Tove Fjeldvig og Anne Golden

AUTOMATISK SPLITTING AV SAMMENSATTE ORD - et lingvistisk hjelpemiddel for tekstsøking

SAMMENDRAG

Sammensatte ord skaper problemer ved ulike former for automatisk analyse av vokabularet i en tekst, f.eks. ved frekvensstudier. Problemet består i at meningsinnholdet i et sammensatt ord i mange tilfeller også kan beskrives i et uttrykk med de tilsvarende usammensatte ordene. I tekstsøking kan f.eks. de sammensatte ordene føre til at man ikke finner de dokumentene man søker etter fordi det ikke er samsvar i ordbruken mellom søkeargumentet og dokumentene. Hvis man f.eks. bare søker på et sammensatt ord uten å dele det opp i de enkelte ledd, vil man ikke finne de tekstene hvor alle leddene i det sammensatte ordet er nevnt, men løsrevet fra hverandre.

På denne bakgrunnen ble det utviklet en metode for automatisk splitting av sammensatte ord. Metoden er basert på et sett med ca. 1000 regler - og ikke et leksikon.

Prosjektet er et samarbeidsprosjekt mellom Institutt for rettsinformatikk (Universitetet i Oslo) og NAVF's EDB-senter for humanistisk forskning. Det er finansiert av NORDINFO med 9 månedsverk.

1. SAMMENSATTE ORD OG TEKSTSØKING

Et spesielt trekk ved de nordiske språkene er den hyppige bruken av sammensatte ord. I en hovedfagsoppgave ved Nordisk Institutt i Oslo ble det vist at ca. 25% av de ulike grafordene i en tekst var sammensatte ord (Munthe 1972). Undersøkelsen var basert både på et skjønnlitterært verk (ca. 127000 løpende ord) og en samling sakprosatekster (ca. 72000 løpende ord).

Med et sammensatt ord menes ethvert ord som er bygd opp på en slik måte at det kan deles i mindre enheter og at disse enhetene selv er ord som kan forekomme isolert. Et eksempel på et sammensatt ord er LESELAMPE som kan deles i enhetene LESE og LAMPE. Begge disse enhetene er selvstendige ord som kan opptre alene.

I tekstsøking skaper de sammensatte ordene to typer problemer. Det ene problemet oppstår når man under selve søkingen bare bruker sammensatte søkeord og det i tekstene (dokumentene) bare er brukt usammensatte ord til å beskrive det aktuelle meningsinnholdet. Disse dokumentene vil da ikke bli funnet. Det samme skjer hvis det i dokumentene

Automatisk splitting av sammensatte ord-et lingvistisk hjelpemiddel for tekstsøking
Tove Fjeldvig, Anne Golden
Proceedings of NODALIDA 1985, pages 73-82

bare er brukt sammensatte ord og i søkeargumentet bare de tilsvarende usammensatte ordene.

For å få nærmere innsikt i effekten av å supplere søkeargumentet med de tilsvarende usammensatte ordene, ble det gjennomført en undersøkelse. Denne var basert på søkeargumenter som var stilt til LOVDATA - et system for søking i juridisk kildemateriale. Undersøkelsen viste at det var få av de sammensatte søkeordene som var supplert med andre uttrykk. Dessuten viste den at av ca. 80 søkeargumenter som bare inneholdt sammensatte ord, ga ca. 70 flere relevante dokumenter når de sammensatte ordene også ble splittet. Resultatet styrker med andre ord antagelsen om at man kan øke søkeeffektiviten ved å supplere et sammensatt søkeord med de tilsvarende usammensatte ordene.

Det andre problemet med sammensatte ord i tekstsøking møter man ved rangering av de funne dokumenene. Her benyttes gjerne kriterier som er basert på søkeordfrekvensen - slik at jo flere søkeord et dokument inneholder, jo høyere opp på resultatlista kommer dokumentet. Hvis man i frekvensberegningen ikke tar hensyn til forekomsten av de usammensatte ordene som de sammensatte ordene består av, vil søkeordfrekvensen gi et galt inntrykk av ordets hyppighet i dokumentet.

I dagens tekstsøkesystemer har man mulighet for å trunkere et søkeord - dvs. at det ikke søkes på hele ordet, men bare et nærmere spesifisert antall tegn i begynnelsen av ordet (høyretrunkering) eller i slutten av ordet (venstretrunkering). Søker man f.eks. på ordet ARV* - hvor "*" er trunkeringssymbolet - vil man også finne de dokumentene som inneholder ARVEAVGIFT, ARVERETT, ARVELOVEN osv. Tilsvarende for *TRYGD hvor BARETRYGD, SYKETRYGD, ALDERSTRYGD osv. vil bli funnet. Trunkering vil derfor kunne avhjelpe problemet med manglende bruk av sammensatte søkeord, men erfaring viser at det er mange som ikke benytter seg av dette hjelpemiddelet (jfr. Fjeldvig 1986).

2. SAMMENSATTE ORD I NORSK

Som nevnt innledningsvis, består et sammensatt ord av minst to usammensatte ord. En splitting av et sammensatt ord vil derfor si å finne fram til de usammensatte ordene som ordet er bygd opp av.

Vanligvis deles et sammensatt ord i to ledd, forleddet og etterleddet. Begge disse leddene kan selv være sammensatt og kan følgelig deles i nye ledd. I ordet LASTEBILSJÅFØR er f.eks. det sammensatte ordet LASTEBIL forleddet og SJÅFØR etterleddet. Betingelsen for at et ledd kan deles videre i underledd, er at det betår av frie morfemer, enten frie, semantiske morfemer (f.eks. KJØRE og BIL) eller frie, grammatiske morfemer (f.eks. OG, SÅ og TIL). Ved siden av de frie morfemene kan leddene inneholde bundne morfemer, dvs. bøyningsmorfemer (f.eks. -ENE og -TE) eller avledningsmorfemer (f.eks. -ING og U-). De forskjellige

bundne morfemene har faste plasser i forhold til det frie morfemet.

Forskjellen på frie, grammatiske morfemer og frie, semantiske morfemer er at de semantiske morfemene har både bøynings- og avledningsmuligheter (f.eks. KJØRE, KJØRTE og KJØRING), mens de grammatiske ikke har noen av delene. Begge kan imidlertid settes sammen med andre frie morfemer og danne sammensatte ord.

Verken de bundne morfemene eller de frie, grammatiske morfemene får tilført nye morfemer ved lån eller nydannelser. Vi kan derfor si at de er endelige grupper.

Leddene i et sammensatt ord vil inneholde følgende morfemtyper:

a) et fritt, semantisk morfem (f.eks LESE i LESELAMPE)
b) et fritt, grammatisk morfem (f.eks INN i INNGANG)
c) en avledning som bare inneholder ett fritt morfem (f.eks. ANBEFALING i ANBEFALINGSBREV),
d) en sammensetning av frie, semantiske morfemer (f.eks. LASTEBIL i LASTEBILSJÅFØR),
e) en sammensetning med kombinasjonen frie, grammatiske og frie, semantiske morfemer (f.eks. INNGANG i INNGANGSBILLETT),
f) en sammensetning av frie, grammatiske morfemer (f.eks. INNTIL i INNTILLIGGENDE),
g) en avledning som inneholder en sammensetning uten frie, grammatiske morfemer (f.eks BOKFØRING i BOKFØRINGSKURS),
h) en avledning som inneholder en sammensetning med kombinasjonen frie, grammatiske morfemer og frie, semantiske morfemer (f.eks UTDANNELSE i LÆRERUTDANNELSE),

I de sammensatte ordene er det etterleddet som er hovedleddet og som angir ordklassen til ordet. Vi finner sammensatte ord i de fleste ordklasser, men enkelte ordklasser har langt høyere frekvens av sammensatte ord enn andre. I Munthe (1972) framkommer det at ca. 75% av alle de sammensatte ordene var substantiv, ca. 15% var verb og ca. 6% var adjektiv. Når det gjaldt ordklassetilhørigheten til forleddet, var ca. 55% substantiv, ca. 26% adverb/preposisjoner og ca. 8% adjektiv. Den hyppigste kombinasjonen var "substantiv + substantiv" som utgjorde ca. 50% av de sammensatte ordene, mens kombinasjonen "adverb/preposisjon + verb" var den nest hyppigste og utgjorde ca. 11%.

Sett fra et nåtidsperspektiv er forleddet i en sammensetning nesten alltid ubøyd, men vi finner noen unntak.

Leddene i sammensetningene kan settes sammen på tre ulike måter:

1) direkte, f.eks. FOTFESTE og INNSETTE
2) med fuge-s, f.eks. DAGSREISE
3) med fuge-e, f.eks. BARNEHAGE

Det er i forbindelser med substantiv som forledd at fugebokstaven settes inn, og det er lydlige forhold som i første rekke bestemmer dette.

Det som står igjen av et ord når bøynings- og avledningsmorfemene er fjernet, kalles gjerne rotmorfemet. Vi vil i det følgende bruke denne betegnelsen i stedet for fritt morfem, fordi enkelte rotmorfemer må ha en endelse for å stå alene og følgelig ikke alltid er identiske med de frie morfemene.

Stavelsesstrukturen i usammensatte ord.
Ethvert ord i norsk og liknende språk er en sekvens av konsonanter og vokaler. Hvis et kluster defineres som 0, 1 eller flere konsonanter, vil alle ord passe inn i formelen:

ini + vok+ (med + vok + med + vok ....) + fin

hvor
ini: initialkluster
vok: vokal eller diftong
med: medialkluster
fin: finalkluster

(se Sigurd 1965 og Brodda 1979).

De aller fleste hjemlige, usammensatte ordene er enstavede når vi fjerner bøynings- og avledningsmorfemene, dvs. roten er enstavet (f.eks. ARV, BÅT og MANN). Et unntak er de såkalte lette tostavelesesordene som har en trykklett E i utlyd, f.eks. HANSKE og JENTE. Disse ordene mister imidlertid E'en når de får lagt til en endelse og de får da samme struktur som enstavelsesordene. Et annet unntak er tostavelsesord som ender på -EL, -EN, -ER, eks. SIRKEL, LAKEN, SKULDER. Disse ordene vil i noen tilfelle kaste E'en når de blir knyttet sammen med et bøynings- eller avledningsmorfem, f.eks SIRK'LER, SKULD'RE (tegne "'" markerer E'ens fjerning). Vi velger heretter å reservere uttrykket hjemlige ord til ord som bare inneholder norske bokstaver og i grunnformen er:

a) enstavet
b) enstavet + E
c) tostavet og ender på -EL, -EN eller -ER

Det er et begrenset antall klustre som kan stå i de forskjellige posisjonene på norsk. At et kluster kan bestå av 0 konsonanter, vil si at ordet kan innledes eller avsluttes med en vokal. Vi får følgende stavelsesstruktur for de tre typene hjemlige ord:

1) **ini + vok + fin**
2) **ini + vok + med** + E
3) **ini + vok + med** + EL/EN/ER

Hvis vi sammenholder disse strukturene, kan vi gi en fellesformel for alle hjemlige, usammensatte ord:

fin
ini + **vok** + (L)
med + E + (N)
(R)

Fremmede ord kaller vi ord som har

a) en annen struktur enn de hjemlige
b) unorske bokstaver (f.eks. C, Q, X eller Z)
c) fremmede avledningsmorfemer (f.eks. KON- og -ITET)

De fremmede ordene har ofte en flerstavet rot. Imidlertid er det mange av stavelsene som er svært produktive i utviklingen av nye ord innen teknologi og vitenskap (f.eks. TELE-, BIO-, HOMO-, -LOG, -FON og -GRAF). Disse stavelsene fungerer som avledningsmorfemer ved at de ikke kan stå alene (unntatt i fagsjargongen), men de skiller seg fra dem ved at de kan føyes direkte til hverandre f.eks. BIOLOG og TELEFON, i motsetning til avledningsmorfemene som må inneholde et fritt morfem,.

At de fremmede ordene ofte har flere enn én stavelse, vil si at de kan ha flere medialklustre i samme ord. De fremmede ordene vil altså ha følgende struktur:

ini + **vok** (+ **med** + **vok** + ...) + **fin**

Selv om alle klustrene som forekommer i hjemlige ord, kan forekomme i fremmede ord, er det langt sjeldnere med konsonantkombinasjoner i de fremmede ordene. De har ofte en veksling mellom én konsonant og én vokal.

Stavelsesstrukturen i sammensatte ord
I et sammensatt ord som består av 2 ledd, vil morfemene ha følgende rekkefølge:

(**pre** (+ **pre** ..)) + **rot** (+ **avl** (+ **avl** ..) (+E/S) +
(**pre** (+ **pre** ..)) + **rot** (+ **avl** (+ **avl** ..) (+ **bøyn** (+ **bøyn**..))

hvor
pre : prefiks
avl : avledningsendelse
E/S : fuge E eller S
bøyn : bøyningsendelse
rot : rotmorfem

Røttene vil ha den samme strukturen som i usammensatte ord (se ovenfor).

3. EN METODE FOR AUTOMATISK SPLITTING AV SAMMENSATTE ORD

Metoden for automatisk splitting av sammensatte ord vil i bruk bestå av et datamaskinsprogram og et sett med regler. Den er selvstendig og generell i den betydning at den for det første kan kobles til et hvilket som helst system - ikke bare tekstsøkesystemer, og for det andre kan anvendes på et hvilket som helst språk som følger overnevnte teori.

Ved bruk i et tekstsøkesystem vil metoden kunne fungere bedre hvis man koblet den til ordboka (søkefilen) i søkesystemet. På den måten kan man sjekke om de foreslåtte leddene forekommer i noen av dokumentene og dermed være bedre i stand til å velge ut den riktige løsningen. Anta f.eks. at vi har BOKSALG som søkeord. Ut fra et teoretisk grunnlag vil dette kunne deles slik:

1) BOK-SALG
2) BOKS-ALG

I ordboka til tekstsøkesystemet finner vi imidlertid ordene BOK, SALG og BOKS, men ikke ordet ALG. Dette taler for at den første løsningen er den riktige.

Resultatet som metoden gir, er bestemt av regelsettet. Dette grunnlagsmaterialet er språkavhengig, og resultatet vil være bestemt av i hvor stor grad det lar seg gjøre å formulere regler av denne type og hvor godt reglene dekker den informasjonen som det er forutsatt at de skal gjøre. Reglene skal gi informasjon om de ulike typer "byggeklosser" som et ord kan være satt sammen av, f.eks. initialkluster, prefikser, suffikser, morfemkombinasjoner, spesialord osv. En regel blir satt i kraft (dvs. at en "byggekloss" er identifisert") når alle betingelsene som knytter seg til regelen, er oppfylt. En avledningsendelse må f.eks. alltid forekomme etter et rotmorfem eller en annen avledningsendelse - og ikke bak et prefiks.

I dette prosjektet har vi utviklet et regelsett for norsk bokmål. En nærmere beskrivelse av regelsettet er gitt i avsnitt 4.

Fremgangsmåte
Metoden starter med de minste enhetene, nemlig vokalene og diftongene som er kjernen i morfemet. Dernest følger konsonantklustrene som omkranser vokalene. Metoden forsøker så å gjenkjenne de morfemene som ordet er bygd opp av, og deretter leddene. Den ender opp med ett eller flere forslag til løsninger. Disse forslagene rangeres ut fra kriterier som vi antar indikerer hvilken løsning som er mest sannsynlig.

Man kan betrakte metoden som bestående av 4 faser:

1) Kartlegge mulige morfemgrenser
2) Gjenkjenne registrerte morfemer (f.eks. bundne morfemer og frie grammatiske morfemer)
3) Gjenkjenne leddene på bakgrunn av mulige morfemkombinasjoner
4) Rangere forslagene

Den første fasen undersøker hvor morfemgrensene kan gå ut fra kjennskap til hvordan klustrene i et morfem er bygd opp. I ordet KJÆRLIGHETSVERSET vil vi finne følgende forslag til morfemgrenser:

KJÆR - L - IG - HET - S - VER - S - ET

Den neste fasen sjekker de mulige morfemene mot en liste med registrerte morfemer. Denne lista består primært av bunde morfer, men vi har også lagt inn de frie, grammatiske morfemene og en begrenset mengde med frie, semantiske morfemer (se avsnitt 4). Antall frie morfemer er begrenset fordi de er relatert til metoden og ikke til teksts-materialet. Det vil derfor ikke være aktuelt å supplere mengden eller eliminere elementer i den ved skifte av tekstmateriale.

I KJÆRLIGHETSVERSET vil de ulike morfemforslagene få følgende markering:

```
                                 rot     rot
                                 VER  -  SET  (1)
                                /
             avl   avl/S/  rot     bøy
            LIG - HETS - VERS -  ET   (2)
rot        /
KJÆR
           \
             avl   avl     rot     rot
            LIG - HET  -  SVER  - SET  (3)
                         \
                          rot     bøy
                          SVERS - ET   (4)

                                 rot    rot
                                 VER  - SET   (5)
                                /
             avl  avl/S /  rot    bøy
            IG - HETS -   VERS - ET    (6)
rot        /
KJÆRL
           \
             avl  avl      rot     rot
            IG - HET  -   SVER  - SET  (7)
                       \   rot     bøy
                           SVERS - ET  (8)

                           rot     rot
                           VER   - SET  (9)
                           rot     bøy
                           VERS  - ET   (10)
                  avl/S    rot
                  HETS -   VERSET       (11)
rot              /
KJÆRLIG
                 \
                  avl      rot      rot
                  HET -    SVER   - SET (12)
                       \   rot     bøy
                           SVERS - ET   (13)
                       \   rot
                           SVERSET      (14)
```

hvor
avl/S betyr avledningsendelse med fuge S.

Den tredje fasen fokuserer på morfemkombinasjonene. I eksempelet ovenfor vil vi få følgende forslag til ledd-grenser:

KJÆRLIGHETS - VER - SET
KJÆRLIGHETS - VERSET
KJÆRLIGHET - SVER - SET
KJÆRLIGHET - SVERSET

Kunnskapen om hvilke avledningsendelser som må ha fuge-s når de står til forleddet, vil imidlertid forkaste to av forslagene og vi står igjen med:

KJÆRLIGHETS - VER - SET
KJÆRLIGHETS - VERSET

Den fjerde fasen stiller de alternative løsningene opp mot hverandre, og rangerer dem. For å avgjøre hvilken løsning som skal foreslås - eller i hvilken rekkefølge løsningene skal presenteres - må vi ta i bruk kriterier som er hentet fra empirien.

Ett av kriteriene vi bruker, er å preferere visse ordklasser framfor andre ut fra kjenskap de ulike ord-klassenes hyppighet som ledd i sammensetninger (se avsnitt 2). Avledningsendelsene bestemmer leddenes ordklasse, og regellista inneholder derfor informasjon om hvilke ord-klasser de forskjellige avledningsendelsene tilhører. Alle unntaksordene har også fått ordklassemarkering. Et annet kriterium for prioritering er antall ledd i et sammensatt ord. I eksempelet ovenfor vil man f.eks. legge vekt på at det er større sannsynlighet for at et sammensatt ord består av to ledd enn av tre ledd. Dessuten kan antall registrerte morfemer og rekkefølgen av dem (f.eks. hjemlige og fremmede morfemer) også kunne brukes i denne fasen. Andre bakgrunnsopplysninger kan f.eks. være:

- statistikk over klustrenes hyppighet
- " prefiksenes "
- " suffiksenes "

4. MER OM REGELSETTET

Regelsettets oppgave er å gjenkjenne morfemets oppbygning og morfemtype. Det kan grovt deles i:

1) morfemregler
2) bokstavregler

Morfemreglene består av tilsammen av ca. 600 regler, og de inneholder:

a) alle bundne morfemer, dvs. prefikser og suffikser
b) alle frie, grammatiske morfemer som forekommer i sammensetninger med frie, semantiske morfemer. Vi behandler dem som om de var bundne morfemer fordi vi pga. tekstsøkingsinteressen ikke ønsker å dele slike ord.

c) de vanligste av de frie, semantiske, hjemlige morfemene som har en struktur som tilsvarer enten

- et prefiks + et initialkluster (f.eks. PREST) eller
- et finalkluster + et suffiks (f.eks. RING)

d) de vanligste av de frie, semantiske morfemene på to bokstaver som kan forekomme i sammensetninger (f.eks. BY, IS og TE).

Bokstavreglene består av tilsammen ca. 400 regler og de inneholder oversikter over:

a) initialklustre
b) finalklustre
c) medialklustre
d) diftonger

Reglene kan inneholde

- betingelser for å tre i kraft
- restriksjoner mot å tre i kraft
- informasjoner om morfemer/klustre
- informasjon om fugebokstaver

Som eksempeler på betingelser, kan vi nevne at noen prefiks krever at rotmorfemet som følger etter, innledes med en bestemt bokstav, f.eks. det fremmed prefikset KOM som må etterfølges av B, M eller P.

Tilsvarende har enkelte prefiks restriksjoner mot enkelte påfølgende bokstaver, f.eks. det fremmede prefikset OB- kan ikke stå foran K, F eller P. Tegnsekvensen OB vil derfor ikke tolkes som prefiks hvis en av disse bokstavene følger etter. Noen suffiks vil forutsette at det ikke kommer flere suffiks foran, f.eks. -ELSE.

Reglene inneholder informasjon om:

a) morfemet er hjemlig, fremmed eventuelt begge deler.
b) det kan følge en affiks (prefiks eller suffiks) av samme type foran eller bak.
c) morfemet egentlig er en fremmed stavelse slik at det ikke nødvendigvis må komme et rotmorfem foran eller bak.
d) suffikset kan tilhøre forleddet.
e) hvilken ordklasse morfemet tilhører.
f) en fuge kan/må forekomme i forbindelse med suffikset når det er del av forleddet.
g) hvilke finalklustre som får fuge-s.

6. RESULTAT

I skrivende stund er prosjektet i avslutningsfasen. Det gjenstår å analysere og systematisere resultatene og kor-

rigere fase 4 (rangeringen) i henhold til disse. Vi har likevel sett nærmere på et mindre antall sammensatte ord (ca. 150 stk.), og resultatene ser lovende ut. De største problemene knytter seg, ikke uventet, til fuge-s (f.eks. BOKSALG som kan gi delingspunktene BOK-SALG og BOKS-ALG). De fremmede ordene skaper også problemer, men mindre enn ventet. Metoden er blitt testet mot et materiale på ca. 50 000 graford, og resultatene vil bli publisert i bl.a. Humanistisk Data i 1986.

Litteratur:

Brodda, Benny: Något om de svenska ordens fonotax och morofotax. Papers from the Institute of Linguistic, University of Stockholm (PILUS) nr 38, dec 1979.

Fjeldvig, Tove: Automatisk trunkering. Vil bli publisert i CompLex-serien (Universitetsforlaget) i 1986.

Fjeldvig, Tove og Anne Golden: Automatisk rotlemmatisering - et lingvistisk hjelpemiddel for tekstsøking. CompLex nr. 9, Universitetsforlaget. Oslo 1984.

Munthe, Synneve Kjuus: Sammensatte ord - En kvantitativ undersøkelse av norsk litteratur- og sakprosa. Hovedoppgave i nordisk, Universitetet i Oslo 1972.

Sigurd, Bengt: Phonotactic Structures in Swedish, Scandinavian University Books, Lund 1965.