

9/84

Tove Fjeldvig og Anne Golden

**AUTOMATISK
ROTLEMMATISERING —
ET LINGVISTISK
HJELPEMIDDEL FOR
TEKSTSØKING**

NORIS (64) og NORIS (58)

COMPLEX

INSTITUTT FOR RETTSINFORMATIKK

UNIVERSITETSFORLAGET



IBM Bergen Dreggsalmenning 10/12 III (05) 31 55 00
IBM Hamar Parkgt 2 III (065) 27 745
IBM Kristiansand S Radhusgt 3 III (042) 29 160
IBM Sarpsborg Kirkegt 63, III (031) 57 036
IBM Stavanger Auglendsdalen #1 III (04) 58 85 00
IBM Trondheim Kongensgt 60, III (075) 30 644
IBM Tønsberg Nedre Langgt 33 III (033) 12 013
IBM Ålesund Parkgt 7 III (071) 24 387
IBM Oslo Dronning Maudsgt 10/11 III (02) 20 54 50

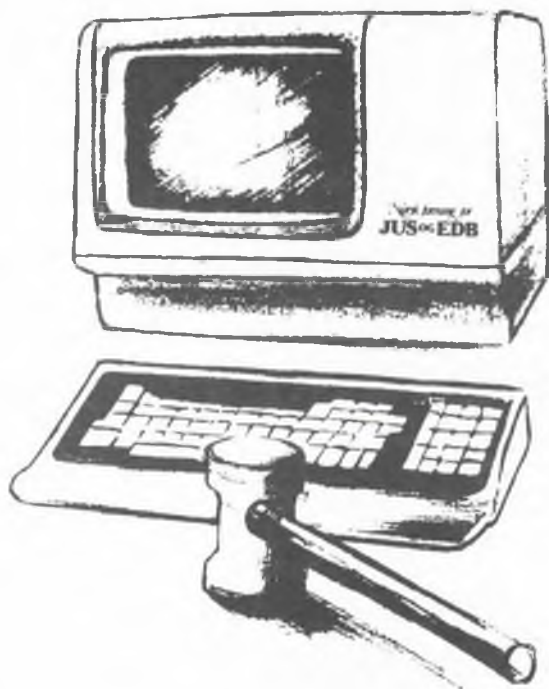


Norsk forening for **JUS og EDB**

Postboks 7557, Skillebekk, OSLO 2

Postgiro: 5 13 96 54 Bankgiro: 8200 42 49727

Postgiro CompLex: 2 12 16 63



Foreningen står bl.a. for
salget av CompLex-heftene
og vedlikeholder abonnements-
ordningene for serien.



CompLex nr. 9/84

Institutt for rettsinformatikk
Niels Juels gate 16
0272 OSLO 2

Tove Fjeldvig og Anne Golden

AUTOMATISK ROTLEMMATISERING — ET LINGVISTISK HJELPEMIDDEL FOR TEKSTSØKING

NORIS (64) og NORIS (58)

Prosjektet er støttet av Norges teknisk-naturvitenskapelige forskningsråd (NTNF), og Statens rasjonaliseringsdirektorat har gitt et bidrag til publisering av denne rapporten.

Universitetsforlaget
Oslo

© Universitetsforlaget 1984

ISBN 82-00-07035-2

Utgivelsene i skriftserien Complex støttes av:

Den norske Bankforening

Bergens Tidende

Norsk Arbeidsgiverforening

Televerket

Norsk senter for informatikk A/S

Den Norske Advokatforening

Norges Forsikringsforbund

Orkla Industrier

Printed in Norway

by GCS A/S, Oslo

INNHOLDSFORTEGNELSE

FORORD	side
1. SYNONYMPROBLEMET I TEKSTSØKING	8
1.1 Tekstsøking	8
1.2 Synonymer i tekstsøking	10
1.3 Hjelpemidler til bruk ved formulering av søkeargumentet	11
1.4 Forskning omkring lingvistiske hjelpemidler	14
1.5 Behovet for en metode for løsning av de morfologiske problemene på norsk	17
2. METODE FOR AUTOMATISK ROTLEMMATISERING	19
2.1 Regelbasert metode	19
2.2 Begrepet "automatisk rotlemmatisering"	19
2.3 Alternative framgangsmåter	22
2.4 Skisse av en regelbasert metode for automatisk rotlemmatisering	29
3. MORFOLOGIEN I NORSK	31
3.1 Innledning	31
3.2 Bøyningsmuligheter	31
3.3 Avledningsmuligheter	33
3.4 Sammenfall i endelser	34
3.5 Åpne og lukkede ordklasser	35
3.6 Analyse av ordklassene	35
(1) Innledning	35
(2) Pronomen, artikler og tallord	36
(3) Adverb	36
3.7 Åpne ordklasser med bøyningsmuligheter	36
(1) Innledning	36
(2) Substantiv	38
(3) Verb	38
(4) Adjektiv	39
3.8 Åpne ordklasser med avledningsmuligheter	40
(1) Innledning	40

	side
(2) Substantiv	40
(3) Verb	40
(4) Adjektiv	40
4. ETABLERING AV REGELSETTET	41
4.1 Eksperimentmaterialet	41
4.2 Oppslagsform	41
4.3 Eksempel på problemer ved rotlemmatisering	42
4.4 Beskrivelse av en regel	46
(1) Innledning	46
(2) Bokstavstreng	46
(3) Betingelse	47
(4) Ordre	47
(5) Typebetegnelse	48
4.5 Regeltyper	48
(1) Tre hovedoppgaver	48
(2) Operasjoner	50
(3) Regler som nøytraliserer både bøynings- og avledningsendelser	53
4.6 Endelige ordgrupper - leksikon	54
5. FORSØK MED AUTOMATISK ROTLEMMATISERING	58
5.1 Hva er riktig rotlemmatisering?	58
(1) Entydig oppslagsform	58
(2) Leksikaliseringsproblemet	59
(3) Gradbøyningsproblemet	59
(4) Homografer	59
5.2 Ord som ble gruppert riktig	62
5.3 De mest problematiske ordformene	65
(1) Oversikt over resultatet	65
(2) Bøynings- og avledningsregler	66
(3) Feiltyper	67
5.4 Regler som førte til feil rotlemmatisering	70
5.5 Regelstatistikk	73
5.6 Resultater med testmaterialet	78

	side
6. AUTOMATISK ROTLEMMATISERING I TEKSTSØKING	79
6.1 Eksperimenter med automatisk rotlemmtisering	79
6.2 Automatisk rotlemmatisering i forhold til trunkering	79
6.3 Bestemmelse av ordenes vekt ved søking	82
6.4 Automatisk rotlemmatisering som ingrediens i andre hjelpemidler	85
(1) Innledning	85
(2) Automatisk trunkering	85
(3) Automatisk splitting av sammensatte ord	87
(4) Automatisk ordklassebestemmelse	87
6.5 Forbedring av metoden	88
(1) Vurdering av dagens metode	88
(2) Reduksjon av eksisterende feil	89
(3) Overgangen til et nytt materiale	90
(4) Endringer i regelsettet ved oppdatering av tekstmaterialet	91
(5) Bruk av metoden	92
7. ANDRE ANVENDELSER AV METODEN	93
7.1 Innledning	93
7.2 Datamaskinstøttet stikkordproduksjon	93
7.3 Lingvistisk hjelpemiddel	94
7.4 Videreføring	96

LITTERATURLISTE



FORORD

Med denne rapporten setter vi sluttstrek for et stort, arbeidskrevende, men oppslukende arbeid innen språklig databehandling.

Sluttproduktet er et resultat av mange års arbeid. Samarbeidet mellom forfatterene har vært en klar forutsetning for fullføringen av prosjektet fordi det krevde både lingvistisk kompetanse og godt kjennskap til tekstsøking og databehandling. Vi ønsker derfor å benytte denne anledning til å fortelle hvor tilfeldig det var at dette samarbeid i det hele tatt kom i gang og hvor viktig det er at man får muligheten til å reise på seminarer, kongresser o.l. der møte med andre innenfor beslektede fagområder ofte spiller en større rolle enn seminarets innhold.

Vi møtte hverandre i 1982 på vei til De nordiske datalingvistikk-dager i Trondheim. Ved en tilfeldighet havnet vi ved siden av hverandre i flyet, og det tok ikke lang tid før vi oppdaget at Tove (Fjeldvig) forsøkte å utvikle det som Anne (Golden) hadde savnet i sitt prosjekt.

Tove arbeidet med tekstsøkesystemer (datamaskinelle systemer for gjenfinning av dokumenter) og var opptatt av å finne fram til en regelbasert metode som kunne kjenne igjen ord med samme rot. I et tekstsøkesystem er ethvert ord i teksten søkbart, og uten informasjon om hvilke ord som er bøynings- og avledningsformer av samme rot, vil systemet oppfatte to ord som forskjellig hvis en av bokstavene er forskjellig eller rekkefølgen av dem forskjellige.

Anne arbeidet med norskundervisning for utenlandske studenter og var engasjert i et prosjekt sammen med amanuensis Anne Hvenekilde der ordforrådet i grunnskolenes o-fagsbøker (orienteringsfag) skulle analyseres. Et viktig ledd i dette studiet var å undersøke hvilke ord som forekom hyppigst og som det var grunn til å legge vekt på i undervisningen av fremmedspråklige elever i grunnskolen. På grunn av materialets størrelse (ca. 830 000 løpende ord) var det en forutsetning at frekvens-

studiene ble gjennomført maskinelt, men grupperingen av ord med samme grunnform måtte gjøres manuelt i mangel av en maskinell rutine. Dette arbeidet krevde at man satte seg inn i ordenes betydning og den betydningsendringen avledningsendelsene førte til. Det var både tidkrevende og lett å gjøre feil i kodearbeidet, fordi materialet var så stort.

Samarbeidet startet umiddelbart og fungerte utmerket - ikke minst på grunn av våre ulike kompetanseområder. Tove står ansvarlig for konstruksjon av metoden, dens tilpasning og anvendelse i tekstsøkesystemer og de maskinelle oppgaver som dannet data-grunnlaget for metodens utvikling og testing. Anne bærer ansvaret for den lingvistiske delen og har analysert det morfologiske systemet i norsk med tanke på en formalisering i regler. Hun har også arbeidet med å kartlegge de semantiske variasjonene som avledningsendelsene kan føre til og har på denne bakgrunn utformet forslagene til reglene. Vi har begge deltatt i etableringen av regelsettet, og bærer sammen ansvaret for det regelsettet som vi sitter igjen med i dag. Denne delen av prosjektet som var den mest tidkrevende, men samtidig den mest besettende som drev det hele videre gjennom mange sene kveldstimer og helger. Arbeidet var preget av mye prøving og feiling, og med et materialet på ca. 24 000 ulike ord var det mange sider å lese gjennom for hver gang noe ble forandret. Denne type arbeid tar egentlig aldri naturlig slutt, ettersom det er umulig å oppnå et 100% riktig resultat. Det var derfor vanskelig å si stopp, fordi vi alltid øynet muligheten for å gjøre det litt bedre.

Nå er imidlertid punktumet satt, til tross for at vi (selvfølgelig) ser feil som vi med letthet kunne ha rettet. I denne rapporten vil vi primært presentere metoden, så får poleringen komme alt etter sammenhengen metoden skal anvendes i.

Prosjektet er blitt til innenfor NTNf-prosjektene STANS og FORT, og er definert som delprosjekt 64 i Institutt for rettsinformatikk's forskningsprogram NORIS.

Til slutt vil vi takke kontorfullmektig Gina Fraas for innsatsen i forbindelse med utgivelsen av denne rapporten.

Oslo 15.12.1984

Tove Fjeldvig Anne Golden



1. SYNONYMPROBLEMET I TEKSTSØKING

1.1 Tekstsøking

Begrepet "tekstsøking" anvendes i dag som en betegnelse på en datamaskinell prosess for gjenfinning av dokumenter. Et dokument kan i prinsippet være et hvilket som helst skrevet materiale, f.eks. et brev, en artikkel, en bok, et skjema eller en tabell. I juridiske informasjonssystemer vil lover, forskrifter og domsavgjørelser være typiske eksempler på dokumenter, og det er i dag mulig å søke maskinelt i disse samlingene.

Tekstsøkesystemer skiller seg fra andre typer søkesystemer ved at de er basert på dokumenter i form av tekster. Det er med andre ord ikke nødvendig å bearbeide dokumentene før de registreres i søkesystemet – man kan benytte den dokumentformen man finner mest hensiktsmessig. En tekst kan i prinsippet beskrives på tre ulike måter; ved sin autentiske form, som et sammendrag eller i form av stikkord. Alle tre formene aksepteres av tekstsøkesystemet, men systemet er ikke i stand til å skille mellom dem. Dokumentene blir – uansett i hvilken form innholdet er presentert – oppfattet som en rekke med ord, og det er ordene som representerer innholdet i dokumentet og som utnyttes ved søking. Dette forutsetter at man på forhånd har gitt systemet informasjon om hva som er et ord og hvor skillet går mellom to dokumenter.

Et tekstsøkesystem er spesielt anvendelig i situasjoner hvor en bruker er engasjert i et problem og ønsker dokumenter som kan belyse dette problemet. En jurist kan f.eks. være interessert i alle rettsavgjørelser som belyser spørsmålet om betinget straff for unge bilførere i promillesaker. Problemet beskrives for søkesystemet i form av ord og uttrykk, og det er viktig for søkeresultatet at brukeren definerer de samme søkeordene som er brukt til å uttrykke dette problemet i dokumentene. Dokumenter som ikke inneholder søkeordene, vil ikke bli funnet. Dette betyr at hvis man i tilfellet ovenfor bare hadde brukt søkeordet

PROMILLE, vil man ikke ha funnet fram til de dokumenter som bare inneholdt ordene PROMILLESAK eller PROMILLEN. Søkesystemet oppfatter to ord som forskjellige så sant en av bokstavene eller rekkefølgen av dem er forskjellig. Ulike bøyningsformer av samme grunnform vil derfor oppattes som forskjellige ord.

Ethvert ord i dokumentene kan i prinsippet anvendes som søkeord. Det er imidlertid slett ikke alle ord som er egnet til det. Enkelte ord må sees i sammenheng med andre ord (f.eks. preposisjoner, konjunksjoner, pronomen, tall, benevelser, osv.), mens andre har så stor spredning i dokumentsamlingen at de ikke er karakteristiske nok for nettopp dette dokumentet. Det finnes også ord som uttrykker et meningsinnhold som det sjelden er aktuelt å søke på, f.eks. ANSE, GJELDE, SYNES, o.l.

For å få en bedre representasjon av dokumentene i søkesystemet, tillegger enkelte tekstsøkesystemer ordene en vekt som skal gjenspeile deres betydning for innholdet i dokumentet. Normalt anvendes ordets frekvens i dokumentet til dette formålet, men det finnes også mer avanserte vektfunksjoner. Dette gjelder ikke ord som preposisjoner, konjunksjoner, artikler, pronomen o.l., for disse blir vanligvis fjernet helt fra søkegrunnlaget fra starten av (de kalles gjerne stoppord). Ved søking blir vektene brukt til å rangere de dokumentene som er funnet, slik at de dokumentene med flest søkeord med høy vekt blir plassert først på resultatlista. Dette skjer ut fra en hypotese om at jo flere søkeord et dokument inneholder og høyere vekt disse har, jo større sannsynlighet er det for at dokumentet berører det meningsinnholdet som søkeordene representerer. Formålet er å få ordnet resultatlista på en slik måte at de dokumentene med størst sannsynlighet for å være relevante, blir presentert først for brukeren. Det bør imidlertid påpekes at ingen tekstsøkesystemer kan garantere et 100% tilfredsstillende resultat i enhver søkesituasjon - dvs. at man finner alle de relevante dokumentene og bare disse. Ettersom relevansvurderingen av et dokument skjer på et subjektivt grunnlag, vil det alltid kunne oppstå uenighet om hvorvidt et dokument er relevant eller ikke. Selv om man var enige, så er språket så tvetydig og mangfoldig at det kan være

vanskelig å forutse hvilke ord som er brukt til å uttrykke et bestemt meningsinnhold (synonymproblemet) og om forekomsten av et bestemt ord uttrykker det meningsinnholdet som det søkes etter (homografproblemet).

1.2 Synonymer i tekstsøking

Forskjellige tekster kan handle om det samme, men allikevel bestå av helt forskjellige ord og uttrykksmåter. De ordene som er brukt i en tekst er et resultat av forfatterens valg, samtidig som de er avhengige av en rekke faktorer: hva teksten handler om, hvem den er skrevet for, i hvilken hensikt den er skrevet, hvorvidt den er en del av en større helhet, forfatterens spesielle stil osv. Ordformene som brukes, avhenger av hvilken funksjon ordet har i setningen, og styres av de grammatiske reglene i språket. Setningenes rekkefølge er heller ikke tilfeldig, de må også følge visse regler for at de skal kunne utgjøre en tekst med mening.

At forskjellige ord kan brukes om det samme meningsinnholdet, og at de ulike ordene har bøynings- og avledningsmuligheter, er nettopp det som vanskeliggjør tekstsøkningen. Det er dette problemet som gjerne kalles synonymproblemet i tekstsøking. Synonymer viser altså ikke bare til ord som betyr akkurat det samme (har samme referent), men til ord som er likeverdige når det gjelder å finne ut hva teksten handler om. I en tekstsøkingssituasjon kan vi derfor stå ovenfor følgende typer synonymer:

- 1) ulike bøyningsformer av samme ord (f.eks. KJØPE - KJØPTE),
- 2) ulike avledninger av et ord (f.eks. KJØPE og KJØP)
- 3) ulike skriftvarianter av samme ord (f.eks. SYKEHUS - SJUKEHUS)
- 4) likeverdige uttrykksmåter av typen PIKE - JENTE, ARVEAVGIFT - AVGIFT PÅ ARV,
- 5) generiske relasjoner av typen ROSE - BLOMST, SAKS - MORDVÅPEN

Bortsett fra de morfologisk betingede variasjoner av et ord (jfr. punkt 1 og 2) og ulike skriftvarianter av samme ord (punkt 3), er det få eksempler på helt kontekstuavhengige synonymer, dvs. ord som kan erstatte hverandre i enhver sammenheng. Frasen "AVGIFT PÅ ARV" og ordet ARVEAVGIFT er et eksempel på dette, og de fleste vil vel også hevde at JENTE og PIKE er kontekstuavhengige synonymer. Graden av kontekstavhengighet varierer, og man kan nærmest tenke seg synonymene plottet inn på en skala fra ren kontekstuavhengighet til ren kontekstavhengighet. Enkelte synonymer kan betraktes som kontekstuavhengige i forhold til en gitt dokumentsamling eller et gitt emneområde. Andre ord derimot, er kun synonyme i den aktuelle søkesituasjonen, som f.eks. ordet BÅTJENTE som var brukt synonymt med DÅRLIG VANDEL i domsavgjørelsen til en farskapssak i vårt eksperimentmateriale.

Den utstrakte bruken av synonymer gjør formuleringen av søkeargumentet spesielt vanskelig for uerfarne brukere og de som ikke kjenner språkbruken i dokumentbasen. Studier av årsaken til effektivitetssvikt i tekstsøking viser at den største grunnen til at relevante dokumenter ikke blir funnet, skyldes nettopp mangler ved søkeargumentet. Brukeren har f.eks. glemt aktuelle synonymer eller grammatikalske bøyninger, eller oppgitt for generelle ord. Dette er en langt vanligere årsak til et dårlig resultat enn de årsaker som kan spores tilbake til selve språket, f.eks. at meningsinnholdet ikke er uttrykt godt nok ved ord i dokumentet, men "gjemt mellom linjene".

1.3 Hjelpemidler til bruk ved formulering av søkeargumentet

For å kunne løse synonymproblemet i tekstsøking vil det være behov for tre ulike typer hjelpemidler.

For det første trengs det et hjelpemiddel som kan behandle de morfologiske variasjonene i språket, slik at brukerne slipper å bekymre seg om hvilke bøynings- og avledningsformer som er brukt i dokumentmaterialet.

For det andre må man kunne ta hånd om det de fleste betrakter som kontekstuavhengige synonymer (f.eks. JENTE - PIKE, BLOMST - ROSE) og andre ord som kan erstatte hverandre innenfor det bestemte dokumentasjonsområdet som dokumentene representerer.

Langt større krav stilles det til et tredje hjelpemiddel som kan hjelpe brukeren i spesifiseringen av de kontekstavhengige synonymer. Dette forutsetter at tekstsøkesystemet kan se søkeordene i sammenheng med det problemområdet som beskrives - noe som kan bety at tekstsøkesystemet må inneholde en beskrivelse av dokumentasjonsområdet. Et slikt hjelpemiddel finnes ikke i dag, men med den forskningen som skjer innenfor området kunstig intelligens og kunnskapsbaserte systemer, ser vi muligheten for et slikt hjelpemiddel i framtiden.

Hjelpemidler til spesifisering av kontekstuavhengige synonymer finnes. Synonymtesaurusen er et eksempel på dette. En synonymtesaurus er en synonymordbok hvor det til hvert ord er knyttet informasjon om hvilke ord som er relatert til dette ordet, f.eks. til ordet HUND finner man det likeverdige ordet BIKKJE, det underordnede ordet COLLIE, det sideordnede ordet KATT og det overordnede ordet HUSDYR.

Det er imidlertid kostbart å utvikle og - ikke minst - vedlikeholde tesauruser. De må utvikles manuelt av personer som er eksperter på området, og de må holdes ajour ved endringer i dokumentsamlingen. Det kan derfor reises tvil om de virkelig gir en effektivitetsøkning som kan svare til kostnadene forbundet med dem. Tesauruser er derfor lite tatt i bruk i tekstsøkesystemer. De fleste eksempler på bruk av tesauruser finner man i Frankrike (f.eks. i systemene MISTRAEL og SPIRIT) og i Italia (f.eks. i systemet ITALGIURE).

Blant operative tekstsøkesystemer finner man få som har løst de morfologiske problemer på en tilfredstillende måte. Med det sikter vi til hjelpemidler som kan føre til at søkeargumentet automatisk blir supplert med alle aktuelle bøynings- og avledningsformer av de ordene som det søkes på. Eksempel på systemer som har et slikt tilbud er LEXIS, RESPONSA, ITALGIURE,

SPIRES og PASAT. I LEXIS er det implementert en regelbasert metode som tar hånd om denne oppgaven, mens i de øvrige systemer skjer det mer eller mindre med utgangspunkt i et lingvistisk leksikon som er satt opp manuelt. Et lingvistisk leksikon vil her si en ordbok hvor det til hvert ord er knyttet nødvendig lingvistisk informasjon som f.eks. morfologisk informasjon. Metoden omfatter vanligvis bare bøyninger til søkeordene og ikke avledninger (f.eks. ARVEN til søkeordet ARV, men ikke ARVING).

Det mest vanlige - og ofte det eneste - hjelpemiddelet i tekst-søkesystemer, er trunkering. Trunkering er et svært enkelt verktøy som går ut på at man kan få supplert søkeargumentet med alle ord som begynner på en gitt tegnstreng (høyretrunkering) eller avslutter med denne tegnstrengen (venstretrunkering). I de fleste tilfeller vil tegnstrengen være lik rot til et ord, og på denne måten kan man enten få alle ord som innledes eller avsluttes med denne roten.

Høyretrunkering er mest vanlig. F.eks. vil søkeargumentet "BIL*" (hvor * er valgt som trunkeringstegn) omfatte alle ord som begynner med bokstavsekvensen "BIL", f.eks. BIL, BILE, BILER, BILENE, BILHOLD, osv. På denne måten får man definert de fleste bøynings- og avledningsformene til søkeordene på en enkel og lite ressurskrevende måte. Trunkering dekker også sammensatte ord (f.eks. BILHOLD, BILVEI, BILAVGIFT etc.) og ofte finner man mange relevante søkeord blant de sammensatte ordene.

Svakheten ved trunkering er at den fanger opp irrelevante ord, f.eks. som BILLION og BILLIG i tilfellet ovenfor. Denne effekten blir ofte forsterket ved at brukeren definerer for korte trunkerte strenger, f.eks. trunkerer BI* istedenfor BIL*. det er også vanlig at en del av de sammensatte ordene er irrelevante. De fleste irrelevante ordene vil imidlertid være så perifere i forhold til aktuelle problemstillingen (f.eks. BILLION i forhold til BIL), at de ikke influerer på søkeresultatet. Anvendes derimot mange trunkerte strenger, vil dette kunne medføre mye støy og følgelig dårlig presisjon. Dessuten vil ukorrekt trunkering kunne føre til søkingen blir svært ressurskrevende.

En annen svakhet ved trunkering er at den ikke fanger opp

ord som endrer roten ved bøyning, som f.eks. uregelmessige bøyninger som MOR - MØDRE, bøyninger med vokalsskifte som TANN - TENNER, konsonantfordoblinger som DAM - DAMMEN og stavelses-sammentrekninger som REGEL - REGLER. Det finnes imidlertid et fåtall av systemer hvor man kan nøye seg med å spesifisere en hvilken som helst del av søkeordet - ikke bare begynnelsen og slutten som ved trunkering. Søkeargumentet "NI*LS*EN" vil f.eks. omfatte alle ord som inneholder de tre bokstavkombinasjonene "NI", "LS" og "EN" i den angitte rekkefølgen som f.eks. NILSEN, NIELSEN, NILSSEN og NIELSSEN. Denne såkalte "maskefunksjonen" kan bidra til å løse en del av de "uregelmessige" bøyningene, men heller ikke den løser god nok de morfologiske problemene.

1.4 Forskning omkring lingvistiske hjelpemidler

Man kan stille seg noe undrende til at tekstsøkesystemer i dag ikke gjør mer bruk av lingvistisk kunnskap. Problematikken i tekstsøking er egentlig et "kommunikasjonsproblem" - eller et språkproblem - ettersom man må gå veien om (skrift)språket for å nå fram til den ønskede informasjon. Med mer kjennskap til sammenhengen mellom skriftspråket og det meningsinnholdet som det uttrykker og utnyttelse av denne informasjonen, ville tekstsøkesystemet lettere kunne analysere og kartlegge innholdet i dokumentene og søkeargumentet og følgelig forbedre kommunikasjonen med brukeren. Språket er regelstyrt, det er f.eks. klare regler som beskriver hvordan ord skal se ut, konstrueres, knyttes sammen til setningsledd og setninger, som man bør kunne utnytte i tekstsøkesystemer.

Det har pågått en del forskning opp gjennom årene som har vært rettet mot bruk av lingvistisk informasjon i tekstsøkesystemer. SMART-prosjektet tok opp spørsmålet alt i begynnelsen av 1960-årene og gjennomførte en del eksperimenter med bruk av syntaktiske analyser i tekstsøking (jfr. Salton 1983:103). Blant annet forsøkte man på denne måten å identifisere fraser og synonymer, samt å vektlegge ordene ut fra deres syntaktiske

funksjon i setningen. Forsøkene førte imidlertid ikke fram, og man valgte i stedet å konsentrere seg om enkle statistiske analyser. Det ble bl.a. gjennomført forsøk på automatisk å forbedre søkeargumentet gjennom å analysere et utvalg av de relevante dokumentene som var funnet ved et søk (jfr. Salton 1968 og 1983). Opplysninger om hvilke dokumenter som var relevante, ble innhentet fra brukeren. Framgangsmåten ga vellykkede resultater, og senere har flere forsøk vært gjennomført med såkalt "relevance feedback". I SMART-prosjektet ble det også utviklet en regelbasert metode for "stemming" - dvs. automatisk gjenkjenning av stammen til et ord. Metoden ble brukt til automatisk utvidelse av søkeargumentet (formulert i naturlig språk) med alle ord i dokumentbasen med samme stamme som søkeordene. Forsøk med metoden viste at dette hadde en positiv effekt på søkeresultatene (jfr. Salton 1983:103).

Det israelske RESPONSA-prosjektet representerer kanskje det mest imponerende forskningsarbeidet som har vært gjennomført omkring bruk av lingvistiske hjelpemidler i tekstsøking (jfr. Choueka 1980). Prosjektets formål er å gjøre den enorme responsalitteraturen tilgjengelig i søkbar form, og all forskning tar derfor sikte på å resultere i produkter som kan anvendes i RESPONSA-systemet. Tekstene er skrevet på hebraisk og arameisk, og det er særegenheter ved disse språkene som er lagt til grunn for den lingvistiske forskningen. Spesielt har de stått ovenfor store oppgaver ved løsning av de morfologiske problemene, og arbeidet har bl.a. resultert i det enorme lingvistiske verktøyet KEDMA som i dag tar hånd om de grammatikalske bøyingsformene til søkeordene. På hebraisk kan det f.eks. forekomme mange tusen bøyninger av samme substantiv, og ordene endrer ofte stamme ved bøyning.

I RESPONSA-prosjektet er det også gjennomført forsøk med automatisk gjenkjenning av synonymer. I likhet med SMART-prosjektet er det gjort forsøk med bruk av "feedback"-teknikker, men i stedet for å innhente informasjon om hvilke dokumenter som er relevante, analyserer systemet alle de funne dokumentene (kalt "local metrical feedback", jfr Attar/Fraenkel 1980). Analysen

går ut på å finne fram til ord som forekommer hyppig og i nærheten av søkeordene, og disse ordene blir lagt fram for brukeren som et forslag til synonymer. Forskningsresultatene har vist at dette er en nyttig framgangsmåte, men metoden er ikke i operativ bruk.

CONDOR-prosjektet var et stort, tysk prosjekt med ambisjoner om å utvikle et avansert tekstsøkesystem basert på bruk av lingvistikk i en rekke ulike forbindelser (jfr. Banerjee 1977). Ved prosjektets nedleggelse i 1980 var hele 10 av de 50 prosjektansatte lingvister. De arbeidet bl.a. med å utvikle av en rutine for syntaktisk og morfologisk analyse av en tekst, som både skulle anvendes på dokumentene og på søkeargumentet. Søkeargumentet var selvfølgelig formulert i naturlig språk. Likeledes arbeidet de med automatiske rutiner for generering av synonymtesauruser, etablering av støyordlister, splitting av sammensatte ord osv. Prosjektet ble nedlagt før det ble avsluttet, og så vidt oss bekjent er ingen av resultatene kommet andre søkesystemer til gode. Det er også usikkert hvilken effekt de lingvistiske hjelpemidlene hadde på søkeresultatet.

SPIRIT er et fransk system som gjør bruk av lingvistisk informasjon bl.a. til å rette skrivefeil, identifisere homografer, finne stammen til et ord og identifisere visse typer fraser (jfr. Fluhr 1981). Systemet er ferdig utviklet og er planlagt tatt i bruk en rekke steder.

Det har også pågått en del forskning omkring bruk av lingvistiske hjelpemidler i systemene STAIRS/TLS (IBM) og MISTRAEL (Honeywell Bull), men vi er verken kjent med hvilke forskningsresultater som er oppnådd eller kjenner til eksempler på operativ bruk av de lingvistiske hjelpemidlene. Felles for de fleste forskningsprosjekter på dette området er at de har utviklet metoder som senere har vist seg for ressurskrevende til å bli satt i verk. Dessuten er det nesten total mangel på eksperimenter som kan vise i hvilken grad effektiviteten øker med disse hjelpemidlene. SMART-prosjektet er det eneste prosjektet vi kjenner til som har

drevet forsøk av denne typen.

1.5 Behovet for en metode til løsning av de morfologiske problemene på norsk

Man kan konkludere med at det er gjort svært lite for å løse synonymproblemet i tekstsøking - i alle fall har ikke forskningen bidratt med effektive og økonomisk forsvarlige løsninger. Selv de mest "elementære" synonymproblemer som f.eks. grammatikalske bøyninger og avledninger, er ikke løst i de fleste tekstsøke-systemene. Det eneste hjelpemiddelet som tilbys er trunkering, og selv om dette er et effektivt og billig redskap, er det ikke godt nok til dette formålet.

Ved IRI pågår det flere aktiviteter som tar sikte på å nå fram til metoder som kan lette synonymproblemet i tekstsøking. I denne rapporten skal vi redgjøre for arbeidet med å utvikle en automatisk metode som kan løse de problemene som knytter seg til at et ord har ulike bøyings- og avledningsformer. En slik metode er nyttig på en rekke ulike områder i tekstsøking, f.eks. utvikling av støyordlister (ord som ikke er egnet som søkeord), identifisering av fraser (f.eks. nominalfraser), vektlegging av ord, splitting av sammensatte ord, identifisering av et ords oppslagsform og ordklasse m.m. Vi kjenner ikke til noen lignende metode på norsk - eller i Skandinavia for øvrig. Et unntak gjelder finsk, men ettersom finsk er så forskjellig fra norsk, er det ikke mye som er overførbart. Det samme gjelder også fra andre utenlandske språk, som f.eks. engelsk (det amerikanske tekstsøkesystemet LEXIS) eller hebraisk (RESPONSA-systemet). Metoden må utvikles på bakgrunn av morfologiske studier, og det er derfor begrenset hvor mye man oppnår ved å studere utenlandske metoder. Enkelte språk har en svært kompleks morfologi som det er vanskelig - og nærmest umulig - å systematisere med henblikk på automatisering (f.eks. hebraisk). Andre språk, som f.eks. engelsk, er langt lettere, og det samme gjelder vel også norsk - selv om problemene her antagelig er større enn på engelsk.

Mange stiller seg noe undrende til hvorfor vi i det hele tatt er opptatt av denne problematikken ettersom de morfologiske problemene i tekstsøkesystemet kan løses ved hjelp av en ordbok med de nødvendige lingvistiske opplysninger. Dette er selvfølgelig en mulig framgangsmåte, og som enkelte tekstsøkesystemer også har valgt. Løsningen er imidlertid svært kostbar. Den forutsetter at ordboka etableres manuelt og - kanskje det viktigste - vedlikeholdes manuelt. Etterhvert som dokumentbasen oppdateres, vil stadig nye ord dukke opp, og det vil aldri ta slutt. Skal man derfor kunne oppnå et 100% tilfredstillende resultat med en ordbok (som det argumenteres for), er det en forutsetning at ordboka oppdateres parallelt med dokumentbasen. Dette er bakgrunnen for at vi fant det mer interessant å undersøke i hvilken grad det er mulig å utvikle en effektiv, regelbasert metode som ikke krever manuelt vedlikehold.

2. METODE FOR AUTOMATISK ROTLEMMATISERING

2.1 Regelbasert metode

Målsetningen for vårt arbeid var å konstruere en regelbasert metode som sørger for at tekstsøkesystemet kan oppfatte ulike bøyninger og avledninger til en rot som like. At metoden skal være regelbasert, betyr at den skal være styrt av et endelig antall regler som er uavhengig av dokumentsamlingen. Det skal med andre ord ikke være nødvendig å endre reglene ved oppdatering eller skifte av dokumentbasen. Dette betyr ikke at reglene ikke kan være avhengig av fagområde eller språkform (riksmål, bokmål eller nynorsk), men at det er mulig å etablere et endelig sett med regler som dekker alle fagområder og norske språkformer. Ved etablering av en dokumentbase kan man så velge ut de regler som er aktuelle for vedkommende dokumentsamling. Det eneste nødvendige tilfellet av oppdatering av regelsettet vil skje ved endringer av språket over tid, f.eks. ved at nye ord med en annen struktur kommer til - eller at det lages nye avledninger som får en innsnevret eller annen betydning enn det ordet det er avledet fra.

I dette prosjektet har arbeidet bestått i å bestemme hvilken informasjon som skal med i reglene, å utvikle et sett med generelle regler for bokmål og konstruere en metode (algoritme) som kan styre reglene. Eksperimentmaterialet besto av tekster fra fagområdene jus, historie, geografi og fysikk, og noen regler gjelder spesielt for disse områdene.

2.2 Begrepet "automatisk rotlemmatisering"

Metoden skulle sammenføre både bøynings- og avledningsformer til samme rot. I enkelte tekstsøkesystemer (f.eks. LEXIS) har man valgt å begrense den bare til bøyningsformene til samme grunnform, men vi ser ingen grunn til å utelukke avledningsformene. I en situasjon hvor man søker etter informasjon som belyser et gitt

problem, er det sjelden at formen på ordet spiller noen rolle. Er man f.eks. interessert i dokumenter som angår beskatning ved salg av aksjer, er både ordet SALG, bøyningsformen SALGET og avledningsformen SELGEREN aktuelle søkeord.

I denne rapporten vil vi skille mellom begrepene rot, stamme og grunnform på følgende måte: Det semantiske morfemet i et graford kaller vi en rot (f.eks. ARV, LES og HØY). En rot plus et avledningsmorfem eller to røtter i sammensetning eller en større kombinasjon av avledninger og sammensetninger, kaller vi en stamme, f.eks. ARVING og FARSARV (jfr. Borgstrøm 1975:74). En stamme blir altså "det man kan henge bøyningsendelsen på". Grunnformen brukes om de formene vi finner i en ordbok, dvs. infinitivsformen av et verb (f.eks. ARVE), ubestemt entallsformen av et substantiv (f.eks. ARV), positivformen av et adjektiv (f.eks. GAMMEL). Med andre ord,

- ARV er et eksempel på en rot, en stamme og en grunnform,
- ARVE er et eksempel på en stamme og en grunnform,
- ARVING er et eksempel på en stamme.

Uttrykket "rotlemmatisering" er nettopp brukt for å markere at metoden omfatter alle ord med samme rot - forutsatt at endelsen ikke endrer betydningen av ordet. Begrepet "lemma" brukes om en samling av bøyningsformer innen en ordklasse. Sture Allen innførte dette begrepet i begynnelsen av 1970-tallet, og han definerte et lemma som (jfr. Allen 1971 bind 2, side XVIII):

"en grupp ordformer inom en ordklass vilka kan hänföras till antingen en och samma flexionsserie eller flera i tal och/eller skrift konvergerande serier vars divergenser visar rent fakultativ variation."

Mens lemmatisering betyr å samle ord innenfor samme lemma, vil rotlemmatisering bety å samle alle lemmaene som kan føres tilbake til samme rot uten at betydningen blir endret. Et rotlemma er med andre ord en gruppe med lemma som har samme betydning når

man ser bor fra den informasjon som ligger i selve bøynings- og avledningsendelsene. Et eksempel på ordformer som tilhører samme rotlemma er:

(ET) KJØP	(EN) KJØPER	(Å) KJØPE	KJØPENDE
KJØPET	KJØPEREN	KJØP	
KJØPENE	KJØPERE	KJØPER	
	KJØPERNE	KJØPTE	
KJØPS		KJØPT	
KJØPETS	KJØPERS		
KJØPENES	KJØPERENS	KJØPES	
	KJØPERES	KJØPTES	
	KJØPERNES		

Vi ser at det er ordformer fra tre forskjellige paradigmer (dvs. bøyningsrekker) med grunnformene KJØP (substantiv), KJØPE (verb), KJØPER (substantiv).

Verbet KJØPE har den samme betydningen som substantivet KJØP - endelsen -E viser bare at det er et verb, dvs. at det dreier seg om en handling. Det samme forholdet har vi med substantivet KJØPER, hvor endelsen -ER bare betegner at det er en person som utfører denne handlingen. Bortsett fra den informasjonen som ligger i selve avledningsendelsene, har ikke disse ordformene noen sær betydning, alt handler om "det å kjøpe noe". De tilhører derfor samme rotlemma. Det er naturlig å gi rotlemmaet oppslagsformen KJØP, siden dette er felles for alle bøyningene og samtidig entydig for rotlemmaet.

En del ord får innsnevret eller forandret betydning når de får lagt til en avledningsendelse. Vi sier at de blir leksikalisert. Eksempler på dette er OPPDRAG og OPPDRAGELSE, og KOMMUNE og KOMMUNIST. Leksikaliserte ordformer vil tilhøre et nytt rotlemma.

2.3 Alternative framgangsmåter

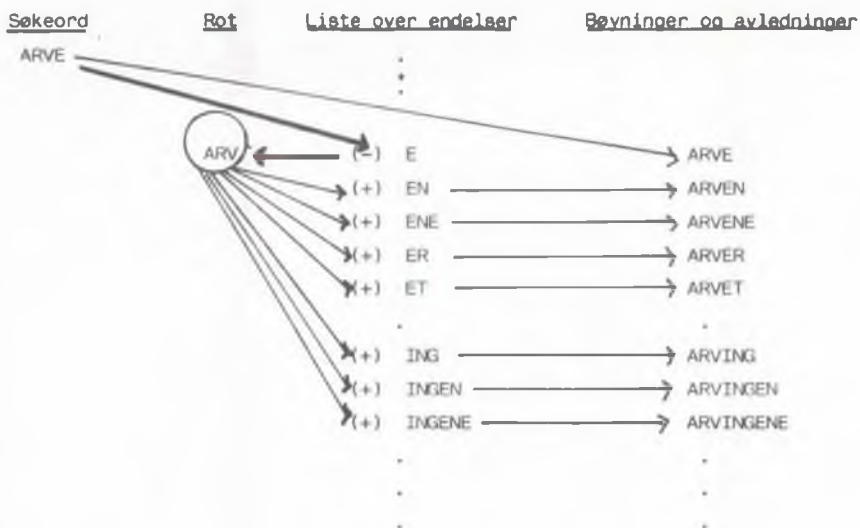
Det er ulike måter å gjennomføre automatisk rotlemmatisering på i et tekstsøkesystem. Siden vår metode skulle anvendes i tekst-søking, var det viktig å finne en løsning som både var effektiv og ikke for ressurskrevende. Nedenfor er det skissert tre prinsipielt forskjellige framgangsmåter; man kan

- (A) finne roten til ordet og konstruere alle bøyning- og avledningsendelser til denne roten ut fra en liste med mulige suffikser (endelser),
- (B) finne roten til ordet og søke i den alfabetiske ordboka til tekstsøkesystemet etter alle ord med denne roten etterfulgt av et gyldig suffiks,
- (C) konstruere en oppslagsform (gjerne roten) til ordet som er felles for alle bøyninger og avledninger av samme ord. Når man deretter slår opp på denne oppslagsformen i en "rot"-ordbok, finner man alle ord med denne oppslagsformen.

Den første framgangsmåten er illustrert i figuren nedenfor. Man konstruerer en liste med mulige suffikser ut fra morfologiske regler, og benytter denne lista til å konstruere mulige bøyninger av et ord. Til ordet ARV får vi f.eks. konstruert ARVEN, ARVENE, ARVER osv.

Begrepene "suffiks" og "endelse" brukes her synonymt, dvs. de viser både til bøynings- og avledningsmorfemer som står etter roten.

Fig. 2.3/1 Eksempel på konstruksjon av bøynings- og avledningsformene til en rot (metode A).



Fordelen ved denne metoden er at den ikke forutsetter at det lagres informasjon i tekstsøkesystemet om hvilke ord som tilhører samme rotlemma. Systemet konstruerer og supplerer søkeargumentet med alle mulige bøyningsformer uten å måtte gå veien om ordboka i svstemet. Metoden er kun basert på regler, men det vil antagelig være nødvendig med to regelsett; ett som kan analysere ordet og finne roten til ordet, og ett som kan anvendes til å konstruere mulige bøyninger av denne roten.

Metoden har ikke vært testet på norsk, og det er vanskelig å si i hvilken grad den ville ha lyktes. Vanskeligheter vil oppstå både ved etableringen av roten til ordet og ved konstruksjon av bøyning- og avledningsformene. Det kan ofte herske tvil om hva som er roten til ordet, og i mange tilfeller vil roten "endres" ved bøyning av ordet (f.eks. REGEL - REGLER, PRODUSERE - PRODUKSJON, FINGER - FINGRE, MOR - MØDRE og BRENNE - BRANT). Det

er også mange ulike typer suffikser, og det kan være vanskelig å avgjøre automatisk hvilke av dem som er aktuelle for det enkelte ord. Metoden vil derfor sannsynligvis resultere i alt for mange ord. Dette vil ikke gå ut over søkeeffektiviteten hvis ordene ikke eksisterer, men det kan ha en negativ effekt på søke- resultatet hvis det genereres ord som har en helt annen betydning enn søkeordet (f.eks. KOMMUNE - KOMMUNIST). Alt for mange søkeord vil dessuten kreve ekstra oppslag i ordboka til tekst- søkesystemet og følgelig økte kostnadene ved søking.

Metode (B) er illustrert i figur 2.3/2. Den har delvis vært testet ved IRI. Det ble utviklet en liste med suffikser som både ble brukt til å finne roten til ordet og til å undersøke om "differansen" mellom roten og ordet i ordboka var en gyldig suffiks. Alle ord som var innledet med denne roten ble sjekket. Dette inkluderte også ord som ikke hadde felles rot med søkeordet og sammensatte ord, f.eks. som for søkeordet ARV nedenfor:

ARV
 ARVE
 AREANDEL
 ARVEATTEST
 ARVEATTESTEN
 ARVEATTESTER
 AVEAVKALL
 ARVEBELØPENE
 ARVEDE
 *
 ARVEN
 *
 ARVID
 ARVING
 *
 *

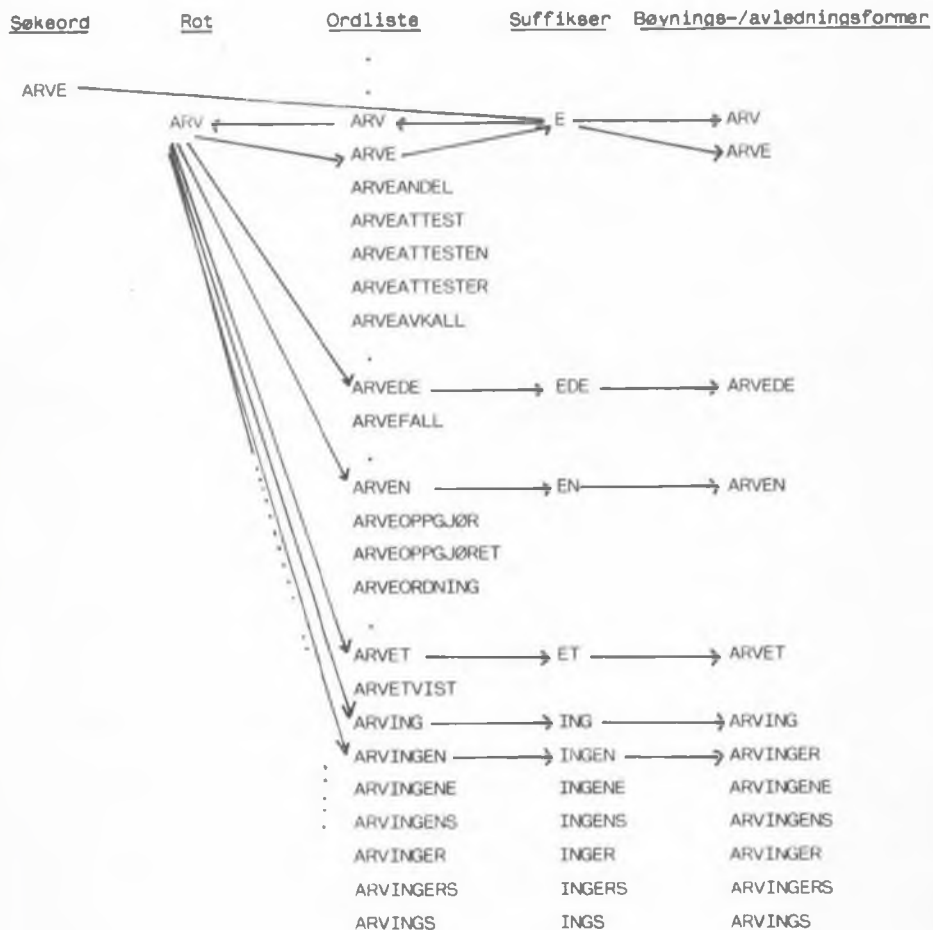
Det oppsto imidlertid problemer i de tilfeller hvor bøynings- og avledningss morfemet grep inn i roten til ordet, f.eks. ved

stavelsessammetrekninger (REGEL/REGLER), uregelmessige substantiv (MOR/MØDRE) og konsonantforenklinger (SYKKEL/SYKLE). I disse tilfellene måtte man sjekke forekomster av begge røttene til ordet (f.eks. REGL og REGEL). Spesielt vanskelig var det hvis den grunnformen som er lik roten til ordet, ikke forekom i den alfabetiske lista og søkeordet var bøydd.

I arbeidet med å utvikle en slik metode ved IRI, vokste metoden etterhvert til å bli så komplisert og ressurskrevende, at vi valgte å se oss om etter andre framgangsmåter.

Metoden likner imidlertid på den Steffan Helberg fra Språkdata (Gøteborg universitet) presenterer i artikkelen "Automatisk lemmatisering - En modell för upprättande av böjningsserier i ett frekvenslexikon" (jfr. Helberg 1971). Der blir problemet med stavelsessammetrekninger og vokalvekslinger løst ved at de lager "egne paradigmer" som tar seg av disse uregelmessighetene. I dette prosjektet er det imidlertid automatisk lemmatisering som er målet, mens vår oppgave var å gruppere ordformer med felles rot.

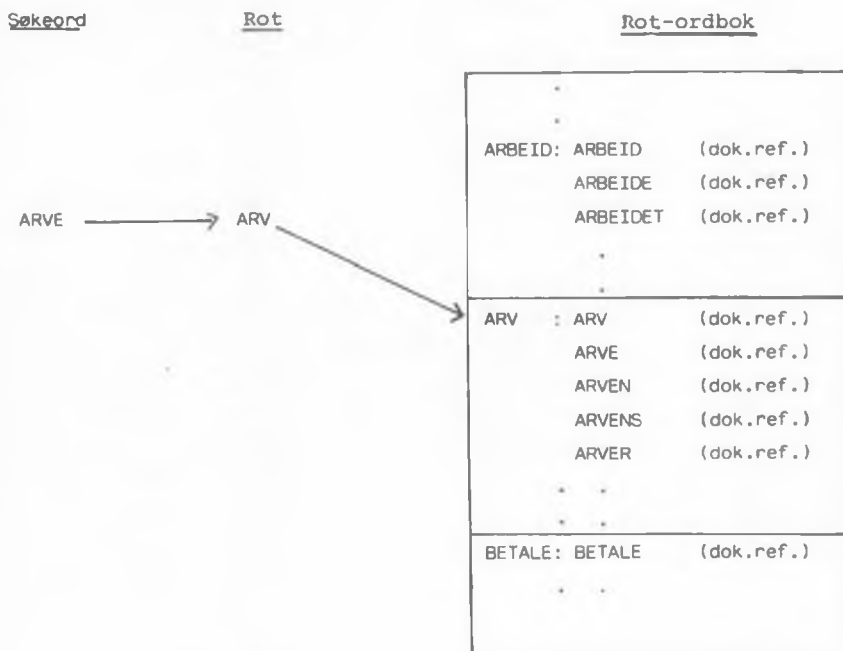
Fig. 2.3/2 Søke i ordboka til tekstsøkesystem etter bøynings- og avledningsformer ut fra en liste med suffikser (metode B).



Metode (C) forutsetter at det etableres lenker mellom ord som tilhører samme rotlemma. Dette skjer ved at hvert ord blir transformert til en oppslagsform som er identisk med oppslagsformen til de andre ordene innenfor dette rotlemmaet. Denne prosessen skjer ved etablering av basen og belaster derfor ikke brukeren ved søking.

Relasjonene mellom ord med samme rotlemma kan lagres på ulike måter. Man kan f.eks. etablere en "rot"-bok hvor hver oppslagsform representerer et rotlemma. Til hver oppslagsform kan man knytte opplysninger om hvilke ord som forekommer innen samme rotlemma og i hvilke dokumenter de forekommer. Ved søking behandles også søkeordene av den samme prosessen og får tildelt en oppslagsform som brukes som inngang i "rot"-ordboka (jfr. fig. 2.3/3).

Fig. 2.3/3 Eksempel på en "rot"-ordbok



Fordelen ved denne modellen er at systemet bevarer informasjon om ordenes opprinnelige form. Ønsker man å søke på et spesielt ord, går man inn i "rot"-ordboka på vanlig måte, men henter kun fram referansene til dette spesielle søkeordet.

For å spare lagringsplass, kan man fjerne informasjonen om hvilke ord som tilhører det enkelte rotlemma. Man kan da ikke søke på den ordformen som står i teksten, f.eks. søkefrasen "ULVEN, LENE" vil også gi tilslag på dokumenter hvor " ...ULVENE LENTE seg framover". Tekstsøkesystemet LEXIS har valgt denne løsningen, men de kritiseres også for det. Dette er likevel en interessant løsning i tilfeller hvor lagringsplassen er begrenset, f.eks. på personlige datamaskiner slik de er i dag.

En alternativ framgangsmåte kunne være å beholde grafordene som oppslagsord (dvs. i den form de forekommer i teksten) og i stedet knytte lenker (pekere) mellom ord innenfor samme rotlemma. På den måten sparer man søketid i og med at søkeordene ikke blir behandlet og transformert til en oppslagsform. Man støter imidlertid på et problem i de tilfeller hvor søkeordet har en form som ikke forekommer i dokumentbasen, men det finnes ulike løsninger også på dette problemet. Man kan f.eks. gå veien om oppslagsformen til ordet som ligger lagret i ordboka som et "hode" til kjeden (pekerstrukturen) (jfr. fig. 2.3/4).

Fig. 2.3/4 Eksempel på en kjedet fil hvor alle ord innenfor samme lemma danner kjeder

<u>Søkeord</u>	<u>Oppslagsform (rot)</u>	<u>Kjedet ordbok</u>	
		Nr.	PEKER
ARVE →	ARV →	50	ARV (dok.ref.) 51
		51	ARVE (dok.ref.) 75
		52	ARVEANDEL (dok.ref.)
		53	ARVEATTEST (dok.ref.) 54
		54	ARVEATTESTEN (dok.ref.) 0
		75	ARVEDE (dok.ref.) 90
		76	ARVEFALL (dok.ref.) 0
		90	ARVEN (dok.ref.) 110
		91	ARVEOPPGJØR (dok.ref.)

2.4 Skisse av en regelbasert metode for automatisk rotlemmatisering

Uten å ta standpunkt til hvordan rotlemmatiseringen bør implementeres i tekstsøkesystemet, valgte vi gå fram etter modell (C) ovenfor. Det ble utviklet en metode hvor et graford blir gitt som inndata og hvor oppslagsformen til ordet kommer ut som resultat. F.eks. vil ordet ARVING få oppslagsformen ARV.

Det var ikke inkludert i målsetningen å finne fram til de ulike ordenes rot. Det viktigste var å få konstruert en oppslagsform som var lik for alle ord innen samme rotlemma og bare disse. Dette har resultert i at oppslagsformen ofte kan virke knapp og noe merkelig, f.eks. ordet UNIVERSITET får oppslagsformen "UNIVERS", mens ordet UNIVERS får oppslagsformen "UNIV". Brukeren vil imidlertid ikke bli konfrontert med oppslagsformen ved søking, og det spiller derfor ingen rolle hvordan oppslagsformen ser ut - bare den er entydig.

Hvis vi hadde valget mellom alternative oppslagsformer, valgte vi selvfølgelig den som liknet mest på ordets rot. Oppslagsformen til et ord blir laget ved å aktivisere regler som passer for ordets slutt. Enkelte regler kan betraktes som hovedregler og gir beskjed om at et bestemt suffiks skal fjernes, f.eks. at bøyningsendelsen -ENE skal fjernes. Andre regler er unntak til hovedreglene fordi man ikke oppnår en entydig oppslagsform bare ved å slette endelsen til ordet, f.eks. ordet REGELEN skal ha samme oppslagsform som REGLER, ordet BRENNE som BRANN og ordet PRODUKSJONEN som PRODUSERT. Noen ganger utgjør unntakene hele ord, mens andre ganger bare bokstavkombinasjoner enten alene eller i forbindelse med en suffiks. Det var en målsetning i seg selv å holde antallet med hele ord lavt, hvis de ikke tilhører til de lukkede ordklassene. Dette gjøres fordi vi ønsker å tilfredstille kravet om at regelsettet skal være generelt og endelig.

Selve behandlingen av et ord skjer ved at ordet blir snudd og

sammenlignet med tegnstrengen i reglene som også er snudd. For at en regel skal aktiviseres, må ofte visse betingelser være oppfylt. Betingelsene kan være av flere typer, f.eks. at regelen gjelder bare for hele ordet eller for ord som bare inneholder et suffiks - eller for ord med et bestemt prefiks. Reglene ble dannet på bakgrunn av en omfattende analyse av eksperiment-materialet og morfologien i norsk.

3. MORFOLOGIEN I NORSK

3.1 Innledning

Det er vanlig å dele ordforrådet i norsk inn i 10 ordklasser.

Disse er:

- verb
- substantiv
- adjektiv
- pronomer
- adverb
- artikler
- konjunksjoner
- preposisjoner
- tallord
- interjeksjoner

I noen av ordklassene opptrer ordene bare i en form, mens det i andre ordklasser finnes forskjellige bøyningsformer av det samme ordet. Noen ordklasser inneholder også ord som er avledet av ord fra andre ordklasser. Bøynings- og avledningsformene dannes som oftest ved bøynings- og avledningsmorfemer som legges til roten av ordet.

Vi skal se nærmere på de forskjellige bøynings- og avledningsformene som finnes i norsk.

3.2 Bøyningsmuligheter.

Følgende ordklasser har bøyningsmuligheter i norsk:

- substantivene
- verbene
- adjektivene
- noen av adverbene
- pronomene
- artiklene

De kategoriene disse ordklassene uttrykker ved bøyning, kan vi lese av skjemaet:

Tabell 3.2/1 Ordklassenes bøyningmuligheter i norsk

	! genus !	! numerus !	! bestemthet !	! kasus !	! tempus !	! modus !	! diatese !	! grad !
!subst. !	x	x	x	x				
!verb !					x	x	x	
!adj. !	x	x	x					x
!adv. !				x*)				x
!pron. !	x	x		x				
!art. !	x	x	(x)					

*) tilsteds/påstedsbøyningen av adverbene kan regnes som en kasusbøyning.

Disse bøyningmulighetene realiseres på forskjellige måter i skrift. Den vanligste måten er at en bøyningsendelse legges til roten. En slik bøyningsendelse vil i noen tilfeller bare uttrykke en kategori av gangen, f.eks. -S som uttrykker genitiv i substantiv, eller -ERE som uttrykker komparativ av adjektiv og adverb. Men de fleste endelser uttrykker flere kategorier på en gang, f.eks. vil -EN uttrykke både hankjønn, bestemt form og entall for et substantiv og -ER vil uttrykke presens og aktiv for et verb.

I norsk har vi også indre bøyning, dvs. de overnevnte kategoriene uttrykkes ved at fonemene forandrer seg. Det vanligste er vokalveksling som forekommer både alene, (f.eks. SYNGE - SANG) og i forbindelse med en endelse, (f.eks. TÅ - TÆR). Vi finner også eksempler på at vi får lagt til en ekstra konsonant i skrift når kvaliteten på vokalen endres, f.eks. SKYTE - SKJØT.

Bøyningen kan også føre til at kvantiteten til vokalen forandres. Enten uttrykkes bøyningen bare ved kvantitetsforskjellen, f.eks. SØT - SØTT, MØTE - MØTTE, eller forekommer kvantitetsforskjellen i forbindelse med en endelse, f.eks. BLÅ - BLÅTT, GÅ - GÅTT (det siste gjelder spesielt for bøyningsendelsen -T). Når kvantiteten på vokalen endres fra lang til kort, vil dette vise seg i skrift ved at den påfølgende konsonanten blir fordoblet.

Til slutt er det noen ord som forandrer roten fullstendig, (såkalt suppletiv bøyning) f.eks. GOD - BEDRE, LITEN - SMÅ, og til denne typen kan kanskje også kasusbøyningen av pronomen regnes, f.eks. JEG - MEG, og påsted/tilstedsbøyning av adverb, f.eks. HER - HIT.

3.3 Avledningsmuligheter

De fleste avledningene innebærer at ordene skifter ordklasse, men det finnes også avledninger innenfor samme ordklasse. De vanligste overgangene ved avledninger er:

Tabell 3.3/1 Ordklasseoverganger ved avledninger

!FRA/TIL !	subst. !	verb !	adjektiv !	adverb !	tallord !	
!-----!	!-----!	!-----!	!-----!	!-----!	!-----!	
!subst. !	x	x	x	-	-	!
!verb !	x	-	x	(x)	-	!
!adjektiv!	x	x	-	x	-	!
!adverb !	-	-	-	-	-	!
!tallord !	-	-	x	-	-	!

EKSEMPEL

Fra substantiv til substantiv:	ET ARBEID	- EN ARBEIDER
	ET FORBRUK	- EN FORBRUKER
	verb: ET ARBEID	- Å ARBEIDE
	EN MARSJ	- Å MARSJERE
adjektiv:	EN STEIN	- STEINETE
	ET HULL	- HULLETE
Fra verb til	substantiv: Å KASTE	- ET KAST
	Å VURDERE	- EN VURDERING
	adjektiv: Å SKRAVLE	- SKRAVLETE
	Å SPISE	- SPISELIG
Fra adjektiv til	substantiv: KJÆRLIG	- EN KJÆRLIGHET
	REDD	- EN REDSEL
	verb: VARM	- Å VARME
	ÅPEN	- Å ÅPNE
adverb:	PEN	- PENT
Fra tallord til	adjektiv: FEM	- FEMTE
	SEKSTEN	- SEKSTENDE

De fleste avledningene dannes ved at en avledningsendelse legges til, eventuelt at en endelse (-E) fjernes. Men det finnes også eksempler på avledninger som dannes ved vokalvekslinger, f.eks. Å FINNE - ET FUNN, Å BRENNE - EN BRANN. Denne avledningstypen finner vi blant de ordene som også har vokalvekslinger i bøyningen, dvs. særlig blant de sterke verbene.

3.4 Sammenfall i endelser

Ved å gjennomgå bøyings- og avledningsendelsene i ordene som har bøyning og avledningmuligheter på norsk, finner man at det er stort sammenfall i måten kategoriene realiseres på (hvilke bokstav-kombinasjoner som forekommer). Suffiksen -ER uttrykker

f.eks. både presens av verb og ubestemt flertall av substantiv, samtidig som det er en avledningsendelse (nomen agentis). Suffiksen -T uttrykker bestemt form av intetkjønnsord for substantivene, perfektum for visse verb og danner adverb av adjektiv.

3.5 Åpne og lukkede ordklasser

Når vi sammenligner ordklassene som har bøyningssmuligheter med de ordklassene som kan avledes av andre, ser vi at substantivene, verbene, adjektivene og adverbene har begge mulighetene. Disse ordklassene hører med blant de åpne ordklassene - eller innholdsordene som de også kalles. Det er disse ordklassene som får nye ord, og det er dem som i første rekke har en mening (i motsetning til funksjonsordene som har som oppgave å angi relasjonene mellom innholdsordene eller referere til innholdsordene). Både pronomene, artiklene og tallordene regnes med blant funksjonsordene. Disse ordklassene er endelige klasser, det vil si at de kan regnes opp og at det ikke dannes nye ord. De kalles derfor lukkede ordklasser.

Til de lukkede ordklassene hører også preposisjoner og konjunksjoner, men disse har verken bøyings- eller avledningsmuligheter og vil altså bare finnes i en form. Interjeksjonene er nok en åpen ordklasse i den forstand at det lages nye ord. De forekommer imidlertid bare i en form og er lite brukt i skriftspråket. Vi ser bort fra denne ordklassen i den videre behandlingen.

3.6 Analyse av ordklassene

(1) Innledning

Kunnskapen om det morfologiske systemet i norsk ønsket vi å få formalisert i et sett med regler som hadde til hensikt å føre alle ordformene tilbake til en form. I de fleste tilfellene ville

det si å fjerne en bøynings- eller avledningsendelse, men i mange tilfeller måtte også andre forandringer utføres med ordformen.

Vi analyserte alle ordklassene som hadde bøynings- og avledningsmuligheter, for å finne fram til hvilke variasjoner vi fant innen de forskjellige ordklassene, og vi satt opp et forslag til regler som skulle dekke disse variasjonene.

(2) Pronomen, artikler og tallord

Både pronomene og artiklene har forholdsvis uregelmessig bøyning, men siden de tilhører de lukkede ordklassene og ordklassene dessuten er små, kan unntakene regnes opp. Disse vil derfor ikke bli drøftet disse nærmere. Det samme gjelder avledningsmulighetene til tallordene. De danner ordenstall, dvs. adjektiv, ved noen få enkle endelser, og skaper ingen problemer når det gjelder å formulere regler som fjerner endelsene.

(3) Adverb

Adverbene er en sammensatt gruppe. En del av dem kunne regnes opp og burde egentlig høre med blant de lukkede ordklassene. Det gjelder både de modale adverbene (JO, VEL, NOK, osv.) som er ubøyelige, og stedsadverbene som har påsteds- og tilstedsbøyning (HER, DER og INNE uttrykker betydningen "på stedet", mens HIT, DIT og INN uttrykker "til stedet"). I de tilfellene hvor bøyningen er uregelmessig, har vi formulert regler for unntakene, men de fleste stedsadverbene er regelmessige, dvs. de legger til -E i ordene med påstedsbøyning, og denne bøyningen er lett å fange opp. Adverbene kan bare dannes av adjektiv, og dette ved å legge til -T, noe som heller ikke skaper problemer for regelformuleringen.

3.7 Åpne ordklasser med bøyningsmuligheter

(1) Innledning

Det er altså i første rekke substantivene, verbene og adjektivene som må analyseres. En oversikt over de regelmessige bøynings-

endelsene som uttrykker de kategorier nevnt i avsnitt 3.2 ville bli slik:

Tabell 3.7/1 Oversikt over regelmessige bøyningsendelser for substantiv, verb og adjektiv.

	genus, numerus, bestemthet	kasus	tempus, modus, diatese	grad
!subst. !	-(E)N, -A, -(E)T	! -S	!	!
!	-(E)R, -E	!	!	!
!	-(E)NE	!	!	!
!		!	!	!
!verb		!	-(E), -(E)R, -(E)T	!
!		!	-A	!
!		!	-TE, -T	!
!		!	-DE, -D	!
!		!	-DEE, -DD	!
!		!	-(E)S, -TES	!
!adj. !	-T, -TT, -E	!		-(E)RE, !
!		!		-(E)ST(E)!

Hovedreglen er at bøyningsendelsene som framgår av skjemaet blir lagt til grunnformen. Når grunnformen er et "lett" tostavelseord, dvs. at det ender på trykksvak -E, vil en eventuell E- som innleder en bøyningsendelse, falle bort (derav parentesen i skjemaet). Men i forbindelse med disse endelsene vil vi også kunne få forandringer i rotens skriftbilde, og vi skal se nærmere på hvilke konsekvenser dette medfører i substantiv, verb og adjektiv.

(2) Substantiv

I forbindelse med substantivbøyningen finner vi:

- stavelsessammentrekning
- konsonantfordobling
- vokalveksling

Stavelsessammentrekning. De fleste substantivene som ender på -EL, -EN, -ER mister E'en i denne stavelsen i forbindelse med flertallsendelsene, f.eks. TEMPEL - TEMPLER. Hunnkjønnsordene mister også denne E'en i bestemt form entall, f.eks. HULDER - HULDRA.

Konsonantfordobling. Alle substantivene som ender på -M i ubestemt form entall, får M'en fordoblet foran en endelse når vokalen foran er kort, f.eks. DAM - DAMMEN.

Vokalvekslinger. Noen substantiv danner flertall ved vokalvekslinger. I enkelte tilfeller er dette den eneste måten flertall uttrykkes på (MANN - MENN), i andre tilfeller kombineres vokalvekslingen med en bøyningsendelse (KNE - KNÆR) eventuelt kombinert med stavelsessammentrekning og konsonantforenkling (DATTER - DØTRE).

(3) Verb

I forbindelse med bøyningsendelsene til verbene finner vi:

- konsonantforenklinger og -fordoblinger
- vokalvekslinger

Konsonantforenklinger og -fordoblinger. På samme måte som for substantivene, er M'en i utlyd alltid skrevet enkelt, selv om den er lang. Når det legges til en bøyningsendelse, realiseres den lange M'en som en dobbelt -M i skrift. Det fører til at imperativformen i slike tilfeller vil ende på enkelt -M, mens de andre formene har dobbelt -M. Verb som har en dobbeltkonsonant i roten, vil få denne forenklet foran endelser som begynner på konsonant, dvs. forenklingen kan forekomme i forbindelse med preteritums- og perfektumendelsene, f.eks. BEGYNNE - BEGYNTE - BEGYNT.

Vokalvekslinger. Vokalvekslingene forekommer hyppig i forbindelse med verbbøyning. De sterke verbene danner preteritum bare ved vokalveksling og perfektum ved vokalveksling pluss endelsen -ET.

(4) Adjektiv

I forbindelse med adjektivbøyningen kan vi finne følgende forandringer av roten:

- stavelsessammentrekning
- konsonantforenkling og -fordobling
- vokalveksling
- suppletisme

Stavelsessammentrekning. Vi finner eksempler på stavelsessammentrekninger i forbindelse med adjektivbøyningen, både ved gradbøyningen, f.eks. DIGER -DIGRERE - DIGREST, og ved bestemthets- og flertallsendelsen -E, f.eks. DIGER - DIGRE.

Konsonantforenkling og -fordobling. Adjektiv som ender på lang - M, vil fordoble denne foran gradbøyningsendelsene -ERE, -EST, f.eks. DUM - DUMMERE - DUMMEST. Adjektiv som ender på dobbeltkonsonant, vil få denne forenklet foran nøytrumsendelsen (-T), f.eks. SUNN - SUNT, GRØNN - GRØNT. Dette skjer gjerne i forbindelse med konsonantforenklinger, f.eks. GAMMEL - GAMLE.

Vokalveksling. Vokalveksling finner vi i forbindelse med gradbøyning av adjektiv, f.eks. STOR - STØRRE - STØRST og UNG - YNGRE - YNGST.

Suppletisme. I forbindelse med gradbøyninga finnes det eksempler på at ikke bare vokalen, men hele roten forandres, såkalt suppletiv bøyning, f.eks. GOD - BEDRE.

Leksikalisering. Gradbøyning av adjektiv kan føre til at betydningsen av ordformen endres.

3.8 Åpne ordklasser med avledningsmuligheter

(1) Innledning

Avledningsendelsene i norsk er tallrike. Gjennom et studie av 39 forskjellige avledningsendelser, fant vi at 31 var av en slik type at de ikke vil forandre betydningen av roten. Denne vurderingen er foretatt på grunnlag av de eksemplene som finnes i eksperimentmaterialet.

(2) Substantiv

Substantiv kan dannes ved å legge avledningsendelser til roten til ordet, f.eks. roten ARV pluss suffiksen -ING gir avledningsformen ARVING. Hvis avledningsendelsen begynner med en vokal, vil en eventuell -E i utlyd falle bort, f.eks. grunnformen HENDE pluss suffiksen ELSE gir ordet HENDELSE. Ved å gjennomgå de substantivene i materialet som var dannet ved avledning, undersøkte vi i hvilken grad avledningsendelsene hadde ført til betydningsendring i forhold til grunnformen. Eksempler på avledninger som ikke forandret betydningen er SYKLING, ARBEIDER, UTVIDELSE. Det var ikke til å unngå at vi enkelte ganger var i tvil, f.eks. ved BIL - BILIST, EKSPLOSIV - EKSPLOSJON og REPRESENTANT - REPRESENTATIV. En del ganger førte avledningsendelsen til at ordet ble leksikalisert (dvs. det fikk en helt spesiell betydning) f.eks. KOMMUNE - KOMMUNISME.

(3) Verb

Mange verb dannes ved at det legges en -E til et substantiv eller et adjektiv, slik at imperativformen av verbet blir lik roten (hvorvidt det er verb som dannes av substantivene og adjektivene eller omvendt, kan for så vidt diskuteres, men det er uinteressant i denne sammenheng). Det er også noen verb som dannes ved å legge endelsen -ERE til et substantiv, f.eks. MARSJ - MARSJERE.

(4) Adjektiv

Det er mange avledningsendelser som danner adjektiv, f.eks. -AKTIG, -BAR, -LIG, -IG, -ISK osv. Alle forekomstene med disse endelsene ble gjennomgått for å finne ut om de førte til leksikaliseringer.

4. ETABLERING AV REGELSETTET

4.1 Eksperimentmaterialet

Reglene som skulle danne grunnlaget for rotlematisering, ble testet og videreutviklet på grunnlag av et eksperimentmateriale sammensatt av tekster hentet fra grunnskolebøker i fysikk, historie, geografi og 2 større juridiske samlinger. Totalt utgjorde materialet ca. 30 000 ulike graford, men etter at alle skrivefeil (999 ord), steds- og egennavn (534 ord), forkortelser (339 stk.), ord med bindestrek (119 ord), tall, gamle skrivemåter (185 ord) og nynorske ord (211 ord) var fjernet, sto vi igjen med et eksperimentmateriale på 23890 ulike graford og ca. 1/2 mill. løpende ord.

Tabell 4.1/1 Oversikt over eksperimentmaterialet før fjerning av skrivefeil, stedsnavn, tall etc. (jfr. ovenfor).

TYPE TEKSTER	Ulike graford	Løpende graford
Grunnskolepensum i fysikk	6883	67919
" geografi	6789	42555
" historie	4101	19301
Lagmannsrettavgj. i familie-,		
skifte- og arverett	10026	188432
Tinglysingsavgjørelser	15057	122226

4.2 Oppslagsform

Reglenes oppgave var å gi alle ordformene innenfor samme rotlemma en identisk og entydig oppslagsform. På denne måten kunne vi samle ordene som tilhørte et rotlemma i en gruppe. Reglene satte

vi opp på grunnlag av den morfologiske analysen nevnt i kap. 3. I første omgang satte vi opp et sett med hovedregler som stort sett ga beskjed om at enten en bøyings- eller en avledningsendelse skulle strykes. I en del tilfeller måtte vi legge inn spesialregler som også forandret deler av roten for at ordformene skulle få samme oppslagsform. Både til hovedreglene og til spesialreglene kunne det forekomme unntak. Reglene var derfor nøye knyttet til hverandre, og forandring av en regel førte ofte til at også flere andre regler måtte endres.

Det var i hovedsak ordformens rot som ble oppslagsformen til ordet, men hvis roten endte på -E, ble -E'en fjernet. I slike tilfeller ble altså oppslagsformen mindre enn roten, f.eks. ble "JENT" oppslagsformen til ordformen JENTE. Vi fjernet altså konsekvent E'en i utlyd og vi gjorde dette for å forenkle reglesettet. På denne måten var det nok med et sett regler for de regelmessige substantivene og vi slapp å definere et ekstra sett som skulle dekke de "lette" tostavelssubstantivene (dvs. de som ender på -E og følgelig bare får bøyingsendelsene -N, -R, -NE osv.).

4.3 Eksempel på problemer ved rotlemmatisering

For å gi et innblikk i hvordan vi gikk fram ved etablering av metoden og - ikke minst - hvilke problemer vi sto ovenfor, har vi nedenfor gitt noen konkrete eksempler.

La oss se nærmere på bøyingsmønsteret til substantiv hannkjønn, f.eks.

GUTT	GUTTEN	GUTTER	GUTTENE
BIL	BILEN	BILER	BILENE

Her er det naturlig å sette oppslagsformen lik grunnformen "GUTT" og "BIL". Dette oppnås ved reglene:

- (1) fjern endelsen -EN
- (2) " " -ER
- (3) " " -ENE

Reglene er i samsvar med bøyningsendelsene for substantiv hankjønn, og gir riktig oppslagsform for størsteparten av hankjønnssubstantivene.

Det finnes allikevel mange unntak. De "lette" tostavellesubstantivene som f.eks. LEGE, BAMSE og HANSKE, ender på -E, og får følgelig bare endelsen -N, -R og -NE. Hvis reglene ovenfor gjelder, vil ikke ubestemt form entall bli gruppert sammen med de andre bøyningsformene av ordet. Den enkleste måten å løse dette problemet på, er å fjerne E'en også i ubestemt entall. Vi legger derfor til regelen

- (4) fjern endelsen -E

En del substantiv får imidlertid stavelsessammentrekning i flertall, f.eks. ord som slutter på -EL, slik som ordene REGEL, FØDSEL, EKSEMPEL og SIRKEL.

Eksempel:

(EN)	<u>REGEL</u>	<u>REGEL</u> <u>EN</u>	<u>REGL</u> <u>ER</u>	<u>REGL</u> <u>ENE</u>
"	<u>FØDSEL</u>	<u>FØDSEL</u> <u>EN</u>	<u>FØDSL</u> <u>ER</u>	<u>FØDSL</u> <u>ENE</u>

Ordene REGEL og REGELN får oppslagsformen "REGEL" etter våre regler, mens bøyningsformene REGLER og REGLENE får oppslagsformen "REGL". En måte å løse dette problemet på er å føre inn den "fortapte" -E'en i det vi stryker flertallsendelsen. Vi innfører derfor reglene:

- (5) fjern endelsen -LER og legg til -EL
- (6) " " -LENE " -EL

Ordet REGEL får nå oppslagsformen "REGEL" i alle bøyningsformene.

Regel (5) og (6) skaper imidlertid problemer for ord som IGLER, SKOLER, LOKALER, AVTALER, SAMTALER etc. Disse blir fanget opp av reglene (5) og (6) og får en ekstra E i flertall. F.eks.

AVTALE og AVTALEN	får oppslagsformen "AVTALE", mens
AVTALER og AVTALENE	"AVTAE"

Reglene (5) og (6) gir derfor alt for mange feil, og en annen løsning på problemet må til. Siden oppslagsformen i seg selv ikke er av betydning i denne sammenheng (bare gruppering blir riktig), kan vi i stedet sette oppslagsformen lik grunnformen i flertall, dvs. "REGL". Vi erstatter derfor (5) og (6) med reglene

(5)'	fjern endelsen -EL	og legg til -L
(6)'	" "	-ELEN " -L

Dette gir følgende resultat:

REGEL (5)'	REGELEN (6)'	REGLER (2)	REGLENE (3)	gir oppslagsformen "REGL"
AVTALE (4)	AVTALEN (1)	AVTALER (2)	AVTALENE (3)	" " "AVTAL"

hvor tallet til høyre for ordet angir hvilken regel som er brukt.

Et annet problem er ord med dobbeltkonsonant i en ordform og enkelkonsonant i en annen. Substantiv som ender på -EL forenkler konsonanten samtidig med at de mister en stavelse når flertallsendelsen legges til, f.eks. SYKKEL - SYKLER og TOMMEL - TOMLER. Med våre regler ville entalls- og flertallsformene ikke bli gruppert sammen, f.eks.

SYKKEL og SYKKELEN	får oppslagsformen "SYKKL", mens
SYKLER og SYKLENE	" " "SYKL"

For at disse ordene skal få felles oppslagsform, må enten en konsonant fjernes i entall eller en konsonant legges til i flertall. Det siste er vanskelig siden det finnes ord som slutter på -KLER og som ikke er av denne typen, f.eks. UTVIKLER.

Vi velger derfor det første alternativet og innfører reglene:

(7)	fjern endelsen	-KKEL	og legg til	-KL
(8)	"	"	-MMEL	"
(9)	"	"	-KKELEN	"
(10)	"	"	-MMELEN	"

SYKKEL og SYKKELEN vil med dette få oppslagsformen "SYKL" og bli gruppert sammen med SYKLER og SYKLENE.

Det er verd å legge merke til hvordan en ny regel kan trekke med seg flere nye regler. F.eks. vil regel (7) medføre at også regel (9) må legges inn osv. Likeledes vil ønsket om å få gruppert genetivsformen til substantivene sammen med de øvrige bøyningsformene (f.eks. GUTTENS, LEGER, SYKLERS, AVTALENS osv.) i prinsippet kreve en ny regel for hver av de 10 eksisterende reglene.

For å unngå denne eksplosive økningen i antall regler, kan vi la enkelte ord behandles av flere regler. F.eks. kan ordet GUTTENS først behandles av en regel som fjerner genetivendelsen (-S), og deretter ved en regel som fjerner endelsen i bestemt form entall (-EN). På den måten kan vi nøye oss med bare en regel for genetivsformene. Likeledes kan vi erstatte reglene (6)', (9)-(10) med en regel (6)'' som sier at ord som slutter på -ELEN skal fjerne -EN og behandles på nytt:

(6)'' fjern endelsen -ELEN, legg til -EL og behandle ordet på nytt

Genetivsregelen blir derfor regel nr. 9'' i dette tilfellet og den sier

(9)'' fjern endelsen -S, og behandle ordet på nytt

Ordet SYKKELENS blir behandlet av regel 9'' og får formen SYKKELEN som

blir	"	6''	"	SYKKEL som
blir	"	7	"	SYKL.

Vi står nå igjen med følgende regler:

- | | | | | |
|-----|----------------|-------|-------------|-------------------------------|
| (1) | fjern endelsen | -EN | | |
| (2) | " | -ER | | |
| (3) | " | -ENE | | |
| (4) | " | -E | | |
| (5) | " | -EL | og legg til | -L |
| (6) | " | -ELEN | " | -EL og behandle ordet på nytt |
| (7) | " | -KKEL | " | -KL |
| (8) | " | -MMEL | " | -ML |
| (9) | " | -S | | " |

som gir resultatet:

GUTT,	GUTTEN	(1)	GUTTER	(2) og	GUTTENE	(3)	oppslagsform	"GUTT"
LEGE	(4)	LEGEN	(1)	LEGER	(2) og	LEGENE	(3)	" "LEG"
REGEL	(5)	REGELEN	(6)"", (5)	REGLER	(2) og	REGLENE	(3)	" "REGL"
AVTALE	(4)	AVTALEN	(1)	AVTALER	(2) og	AVTALENE	(3)	" "AVTAL"
SYKKEL	(7)	SYKKELEN	(6)"", (7)	SYKLER	(2) og	SYKLENE	(3)	" "SYKL"

4.4 Beskrivelse av en regel

(1) Innledning

Regelsettet består av tilsammen 670 regler. Reglene undersøker de ulike grafordene for å finne ut om oppslagsformen bare skal være en kopi av grafordet eller om det må en forandring til, og de utfører den eventuelle forandringen.

En regel kan deles inn i en bokstavstreng, en betingelse, en ordre og en typebetegnelse.

(2) Bokstavstreng

Bokstavstrengen inneholder hovedsaklig et suffiks eller et fritt morfem (helt ord). I en del tilfeller har suffikset en eller

flere bokstaver foran, og vi har også noen bokstavstrenger som bare består av en eller flere bokstaver. Når rotlematiseringen utføres, er det de ulike reglenes bokstavstrenger som blir sammenliknet med grafordet. Eksempler på bokstavstrenger er -ENE, -SETT, -VD og -STR.

(3) Betingelse

Betingelsen angir regelens begrensninger. Enkelte regler vil f.eks. bare tre i kraft hvis bokstavstrengen er identisk med grafordets endelse (f.eks. -ENE, -ER o.l.), mens andre krever at identiteten gjelder hele grafordet (f.eks. HADDE og MEN).

Enkelte endelser kan være etterfulgt av andre endelser, mens andre kan det ikke. Noen regler stiller derfor som betingelse at ordet ikke har vært behandlet tidligere. Vi sier at regelen er lukket. Nødvendigheten av dette ser vi ved behandlingen av grafordet SETTE. Det får først tilslag på regelen for -E som fjerner -E'n og gir ordre om at ordet skal behandles på nytt. Oppslagsformen blir da SETT. Men vi har også en regel som sier at SETT skal fjerne -TT slik at rotlemmaet blir SE. For at grafordet SETTE ikke skal få tilslag på regelen for SETT, er det nødvendig å lukke SETT-regelen.

Noen regler forutsetter at det som følger foran bokstavstrengen er et prefiks. Vi har f.eks. en regel for det sterke verbet LA, hvor betingelsen er at hvis det står noen bokstaver foran -LA, må disse utgjøre en preposisjon. På denne måten vil grafordet PÅLA bli behandlet av regelen, men ikke SOLA fordi SO- ikke er et prefiks i motsetning til PÅ-.

(4) Ordre

Ordren angir hvordan man skal komme fram til oppslagsformen på ordet, dvs. hvilke bokstaver som skal fjernes og hva som eventuelt skal legges til. Enkelte regler er, som nevnt, bare temporære i den forstand at behandlingen av ordet ikke er avsluttet (f.eks. regelen for genitiv -S).

(5) Typebetegnelse

Typebetegnelsen angir - i den grad det er mulig - hvilken ordklasse de ordene som dekkes av regelen, tilhører. Regelen for -ENE er f.eks. merket som substantiv. Denne informasjonen er lagt inn med tanke på bruk i automatisk indeksering (jfr. 6.4 (4)). Dessuten er en del funksjonsord lagt inn med en spesiell typebetegnelse, slik at de ikke grupperes sammen med innholdsord med samme rot (se neste side).

4.5 Regeltyper

(1) Tre hovedoppgaver

Automatisk retlemmatiseringen innebærer følgende tre oppgaver:

- a) nøytralisering av bøyningsmorfemer
- b) nøytralisering av avledningsmorfemer
- c) nøytralisering av skriftvarianter

Den siste oppgaven - nøytralisering av skriftvarianter - gjelder få ord, og den har vi derfor løst ved å definere egne regler for de ord det gjelder (f.eks. FRAM - FREM, DATO - DATUM osv.).

Problemene knytter seg til nøytralisering av bøynings- og avledningsmorfemene. For å kunne løse disse to oppgavene, er det nødvendig med regler som utfører følgende operasjoner:

- fjerner endelse
- nøytraliserer stavelssammentrekning
- nøytraliserer vokalskifte
- nøytraliserer konsonantforenklinger og -fordoblinger

I tillegg var det nødvendig med en del regler som tar hånd om funksjonsord som er identisk med oppslagsformen til andre ord, f.eks. MEN og MENE. For å hindre at et funksjonsord blir gruppert sammen med andre ord, har vi definert egne regler som gir dem typebetegnelser funksjonsord.

Tabellen nedenfor viser hvor mange av reglene som er involvert i nøytralisering av bøynings- og avledningsmorfemer fordelt på de ulike operasjonene:

Tabell 4.5/1 Oversikt over regelsettet

OPERASJON (A)	Antall Bøynings- regler (B)	Antall avlednings- regler (C)	Antall bøynings- avledningsregler (D)	Sum (E)
Fjerne endelsen(e)	388	152	45	585
Stavelsessammentrekning	8	6	13	27
Vokalskifte	150	0	8	158
Konsonantforenkling/-fordobl.	24	10	4	38
Ordmarkering			10	10
Skriftvarianter				3

Mange regler er både rettet mot nøytralisering av bøynings- og avledningsmorfemer, og disse er skilt ut i egen kolonne (D) i tabellen. Det betyr at reglene i kolonne (B) kun har til oppgave å nøytralisere bøyningsmorfemene, og reglene i kolonne (C) kun å nøytralisere avledningsmorfemene.

Enkelte regler vil utføre flere operasjoner, f.eks. både å fjerne en endelse og foreta et vokalskifte. Dette framgår ikke av denne tabellen. Her inkluderer tallene alle regler som utfører en bestemt operasjon uavhengig av om regelen også utfører andre operasjoner. Summerer man derfor summen vertikalt, får man et tall som er langt høyere enn 670 som er det totale antall regler i regelsettet. I pkt. (3) nedenfor vil vi komme tilbake til de regler som utfører flere operasjoner.

(2) Operasjoner

Vi skal se nærmere på hva som utføres med de forskjellige

(2) Operasjoner

Vi skal se nærmere på hva som utføres med de forskjellige operasjonene.

Fjerning av endelser. Når vi legger inn generelle regler som fjerner bøyings- og avledningsendelsene, vil også deler av grunnformen bli fjernet hvis den tilsvarende en endelse i form. En slik bokstav eller bokstavkombinasjon som er en del av roten, men som har en form som er identisk med en bøyings- eller avledningsendelse, kalles en falsk endelse (f.eks. LAKEN, METER, PRIS). Det finnes falske endelser som tilsvarende alle bøyings- og avledningsendelsene, men det er stor forskjell på hvor hyppig de forekommer. I vårt arbeid har det ingen betydning om vi også fjerner de falske endelsene så lenge oppslagsordet for rotlemmet blir entydig, slik tilfellet er med METER som får oppslagsordet "MET".

Alle avledningsendelser som ikke fører til leksikalisering blir fjernet. Noen avledningsendelser gir for mange feil når vi fjerner endelsen generelt (f.eks. -LIG). Vi har derfor på statistisk grunnlag definert spesialregler som angir nærmere hvilken bokstavkombinasjon som endelsen skal følge etter i de tilfeller hvor den skal fjernes (evt. beholdes).

Stavelsessammentrekninger. Vi finner stavelsessammentrekninger i forbindelse med bøyingsendelsene til substantivene (TEMPEL - TEMPLER), adjektivene (DIGER - DIGRE) og dessuten ved avledninger som f.eks. fra substantiv til verb (STEMPEL - STEMPLE)

Dette fører til at røttene blir forskjellige for disse ordformene når bøyingsendelsen blir fjernet, f.eks. TEMPEL - TEMPL og DIGER - DIGR. For at disse ordene skal få en felles oppslagsform, må enten E' er fjernes i ordformene med fullform, eller legges inn der stavelsessammentrekningen har funnet sted. Et annet alternativ er å fjerne både fullstavelsen (-EL, -EN, -ER) i de formene der det er aktuelt, og den enslige konsonanten (-L, -N, -R) etterfulgt av de

bøyningsendelsene som fører til sammentrekningen. Siden stavelsene -EN og -ER har samme bokstavkombinasjon som bøyningsendelsene, valgte vi det siste alternativet som spesialregel i disse tilfellene, mens vi valgte å fjerne E'en i de ordene som ender på -EL.

Vokalvekslinger. Vokalskifte forekommer i forbindelse med bøyninger av sterke verb (FINNE - FANT - FUNNET) og uregelmessige substantiv (MANN - MENN). Siden sterke verb og uregelmessige substantiv utgjør en avgrenset gruppe, er de lagt inn på regellista. Stort sett har vi lagt inn alle disse bøyningene i full form, dvs. at det står på regellista at SANG skal erstattes med SYNG.

Det viste seg imidlertid at flere av de sterke verbenes preteritumsformer bare besto av en stavelse på 2-3 bokstaver, f.eks. LA, GA, SÅ, LOT, BAR. Når vi da hadde lagt inn regler som forandret disse formene til imperativsformer, og samtidig åpnet muligheten for at de kunne være sammensatt, ble mange andre ord også oppfattet som sterke verb (særlig bestemt form av substantivet, hunkjønn), f.eks. SOLA. Ved å gjennomgå alle de sammensatte sterke verbene som vi jo ønsket å dekke, fant vi ut at de stort sett var sammensatt med preposisjoner, dvs. de hadde formen preposisjon + sterkt verb. Det var svært få verb som var sammensatt med ord fra andre ordklasser, slik som f.eks. PLANLEGGE, SKATTLEGGE. Vi fant derfor ut at vi kunne dekke de verbene som var sammensatt med en preposisjon ved å legge inn de aktuelle preposisjoner (som jo består av et begrenset antall) på en egen liste. Reglene måtte da undersøke om bokstavkombinasjonen foran et stekt verb svarte til ett av ordene på denne lista. Bare i disse tilfellene ble ordet behandlet som sterkt verb.

Konsonantfordoblinger og -forenklinger. Vi finner konsonantforenklinger i forbindelse med bøyningsendelsene til substantiv (SYKKEL - SYKLER), verb (BEGYNNER - BEGYNTE) og adjektiv (GRØNN - GRØNT) og dessuten ved avledninger som fra substantiv til verb (SYKKEL - SYKLE) og adjektiv til adverb (SKJØNN - SKJØNT). For at disse ordformene skal få felles oppslagsform, må konsonantforenklingen som har skjedd i forbindelse med endelsen, nøytraliseres. Konsonantforenklinger forekommer ofte i forbindelse

med en stavelsessammentrekning, og vi gjennomgikk derfor alle dobbeltkonsonantene etterfulgt av -EL, -EN, -ER og fant ut hvilke konsonanter som faktisk forekom i denne posisjonen. Teoretisk ville denne posisjonen være mulig for alle de norske konsonantene med unntak av 'H', 'J' og 'V'. I praksis viste det seg at det bare var 'B', 'K' og 'M' som hadde flere enn en forekomst i vårt materiale, og derfor ble bare regler som forenklet disse konsonantene, lagt inn på regellista. En slik regel førte til at oppslagsordet til f.eks. SYKKEL - SYKLER ble SYKL. De aktuelle konsonantene som innleder en bøyningsendelse for verb, er 'D' og 'T'. Vi undersøkte derfor de samme konsonantene som for substantivene, og fant ut hvilke som faktisk forekom i slutten av roten på verbene. I disse tilfellene måtte vi legge inn regler som fordoblet konsonanten når bøyningsendelsen ble fjernet.

Konsonantfordoblinger forekommer i forbindelse med kort vokal etterfulgt av 'M'. Vi valgte enkelt 'M' som oppslagsform og la inn regler som dekket "-MM-" pluss alle bøyningsendelsene som begynner på vokal.

Ulike ord med felles oppslagsform. En del ord får felles oppslagsform selv om de tilhører ulike rotlemma. I mange tilfeller dreier det seg om et innholdsord som får samme form som et funksjonsord, f.eks. MEN som både er en konjunksjon og oppslagsordet (imperativformen) til Å MENE. I slike tilfeller (hvor imperativformen sannsynligvis aldri forekommer eller forekommer svært sjelden) har vi lagt inn regler som merker de aktuelle funksjonsordene (de forekommer i et endelig antall). I tekstsøking defineres funksjonsord vanligvis som stoppord, og det er derfor viktig at disse ordene ikke blir gruppert sammen med innholdsmettede ord.

Ulike skriftvarianter. Det finnes en del ord som har ulike skriftvarianter, f.eks. FRAM - FREM, ETTER - EFTER, NÅ - NU og SJUK - SYK. Disse er lagt inn på regellista bare i den grad de har vist seg å vært nødvendige.

(3) Regler som nøytraliserer både bøynings- og avledningsmorfemer

Mange av bøynings- og avledningsreglene utfører to operasjoner samtidig. Et eksempel er regelen for det sterke verbet SUNGET som gir ordet oppslagsformen SYNG, dvs. den både fjerner endelsen -EN og forandrer vokalen til Y. Andre regler kombinerer det å fjerne en bøynings- og avledningsendelse, som f.eks. regelen for ord som slutter på -ATIVT som både fjerner avledningsendelsen -ATIV og bøyningsendelsen -T (f.eks. i ordet ADMINISTRATIVT). Slike regler er tatt med under kolonne (D) i tabellen 4.4/1.

Vi har også regler som lokaliserer en avledningsendelse etterfugt av en bøyningsendelse, men som bare fjerner bøyningsendelsen og gir ordre om at ordet skal behandles på nytt. F.eks. regelen for ord som slutter på -INGENE, fjerner -ENE og lar ordet bli behandlet. På denne måten oppnår vi at ulike bøyningsformer av en avledningsform (f.eks. ARVING, ARVINGEN, ARVINGER og ARVINGENE) får den samme behandlingen (dvs. de blir behandlet av de samme reglene) etter at bøyningsendelsene er fjernet. Disse reglene er tatt med både under bøyningsregler og avledningsregler (dvs. (D)).

Det samme gjelder for de reglene som først trer i kraft etterat en endelse er fjernet. Et eksempel er reglen for ord som slutter på -DL som erstatter -L med -EL.

Tabellen nedenfor (4.4/2) viser kombinasjonene av alle operasjonene som blir utført for å nøytralisere bøynings- og avledningsmorfemer, men i denne tabellen skiller vi ikke mellom bøynings- og avledningsendelser.

Tabell 4.5/2

OPERASJON	ANTALL REGLER
Fjerne endelse(r)	288
komb. med stavelsessammentrek.	14
komb. med vokalskifte	99
komb. med konsonantforenkl./-fordobl.	34
unntaksord (leksikon)	150
Stavelsessammentrekning	9
komb. med fjern endelse (jfr. ovenfor)	
komb. med konsonatforenkl./-fordobl.	4
Vokalskifte	59
komb. med fjern endelse (jfr. ovenfor)	
Konsonantforenkl./-fordobling	
komb. med fjern endelse (jfr. ovenfor)	
komb. med stavelsessam. (jfr. ovenfor)	
Ordmarkering	10
Skriftvarianter	3
Totalt antall regler	670

4.6 Endelige ordgrupper - leksikon

Selv om utgangspunktet for arbeidet med rotlemmatiseringen var å utvikle en regelbasert og ikke en leksikonbasert metode, kan regelsettet allikevel inneholde ord (frie morfemer). Ordene må imidlertid være av en generell karakter og tilhøre en lukket (begrenset) gruppe.

Av det totale antall regler på 670

- var 68 regler formulert på grunnlag av endelser
- var 192 regler formulert på grunnlag av endelser sammen med er

bokstavkombinsjon (f.eks. regelen for -ELT som erstatter -T med en -L som i STELT - STELL(E) og FELT - FELL(E).

- besto 10 regler bare av en bokstavkombinasjon (de har med stavelsessammentrekninger å gjøre, f.eks. -DL erstattes med -DEL)
- var 400 rgler hele ord.

At en regel består av et helt ord, vil enten si at den bare dekker dette ordet eller også ord hvor dette ordet utgjør siste ledd i en sammensetning. Denne regelgruppa ønsket vi å begrense til å innholde ord fra endelige ordgrupper.

Det ble lagt fire forskjellige kriterier til grunn ved vurdering av hva som er endelige ordgrupper:

- a) Lingvistiske kriterier; omfatter ord fra lukkede ordklasser og subklasser, dvs. hjelpeverb, funksjonsord, tall, (korte) modale adverb, osv.
- b) Semantiske kriterier; omfatter ord fra lukkede semantiske grupper, slik som ord for farger, årstider, himmelretninger o.l.
- c) Formelle kriterier; omfatter ord fra ordgrupper som kunne avgrenses på grunnlag av formen, f.eks. verb og substantiv med vokalforandringer i bøyningssystemet, og den gruppen med avledninger som er dannet av tidligere "langformer" av verb, f.eks. "TAKER - TAKER" av langformen (Å) TAGE og "DRAGER" av langformen (Å) DRAGE, ord med falsk endelse i kombinasjon med en konsonant eller et funksjonsord. .
- d) Logiske kriterier; omfatter ord som hadde en grunnform på bare en stavelse (eventuelt med prefiks) av typen konsonantgruppe + vokal f.eks. BO, FRI, SPÅ og GRO (de fleste er verb).

I tabell 4.4/3 har vi gitt en oversikt over hvor mange regler som

utgjorde frie morfemer og hvilke type frie morfemer dette var.

Tabell 4.4/3

TYPE FRIE MORFEMER	ANT. REGLER	ULIKE ORD	LØPEN. ORD
a)			
Hjelpeverb	6	8	4447
Funksjonsord, div. adverb	67	122	51186
Tallord	2	35	92
b)			
Farger, himmelretn., mål	11	54	381
c)			
Sterke verb, uregelm. subst.	198	882	44969
Langformer	27	239	1560
Konsonant/suffiks + suffiks	31	255	2951
Skriftvarianter	1	3	3
d)			
Enstavelsesord (evnt. + prefiks)	24	54	531
Unntak			
Vanlige	13	267	4124
Fagord	20	286	4419
SUM	400	2205	114653

Det er altså 33 regler som er "rene" unntak. Disse ordformene har stort sett en finalsekvens som er identisk med en bøynings- eller avledningsendelse (dvs. en "falsk endelse").

De 400 reglene som utgjør frie morfemer (selvstendige ord) er

egentlig et lite leksikon. Leksikonet er ikke endelig i den forstand at det dekker absolutt alle ordene i gruppene ovenfor. Det dekker kun de ordene fra endelige ordgrupper som forårsaket feil ved rotlemmatiseringen i vårt eksperimentmateriale. Ved bruk av regel-lista i tekstsøking kan det være ønskelig med en komplett liste med funksjonsordene, da dette er ord som vanligvis er uegnet som søkeord og som vanligvis fjernes fra søkegrunnlaget. I et operativt system vil det være ønskelig å gjøre regellista så fullstendig og uavhengig av tekstmaterialet som mulig, og man burde da vurdere om den ikke burde omfatte alle sterke verb og uregelmessige bøyninger (dette er en endelig ordgruppe). Adjektiv og adverb finnes det uendelig mange av, og regelista omfatter bare de som tilhører endelige subkategorier innenfor disse ordklassene og som er i bruk i eksperimentmaterialet, f.eks. farger, gradsadverb, "korte" modale adverb, osv. De øvrige 33 ordene i "leksikonet" er tatt med fordi de forekommer så hyppig både alene og i sammensetninger, at de ville ha forårsaker mange feil i rotlemmatiseringen hvis de ikke ble behandlet særskilt. Disse ordene avhenger i en viss grad av eksperimentmaterialet og de forekommer ikke i et endelig antall.

5. FORSØK MED AUTOMATISK ROTLEMMATISERING

5.1 Hva er riktig rotlemmatisering?

(1) Entydig oppslagsform

Vår målsetting var å gruppere alle ordformene som tilhørte samme rotlemma (se definisjon av rotlemma pkt. 2.2). Hvilken oppslagsform de forskjellige rotlemmaene har fått, er derfor uten betydning så lenge de er entydige for det enkelte rotlemma.

Det er viktig å understreke at vi med entydig oppslagsform mener entydig innenfor et gitt tekstmateriale. I vårt eksperimentmateriale var det flere eksempler på at rotlemmatiseringen ville ha blitt gal hvis alle bøyings- og avledningsformene til ordet hadde forekommet. Men fordi oppslagsformen var entydig for de ordformene som forekom, ble medlemmene av rotlemmaet samlet i en gruppe og rotlemmatiseringen ansett som vellykket. Eksempel på denne type ord er

TUNGVINT	som har fått oppslagsformen	"TUNGVIN"
SPIRET og SPIRE	- " -	"SPI"
ASKE	- " -	"ASK"

I det første tilfellet forekom bare ordet TUNGVINT, og ettersom oppslagsformen "TUNGVIN" var entydig, ble den akseptert. Bøyingsformen TUNGVINTE ville derimot ha gitt oppslagsformen "TUNGVINT", og hadde begge formene forekommet, ville det siste ha blitt foretrukket. Det samme gjelder bøyingsformene SPIRET og SPIRE som begge fikk oppslagsformen "SPI". Hadde derimot bøyingsformene SPIRTE og SPIRENDE forekommet, ville disse ha fått oppslagsformen "SPIR", og ordene SPIRET og SPIRE ville ha blitt registrert som feil.

I det siste eksempelet er oppslagsformen identisk med et annet ord, men siden ordet ASK ikke forekommer i tekstmaterialet, ble oppslagsformen akseptert.

(2) Leksikaliseringsproblemet

Det var ofte vanskelig å avgjøre om en avledningsendelse førte til en leksikalisering av ordformen eller ikke. Det erfarte vi både da vi satte opp forslagene til reglene og - ikke minst - da vi skulle vurdere resultatet. Det var mange avledningsformer som falt mellom ytterpunktene "samme betydning" og "ny betydning", og det var vanskelig å avgjøre om rotlemmatiseringen av disse ordene hadde vært vellykket eller ikke. Vi foretok en skjønnsmessig vurdering av rotlemmatiseringen og forsøkte å være så konsekvente som mulig. Nedenfor er det gitt eksempler på ordpar som vi var i tvil om, men som vi besluttet å holde hver for seg:

BIL - BILIST

BARN - BARNDOM

BLOD - BLØDNING

AKTUELT - AKTUALISERE

ADRESSE - ADRESSAT

EKSPLOSIV - EKSPLOSJON

REPRESENTERE - REPRESENTANT

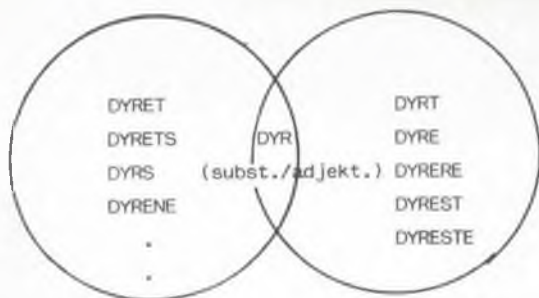
UNNTAKSVIS - UNNTAK

(3) Gradbøyningsproblemet

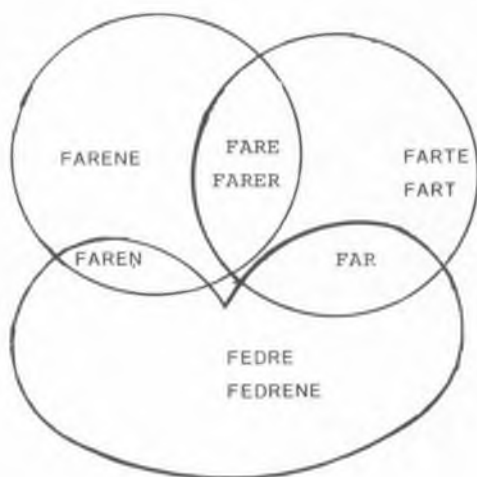
Mange adjektiv skapte også tvil. De fleste av de uavledete adjektivene som har med dimensjon, alder, vurdering, fysisk egen-skap og hastighet å gjøre, vil ofte ha en komparativ- og en superlativform som har liten eller ingen sammenheng med betydningen av positivformen. Som eksempel kan vi nevne at det å være "større" eller "størst" slett ikke alltid betyr å være "stor", eller det å gjøre noe "fortere" eller "forrest" ikke må bety at det gjøres fort. I slike tilfeller har vi bestemt at den positive formen skal tilhøre et eget rotlemma, mens den komparative og den superlative skal tilhøre et annet.

(4) Homografer

Siden vi ikke hadde gjort noe forsøk på å skille homografer, anså vi rotlemmatiseringen som riktig så sant ordformen ble gruppert sammen med en av betydningsvariantene. Eksempel på denne type ord er:



og



I det første tilfellet er ordet DYR en homograf. Siden dette er grunnformen til begge homografkomponentene, er det naturlig at det nettopp er den formen som blir valgt til oppslagsform. Problemet er da at rotlemmaene vil bli gruppert sammen, og ordet DYRET vil f.eks. bli betraktet som om det tilhørte det samme rotlemmaet som DYRESTE. Dette ble allikevel akseptert i forsøket.

Det andre tilfellet er et eksempel på at medlemmer av ulike rotlemmaer kan være identiske. Bøyningsformen FAREN er både bestemt form av substantivet (EN) FAR og substantivet (EN) FARE. Tilsvarende er ordet FARER både flertallsformen til substantivet (EN) FARE og presensformen til verbet (Å) FARE. Ordet FAR kan

bety en far og imperativformen av verbet (A) FARE.

Det er likevel mulig å gjøre rotlemmatiseingen bedre ved å legge inn flere betingelser til noen av reglene. Det er gjerne slik at substantiv og adjektiv som er homografer, ikke har noe med hverandre å gjøre, og skal følgelig ikke grupperes sammen (f.eks. adjektivet DYR (kostbart) og substantivet (ET) DYR - eller adjektivet REIN og substantivet (EN) REIN. Hvis et adjektiv og et substantiv tilhører samme rotlemma, vil vi finne en avledningsendelse enten i substantivet eller i adjektivet, slik som i REIN - REINHET, DYR - DYREAKTIG. Denne kunnskapen kan vi legge inn i regellista, slik at de bøyde formene ikke blir gruppert sammen (f.eks. DYRERE og DYREAKTIG). Homografene kan vi plassere i begge rotlemmaene.

Ved rotlemmatisering vil man likevel stå ovenfor langt færre homografer enn det man gjør når man sammenligner ordformer. De fleste homografer er varianter av samme rotlemma og skal grupperes sammen, f.eks.

<u>(verb)</u>	<u>(substantiv)</u>
ARBEIDE	
	ARBEIDER
ARBEIDET	ARBEIDEREN
ARBEIDES	ARBEIDERE
ARBEIDENDE	ARBEIDERNE
	ARBEIDERS
	ARBEIDERENS
	ARBEIDERES
	ARBEIDERNES

hvor ordet ARBEIDER er en homograf

5.2 Ord som ble gruppert riktig

Ved testing av reglene mot eksperimentmaterialet viste resultatet at ca. 97,7% av grafordene ble gruppert riktig. Disse utgjorde ca. 98,8% av det totale antall løpende ord i materialet. At prosenten er høyere for antall ulike ord enn for antall løpende ord, skyldes at de ordene som ble feil gruppert, forekom mindre hyppig enn gjennomsnittet i tekstmaterialet.

Blant feilene finner vi 45 graford som er navn på verdensdeler og land, nasjonalitetsbetegnelser, etc. Disse utgjorde 22 rotlemma, og som et rotlemma regnet vi f.eks. ordene SVERIGE, SVENSK og SVENSK. Disse feilene lar seg løse ved spesialregler, men slike regler ble ikke lagt inn i vårt regelsett fordi de dekket for få ord. Vi vil heller, ved en senere anledning, se på dette området spesielt. I den videre vurderingen av resultatet vil vi derfor se bort fra disse feilene.

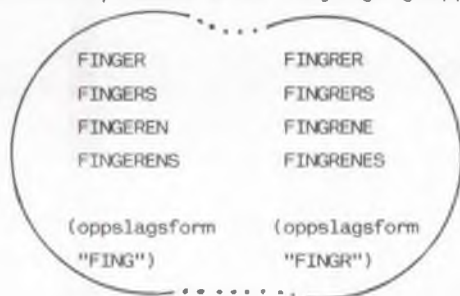
Eksperimentmaterialet besto av 13624 rotlemmaer. Av disse ble 13358 samlet (98%), men 82 av dem inneholdte også ord fra et annet rotlemma. De resterende 184 rotlemmaene (2%) ble splittet, slik at ordene fra et rotlemma ble spredt på flere grupper og i enkelte tilfeller sammen med ord fra andre rotlemmaer.

Tilsammen ble det etablert ca. 13900 ordgrupper. Av disse var det 624 som enten inneholdt deler av et rotlemmaet eller en blanding av ord fra flere rotlemma. Antall slike "ufullstendige" eller "overfylte" grupper er nesten dobbelt så stort som antall graford som ble feil gruppert, fordi feil plassering fører til at rotlemmaet blir splittet.

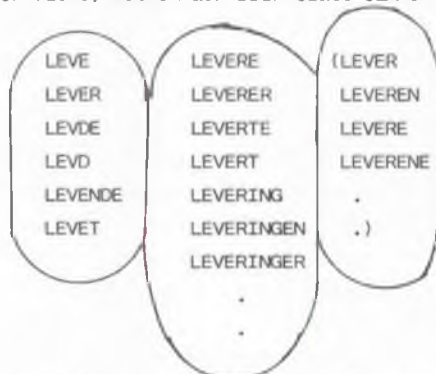
At et ord blir feil gruppert vil kunne føre til tre ulike feiltyper som er illustrert i figur 5.2/1 nedenfor.

Fig. 5.2./1 Konsekvensene av feil rotlemmatisering

(A) Rotlemmat blir splittet i to entydige grupper

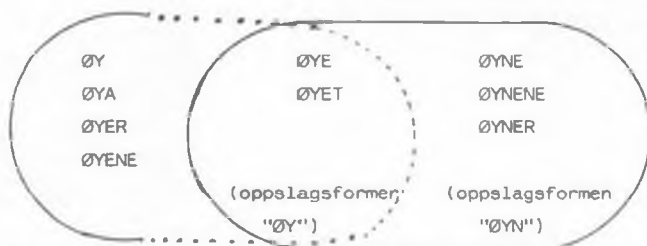


(B) To (eller flere) rotlemmaer blir slått sammen



(alle har fått oppslagsformen
"LEV")

(C) Deler av rotlemmaet er gruppert sammen med et annet rotlemma



I det første tilfellet har ordene fått ulik oppslagsform, og rotlemmaet er blitt fordelt på to (eller flere) grupper. Ingen av gruppene består av ord fra andre rotlemmaer. Totalt var det 234 rotlemmaer som ble splittet på denne måten, og i 168 av tilfellene var det bare ett ord som skilte seg ut og dannet en egen gruppe. Tilsammen resulterte denne feiltypen i at det ble etablert 478 grupper som inneholdt deler av et rotlemma (Jfr. tabell 5.2/2).

Den andre feiltypen (B) er et eksempel på at alle ordene innenfor et rotlemma har fått en oppslagsform som er identisk med oppslagsformen til et annet rotlemma. I motsetning til tilfellet foran, ble nå hele rotlemmaet gruppert sammen med et annet rotlemma. Dette gjaldt totalt 82 rotlemmaer, hvorav ca. 1/3 besto bare av ett ord.

Den siste situasjonen (C) er en blanding av de to foregående. Her er også rotlemmaet blitt splittet, men i dette tilfellet har de ordene som har "brutt" ut, blitt slått sammen med ord fra et annet rotlemma. Resultatet er at vi står igjen med en gruppe som inneholder deler av et rotlemma, og en som inneholder ord fra ulike rotlemmaer. I alt var det 32 rotlemmaer som fikk denne skjebnen, og de resulterte i at det ble etablert 64 "ufullstendige" eller "overfylte" grupper.

I tabellen nedenfor har vi oppsummert resultatet.

Tabell 5.2/2 Oversikt over ulike feiltyper

!Feiltype	!	graford	!	rotlemma	!	grupper	!
!Splittet i entydige grupper	!	353	!	234	!	478	!
!Rotlemmaer slått sammen	!	194	!	84	!	84	!
!Splittet i tvetydige grupper	!	54	!	32	!	64	!
!Sum	!	601	!	350	!	626	!

5.3 De mest problematiske ordformene?

(1) Oversikt over resultatet

For å få bedre innsikt i hva som skapte størst problemer ved rotlemmatiseringen, studerte vi resultatet i lys av de tre hovedoppgavene som rotlemmatiseringen skulle løse (jfr. pkt. 4.5):

- nøytralisering av bøyingsmorfemer
- nøytralisering av avledningsmorfemer
- nøytralisering av skriftvarianter

Nøytralisering av skriftvarianter ble løst 100% tilfredstillende ved at det ble lagt inn særskilte regler som dekket ord med ulike skiftvarianter. I vårt materiale var det bare tre ord av denne typen; EFTER/ETTER, FREM/FRAM og DATUM/DATO.

Feilene oppsto ved rotlemmatisering av bøyings- og avledningsformene, og hvordan de fordelte seg på disse to kategoriene er vist i tabell 5.3/1 nedenfor.

Tabell 5.3/1 Oversikt over bøyings- og avledningfeil

KATEGORI	ANTALL FEIL	
	Ulike ord	Løpende ord
Bøyingsformer	194	1941
Avledningsformer	372	3211
Bøyings- og avledningsf.	35	534
Sum	601	5686

Blant de 601 grafordene som ble feil rotlemmatisert, var det 194 (32%) "bøyningsfeil" og 372 (62%) "avledningsfeil". Det betyr at det var 194 graford som ble gruppert feil i forsøket på å nøytralisere bøyningsmorfemene og 372 i forsøket på av nøytralisere avledningsmorfemene. Det betyr ikke at det 372 var avledningsformer som ble feil, men at det var 372 graford som ble feil på grunn av reglene som skulle nøytralisere avledningsmorfemene - eller mangel på sådanne. F.eks. fikk ordet FARSKAPSBEVIS oppslagsformen FARSKAPSBE fordi -VIS ble identifisert som en avledningsendelse og fjernet.

Av tabellen framgår det også at 35 graford ble feil rotlemmatisert på grunn av regler som både skulle nøytralisere et bøynings- og et avledningsmorfem. Et eksempel er ordet LEVEREN som fikk oppslagsformen "LEV" på grunn av regelen for -ER. Denne regelen er både rettet mot bøyningsendelsen -ER som i GUTTER, og avledningssuffiksen som i ARBEIDER.

Vi ser at det var nesten dobbelt så mange "avledningsfeil" som "bøyningsfeil". Skal man imidlertid få et riktig bilde av den prosentvise andelen feil for de to kategoriene, må tallene sees i relasjon til hvor mange bøynings- og avledningsformer det var totalt i eksperimentmaterialet. Disse tallene er det tidkrevende å framskaffe, fordi det forutsetter en kartlegging av hvert enkelt ord i hele tekstmaterialet.

(2) Bøynings- og avledningsregler

I tabellen 5.3/2 har vi derimot gitt en oversikt over frekvensen til bøynings- og avledningsreglene og hvor mange feil de gir. Disse tallene gir et tilnærmet bilde av den prosentvise andelen bøynings- og avledningsfeil - og et inntrykk av hvor vellykket reglene har fungert. Oversikten er tilnærmet fordi tallene inkluderer antall tilfeller av "falske endelser" - dvs. tilfeller hvor finalsekvensen av roten er identisk med en suffiks, f.eks. -ING i RING og FINGER (jfr. avsnitt 4.5 pkt. 2). For det andre er, som nevnt, enkelte regler rettet både mot bøyningsmorfemer og avledningsmorfemer, og vi har ikke eksakte tall for fordelingen mellom dem. I tabellen har

vi derfor skilt ut disse reglene.

Tabell 5.3/2

! TYPE REGEL	TOTALT			FEIL		
	!Ant.regler	!Ulike ord	!Løp. ord	!Ant.regler	!Ulike ord	!Løp.ord
Bøyning	440	13364	209338	43	156	1732
Avledning	110	5261	42427	40	214	2279
Kombinasjon	117	9588	114126	6	103	988

Tabellen viser at det er langt flere regler for bøyningsformer enn for avledningsformer, og bøyningsreglene dekker mer enn dobbelt så mange ulike graford. Den prosentvise andelen feil for avledningsreglene (3.5%) er imidlertid tre ganger så stor som den prosentvise andelen feil for bøyningsreglene (1.3%), men begge prosentene er små. Det er derfor vanskelig å trekke noen konklusjon om hvilke typer ordformer som er de mest problematiske, ettersom bøyningsmorfemene er representert ved flere regler og avledningsreglene gir mest feil.

(3) Feiltyper

Det største problemet med automatisk rotlemmatisering er at sufiksene består av bokstavkombinasjoner som også kan forekomme finalt i en rot (såkalte falske endelser), f.eks.

bøyningsendelsen	-EN	og finalekvensen i roten	VERDEN
"	-ET	"	MAGNET
"	-ENE	"	ALENE
"	-ER	"	LEVER (substantiv)
"	-T	"	MAT
"	-S	"	HUS

avledningsendelsen	-ERE	og finalsekvensen i roten	FLERE
"	-ING	"	RING
"	-MENT	"	ELEMENT
"	-SJON	"	NASJON
"	-VIS	"	BEVIS

I prinsippet er det to måter å gå fram på for å løse dette problemet. Den ene måten er å definere generelle hovedregler som fjerner bøyings- og avledningsendelsene, og så legge inn nødvendige unntak. Den andre måten er å gå den motsatte veien, dvs. at hovedregelen sier at bokstavkombinasjonen skal beholdes, mens spesialreglene tar seg av bøyings- og avledningsendelsene. I regelsettet benytter vi begge framgangsmåtene. Vi har f.eks. regler som fjerner -TE, -A, -DD, -ITET, -IG og -NING i slutten av et ord - og dette fører til at følgende "falske endelser" blir fjernet:

-TE	i HJERTE,	og ordet får oppslagsformen	HJER
-A	i PAPPA,	"	PAPP
-DD	i REDD,	"	RE
og			
-ITET	i BEITET,	"	BE
-IG	i FATTIG,	"	FATT
-NING	i SETNING,	"	SETT

Nå vil imidlertid fjerning av en "falsk endelse" ikke gi feil hvis alle ordene innenfor rotlemmaet får fjernet denne endelsen. Oppslagsformen blir da den samme for alle ordene, og de blir gruppert sammen.

Den andre framgangsmåten går ut på at man definerer hvilke bokstavkombinasjoner som må stå foran endelsene for at de skal fjernes. F. eks. blir endelsen -T fjernet bare i helt spesielle tilfeller, fordi det finnes så mange ord som slutter på -T i roten. Dette resulterte bl.a. i at det var 53 ord i vårt materiale som ikke fikk fjernet bøyingsendelsen -T. Disse tilfellene utgjorde hele 85% av det totale antall bøyingsfeil av denne typen. Eksempel på feil er:

SPIST som får oppslagsformen		SPIST
KOKT	"	KOKT
PENT	"	PENT
TYNT	"	TYNT

Ord som ikke får fjernet endelsene vil nesten alltid gi feil rotlematisering fordi bøyningsformene får forskjellig oppslagsform fra avledningsformene. Eksempel på avledningsfeil av denne typen er:

AKSEPTABLE som får oppslagsformen		AKSEPTABL
KOMMUNAL	"	KOMMUNAL
DEFINISJON	"	DEFINISJON
KATOLIKK	"	KATOLIKK

Generelt kan man si at jo lengre suffikset er, eller jo flere suffikser man behandler samtidig, jo større sannsynlighet er det for at man ikke vil finne den samme bokstavsekvensen som finalsekvens i en rot. I eksperimentmaterialet var det f.eks. svært sjeldent at de reglene som var rettet mot kombinasjoner av suffikser (f.eks. -SJONENE), ga feil. Likeledes var det langt flere feil blant de generelle bøyningsreglene, som ofte er korte, enn blant avledningsreglene.

I tabell 5.3/3 nedenfor har vi gitt en oversikt over de ulike feiltyper fordelt på bøynings- og avledningsfeil.

Tabell 5.3/3

FEIL TYPE	Bøyningsfeil	Avledningsfeil
"Falsk endelse"	100	101
Endelsen ikke fjernet	68	112
Leksikalisering		139

Den siste raden i tabellen viser antall leksikaliseringsfeil. Leksikaliseringsfeil opptrer bare ved nøytralisering av avledningsmorfemer og betyr at ordet har fått endret sin betydning som følge av endelsen. Av hensyn til denne feiltypen måtte vi vurdere hva som ga flest feil ved innføring av hovedregler eller ved innføring av spesialregler. Eksempel på leksikaliseringsfeil er:

BORGERLIG	som får oppslagsformen	"BORG"	(samme som (EN) BORG)
TYDELIG	"	"TYD"	(samme oppslagsform som TYDE)
ÅPENBART	"	"ÅP"	(samme oppslagsform som ÅPEN)

Mange stilte seg tvilende til om det i det hele lot seg gjøre å nøytralisere avledningsmorfemene på grunn av leksikaliseringsproblemet. Av tabellen framgår det at det var 139 "avledningsfeil" av denne typen, dvs. ca. 1/3 av det totale antall "avledningsfeil". Dette tallet tyder på at leksikaliseringsproblemet ikke er så stort som man kanskje antar. Sett i forhold til det totale antall feil utgjør denne gruppen ca. 23%, dvs. under en halv prosent i forhold til det totale antall ulike graford i materialet. Et mer presist bilde av leksikaliseringsproblemet er det vanskelig å gi uten å se det i relasjon til det totale antall avledningsformer i materialet og hvor mange av dem som er leksikalisert. Denne informasjon tok vi ikke vare på i undersøkelsen av avledningssuffiksene.

5.4 Regler som førte til feil rotlemmatisering

Årsaken til feil rotlemmatisering skyldes i de fleste tilfeller at ordet er fanget opp av feil regel. Blant feilen finner vi allikevel 122 graford som er blitt feil fordi det ikke eksisterer noen regel som passer for ordet. Disse utgjør ca. 20% av feilene, og de fleste er avledningsformer hvor det har skjedd en stavelsessammentrekning eller konsonantforenkling når en avledningssuffix er lagt til. Disse ordene er såpass spesielle, at det er umulig å definere generelle regler uten at også andre ord vil bli berørt. Eksempel på denne type ord er :

GARANTI som burde grupperes sammen med GARANTERE, GARANTIST, osv.

SLOKNE - " -

SLOKKE, SLOKKET, osv.

SÆREGNE - " -

SÆREGEN

EPILEPSI - " -

EPILEPTISK, EPILEPTIKER,

PRINSIPIELL - " -

PRINSIPP, osv.

I det følgende skal vi se bort fra disse ordene og i stedet rette oppmerksomheten mot de reglene som førte til at 479 grafordene ble feil rotlematisert. De vil kunne gi oss informasjon om hvilke regler som var de mest problematiske og hvor ofte disse reglene ble anvendt i eksperimentmaterialet.

Totalt inneholdt regelsettet 670 regler, og 91 av disse ga en eller flere feil. Den gjennomsnittlige feilprosenten for disse 92 reglene lå imidlertid på 2.3%, dvs. det var 479 feil blant de 20323 graford som ble behandlet av dem. Tallet 20323 virker kanskje noe høyt, men vi minner om at flere ord ble behandlet av flere regler fordi det f.eks. var sammensatt av flere suffikser.

Eksempler på denne type ord er vist i tabellen nedenfor.

Tabell 5.4/1

	! 1. regel	! 2. regel	! 3. regel	! 4. regel	!
! Graford	! fører til:	! fører til:	! fører til:	! fører til:	!
! ARVINGERS	! ARVINGER	! ARVING	! ARV	!	!
! BEKREFTELSENES	! BEKREFTELSENE	! BEKREFTELSE	! BEKREFT	!	!
! VANSKELIGHETENS	! VANSKELIGHETEN	! VANSKELIGHET	! VANSKELIG	! VANSK	!
! DOBBELTE	! DOBBELT	! DOBBEL	! DOBL	!	!
! UTVALGTE	! UTVALGT	! UTVELG	!	!	!

Hvor mange regler som har gitt feil, sier likevel lite om hvor godt regelsettet har fungert. For det første er bokstavstrengen i flere regler fri morfemer – og 98 av disse regler vil aldri gi feil fordi de dekker bare ett ord. De resterende vil fange opp bøynings- og avledningsformer av ordet eller sammensatte ord hvor det frie morfemet inngår som siste ledd. Disse reglene kan teoretisk gi feil, men det er få som gjør det.

For det andre er feilene ulikt fordelt på de 91 reglene, og skal man få et inntrykk av hvordan regelen har fungert, bør man også se på antall feil i forhold til hvor mange ganger regelen er anvendt.

De fleste feilene finner vi blant hovedreglene, dvs. de som konsekvent fjerner en bøynings- eller avledningsendelse (f.eks. fjern -ENE). Av tabellen nedenfor ser vi at regelsettet består bare av 68 slike regler, men de behandlet hele 60% av ordene i eksperimentmaterialet. Av disse ga 29 regler en eller flere feil. Det er faktisk over halvparten av de feilene som kan føres tilbake til disse reglene, men fordi reglene anvendes så hyppig, er feilprosent nå under 2%.

En del feil finner vi også blant spesialreglene, dvs. regler som også angir hvilken bokstav eller bokstavkombinasjon som må stå finalt i roten, for at suffiksen kan fjernes (f.eks. ord som slutter på -MMELT får fjernet -T), eller hvor flere operasjoner må gjennomføres for at bøynings- eller avledningsmorfemet skal bli nøytralisert (f.eks. stavelsessammentrekning eller konsonantforenkling). I alt omfatter regelsettet 202 regler av denne typen og 54 av disse har gitt feil. Likevel gir disse reglene mindre feil enn hovedreglene, fordi de er mer spesifikke og forekommer mindre hyppig.

De fleste reglene er unntaksord, og av tabell 5.4/2 fremgår det at bare 9 av disse 400 reglene feil. Bokstavstrengen i disse reglene utgjør hele ord, og ved å stille betingelser til de ordene som reglene fanger opp, er det sjeldent at det opptrer feil. Et eksempel på en feil av denne type er ordet RELEVANT som fikk oppslagsformen RELEVINN, fordi det sterke verbet VANT får oppslagsformen VINN. Ordet RELEVANTE derimot fikk bare fjernet -E fordi dette ordet ikke ble fanget opp av regelen.

Konklusjonen på dette blir at ca. 30% av hoved- og spesialreglene ga feil ved rotlemmatisering av eksperimentmaterialet. Den gjennomsnittlige feilprosenten for hoved- og spesialreglene var likevel bare på 1.7%, og disse utgjorde 1.3% av den totale frekvensen for disse ordene (løpende ord). Antall feil var størst for de regler som ble brukt mest, men feilprosenten var størst for de som ble brukt minst.

Tabell 5.4/2

! TYPE REGEL	! TOTALT I REGELSETTET			! FEIL		
!	!	!	!	!	!	!
!	!Ant.regler	!Ulike ord	!Løp.ord	!Ant.regler	!Ulike ord	!Løp.ord
!Hovedregler	! 68	! 17273	! 172646	! 29	! 274	! 3217
!Spesialregler	! 202	! 8734	! 82610	! 54	! 187	! 1773
!Unntaksord	! 400	! 2212	! 114581	! 9	! 18	! 51
! SUM	! 670	! 28219	! 369837	! 91	! 479	! 5041

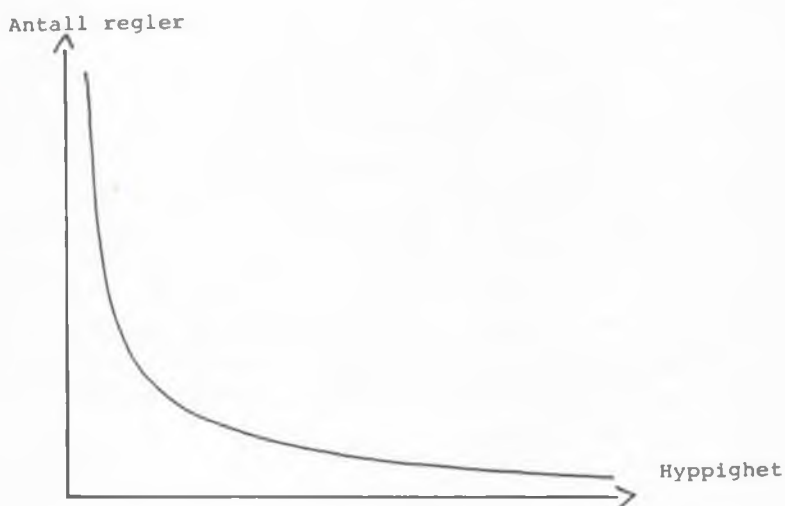
5.5 Regelstatistikk

Informasjon om hvor hyppig reglene blir anvendt - dvs. regel-frekvensene - følger i dag som en del av resultatet ved den automatiske rotlemmatiseringen. Denne informasjon var i utgangspunktet ment å være et hjelpemiddel ved etablering av regelsettet, men siden har den vist seg å være nyttig også i andre sammenhenger. Den gir bl.a. en oversikt over hyppigheten til de forskjellige suffiksene i tekstmaterialet, og gir f.eks. svar på hvor ofte det var aktuelt med en stavelsessammentrekning eller hvor mange sterke verb tekstmaterialet inneholder. Regelfrekvensen kan dessuten oppfattes som et uttrykk for hvor viktig en regel er for rotlemmatiseringen. En regel som f.eks. behandler 1000 ord vil ha

langt større innflytelse på rotlemmatiseringen enn en regel som bare dekker 10 ord. I tekstsøking vil imidlertid regelens verdi også avhenge av hvilke ord regelen er rettet mot, siden det slett ikke er alle ordene som egner seg som søkeord. Nedenfor skal vi se litt nærmere på forholdet mellom antall regler og resultatet av rotlemmatiseringen.

Ettersom enkelte suffikser forekommer langt mer hyppig enn andre, vil man med relativt få regler kunne dekke mange graford. Grovt sett kan man si at forholdet mellom antall regler og hyppigheten til reglene er som i figuren nedenfor.

Figur 5.5/1



Det fremgår av figuren at noen få regler brukes svært hyppig, mens mange regler har en lav hyppighet. Et spørsmål som vi derfor ofte er blitt stilt overfor, er om man kan oppnå nesten like gode resultater ved å fjerne de minst hyppig anvendte reglene. Så enkelt er imidlertid ikke forholdet mellom hyppighet, antall regler og kvaliteten på resultatet. En oversikt over hyppighetsfordelingen til reglene er gitt i tabell 5.5/2. Tallene i parentes angir antall feil.

Av tabellen framgår det at de 171 hyppigst anvendte reglene (alle med frekvens større enn 16, jfr. nederst i tabellen) fanger opp ca. 92% av antall ulike graford. Tilsammen gir disse 446 feil. Det betyr at hvis vi hadde utelatt resten av reglene, ville vi minst ha oppnådd at ca. 90 - 91% av grafordene ville ha fått "riktig" behandling. Disse tallene sier imidlertid lite om resultatet av rotlemmatiseringen, og det er spesielt tre argumenter som taler for at den antagelig vil bli langt dårligere enn hva denne prosenten tilsier.

For det første må alle ordene i et rotlemma være samlet for at rotlemmatiseringen skal karakteriseres som vellykket. Siden reglene i stor grad behandler ulike bøynings- og avlednings-morfemer, vil utelatelsen av en regel føre til feil behandling av alle de ordene som har den strukturen, og følgelig risikerer vi at også alle rotlemmaene som disse ordene er medlem av, ikke blir samlet. Hvis vi f.eks. slettet regelen for substantivendelsen -ENE, ville alle rotlemmaer som omfatter et substantiv med denne endelsen, bli mangelfulle. Det betyr at hvis vi begrenset regelsettet til de 171 mest hyppig anvendte reglene, kunne vi få like mange ufullstendige rotlemmaer som det antallet ord disse reglene skulle behandle.

For det andre så er de minst hyppige reglene (målt i antall graford) rettet mot andre typer ord enn de øvrige reglene. De fleste består av frie morfemer som sterke verb, uregelmessige substantiv, funksjonsord og fagord. Det betyr at hvis man fjerner dem fra regellista, vil vi ikke få samlet disse ord-

gruppene. Spørsmålet om de skal slettes eller beholdes er derfor et spørsmål om hvor viktig det er at de blir riktig rotlemmatisert. I tekstsøking vil mange av dem være viktige, fordi de er typiske søkeord.

For det tredje vil ofte de minst hyppig anvendte reglene dekke de mest frekvente ordene. Feil plassering av disse ordene kan derfor få langt mer alvorlige konsekvenser enn ord som forekommer sjelden. De reglene som er merket som funksjonsord, er av denne typen. I tekstsøking er det svært viktig at disse ordene ikke grupperes sammen med innholdsordene. Også i forbindelse med frekvensstudier er ordenes løpende frekvens av stor betydning.

Vi legger forøvrig merke til i tabellen at hele 39 regler har en frekvens på null - dvs. at de ikke ble brukt ved rotlemmatiseringen av eksperimentmaterialet. Disse reglene kunne vi forsåvidt ha fjernet fra regelsettet uten at det ville ha gått ut over resultatet. Grunnen til at vi allikevel har valgt å beholde dem, er at de er såkalte "konsekvensregler" som følger av andre regler, f.eks. var det flere av de sterke verbene som bare forekom i en bøyningsform.

Tabell 5.5/2 Bruksfrekvensen til reglene

HYPPIGHET		TOTALT ANTALL	
!(ant. graford)	! REGLER !	ULIKE GRAFORD	! LØPENDE GRAFORD !
0	39		
1	147	147	36134
2	59	118	25514
3	51	153 (2)	2596 (3)
4	34	136 (1)	10671 (13)
5	22	115 (2)	4069 (111)
6	27	162 (5)	2942 (38)
7	20	140	2064
8	19	152 (2)	3784 (125)
9	14	126	2636
10	10	100 (8)	1507 (757)
11	17	187	961
12	12	144 (7)	3091 (66)
13	12	156 (2)	4801 (5)
14	11	154 (3)	1454 (39)
15	6	90 (1)	732 (1)
16 - 25	45	1300 (22)	12973 (87)
26 - 35	29	856 (30)	6831 (241)
36 - 45	17	677 (29)	7436 (79)
46 - 55	13	514 (15)	10775 (138)
56 - 65	7	420 (10)	4617 (21)
66 - 3448	60	22031 (340)	219534 (4304)
SUM	670	28211 (479)	369783 (5041)

5.6 Resultatet med testmaterialet

For å undersøke nærmere om regelsettet var spesialsydd for vårt eksperimentmateriale, testet vi det på et tilfeldig valgt materiale. Testmaterialet besto av barneboka Ole Brumm og var altså av en helt annen type. Boka inneholder ca. 2100 ulike graford som tilsvarer ca. 1200 rotlemma.

Resultatet viste at 3.6% av grafordene ble feil rotlemmatisert som utgjorde 4.4% av rotlemmaene. Feilprosentene er noe høyere enn for eksperimentmaterialet, men en del av feilene skyldes at regelsettet ikke inneholder alle ord fra de lukkede ordgruppene, f.eks. sterke verb som KRØP, SPRAKK og GNE (jfr. 4.6). Vi kan forøvrig tilføye at det var flere skrivefeil i materialet (79 feil) enn ord som ble feil rotlemmatisert.

6. AUTOMATISK ROTLEMMATISERING I TEKSTSØKING

6.1 Eksperimenter med automatisk rotlemmatisering

Det finnes få eksempler på studier som har undersøkt økningen i søkeeffektiviteten ved bruk av automatisk rotlemmatisering. Med effektivitetsøkning sikter vi her til forbedringen av søkeresultatet målt i gjenfinningsgrad og presisjon. Gerhard Salton gjennomførte noen forsøk alt tidlig på 60-tallet hvor effekten av det han kalte "automatic stemming" ble testet mot bruk av en synonymtesaurus eller ingen bruk av hjelpemidler (jfr. Salton 1968). "Automatic stemming" betyr at ordene blir redusert til en stamme ved at endelsen til ordet fjernes. Ikke uventet viste resultatet at bruk av en synonymtesaurus ga best resultat, men overraskende var det at forskjellen til det å bruke "automatic stemming" bare var marginal.

Ved Institutt for rettsinformatikk er det gjennomført forsøk med bruk av automatisk rotlemmatisering. Resultatene viste at vi ved å utvide søkeargumentet med ord som automatisk ble identifisert til å tilhøre samme rotlemma som søkeordene, både økte gjenfinningsgraden (ved at flere relevante dokumenter ble funnet) og presisjonen (ved at rangeringen av de funne dokumentene ble bedre) til søkeresultatet.

I begge disse tilfellene var spørsmålene til systemet formulert i naturlig språk. Det er grunn til å tro at behovet for automatisk rotlemmatisering er enda større i slike systemer enn i tradisjonelle søkesystemer hvor man tross alt har muligheten til å spesifisere eksplisitt de bøyings- og avledningsformer som søkeordet har.

6.2 Automatisk rotlemmatisering i forhold til trunkering

For å få et reelt bilde av effekten på søkeresultatet ved å innføre automatisk rotlemmatisering i de tradisjonelle tekstsøkesystemer, bør resultatet sammenlignes med det man får ved å utnytte eksisterende hjelpemidler som f.eks. trunkering

(jfr. avsnitt 1.3).

I litteraturen finnes det lite informasjon om effekten av trunkering. Trygve Harvold gjorde et forsøk tidlig på 70-tallet hvor han bl.a. undersøkte hvor mange av de kontekstuavhengige synonymer som ble fanget opp ved en trunkering av søkeordene (jfr. Harvold 1974). Som kontekstuavhengig synonym regnet han også ordenes bøynings- og avledningsformer, og resultatet viste at trunkeringen omfattet ca. 75% av de kontekstuavhengige synonymene. Forsøket ga derimot ikke informasjon om hvor mange bøynings- og avledningsformer som ble fanget opp av trunkeringen, eller hvor mye støy den ga. For å få nærmere innsikt i dette, ble det som et ledd i prosjekt FORT (NORIS (58)) ved IRI gjennomført et eksperiment hvor bruk av trunkering ble sammenlignet med bruk av automatisk rotlemmatisering.

To personer, begge med bred erfaring i tekstsøking, ble bedt om å trunkere alle aktuelle søkeord i hvert sitt spørsmålsett på ca. 10 spørsmål. Spørsmålene omfattet juridiske problemstillinger og var rettet mot en dokumentbase med lagmannsrettsavgjørelser i familie-, skifte- og arverett. De var alle formulert i naturlig språk og definert av de samme personene som var med i forsøket. For hvert spørsmål ble de trunkerte ordene matet inn i datamaskinen, og ut fikk vi en liste over alle de ordene i dokumentbasen som tilfredstilte trunkeringen. De samme personene ble bedt om å merke av de ordene som var relevante i forhold til det aktuelle spørsmålet, og på denne måten fikk vi et grunnlag som kunne sammenlignes med den automatiske rotlemmatiseringen.

Resultatet av den automatiske rotlemmatiseringen ble framskaffet på tilsvarende måte. Hvert ord i søkeargumentet ble supplert med alle ordene i dokumentbasen som hadde fått felles oppslagsform, og disse ble gjennomgått på samme måte av de samme personene. Vi brukte her den metoden for rotlemmatisering som er beskrevet i denne rapporten.

En sammenligning av resultatene viste at i 13 av 20 spørsmål var det overlapping mellom trunkeringen og den automatiske

rotlemmatiseringen. Eksempler på tilfeller hvor det ikke var overlapping, fantes ved bruk av ord hvor roten endret seg ved bøyning. F.eks. ordet STELLE ble trunkert etter STELL og bøyningsformene STELT og STELTE kom derfor ikke med. I noen tilfeller hadde de løst dette problemet ved å trunkere etter den delen av stammen som ikke endret seg ved bøyning, f.eks. UTLAGT ble trunkert etter "UTL". Dette førte til at mange ikke-relevante ord ble funnet. I andre tilfeller hadde de glemt å sløyfe avledningsendelsen før trunkering og fikk følgelig ikke med de andre ordene med samme rot, f.eks. SEPARASJON ble trunkert etter SEPARASJON og man mistet derfor bøyningsformen av (Å) SEPARERE. Totalt sett omfattet allikvel trunkeringen 94% av ordene som kom med ved en automatisk rotlemmatisering .

Interessant var det imidlertid å merke seg at 65% av ordene som ble fanget opp ved trunkeringen, ikke var bøynings- og avledningsformer til søkeordene. De fleste var sammensatte ord, men det var også flere ord som hadde en helt annet grunnform enn søkeordet, f.eks. ordet MOR trunkert fanget opp ordene MORAL, MORALSK og MORALSKE. Av de ordene som ikke var medlem av samme rotlemma som søkeordet, var det ca. 66% som ikke var relevante for spørsmålene. Sagt på en annen måte; ca. 40% av ordene som ble fanget opp av trunkering var ikke relevante. Det er viktig å påpeke at vi med "ikke relevant" mener ord som ikke er relevant i forhold til det spørsmålet som søkeordet representerte.

Av dette kan vi konkludere at automatisk rotlemmatisering absolutt har sin berettigelse i forhold til trunkering - enten som en erstatning eller som et supplement. Hvis presisjonen til søke-resultatet er det viktigste for brukeren, synes det klart at automatisk rotlemmatisering er et mer effektivt hjelpemiddel enn trunkering. Hvis derimot gjenfinningsgraden er den viktigste, vil en kombinasjon av trunkering og automatisk rotlemmatisering være det mest effektive. Dette vil kunne bidra til at alle bøynings- og avledningsformer blir lagt til søkeargumentet og andre relevante ord (f.eks. sammensatte ord) som trunkeringen fører med seg.

6.3 Bestemmelse av ordenes vekt ved søking

Automatisk rotlemmatisering er også et interessant hjelpemiddel i forbindelse med vektlegging av ord. I tekstsøking kan vektlegging av ord være aktuelt både ved fastlegging av hvor godt et ord er egnet som søkeord og hvor karakteristisk det er for innholdet i det enkelte dokumentet. Vekten anvendes som et uttrykk for hvor "viktig" ordet er.

Som vi har vært inne på tidligere, er det slett ikke alle ordene i en tekst som er egnet som søkeord. Enkelte vil nesten alltid være uaktuelle (f.eks. GJELDE, ANSE, SYNES, HA, MÅTTE, DET, DERE, MEN og FORDI), mens andre er interessante bare hvis de anvendes i kombinasjon med andre ord (f.eks. preposisjonen I i frasen "LOV OM BARN I EKTESKAP"). Noen ord er godt egnet som søkeord i en dokumentsamling, men totalt uinteressante i en annen. F.eks. vil ordet FARSKAP i en samling med bare farskapssaker være uaktuelt som søkeord, fordi det er gitt at dokumentene handler om farskap. En søking på FARSKAP vil derfor kunne resultere i at nesten alle dokumentene vil bli foreslått. Ved å tillegge disse ordene liten vekt, vil man hindre at de ødelegger søkeresultatet for brukeren hvis de allikevel skulle bli brukt som søkeord. Dette er spesielt viktig i systemer basert på spørsmål i naturlig språk, hvor systemet selv må ta stilling til hvorvidt et ord er egnet som søkeord eller ikke. Ved å fjerne dem helt fra søkegrunnet vil man kunne spare mye lagringsplass siden ca. 60-70 % av ordene i en løpende tekst er av denne typen. Dette er aktuelt ved bruk av tekstsøking på mindre datamaskiner, f.eks. personlige datamaskiner.

Flere tekstsøkesystemer gir muligheten for å rangere de funne dokumentene ut fra deres "likhet" med spørsmålet. Til dette formål anvendes en rangeringsalgoritme som tar utgangspunkt i søkeordenes vekt i det enkelte dokument. Vekten er et uttrykk for hvor godt ordet beskriver innholdet i dokumentet, og den settes ofte lik ordets frekvens i dokumentet.

I begge disse tilfellene vil det mest riktige være å la rotlemmaet danne utgangspunktet for vektleggingen. Så lenge det ikke er noen grunn til å skille mellom bøynings- og avledningsformene til et ord, er det heller ingen grunn til å la den få forskjellige vekt. Dessuten vil man få en helt feil vurdering av ordet hvis man bare tar utgangspunkt i hyppigheten til det enkelte graford, og ikke hyppigheten til rotlemmaet. Dette framgår tydelig hvis man sammenligner en frekvensordliste sortert på grafordenes frekvens og en liste sortert på rotlemmaenes frekvens. I tabellen nedenfor har vi gitt et eksempel på to slike frekvensordlister basert på en fysikktekst bestående av ca. 8450 løpende ord og ca. 1650 ulike graford. Vi har strøket funksjonsordene og står følgelig igjen med bare innholdsord. Kolonnen til venstre viser de 30 hyppigste rotlemmaene, mens kolonnen til høyre viser de 38 hyppigste grafordene. Det fremgår at 8 av grafordene er en bøynings- eller en avledningsform av andre ord på lista, derfor måtte vi ha 38 graford for å dekke 30 rotlemmaer også i denne kolonnen.

Tabell 6.3/1 De 30 hyppigste rotlemmaene (eksklusiv funksjonsord)
med utgangspunkt i rotlemmaets og grafordets frekvens.

ROTLEMMÅ		GRAFORD	
	Frekvens	frek.	frek.
VANN	134	78 LUFTA	13 VARME
LUFT	119	70 VANNET	12 DAG
VARM	91	59 VANN	12 <u>KALLER</u>
SE	42	41 LUFT	12 <u>KJELEN</u>
STOR	38	22 MOLEKYLENE	11 <u>LIKE</u>
IS	37	22 VARMEN	11 <u>VISER</u>
BRUK	32	21 GRADER	
MOLEKYL	29	21 SAMMEN	
VANNDAMP	29	19 IS	
<u>FINNE</u>	28	18 KOMMER	
<u>TA</u>	27	18 SER	
<u>STED</u>	26	17 VARM	
KOMME	25	16 BRUKER	
GRAD	24	16 FRYSER	
<u>KALD</u>	24	16 ISEN	
<u>LITEN</u>	24	16 SE	
FRYSE	23	16 SETT	
<u>STOFF</u>	23	16 <u>TING</u>	
<u>LAGE</u>	22	16 VARMT	
<u>LANG</u>	22	15 SIER	
SMELTE	22	15 STORE	
<u>DEL</u>	21	15 VANNDAMPEN	
SAMMEN	21	14 <u>FJELLET</u>	
DAG	20	14 <u>KARTET</u>	
<u>GOD</u>	20	14 SMELTE	
<u>KOKE</u>	20	14 <u>SOLA</u>	
SI	20	14 <u>STIGER</u>	
<u>GJØRE</u>	19	14 <u>STØRRE</u>	
<u>SETTE</u>	19	14 VANNDAMP	
<u>TERMOMETER</u>	19	13 <u>JERN</u>	
		13 <u>JORDA</u>	
		13 <u>LIGGER</u>	

Ordene som er uthevet, forekommer bare på den ene lista. Tabellen viser at blant de 30 mest frekvente rotlemmaene, finner vi igjen bare 16 blant de hyppigste grafordene, dvs. bare litt over halvparten er felles for begge listene.

6.4 Automatisk rotlemmatisering som ingrediens i andre hjelpemidler

(1) Innledning

I de foregående avsnitt pekte vi på to viktige anvendelsesområder hvor selve rotlemmatisering sto sentralt. Det finnes imidlertid også en rekke anvendelser hvor metoden for rotlemmatisering kan inngå som et ledd i andre oppgaver. Vi skal se nærmere på tre slike områder.

(2) Automatisk trunkering

I avsnitt 6.2 var vi inne på forskjellen mellom rotlemmatisering og trunkering, og konkluderte med at trunkering kan være ønskelig ved siden av automatisk rotlemmatisering. En forutsetning for denne påstanden er at trunkeringen blir utført riktig.

Erfaringer viser at denne forutsetning ofte ikke er oppfylt. Spesielt blant uerfarende brukere - men også blant eksperter - finner vi mange eksempler på at man ikke tenker godt nok gjennom konsekvensene av trunkeringen før man trunkerer. Vi har også vært vitne til at det trunkeres bare etter et par bokstaver for å dekke alle bøyings- og avledningsformer. Konsekvensen av dette er at resultatet blir dårlig og søkingen nokså kostbar.

På denne bakgrunn ble det som et ledd i prosjekt FORT gjennomført et forsøk med automatisk trunkering hvor metoden for rotlemmatisering ble lagt til grunn.

Forsøket gikk ut på at man lot søkesystemet trunkere den oppslagsformen som ordet fikk ved automatisk rotlemmatisering. Dette skjedde uten å tilpasse oppslagsformen spesielt til dette formålet eller ved å la systemet holde kontroll med trunkeringen.

Resultatet av trunkeringen ble vurdert av de samme personer som var med på forsøket med den manuelle trunkeringen (jfr. ovenfor), og til slutt ble det sammenlignet med resultatene fra den manuelle trunkeringen og den automatiske rotlemmatiseringen.

Resultatet av sammenligningen var oppsiktsvekkende - det viste at den automatiske trunkeringen faktisk ga litt bedre resultater enn den manuelle trunkeringen. Umiddelbart synes dette for utrolig, spesielt fordi oppslagsformen ikke var tilpasset dette formålet. En analyse av resultatene ga oss imidlertid en forklaring som bekreftet at resultatet var bedre.

I de tilfellene hvor ordet var sammensatt av en grunnform etterfulgt av en eller flere suffikser, var oppslagsformen identisk med grunnformen til ordet. Det er også denne formen av ordet man normalt vil trunkere, og her ga den automatiske trunkeringen maksimalt resultat. Ved den manuelle trunkeringen forekom det derimot eksempler på at deltagerne trunkerte en mindre del av ordet for å rasjonalisere skrivingen. Dette ga ofte flere irrelevante dokumenter uten at antall relevante økte.

Problemene oppsto i de tilfeller hvor ordet var mer komplekst sammensatt, f.eks. ved uregelmessige bøyninger (MANN - MENN), stavelsessammentrekninger (REGEL - REGLER), konsonantforenklinger (SYKKEL - SYKLER) eller -fordoblinger (DAM - DAMMEN). I slike tilfeller må man enten trunkere en forholdsvis liten del av ordet, slik at alle bøyninger og avledninger blir fanget opp, eller trunkere flere ordformer. Ved den manuelle trunkeringen valgte man ofte det første alternativet - noe som førte til mange irrelevante ord. Ved automatisk trunkering ble bare en av ordformene trunkert (f.eks. SYKL) og man mistet dermed en del relevante bøynings- og avledningsformer. Totalt sett ga den automatiske trunkeringen 1% flere relevante ord og 8% færre irrelevante ord.

Den automatiske trunkeringen kan imidlertid gjøres langt mer effektiv enn det som ble testet i dette forsøket. Man kan f.eks. sørge for at alle "varianter" av roten blir trunkert (f.eks.

SYKKEL og SYKL), at ord som ikke bør trunkeres heller ikke blir det (f.eks. flertallsformene av uregelmessige substantiv eller sterke verb) og at oppslagsformen blir bedre tilpasset dette formålet. På denne måten vil man få et hjelpemiddel som er bedre enn dagens trunkering både når det gjelder gjenfinningsgrad og presisjon.

(3) Automatisk splitting av sammensatte ord

Sammensatte ord representerer et problem i tekstsøking fordi flere sammensatte ord også kan skrives om til en frase, f.eks. ARVEAVGIFT kan formuleres som AVGIFT PÅ ARV. En metode for automatisk identifisering og splitting av sammensatte ord vil derfor kunne være et nyttig hjelpemiddel ved formulering av søkeargumentet i tekstsøking. Man kan f.eks. tenke seg at systemet gjør brukeren oppmerksom på sammensatte ord eller foretar en splitting automatisk hvis dette er ønskelig.

Både ved identifisering og splitting av sammensatte ord vil metoden for automatisk rotlemmatisering komme til nytte. De studier av sammensatte ord, som har vært gjennomført i prosjektet FORT, viser at man ved å rotlemmatisere ordene kan skille ut de sammensatte med et godt resultat. Splittingen av det sammensatte ordet i de enkelte leddene vil kreve at man kartlegger ordets morfologiske og fonologiske oppbygging, og deler av regelsettet ville derfor kunne benyttes.

(4) Automatisk ordklassebestemmelse

Informasjon om hvilken ordklasse et ord tilhører kan være nyttig i flere sammenhenger i tekstsøking. I prosjektet STANS ble denne informasjonen bl.a. benyttet til å kartlegge hvilke ord som var egnet som søkeord i spørsmålene. F.eks. vil substantivene ofte være mer aktuelle som søkeord enn verbene, og de ordklassene som funksjonsordene hører til, vil sjelden være aktuelle som søkeord hvis de ikke sees i sammenheng med andre ord. Likeledes vil informasjon om ordklassen være nødvendig ved analyse av søkeargumenter i naturlig språk og generelt ved kartlegging av fraser i en tekst, f.eks. nominalfraser.

Regelsettet i dag bringer oss langt på vei mot å kunne identifisere ordklassen til et ord automatisk. Enkelte regler inneholder all informasjon om hvilke ordklasser reglene er typiske for, f.eks. er suffiksen -ENE en typisk bøyningsendelse for substantiv flertall. En bearbeiding og utvidelse av regelsettet er imidlertid nødvendig hvis metoden skal kunne anvendes til å identifisere ordklassen til et ord. Selv om man skulle ha problemer med å bestemme ordklassene ut fra suffiksene (f.eks. -ER), er det ikke sikkert at dette har så stor betydning for den bruken av ordklasse-informasjonen som man tenker seg i forbindelse med tekstsøking. I noen tilfeller vil det være tilstrekkelig å presentere de alternativene som foreligger. Dessuten kan man kombinere metoden med andre framgangsmåter, og skille ut det riktige alternativet på denne måten. Gunnel Kallgren har gått fram på en liknende måte i sin automatiske setningsanalysemodell. Hun kjører forskjellige programmer sekvensielt og lar hvert program utnytte analysen i de foregående. På denne måten reduserer hun alternativene for hver gang.

6.5 Forbedring av metoden

(1) Vurdering av dagens metode

De sentrale spørsmålene i diskusjonen om bruk av automatisk rotlemmatisering i tekstsøking, er hvorvidt metoden er god nok eller eventuelt kan gjøres bedre - og hva som vil skje ved oppdatering eller ved overgang til et nytt tekstmateriale.

Om metoden er god nok får erfaring vise, men det avhenger selvfølgelig også av hvilke alternativer man har. Som nevnt tidligere, er trunkering ofte det eneste hjelpemiddelet, og eksperimentet med trunkering synes å tyde på at den automatiske rotlemmatiseringen absolutt har sin berettigelse. Før vi forlater dette spørsmålet skal vi kort se nærmere på de feilene som oppsto ved automatisk rotlemmatisering av eksperimentmaterialet, og hvilke konsekvenser de vil ha for tekstsøking.

Automatisk rotlemmatisering av eksperimentmaterialet viste at 2 - 3% av grafordene ble feil plassert og ca. 2,5 % av rotlemmaene var enten "ufullstendige" eller "overfylte". Ca. 75% av feilene førte til at rotlemmaet var blitt splittet, slik at man ved utvidelse av søkeargumentet ville ha utelatt bøyings- og avledningsformer. I ca. halvparten av disse tilfellene var det imidlertid bare ett ord som skilte seg ut (ofte en bøyordform), slik at man ved å bruke de andre ordene innen rotlemmaet bare ville ha mistet dette ene søkeordet. I de resterende 25% av tilfellene ville feilene ha ført til at irrelevante ord ble lagt til søkeargumentet, dvs. ord fra andre rotlemmaer.

Den alvorligste feiltypen er en kombinasjon av disse to, dvs. at utvidelsen av søkeargumentet ville ha ført både til at irrelevante ord ble lagt til søkeargumentet og at relevante bøyings- og avledningsformer ble utelatt. Denne feiltypen utgjorde heldigvis bare 10% av feilene.

Av dette følger at i overkant av 1% av grafordene (brukt som søkeord) ville ha ført til at søkeargumentet ikke ble utvidet med alle bøyings- og avledningsformer. I underkant av 1% ville ha ført til at irrelevante ord ble lagt til søkeargumentet.

(2) Reduksjon av eksisterende feil

Det finnes flere tiltak som kunne ha vært gjennomført for å redusere antall feil. Blant annet kunne man merke reglene slik at man unngår at ord som er behandlet av visse typer regler blir gruppert sammen. Dette gjelder spesielt de reglene som tar hånd om unntaksordene og enkelte andre spesialregler. På den måten vil man f.eks. kunne hindre at ordet TÅ blir gruppert sammen med ordet TÆRE, ved å merke regelen for TÆR. En liknende ordning er allerede innført for funksjonsord.

Ca. 1/3 av feilene er graford som ikke har fått tilslag på noen regler. Disse feilene har vi systematisert, og det er klart at noen feil kan vi unngå ved å definere nye regler. At disse reglene ikke er blitt lagt inn i regelsettet, kommer av at bruks-

frekvensen er for lav i vårt materiale, og at reglene ville ha blitt for spesielle i forhold til vår generelle målsetning.

En forbedring av regelsettet på dette nivå vil i stor grad være et spørsmål om hvor mange regler man er villig til å ha i regelsettet. Jo flere regler man har, jo mer ressurskrevende vil metoden være. Nå ser vi imidlertid ikke bort fra at det er mulig å redusere antall regler i dagens regelsett ved en ytterligere bearbeiding av regelsettet.

(3) Overgang til et nytt materiale

I arbeidet med regelsettet har vi lagt vekt på at reglene skal være så generelle som mulig. De fleste reglene er også satt opp på bakgrunn av norsk morfologi, men enkelte av spesialreglene er definert på statistisk grunnlag. I dag har vi for liten erfaring med andre tekstmaterialer til å kunne garantere at disse spesialreglene er tilstrekkelig for et hvilket som helst materiale. Eksperimentmaterialet besto av tekster fra fire forskjellige faggreiner (se avsnitt 4.1). Da vi utviklet reglene, tok vi ikke vare på informasjonen om reglenes anvendelse i de ulike fagene - bare bruksfrekvensen totalt. Når vi nå har skrevet rapporten har vi savnet denne informasjonen, spesielt for å kunne avgjøre hvor generell vi har klart å gjøre metoden. Vi mener jo at reglene er generelle og de er definert med dette formålet for øyet. Likevel skulle vi gjerne ha kunnet vist det med tall.

Vi skal imidlertid ikke se bort fra at det vil melde seg et behov for nye spesialregler ved overgangen til et nytt materiale, spesielt hvis dette materialet faller inn under et helt annet fagområde enn eksperimentmaterialet. F.eks. innen oljesektoren anvendes mange engelske termer, og for å kunne behandle disse vil det antagelig være nødvendig med nye regler.

I dag inneholder regelsettet 20 fagord. Dette er ord som ellers ville ha blitt feil gruppert, men som på grunn av sin betydning og hyppighet (bl.a. i sammensetninger) var berettiget en plass på

lista. Ved overgangen til et nytt materiale innenfor ett annet fagområde, vil disse reglene antagelig være overflødige. Man kan da legge inn nye regler som er aktuelle for det nye fagområde ved f.eks. å gå gjennom de mest frekvente fagordene i en frekvensordliste. Det er ingen grenser for hvor mange ord man kan legge inn i regelsettet, bare et spørsmål om hvor mange regler man ønsker.

Når det gjelder de øvrige unntaksordene i regellista, tilhører disse lukkede ordklasser, f.eks. sterke verb, uregelmessige substantiv, visse typer adjektiv og adverb. Vi har ikke tatt mål av oss å legge inn alle ordene innenfor de lukkede ordklasser, og i dag består regelsettet kun av de ordene som forekommer i eksperimentmaterialet.

(4) Endringer i regelsettet ved oppdatering av tekstmaterialet

Hvis rotlematiseringen fungerer tilfredstillende for et forholdsvist stort materiale, er det ingen grunn til å tro at resultatene vil bli noe vesentlig dårligere ved oppdatering av materialet. Men vi skulle gjerne ha undersøkt hva som skjer med regelfrekvensen når tekstmaterialet øker, dvs. når nye dokumenter legges inn i basen. Vi skulle gjerne ha fått bekreftet at hovedreglene og spesialreglene vil dekke de fleste nye ordene, men vi ville gjerne se om unntakene måtte utvides utover det å fullstendig gjøre de endelige ordgruppene som vi allerede har lagt inn. Materialet burde derfor ha vært delt opp i mindre deler, slik at vi kunne ha undersøkt problemet med tekstutvidelse. Men her møter man det problemet at en tekst må være av en viss størrelse før man i det hele tatt kan gjøre noen verdifulle analyser når det gjelder ordforråds- og ordstrukturstudier. Ordforrådet ekspanderer jo voldsomt så lenge tekstmengden er liten, men økningen blir langt mindre etter hvert (jfr. Harvold 1980). Variasjonen i ordstrukturen kan man anta vil øke i takt med (om ikke så raskt som) antall ulike ord i begynnelsen, men sannsynligvis vil denne økningen avta mye fortere enn økningen av ulike ord. Fordi man må over en viss "grense", vil det si at man må arbeide med så store materialmengder at det å teste ut

resultatene når tekstene øker, er svært tidkrevende. Over tid vil det selvfølgelig komme inn nye ord i språket, men strukturen vil som oftest være kjent, dvs. ordene vil følge de reglene vi allerede har. De største problemene vil skapes av fremmedord og av ord som blir leksikalisert. Det er imidlertid rimelig å anta at disse ordene i første rekke vil være fagord og derfor ikke skape generelle problemer.

I denne sammenheng hadde det vært nyttig med en tilleggsmodul til systemet som kunne overvåke rotlemmatiseringen og selv fremme forslag til regler. F.eks. kunne den holde øye med hvor mange graford som ikke ble behandlet av regelsettet, systematisere disse og motta informasjon fra systemadministratoren om hvilke andre ord som var relatert til disse. Man kunne også gi brukeren mulighet til å registrere feil eller la systemet registrere "sannsynlige" feil ut fra brukerens handlingsmønster. Denne informasjon kunne så videre blitt brukt til å fremme forslag om endringer av regelsettet. Dette kommer vi tilbake til i avsnittet nedenfor.

(5) Bruk av metoden

I de fleste søkesituasjoner vil det være ønskelig at søkeargumentet ble supplert med alle bøynings- og avledningsformer til søkeordene. Dette taler for at systemet automatisk burde supplere søkeargumentet med alle ordformer uten at brukeren ga beskjed om det. Brukerne bør imidlertid få muligheten til å slå av denne funksjonen, og ved bruk av fraser, avstandsoperatorer o.l. bli spurt om den skal stå på.

Brukeren bør også kunne få se hvilke ord søkeargumentet blir supplert med. I tillegg kunne man tenke seg at systemet automatisk viste brukeren de rotlemmaene som det var knyttet usikkerhet til, f.eks. som var behandlet av regler som vi viste ga en del feil. På en enkel måte kunne brukeren få muligheten til å fjerne eller supplere søkeargumentet med bøyningsformer. Denne "feedback"-informasjonen kunne man så registrere og utnytte ved senere forbedringer av metoden for automatisk rotlemmatisering.

7. ANDRE ANVENDELSER AV METODEN

7.1 Innledning

Tidligere i denne rapporten har anvendelsen av automatisk rotlemmatisering kun vært diskutert i relasjon til tekstsøking, og det var også med tanke på tekstsøking at vi besluttet å utvikle en slik metode. Det finnes imidlertid en rekke andre områder hvor en slik metode vil være et velkomment verktøy. Vi kan egentlig si det så generelt at i enhver form for analyse av ordforrådet i en tekst, hvor ordenes innhold er av interesse, vil det eksistere et behov for å gruppere ord med felles rot eller grunnform. Nedenfor skal vi se nærmere på et par andre tilfeller hvor en metode for automatisk rotlemmatisering kommer til stor nytte.

7.2 Datamaskinstøttet stikkordproduksjon

I de fleste av dagens fagbøker og oppslagsverk finner man et stikkordregister bakerst i boka som angir hvor i boka f.eks. et bestemt emne eller et begrep er omtalt. Produksjon av slike registre er i dag i liten grad automatisert. Det finnes eksempler på at man har produsert dem ved å framstille en alfabetisk sortert liste over alle ulike ord i dokumentasjonsmaterialet, og så gjennom en postediting (eventuelt assistert av et særskilt programsystem) laget et register. Dette representerer en betydelig effektivisering, men fremdeles ligger hovedvekten på den manuelle innsatsen. En vanlig framgangsmåte er at forfatteren eller forlaget går gjennom teksten og merker stikkordene i teksten, og deretter framstiller et register på grunnlag av de merkede ordene. Denne delen av produksjonen lar seg lett automatisere, og enkelte tekstbehandlingssystemer gir også mulighet for dette.

Et slikt register har mye til felles med de rent

datamaskinbaserte registrene som man finner i tekstsøkesystemer, og det er derfor grunn til å tro at man kan utnytte de samme metodene som ved automatisk indeksering. En viktig forskjell mellom dem er at et bak-i-boka register ikke kan være for langt, ettersom det skal anvendes manuelt og trykkes bakerst i en publikasjon. Man må derfor stille strengere krav til utvelgelsen av stikkordene, og ordene må få en hendig oppslagsform.

Skal man produsere et bak-i-boka register automatisk, kreves det mer enn en metode for automatisk rotlemmatisering. Allikevel vil en slik metode kunne være til nytte med de enkle hjelpemidler man har i dag. Man kunne f.eks. ha implementert det i et tekstbehandlingsprogram, slik at man etter å ha merket av de aktuelle stikkordene, fikk presentert en liste over de rotlemmaene som var valgt ut, representert ved en oppslagsform. Her vil det ikke være tilstrekkelig at oppslagsformen er entydig for rotlemmaet - det kreves også at den tilsvarer den formen vi forventer å finne. I mange tilfeller er det naturlig å anvende substantiv som oppslagsord, og dette skulle kunne la seg gjøre med et relativt godt resultat forutsatt at man utvidet metoden for rotlemmatisering.

For øvrig bør det tilføyes at det med den teknikken vi har i dag for automatisk indeksering, er mulig å utvikle mer effektive datamaskinstøttede hjelpemidler for stikkordproduksjon enn bare bruk av automatisk rotlemmatisering. Fullstendig automatisering vil være vanskelig - ja, kanskje umulig - på grunn av homograf-problemet og de krav som det ofte stilles til en hierarkisk oppbygging av et register.

7.3 Lingvistisk hjelpemiddel

En metode for automatisk rotlemmatisering vil være nyttig i mange lingvistiske studier. Vi nevnte i forordet at man i prosjektet LÆREBOKSPRÅK hadde rotlemmatisert ordene manuelt for å finne fram til hvilke som ble hyppigst brukt i visse lærebøker. Dette arbeidet var tidkrevende og forutsatte en stor grad av presisjon

fra dem som utførte det, men det ville ha vært meningsløst å bygge et slikt studium på de forskjellige ordformenes frekvens. En del ord ville da bli regnet som lavfrekvente fordi de opptrådte i mange forskjellige bøynings- og avledningsvarianter som hver hadde lav frekvens, mens de ordene som kun hadde en form i materialet ville ha bli favorisert med hensyn til frekvensen (jfr. tabell 6.3/1).

Hvis man i et lingvistisk studium var interessert i å analysere frekvensen av ordforrådet i en tekst ut fra ordenes etymologi, ville man kunne bruke den samme fremgangsmåten, men reglene måtte da justeres til det nye formålet. I det tilfellet ville f.eks. bøyningsreglene være de samme som i vårt regelsett, mens avledningsreglene ikke behøver å ta hensyn til om ordene ble leksikalisert eller ikke.

Den samme framgangsmåten, men med en justering og tildels forandring av reglene ville åpne for en rekke lingvistiske studier. Vi har allerede nevnt at reglene slik de foreligger i dag, langt på vei gir opplysninger om ordets ordklasse-tilhørighet. En justering av regelsettet ville kunne fullføre denne analysen.

Man kunne også tenke seg at bruksfrekvensen til de forskjellige reglene kunne gi interessante opplysninger om teksten. Hvis man laget et regelsett som i første rekke konsentrerte seg om å kjenne igjen de forskjellige tempusformene til verbene (presens og preteritum), kunne man på grunnlag av regelfrekvensene få en indikasjon på om teksten var fortellende eller konstaterende i sin form, osv. I det hele tatt vil et hvert studium som gjør bruk av ordenes "utseende", kunne ha nytte av framgangsmåten til den automatiske rotlemmatiseringen.

7.4 Videreføring

Som en følge av de erfaringene vi selv har ervervet oss og den responsen vi har fått på de mulighetene vi øyner for metoden, ser vi tre forskjellige retninger hvor vi ønsker å arbeide videre:

- 1) teste ut metoden, slik at vi får svar på de ubesvarte spørsmål som er blitt stilt.
- 2) gjøre metoden mer fleksibel, slik at den lettere kan adapteres til spesialtekster.
- 3) benytte framgangsmåten i andre oppgaver.

Pkt. 1. Som tidligere nevnt, har vi ikke har fått tallfestet eller kontrollert at alle spesialreglene virkelige er generelle og faguavhengige. Heller ikke om de er tilstrekkelige i den forstand at det ikke er nødvendig lage nye regler til et nytt og annerledes materiale. Det er derfor ønskelig å studere reglene nærmere i forhold til et annet materiale, f.eks. hvordan frekvensen blir og hvor dekkende reglene er. Vår hypotese er at de er generelle nok slik de foreligger i dag, og testmaterialet bekrefter også at de fungerte bra for en helt annen type materiale. Vi vurderte imidlertid ikke dette resultatet med hensyn til regel- og feiltyper. Det vil derfor være interessant å sammenlikne regelfrekvensene og regelanalysene i vårt eksperimentmateriale med frekvensene i andre materialer fra andre fagområder og av forskjellig omfang. Det siste ville kunne gi oss svar på hva som skjer ved oppdatering av tekstmengdene - et spørsmål vi har blitt stilt flere ganger.

Pkt. 2. Vi har lagt vekt på å lage en generell metode, men det kan også være aktuelt å skreddersy et regelsett til et spesielt fagområde eller stiltype. For at dette skal la seg utføre i hvert enkelt tilfelle, ville det være hensiktsmessig å dele reglene i forskjellige pakker. Man kan f.eks. tenke seg en basispakke som var "grunnsten" i regelsettet, dessuten diverse spesialpakker som man kunne velge i tillegg til denne, alt etter formålet med rotlematiseringen og tekststypen man arbeidet med.

Hvis man arbeider med et tekstgrunnlag skrevet på riksmål, behøver man f.eks. ikke ha med regler som behandler hunnkjønnsord - eller hvis tekstgrunnlaget er moderne skolebøker, er det unødvendig med regler for forskjellige skriftvarianter utover de ordformene som er sidestilt i normalplaner for skoleverket.

Pkt. 3. Vi har allerede nevnt flere områder innen tekstsøking hvor framgangsmåten for automatisk rotlemmatisering vil være et ledd i løsningen av diverse oppgaver. Vi tenker nå på oppgaver som gjenkjenning og splitting av sammensatte ord, automatisk ordklassebestemmelse og registerproduksjon. I tillegg finnes det andre oppgaver innen tekstbehandling og ordanalyse hvor en morfologiske analyse av ordene vil kunne utnyttes, slik som retting av ortografiske feil, automatisk orddeling osv. Til disse oppgavene må nok regelsettet forandres og andre komponenter legges inn. Selve framgangsmåten, den erfaringen vi har ervervet når det gjelder ordenes struktur og ikke minst det store materialet vi har samlet og kategorisert, vil imidlertid kunne brukes i løsningen av disse oppgavene.

LITTERATURLISTE

- Attar, R/Fraenkel A.S (1980) Experiments in Local Metrical Feedback in Full-Text Retrieval Systems", Information Processing and Management, Vol. 17, pp 115-126, 1981.
- Allen, Sture (1971) Nusvensk frekvensordbok 2.
- Banerjee, Nadakishore (1977) "CONDOR - Communication in Natural Language with Dialogue Oriented Retrieval Systems"; Schneider/Sågvall Hein 1977:163-175.
- Bing, Jon/Fjeldvig, Tove, Flataker, Ole B./Harvold, Trygve (1980) Lovspråk og juristspråk. Skriftserien Jus og EDB nr. 41. Oslo
- Bing, Jon/Selmer, Knut S. (1980) A Decade of Computers and Law. Universitetsforlaget. Oslo.
- Borgstrøm, Carl Hj. (1975) Innføring i sprogvitenskap. Universitetsforlaget. Oslo.
- Ghoueka, Yaacov (1980) "Computerized Full-Text Retrieval Systems and Research in the Humanities: The Responsa Project", Computers and the Humanities volume 14, No. 3, 1980.
- Fjeldvig, Tove (1984) Tekstsøking - Teori, metoder og systemer. Complex 5/84. Oslo.
- Fjelovig, Tove/Golden, Anne (1983) "Automatisk rotlematisering"; Hein 1983:84-103.
- Fluhr, C. (1981) "Linguistic and Statistical Processing of texts and its Application in the Field of Legal Documentation"; Proceedings of the Sixth Symposium on Legal Data Processing in Europe 1982:42-64.

- Harvold, Trygve (1974) "Belysning av synonymproblemet i norske formuerettslige lover"; Bing/Fjeldvig/Flataker/Harvold 1980:133-146.
- Harvold, Trygve (1980b) "Performance of Text Retrieval Systems"; Bing/Selmer 1980:172-219.
- Hein, Anna Sågvall (1983) Foredrag vid De nordiske datalingvistik dagarna 1983. Uppsala universitet.
- Hellberg, Staffan (1971) "Automatisk lemmatisering - En modell for upprättande av bojningsserier i ett frekvenslexikon". Språkdata, Goteborg.
- Salton, Gerard (1968) Automatic Information Organization and Retrieval. McGraw-Hill computer series.
- Salton, Gerard/McGill, Michael J. (1983) Introduction to modern Information Retrieval. McGraw-Hill coputer series.

Tidligere utgitt i skriftserien Complex

Complex 1/81

John Hansen

Et EDB-system for analyse av rettslige avgjørelser

Complex 2/81

John Hansen (red)

Notater om deontiske systemer

Complex 3/81

Vidar Sørensen

Informasjonssystem mm for Norges Geotekniske Institutt

Complex 4/81

Kjetil Johnsen

Systemtekniske konsekvenser av persondatalovgivningen

Complex 6/81

Harit Thorvaldsen

Teledata og rettsinformasjon

Complex 8/81

Anne Kirsti Brække

Postmonopolet

Complex 7/81

Knut S. Selmer (ed)

The LAWDATA Papers

Complex 1/82

Norsk forening for jus og edb & Svenska föreningen för rätts och juridik

Nordiske personregisterlover, Council of Europe Convention and OECD Guidelines

Complex 2/82

Harald Brønd og Vidar Lørensen

Oljevern: brukerbehov og kildemateriale

Complex 3/82

Jon Bing og Dag Frøystad

Rettskildebruk

Complex 4/82

Thomas Prebensen Steen

Post- og televerkets regulerte ansvar

Complex 5/82

Datatilsynet

Arsmelding 1981

Complex 6/82

Birger Eckhoff

i samarbeid med Jon Bing, Dag Frøystad og Anja Oskamp

CATTS: Computerized program for teaching text retrieval systems

Complex 7/82

Jon Bing

Informasjonssystemer for Trygderettens kjennelser

Complex 8/82

de Mulder, Oskamp, van der Heyden og Gubbj

Sentencing by computer: An experiment

Complex 9/82

Colin Tapper

An experiment in the use of citation vectors in the area of legal data

Complex 1/83

Arve Føyen

Utredning om endringer i personregisterloven

Complex 2/83

Stein Schjølberg

Computers and penal legislation

Complex 3/83

John S Gulbrandsen og Terje Hoffmann

Rettigheter i idrettsarrangementer

Complex 4/83

Sally Moon and Anja Oskamp

The Law of Legal Information Systems: Two essays

Complex 5/83

Datatilsynet

Arsmelding 1982

Complex 6/83

Olav Torvund

Betalingsformidling

Complex 7/83

Else Ryen

Lov og lovmottaker

Complex 8/83

Daniel Stripinis

Probability theory and circumstantial evidence

Complex 9/83

Frede Cappelen

Edb/basert informasjonssystem for forvaltningens praksis

Complex 10/83

NORDIPRO

Legal acceptance of international trade data

Complex 11/83

Jørgen Hafstad og Thomas Prebensen Steen
Teleksrett og merverdiavgift på programvare

Complex 12/83

Kristin Kjelland-Mørde
Om forenkling av regler

Complex 13/83

Jon Bing
Edb: muligheter og problem ved forenkling av regelverk

Complex 14/83

Tarjei Stensaasen (red)
Utvalgte emner i jus og edb (3. utg)

Complex 15/83

Anette Klafstad og Ulf Alex. Samer
Plan for et forsikringsrettslig informasjonssystem

Complex 1/84

Jon Erling Skjørshammer
Kabelnett: Bygnings- og ekspropriasjonslov
TERESA (33)

Complex 2/84

Tore Andreas Hauglie og Dag Wiese Schartum
Forslag til et helserettslig informasjonssystem

Complex 3/84

Justisdepartementet
Den elektroniske grunnbok

Complex 4/84

Lututilsynet
Årsmelding 1983

Complex 5/84

Tove Fjeldvig

Tekstsøking: Teori, metoder og systemer

Complex 6/84

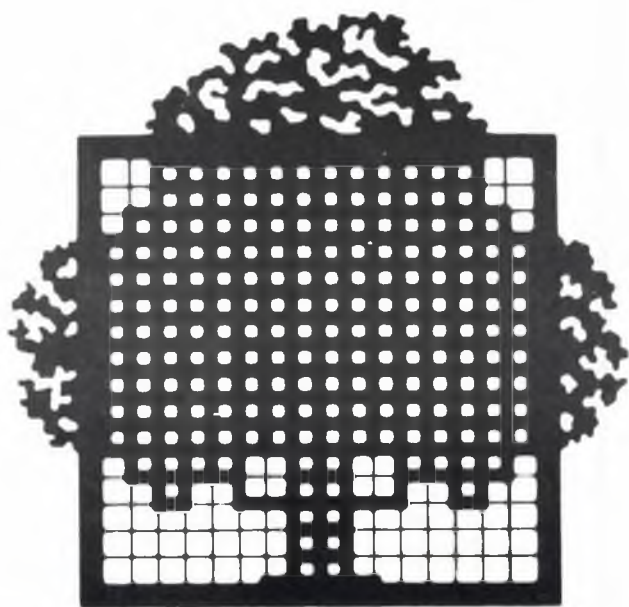
Jon Bing

Offentlighetsloven og edb

TERESA (30)



DSA – nettverksarkitektur for fremtiden



Honeywell Bull

Tollbugt 32 – Oslo 1 Tlf. (02) 41 80 30
Avd kontor: Sandnes – Bergen – Ålesund – Trondheim – Harstad – Tromsø



Norsk Data

ISBN 82-00-07035-2

GCS, Oslo