ut som sjølstendige enheter, men er sammenkoblet med resten av stammen med tegnet \*. Dersom dette tegnet kan tenkes å virke forstyrrende på leseren, kan det om ønskelig sløyfes i utskriftsversjonen. Men da vil det i en del tilfelle komme andre ordtyper innimellom ord med samme prefiks, f.eks.:

Uten prefiks-markering

rebell  
redd hare  
redus er e

Med prefiks-markering

re\*bell  
re\*dus er e  
redd hare

I enkelte tilfelle kan prefikser som pro- og re- betraktes som regulære morfemer i og med at de kan skilles ut ved null-kommutasjon, f.eks. prorektor. Disse blir som andre regulære morfemer markert med "blanktegn" og vil dermed komme umiddelbart foran de ukommutable beleggene med samme prefiks. f.eks.

re engasj er e  
re vurd er e  
re\*form  
re\*prise  
re\*volt

Leseren har dermed grei oversikt over kommutabiliteten til de enkelte innlånte prefikser og kan raskt sjekke om han/hun er enig i kommutasjonene.

Da det i en del tilfelle vil være av interesse å sondre mellom prefiks, rot og suffiks i det længivende språket, kunne vi tenke oss å bruke spesialtegn for dette formålet:

*	det foregående segmentet er et prefiks	pro*test
+	det etterfølgende segmentet er et suffiks	leg+al
.	segmentet foran og etter er ei rot	akva.duk+t

Dette innebærer at ei rot i det længivende språket kan være markert på flere måter. (Tegnet u brukes her til å markere "blank").

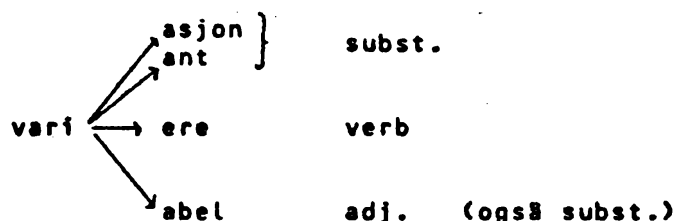
1.	* u	f.eks. *testu	i	pro*testu
2.	u +	uleg+	i	uleg+al
3.	u .	uakva.	i	uakva.duk+t
4.	. +	. duk+	i	"

-----

De innlånte suffiksene spiller på flere måter en viktigere rolle enn de innlånte prefiksene i det norske ordlagings-systemet. Dette viser seg bl.a. ved at suffiksene har:

1. Ordklasseavledende funksjon
2. Større kommutabilitet

Suffiksene har en ordklasseavledende funksjon i den forstand at de tilknyttet samme rot kan danne forskjellige ordklasser, f.eks.



Når det gjelder kommutabilitet, foreligger det tilsynelatende innbyrdes kommutasjon mellom f.eks. de-, in-, ob-, pro- og re- i dedusere, indusere, obdusere, produsere og redusere. Men da verken prefiksene eller det gjenværende segmentet -dusere tilfredsstiller det vanlige kravet ved kommutasjon, nemlig at de utskilte segmentene skal ha betydning, blir ikke prefiksene skilt ut som morfemer i disse orda. De får altså markering med \*, f.eks. produsere - i motsetning til de få forekomstene hvor prefikset er kommutabelt, f.eks. provarerbisk.

Suffiksene derimot kan skilles ut som morfemer i en rekke innlånte ord. Dette kan skje ved to former for kommutasjon:

1. Nullkommutasjon      lakkere  
                                  lakk &
2. Innbyrdes kommutasjon    intervensjon  
                                  intervenere

En rekke innlånte ord av romansk og gresk opphav har innbyrdes kommutasjon, men ikke nullkommutasjon. I en del tilfelle foreligger det ingen kommutasjonsmulighet, f.eks. i ingenigr hvor -ør ikke kan kommuteres mot -ere, som i f.eks. gravggr.

Vi vil først ta for oss de romanske fremmedorda, hvor det kan stilles opp et system med innbyrdes kommutasjon av suffikser. Sentralt i dette kommutasjonssystemet står det verbale -ere. Dette kan illustreres med denne skissen:

Noen subst.suffikser  
(gruppert etter type)

ANT	emigr-	}	ERE
ENT	assist-		
ATOR	administr-	}	
ATØR	observ-		
ITØR	eksped-		
ØR	mass-		
ITRISE	eksped-	}	
ØSE	suffl-		
ASJON	navig-	}	
ISJON	repet-		
SJON	interven-		
AS	disput-		
ATUR	arm-	}	
ITUR	pol-		
<b>Adjektiv</b>			
ABEL	irrit-	}	
IBEL	dispon-		
ANT	toler-	}	
ENT	kongru-		
ATIV	administr-	}	
AT	moder-		
ITT	implis-		

Andre subst.suffikser  
(i alfabetisk rekkefølge)

ADE	promen-
AMENT	incit-
EMENT	raffin-
ANSE	konfer-
AR	honor-
ASJE	mass-
AT	postul-
ATORIUM	observ-
END	divid-
ENS	tend-
ITT	konvert-
O	kommand-

I tillegg til kommutasjon med -ere har en rekke suffikser også kommutasjon med andre suffikser, f.eks.:

-asjon	emigr-asjon	-ent	kongru-ent
-ant	emigr-ant	-ens	kongru-ens

I noen tilfelle forekommer det at -ere-kommutasjon ikke er mulig, mens andre suffikser kan kommuteres, f.eks.:

arrog-ant  
arrog-anse

Blant ord av overveiende gresk opphav har suffikset -isk en sentral plass som ordklasseavledende suffiks:

-i	iron-i	-ikk	polit-ikk
-isk	iron-isk	-isk	polit-isk

Det kunne komme på tale å markere ord med innbyrdes kommutasjon med et spesialtegn, f.eks. #. I så fall ville en i noen tilfelle få en tre-delt oppføring av ord med samme suffiks:

ambassad ør	}	1. null-kommutasjon
im*port ør		
sjarm ør	}	2. innbyrdes kommutasjon (med -ere)
fris#ør		
grav#ør		
mont#ør		
fav+ør	}	3. ingen kommutasjon
in*geni+ør		
vig+ør		

Ved å definere # og + som "skilletegn" under sorteringen (på linje med u i sjarmu.ør) oppnår en at beleggene med samme suffiks - men med forskjellig kommutabilitet - kommer gruppert umiddelbart etter hverandre. Leseren kan da lett sjekke om han/hun er enig i kommutasjonene.

For leseren ville det for oversiktens skyld være en fordel at også ør-ord uten suffiksstatus kom på samme sted. Dette gjelder f.eks. malør (fra fransk malheur). Dermed ville en få en fir-delt oppføring av ord med samme element -ør. Dette kunne oppnås ved baklengssortering (fra høyre mot venstre fra og med "nøkkelmorfem"). Et ord som malør ville da komme umiddelbart etter ør-ord med spesialtegn, dvs. med suffiksstatus.

I sorteringen ovafor ser en at skillet mellom kommutable og ukommutable suffikser gir et semantisk biresultat: Substantiv med kommutable suffikser er personbetegnelser, mens de ukommutable suffiksene står i en mer uensartet gruppe (inkluderer også abstrakter som favør).

Det viktigste suffikset i de innlånte orda er nok det verbale -ere. Med utgangspunkt i slike verb kan det på norsk grunn bli dannet nye substantiv og adjektiv med romanske suffikser. Et eksempel på dette er sondør (avledet av sondere) som har fått betydningen "person som på vegne av sitt parti undersøker mulighetene for samarbeid med andre partier".

Det dynamiske elementet i ordlagingen innebærer at markeringene i ordboka kan bli foreldet. Hvis det f.eks. skulle oppstå et ord som spasør (person som spaserer), måtte spasere få markeringen for innbyrdes kommutasjon: spas#ere.

Det markeringssystemet som vi har antydnet her, kan kanskje bli vel komplisert. En kan også være i tvil om et segment skal skilles ut eller ikke. Semantikken kommer inn som et problem ved en rekke ord, f.eks. president - presidere. Er betydnings-

felleskapet av en slik art at -ent og -ere skal skilles ut som morfemer? Eller skal -al skilles ut i substantivet kolonial (= "forretning som fortrinnsvis fører matvarer"). En eldre språkbruker ser nok forbindelsen til koloni, hvor en stor del av matvarene kom fra. Men nå er koloniveldet avviklet, og en yngre språkbruker ser ikke nødvendigvis denne forbindelsen i kolonial.

Etter det jeg har forstått, behandles fremmedorda i den svenske morfemordboka på en enklere og mer "mekanisk" måte. En skiller således mellom to typer suffikser:

1. Svake suffikser (2 til 4 kommutasjoner)
2. Sterke suffikser (minst 5 kommutasjoner)

Når et suffiks har minst 5 kommutasjoner, ansees suffikset som "så væletablerat at det kan urskiljas også når restled oppkommer" (Allén 1977,6). Et eksempel på et slikt sterkt suffiks er -in, som bl.a. kan kommuteres i balsamin og blondin. Fordi suffikset klassifiseres som sterkt, blir -in også skilt ut i f.eks. gillotin (som skriver seg fra personnavnet Guillotin).

Det er mulig at det er språkpsykologiske betraktninger som ligger bak sontringen mellom svake og sterke suffikser. Ordet "væletablerat" kan tyde på dette, dvs. veletablert i språkbevisstheten. Men det er vanskelig å uttale seg om språkbrukeren ser samme suffiks i såvidt forskjellige ord som blondin, margarin, appelsin, kanin osv.

Et vesentlig poeng med den svenske morfemordboka er at "ordleddene" behandles på flere nivåer. Således kan et element som er skilt ut på et nivå, f.eks. -in i gillot-in og kre- i kre-atur, sammenføres med tilstøtende ordledd på et annet nivå. Dette innebærer at f.eks. kre- i kre-atur ikke opprettholdes som ordledd på det semantiske nivået, mens derimot kre- blir stående på dette nivået i kre-ativ - med betydningskjernen "skape" i kre-.

#### Litteratur:

- Allén, Sture: Morf-o-log-i och poly-sem-i (Göteborg 1977)  
Worth, D.S. et al.: Russian Derivational Dictionary (N.Y. 1970).