Tove Fjeldvig
Statens Datasentral A/S
Oslo.

Anne Golden
Institutt for norsk som fremmedspråk
Universitetet i Oslo

# BRUK AV SPRÅKBASERTE HJELPEMIDLER I INFORMASJONSSØKING

## 1. Dagens informasjonssøkesystemer

Med et informasjonssøkesystem sikter vi her til et system som kan håndtere uformaterte, tekstlige dokumenter. At det også kan håndtere strukturert, feltorganisert informasjon, er mindre interessant for denne sammenhengen.

Gjenfinning av dokumenter er basert på ordene i dokumentene, og i prinsippet kan alle ord anvendes som søkeord. Dette gir muligheten til å stille fleksible søkeargumenter, og det finnes ingen grenser for hva man kan søke på. Resultatet vil avhenge av hvilke dokumenter som inneholder disse søkeordene.

Samtidig stiller denne form for tekstsøking store krav til valget av søkeord. Skal man finne fram til et relevant dokument, må søkeordene finnes blant de ordene forfatteren har brukt til å uttrykke det aktuelle søkebegrepet. Dette byr ofte på problemer fordi et begrep kan uttrykkes ved ulike ord.

Dagens informasjonssøkesystemer er ikke i stand til å likestille ord som innholdsmessig gir uttrykk for det samme i et dokument. Selv ikke ord som er bøyd eller avledet av samme rot, blir forbundet med hverandre. Søker man f.eks. på ordet *mord*, finner man ikke de relevante dokumentene hvor bare formen *mordet* er brukt.

Enkelte systemer gir riktignok muligheten til å kalle opp en synonymtesaurus ved formuleringen av søkeargumentet som kan hjelpe brukeren til å finne synonyme søkeuttrykk. Her kan man definere ulike semantiske relasjoner mellom ord - og man kan også likestille ord med felles rot. Grunnen til at denne type hjelpemiddel likevel er lite brukt, er kostnaden forbundet med etablering og vedlikehold av dem. De fleste databaser vokser over tid, og skal en tesaurus være ajour med databasen, må den oppdateres når databasen oppdaters. For mange er dette en nærmest uoverkommelig oppgave.

Det eneste hjelpemiddelet som er vanlig i dagens informasjonssøkesystemer, er trunkering. Dette er en primitiv liten algoritme som gjør det enkelt å utvide søkeargumentet med alle ord som innledes (ev. avsluttes) med en gitt

Bruk av språkbaserte hjelpemidler i informasjonssøking
Tove Fjeldvig, Anne Golden
Proceedings of NODALIDA 1987, pages 217-228

tegnstreng. Ved å trunkere et søkeord vil man kunne få fanget opp ord som er bøyd, avledet eller sammensatt av søkeordet. Man vil også kunne få med en del ikke-relevante ord, men dette trenger nødvendigvis ikke påvirke søkekvaliteten. Undersøkelser viser at problemet med trunkering er heller at mange brukere ikke gjør benytter seg av den. Enten glemmer de å trunkere, eller så har de ikke forstått hvor viktig det er (jfr. Fjeldvig 1987:43-52).

Vi har derfor stor tro på at en tilføring av nye hjelpemidler i dagens informasjonssøkesystemer som kan bistå brukeren i formuleringen av søkeargumentet, vil kunne øke søkekvaliteten for mange. Gjenfinningsgraden[1] vil kunne øke som en følge av at søkeargumentet blir supplert med flere adekvate søkeord. Likeledes vil presisjonen[2] kunne øke som følge av en bedre rangering av de funne dokumenter.[3]

## 2. Utvikling av språkbaserte hjelpemidler for tekstsøking

For å bøte på denne mangelen, ble det i 1980 satt i gang et prosjekt ved Institutt for Rettsinformatikk, Universitetet i Oslo, som bl.a. hadde til formål å utvikle metoder for automatisk rotlemmatisering og for automatisk gjenkjenning og splitting av sammensatte ord.

Metoden for automatisk rotlemmatisering skulle sørge for å likestille ord som var bøyd eller avledet av samme rot. Et slikt hjelpemiddel vil stille brukeren helt fritt til å velge formen på søkeordet, for systemet vil sørge for at alle bøynings- og avledningsformer kommer med.

På tilsvarende måte skulle metoden for automatisk splitting av sammensatte ord ta hånd om de sammensatte søkeordene og supplere disse med et uttrykk som også dekket eventuelle omskrivinger av disse, f.eks. *knivtrussel* vil bli splittet i *kniv* og *trussel* som vil dekke omskrivingen *"trussel med kniv"*.

I tillegg ønsket vi å se nærmere på muligheten for automatisk (høyre)trunkering. Ofte vil dette kunne være et bedre alternativ enn automatisk rotlemmatisering, bl.a. fordi en trunkering også fanger opp ord som er sammensatt av søkeordet, f.eks. *mord** (hvor * er brukt om trunkeringssymbol) dekker ordene *mordvåpen*, *mordoffer* og *mordkveld*.

---

1. Gjenfinningsgraden (eng. recall) er et uttrykk for hvor stor andel av de relevante dokumentene som er funnet i et søk.

2. Presisjonen er et uttrykk for hvor stor andel av de funne dokumentene som er relevante.

3. Formålet med en rangering er presentere først for brukeren de dokumentene som har størst sannsynlighet for å være relevante. I slike tilfeller vil rangeringskriteriet ofte være basert på den totale søkeordfrekvensen. Dette bygger på en hypotese om at jo flere ganger et ord er nevnt i teksten, jo større sjanse er det for at ordet reflekterer innholdet i teksten. Hvis man her utelater aktuelle søkeord, f.eks. angir mord som søkeord, men ikke mordet, vil man kunne få plassert et dokument som inneholder ordet mord tre ganger foran ett som inneholder mord to ganger og mordet 20 ganger.

Det var en forutsetning at metodene skulle baseres på et sett med regler - og ikke en forhåndsdefinert ordbok. Dette er viktig, fordi de fleste institusjoner som anskaffer et informasjonssøkesystem av denne type, har store - og ofte voksende - databaser. Med en regelbasert metode mener vi her en metode som er basert på et sett med regler som inneholder generell informasjon om ordenes bøynings- og avlednings-muligheter. Dette vil være en språkavhengig metode, men den vil ikke være avhengig av den enkelte database.

### 3. Kort om metoden for automatisk rotlemmatisering

Metoden for automatisk rotlemmatisering grupperer alle ord som tilhører samme rotlemma [4] ved å gi disse ordene en felles oppslagsform. Oppslagsformen kommer fram ved at de enkelte grafordene sammenliknes bakfra med de ulike bokstavstrenger på en regelliste. Hvis grafordet helt eller delvis overlapper bokstavstrengen og alle eventuelle betingelsene oppfylles, utføres visse ordrer. Den vanligste ordren er at de bokstavene som utgjør en endelse, skal strykes, men den kan også være at ordet skal behandles på nytt, eller at visse bokstaver skal legges til.

Oppgavene man står ovenfor ved rotlemmatisering av norske graford, kan føres tilbake til følgende tre punkter, identifisere bøyningsendelser, eks. *-en* i *arven*, identifisere avledningsendelser, eks. *-ing* i *arving*, sammenføre røtter som er realisert som ulike rotformer, eks. *far* og *fedre*.

Eksempler: Grafordene *sykkel, sykkelen, syklene* blir behandlet fordi strengene KKEL, ELEN, og ENE finnes på regellista. Alle disse strengene stiller som betingelse at det står noe foran strengen. Ordrene er forskjellige for disse reglene:

ENE - fjern 3 tegn
KKEL - fjern 4 tegn, legg til tegnet L
ELEN - fjern 2 tegn, så ny behandling

Denne behandlingen fører til at alle *sykkel, sykkelen, syklene* får oppslagsformen SYKL.

En utførlig beskrivelse av metode for automatisk rotlemmatisering er gitt i Fjeldvig/Golden 1986 og Fjeldvig 1987:65-98.

***Reglene***
Regellista gir en oversikt over bøyningsendelsene og avledningsendelsene, dvs. de bundne morfemene i norsk. I tillegg gir den en oversikt over strenger som blir sammendratt ved bøyning og en del ord som har uregelmessig bøyning så sant disse ordene kan avgrenses til lukkede grupper. Tilsammen er det 684 regler på regellista.

---

4. Et rotlemma er en samling av alle ord som kan føres tilbake til samme rot uten at betydningen til ordet blir endret. Oftest er altså et rotlemma det samme som en stamme uten avledningsendelser. I noen tilfeller har imidlertid avledningsendelsene ført til at den nye stammen er blitt leksikalisert, dvs fått en spesiell betydning i forhold til grunnordet. Da utgjør den nye stammen et eget rotlemma.

*Resultat*
Metoden ble testet på et tekstkorpus som var satt sammen av tekster fra juridisk materiale og skolebøker i fysikk, geografi og historie. Tekstkorpuset inneholdt ca 1/2 millioner løpende ord og i overkant av 23000 graford. Resultatet viste at 97,7 % av alle grafordene fikk en oppslagsform som førte til at det kom i riktig rotlemma.

*Feilenes betydning for informasjonssøking*
2,3 % av alle grafordene ble ikke samlet i ett og bare ett entydig rotlemma. Det var tre feiltyper som var mulige. Feiltype a) besto i at grafordene som egentlig tilhørte samme rotlemma, ble delt i to eller flere grupper. Det var 1,4% av grafordene som ble feilplassert på denne måten. Hvis ett av disse grafordene ble valgt som søkeord, ville ikke søkeargumentet bli utvidet med alle de andre bøynings- og avledningsformene til ordet. Denne feilen kan altså føre til at man ikke finner alle de relevante dokumentene, dvs. gjenfinningsgraden kan altså bli dårligere.

Feiltype b) besto i at grafordene som tilhørte ulike rotlemmaer, ble slått sammen til ett rotlemma. Det var 0,8% av grafordene som ble feilplassert på denne måten. Hvis ett av disse grafordene ble brukt som søkeord, ville søkeargumentet bli utvidet med alle bøynings- og avledningsformene til ordet, men også andre irrelevante ord. Denne feilen kan altså føre til at man finner flere irrelevante dokumenter, dvs. presisjonen blir dårligere.

Feiltype c) besto i at grafordene som egentlig tilhørte samme rotlemma fordelte seg på flere grupper som også inneholdt andre rotlemma. Det var 0,2% av grafordene som ble feilplassert på denne måten. Hvis ett av disse grafordene ble brukt som søkeord, vil ikke søkeargumentet bli utvidet med alle bøynings- og avledningsendelsene til ordet, men derimot kan det bli utvidet med andre irrelevante søkeord. Både gjenfinningsgraden og presisjonen kan altså bli dårligere.

## 4. Kort om metoden for automatisk trunkering

Den automatiske trunkeringen ble utviklet som et alternativ til den automatiske rotlemmatiseringen. Dette er et nyttig alternativ i tilfeller hvor dokumentbasen ikke er tilgjengelig for bearbeiding for soking. Egentlig dreier ikke dette seg om en egen metode, men snarere en annen anvendelse av rotlemmatiseringen i informasjonssøking. Systemet står selv for trunkeringen ved først å rotlemmatisere søkeordet og så trunkere både oppslagsformen og søkeordet. I de fleste tilfeller er oppslagsformen identisk med grunnformen, slik som oppslagsformen *arv*. I tilfeller hvor det forekommer ulike rotformer innen et paradigme, har vi gitt rotlemmaet den synkoperte rotformen som oppslagsform, f.eks vil det rotlemmaet som *regel* (rotform: *regel*) og *regler* (rotform: *regl*) tilhører, ha oppslagsformen *regl*. Den automatiske trunkeringer vil imidlertid sørger for at begge rotformene blir trunkert.

En nærmere beskrivelse av metode for automatisk trunkering er gitt i Fjeldvig 1987:149-170.

## 5. Kort om metoden for automatisk gjenkjenning og splitting av sammensatte ord.

Denne metoden skiller først ut de enstavede usammensatte ordene (dvs. ord med en stavelse i roten), for så å identifisere de ulike morfemene i de resterende ordene.

Et hvert norsk graford ***(O)*** vil passe inn i formelen (1):

(1) ***O: pre* rot avl* ((E/S) pre* rot avl*)* bøyn****

hvor

***pre**** står for 0, 1 eller flere prefikser
***rot*** står for rot
***avl**** står for 0, 1 eller flere avledningsendelser
***(E/S)*** står for mulig fuge-E eller fuge-S
***bøyn**** står for 0, 1 eller flere bøyningsendelser
( )* står for 0,1 eller flere forekomster av det som står inni parantesen

Stavelsesstrukturen i norsk kan beskrives ved hjelp av formel (2) som viser strukturen i enhver morf ***(M)***:

(2) ***M: ini vok (med vok)* fin***

hvor

***ini*** betyr initialkluster dvs. 0, 1 eller flere konsonanter som forekommer initialt i morfen
***med*** betyr medialkluster dvs. 0, 1 eller flere konsonanter som forekommer medialt i morfen
***fin*** betyr finalkluster dvs. 0, 1 eller flere konsonanter som forekommer finalt i morfen

De bundne morfemene, dvs prefiksene, avledningsendelsene og bøyningsendelsene finnes bare i et begrenset antall, og det er svært sjelden at det kommer nye medlemmer inn i disse gruppene slik at vi kan regne dem for lukkede. Disse morfene har vi derfor oversikt over. Det samme gjelder de ulike konsonantklustrene.

I prinsippet består oppgaven i først å finne fram til alle mulige morfemgrenser ut fra formel (2), for så å redusere dette løsningsforslaget ut fra formel (1). Deretter rangeres de ulike forslagene.

En mer utforlig beskrivelse av denne metode er gitt i Fjeldvig 1987:99-148 og Fjeldvig/Golden 1987.

***Reglene***
Ved utviklingen av denne metoden tok vi utgangspunktet i den samme oversikten over de bundne morfemene som vi brukte til metoden for automatisk rotlemmatisering. Men informasjonen, betingelsene, og kravene er annerledes på denne regellista. I tillegg gjør vi bruk av en liste over de ulike konsonantklustrene og deres plasseringsmuligheter i en morf. Framgangsmåten går ut på å sammenlikne bokstavstrengene i grafordet med disse regellistene.

*Resultat*
For å teste hvor vellykket metoden var når det gjaldt å kjenne igjen sammensetninger, ble et tilfeldig utvalg på 1019 graford testet, 149 av disse var sammensatte. 1018 av grafordene ble riktig vurdert. Den ene feilen var et usammensatt graford som ble behandlet som et sammensatt. Resultatet var altså tilnærmet lik 100% riktig.

Når det gjaldt splittingen, ble den testet på et tilfeldig utvalg på 160 graford. Her ga metoden alternative løsninger som ble rangert. For 144 av de sammensatte grafordene (90%) kom den riktige løsningen på 1. plass. I 6 tilfeller kom den riktige løsningen på delt 1. plass, i 6 tilfeller på 2. plass, i 1 tilfelle på 3. plass, i 2 tilfeller på 4. plass og i 1 tilfelle på 5. plass. I 97,5% kom altså den riktige løsningen på 2. plass eller bedre.

*Feilenes betydning for tekstsøking*
I vårt lille eksperimentmateriale ble alltid den riktige sammensetningsgrensen funnet, selv om denne løsningen i 10% av tilfellene ble rangert lavere enn andre uriktige forslag. Selv om disse ordene ble brukt som søkeord, vil ikke dette føre til noe problem, siden én feil deling vil gi ord som ikke eksisterer i norsk. Man kan da først undersøke om den delingen man får, inneholder ord som forekommer i databasen. Hvis de ikke gjør det, kan man la systemet gå videre med neste forslag og undersøke om man får tilslag her på de enkelte leddene. Hvor mange forslagene man skal undersøke, er et spørsmål om hvor store ressurser man er villig til å bruke. I følge vår undersøkelse vil det være rimelig å stoppe etter de 2-3 første forslagene, fordi den riktige løsningen finnes som oftest bl.a. disse.

## 6. Forsøk med metodene i tekstsøking

Som et ledd i arbeidet med disse metodene, ble det gjennomført flere forsøk med formål å undersøke hvilken effekt de vil ha på søkeresultatet. Ett av dem gikk ut på å undersøke hvor mange flere dokumenter som ble funnet når man søkte på alle bøyning- og avledningsformer til et ord i stedet for bare grunnformen til ordet. (Det er grunnformen som oftest bli anvendt i søkeargumentet.) Forsøket omfatte 121 søkeord, og søkingen var rettet mot en juridisk database som omfattet ca. 14-15000 sammendrag av høyesteretts-avgjørelser (én av LOVDATA's databaser). Resultatet viste en gjennomsnittlig økning i antall funne dokumenter på hele 215%.

Et annet forsøk var rettet mot sammensatte søkeord, og tok sikte på å undersøke i hvor stor grad man var i stand til å fange opp omskrivinger av de sammensatte søkeordene ved å søke på uttrykk som besto av en kombinasjon av de enkelte leddene i det sammensatt ordet (eller bøynings- og avledningsformer av disse). Dette forsøket omfattet 54 sammensatte ord, og resultatet viste at i 49 av disse tilfellene ble omskrivinger fanget opp. Det ble stilt som krav at de enkelte (usammensatte) ordene skulle forekomme i samme setning.

Disse undersøkelsene viser at både metode for automatisk rotlemmatisering og metode for automatisk splitting av sammensatte ord vil kunne bidrar til mer fullstendige søkeargumenter. Derimot gir de ikke informasjon om den endelige effekten av disse metodene på søkeresultatet da dette vil avhenge av både

brukerens informasjonsbehov, den totale mengden med søkeord og søkestrategien, samt den aktuelle databasen. Det er samspillet mellom disse faktorene som er bestemmende for søkekvaliteten.

For derfor å få nærmere innsikt i effekten av bruk av disse metodene på søkeresultatet, ble det gjennomført en serie med kontrollerte forsøk i tekstsøking. Lignende forsøk med automatisk rotlemmatisering har vært gjennomført tidligere for engelsk (Salton 1968) og tysk (Niedermair/ Thurmair/Bütler 1984), og begge ga positive resultatet.

***Beskrivelse av et kontrollert forsøk***
Et kontrollert forsøk i tekstsøking går i korte trekk ut på at man med utgangspunkt i en gitt dokumentsamling og et sett med spørsmål, sammenligner resultatet av en maskinell søking med resultatet av en manuell gjennomlesing av hele dokumentsamlingen. Resultatet uttrykkes ofte ved bruk av effektivitetsmålene gjenfinningsgraden (G) og presisjon (P). Ved å dele inn resultatlista i rangsett - noe som forutsetter en søkestrategi som rangerer de funne dokumentene - kan resultatet av et søk presenteres i en GP-kurve. På bakgrunn av de individuelle GP-kurvene kan man så beregne en gjennomsnittlig GP-kurve som gir uttrykk for søkekvaliteten for dette søkesettet. Ved å gjennomføre flere slike forsøk for ulike typer søkeargumenter (eller søkestrategier), vil man ved å sammenligne de gjennomsnittlige GP-kurvene kunne si noe om hvilke type søkeargumenter som gir best resultat. Man vil så kunne gjennomføre en nærmere undersøkelse av de enkelte søk for å få innsikt i hvorfor resultatet er som det er.

Våre forsøk var basert på et eksperimentmateriale som omfattet ca. 1300 sammendrag av domsavgjørelser i familie-, skifte- og arverett.[5] Det ble utformet 22 spørsmål av varierende kompleksitet og lengde. En mer utførlig beskrivelse av eksperimentmaterialet og forsøkene er gitt i Fjeldvig 1987:171-200.

Den manuelle gjennomlesingen av dokumentsamlingen ble foretatt av de samme personene som hadde stilt spørsmålet. Hvert spørsmål ble behandlet for seg, og de relevante dokumentene ble notert på en liste. Denne listen kalles ofte fasiten til spørsmålet.

For hvert spørsmål ble det konstruert et søkeargument som bare omfattet ord som forekom i spørsmålet. Det var gjennomsnittlig 5 søkeord pr. søkeargument. Ved søking ble alle dokumenter som inneholdt minst ett av disse søkeordene, valgt ut. Deretter ble de funne dokumentene rangert ut fra hvor mange ulike søkeord de inneholdt.

Det første forsøket ble gjennomført med søkeargumenter som bare inneholdt grunnformen til søkeordene. Den gjennomsnittlige GP-kurven som dette ga, er representert ved den heltrukne kurven i fig. 1. Denne kurven gir altså uttrykk for de søkeresultatene vi fikk uten å ta i bruk noen av våre metoder.

Eksempel: Ett av våre spørsmål lød omtrent som følger: *Har et barns handicap betydning for oppfostringsbidragets størrelse?* Søkeargumentet besto i dette tilfellet av følgende ord:

---

5. Totalt antall ord var ca. 190 000 og antall ulike graford ca. 11 000.

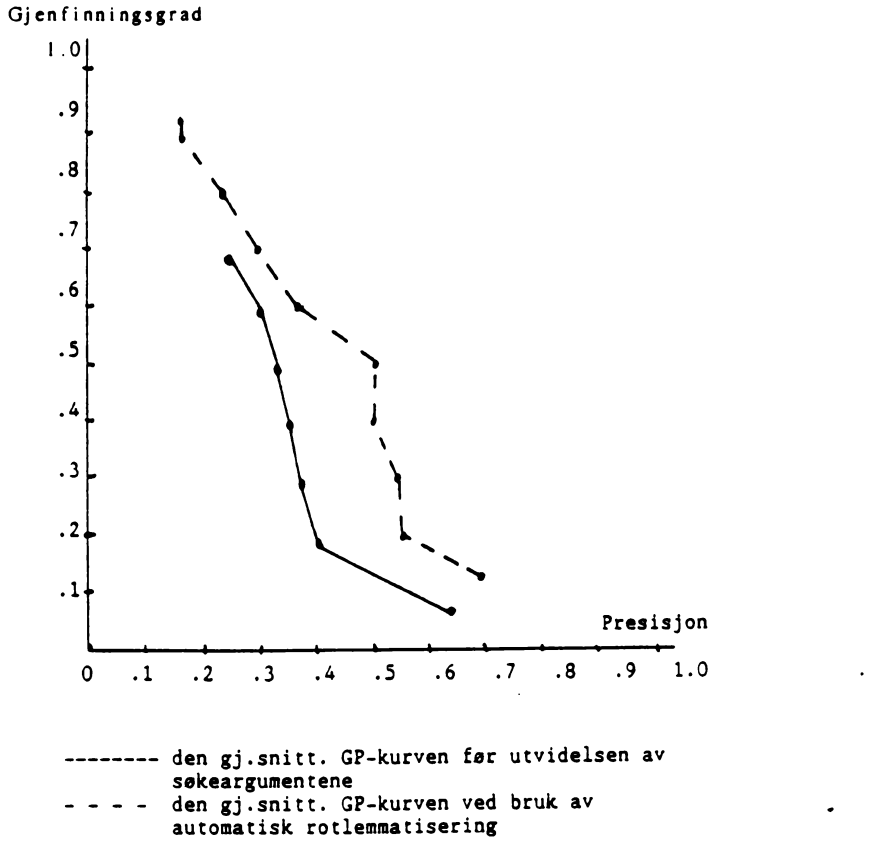
*barn*
*handicap*
*oppfostringsbidrag*

***Automatisk rotlemmatisering***
Deretter erstattet vi hvert søkeord med en gruppe med ord som inneholdt alle de bøynings- og avledningsformene til dette ordet som forekom i databasen. Søkeargumentet i eksempelet ovenfor omfattet nå følgende ord:

*(barn, barns, barnet, barnets, barna, barnas)*
*(handicap, handicapet, handicapede, handicapedes)*
*(oppfostringsbidrag, oppfostringsbidraget, oppfostringsbidragets)*

Rangeringen av de funne dokumentene ble nå foretatt på bakgrunn av hvor mange slike grupper som var representert i dokumentet. Resultatet er representert ved den stipplete kurven i fig. 1.

*Fig. 1 Effekten av å utvide søkeargumentene med bøynings- og avledningsformer*



Sammenligner vi nå disse to kurvene, ser vi at vi har oppnådd et betydelig bedre resultat med bruk av metoden for automatisk rotlemmatisering. En sammenligning av resultatene for hvert enkelt søk viste også en forbedring av søkekvaliteten i over halvparten av tilfellene. Den gjennomsnittlige økningen i gjenfinningsgraden var på hele 24%.

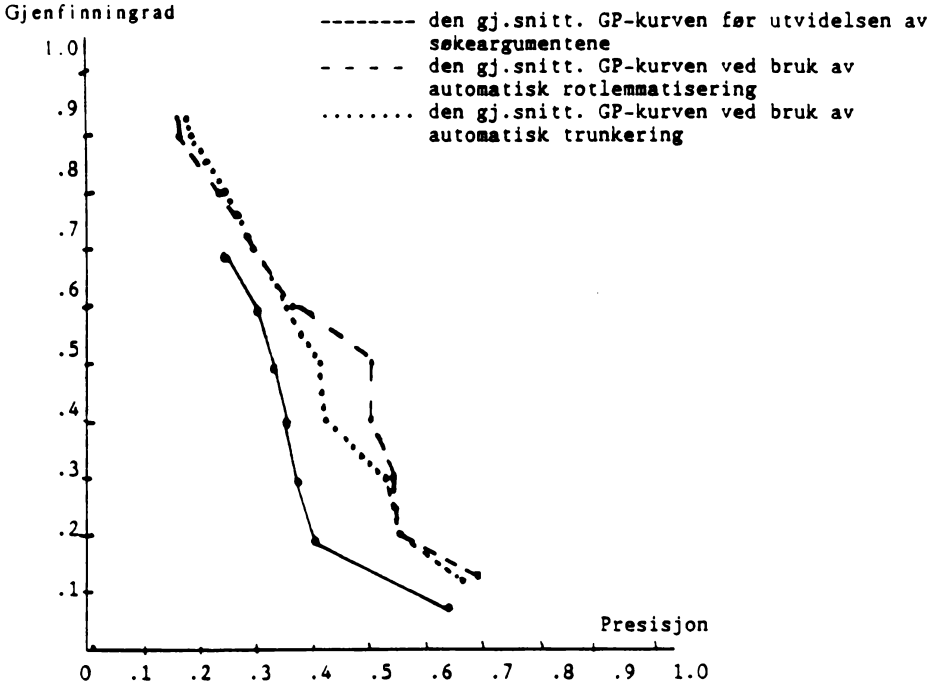
***Automatisk trunkering***
På lignende måte ble det gjennomført et kontrollert forsøk med automatisk trunkering. Her ble hvert (opprinnelig) søkeord automatisk trunkert.

Eksempel:

*barn**
*handicap**
*oppfostringsbidrag**

Resultatet er gjengitt ved den prikkete linjen i fig. 2. nedenfor. De to øvrige kurvene i figuren er de samme som i fig. 1.

*Fig. 2 Sammenligning av resultatet ved automatisk trunkering og automatisk rotlemmatisering*




En sammenligning av kurven for automatisk trunkering med den vi fikk ved eksplisitt å supplere søkeargumentet med alle bøynings- og avledningsformer.

viser helt klart at den automatisk trunkering absolutt er et aktuelt hjelpemiddel ved tekstsoking. Forskjellen mellom disse to kurvene er helt marginal, og det er vanskelig å si om den ene kurven er bedre enn den andre.

En nærmere undersokte av de enkelte søkene viste at den automatiske trunkeringen i et par av tilfellene hadde ført til at flere relevante dokumenter. De fleste resultatlistene inneholdt nå flere irrelevante dokumenter, men disse dokumentene havnet som oftest lenger ned på lista fordi de inneholdt relativt få søkeord.

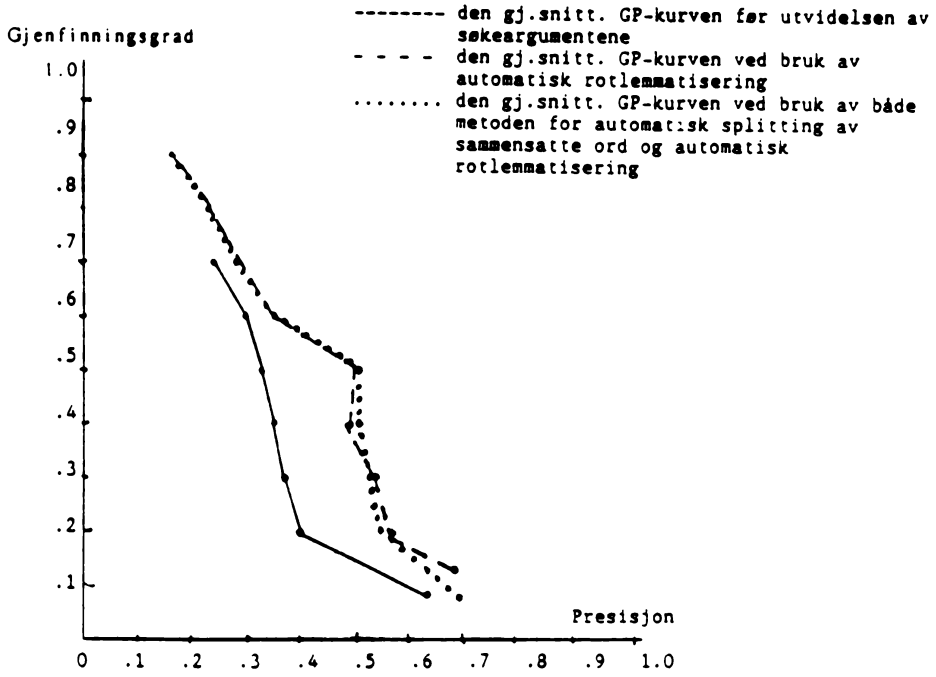
***Automatisk gjenkjenning og splitting av sammensatte søkeord.***
I neste forsøk ble i tillegg alle sammensatte søkeord splittet og supplert med et uttrykk som inneholdt de enkelte leddene i det sammensatte ordet, f.eks. ordet *oppfostringsbidrag* ble supplert med uttrykket

*(oppfostre, oppfostrer, oppfostret, oppfostring, oppfostringen)*
SETN (*bidra, bidrar, bidratt, bidraget, bidragets, bidragene, bidragene)*

Operatoren SETN, også kalt setningsoperatoren, stiller som krav at ett av ordene i den første parantesen må forekomme i samme setning som ett av ordene i den andre parantesen.

Resultatet av dette forsøket er gjengitt ved den prikkete kurven i fig. 3. Den stipplete kurven er den samme som i fig. 2, dvs. resultatet uten at de sammensatte søkeordene er behandlet.

*Fig. 3 Effekten av å splitte de sammensatte søkeordene*



Som det fremgår av figuren, ga denne utvidelsen av søkeargumentet lite utslag på det gjennomsnittlige søkeresultatet. Den førte ikke til at det ble funnet flere relevante dokumenter, fordi de fleste relevante dokumentene var alt funnet på bakgrunn av de øvrige søkeordene. Rangeringen ble bedre i noen av tilfellene, men i de fleste tilfellene var økning i antall søkeord i de relevante dokumentene for liten til å kunne endre rekkefølgen i resultatlista.

Jevnt over førte denne utvidelsen til at det ble funnet flere irrelevante dokumenter. Dette hadde heller ikke særlig innflytelse på resultatet fordi disse dokumentene inneholdt relativt få søkeord og ble derfor plassert i de bakerste rangsett.

En nærmere undersøkelse av hvert enkelt søk viste imidlertid at denne behandlingen av de sammensatte søkeordene i ca. halparten av tilfellene hadde ført til at søkeargumentet nå inneholdt søkeuttrykk som fanget opp omskrivinger av de sammensatte søkeordene. I de resterende tilfellene utgjorde de sammensatte ordene typiske juridiske faguttrykk (f.eks. *ektepakt* og *halvpart*) eller leksikaliserte ord (f.eks. *pleiedatter* og *lønnsbringende*). Det var en langt høyere andel av denne type ord enn det vi hadde oppnådd i andre sammenhenger. Likevel er det sjelden at en splitting av denne type ord vil få negativ effekt på søkeresultatet, fordi de enkelte usammensatte ordene ikke vil forekomme i samme setning.

## 7. Konklusjon

Resultatene av disse forsøkene viser helt klart at både metoden for automatisk rotlemmatisering og metoden for automatisk trunkering gir en betydelig forbedring av søkekvaliteten. Hvilke av de to fremgangsmåtene man bør satse på, vil avhenge av ulike hensyn både av systemmessig og ressursmessig art.

Forsøket med automatisk splitting av sammensatte ord viste at dette hadde liten innflytelse på søkekvaliteten. Likevel har vi tro også på en slik metode i informasjonssøking, fordi forsøket tross alt viste at vi på denne måten får utvidet søkeargumentene med adekvate søkeuttrykk. At resultatet ikke ble bedre i dette forsøket, mener vi skyldes ulike egenskaper ved eksperiment-materialet.

Tatt i betraktning at svært mange brukere (typisk uøvde og sporadiske brukere) verken trunkerer søkeordene eller på annen måte sørger for å få med de ulike bøynings- og avledningsformene til søkeordene (jfr. Fjeldvig 1987:53-52), anser vi disse metodene for å være et vesentlig bidrag til løsningen av synonymproblemet i tekstsøking.

**LITTERATURLISTE**

Fjeldvig, Tove/Golden, Anne (1984) *Automatisk rotlemmatisering - et lingvistisk hjelpemiddel for tekstsøking.* CompLex 9/84, Universitetsforlaget. Oslo.

Fjeldvig, Tove (1986) *Tekstsøking - teori, metoder og systemer.* Universitetsforlaget. Oslo.

Fjeldvig, Tove/Golden, Anne (1986) "Automatisk splitting av sammensatte ord - et lingvistisk hjelpemiddel for tekstsøking"; Karlsson 1986:73-82.

Fjeldvig, Tove (1987) *Effektivisering av tekstsøkesystemer. Utvikling av språkbaserte metoder.* Universitetsforlaget. CompLex nr. 13/87. Oslo.

Gavare, Rolf (1979) "Automatisk lemmatisering utan stamlexikon - Några synspunkter tio år efteråt". Maegaard 1979: 123-131.

Hellberg, Steffan (1971) "Automatisk lemmatisering - En modell for upprättande av böyningsserier i ett frekvenslexikon". Språkdata. Gøteborg.

Karlsson, Fred (ed) (1986) *Papers from the 5th Scandinavian Conference of Computation and Linguistics.* University of Helsinki, Department of General Linguistics.

Källgren, Gunnel (1985) *En algoritme för deling av sammensatte ord i svenskan.* Institutionen för lingvistik. Stockholms Universitet. Stockholm.

Maegaard, Bente (1979) *Nordiske datalingvistikdage i København 6.-10. oktober 1979.* Foredrag. Institut for anvendt og matematisk lingvistik, Københavns Universitet 1979. København.

Munthe, Synneve Kjuus Munthe (1972) *Sammensatte ord. En kvantitativ undersøkelse av norsk litteratur og sakprosa.* Hovedfagsoppgave ved Nordisk institutt, Universitetet i Bergen og Oslo.

Niedermair, G.T./Thurmair G./Büttel I. (1984) "MARS - a retrieval tool on the basis of morphological analysis", van Rijsbergen 1984:369-382.

Salton, Gerard (1968) *Automatic Information Organization and Retrieval.* McGraw-Hill computer series.

van Rijsbergen C.J. (1984) *Research and Development in Information Retrieval.* Proceedings of the third joint BCS ans ACM symposium King's College, Cambridge 2.6 July 1984. British Computer Society Workshop Series.