Forsk.stip. Tove Fjeldvig og cand.philol. Anne Golden
Institutt for rettsinformatikk
Universitetet i Oslo
Niels Juelsgt. 16
Oslo 2

Oslo, 1. juni 1983

# AUTOMATISK ROTLEMMATISERING

## 1. Prosjekt for automatisk rotlemmatisering

Institutt for rettsinformatikk (IRI) (tidligere Institutt for privatretts avdeling for EDB-spørsmål) har i mange år drevet forskning omkring tekstsøkesystemer. I ett av disse prosjektene har vi spesielt tatt opp ulike lingvistiske aspekter knyttet til denne type systemer. Foreløpig har arbeidet vært konsentrert omkring utvikling av en metode for gruppering av ord med felles rot på tvers av ordklassene. Prosessen har fått navnet "automatisk rotlemmatisering", bl.a. for å skille den fra den mer vanlige lemmatiseringsprosessen som opererer innenfor de tradisjonelle ordklassegrensene.

> Et lemma kan sammenlignes med et slags stikkord eller et oppslagsord i en ordbok. At to ord tilhører samme lemma, betyr enten de er ulike bøyningsformer av samme grunnform (leksem) eller at de er to ulike skriftvarianter av samme leksikalske ord (f.eks. fram og frem). Et rotlemma vil derfor være en betegnelse på ord som har samme rot og samme semantiske betydningen når man ser bort fra den informasjon som ligger i selve bøynings- og avledningsendelsen.



Automatisk rotlemmatisering
Tove Fjeldvig, Anne Golden
Proceedings of NODALIDA 1983, pages 84-103

## 2. Bakgrunn

Arbeidet med gruppering av ord med felles rot ble allerede påbegynt i 1979 som en aktivitet under prosjekt NORIS (34) ved IRI. Dette prosjektet, som var finansiert av Norges Teknisk-Naturvitenskapelige Forskningsråd og ledet av cand.mag. Tove Fjeldvig, tok sikte på å undersøke muligheten for enkle strategier for tekstsøking basert på argumenter i naturlig språk.

Blant de problemer man ønsket å belyse i dette prosjektet, var muligheten for automatisk utvidelse av søkeargumentet med alle aktuelle bøyningsformer til søkeordene. Også avledningsformene var aktuelle forutsatt at ordene representerte det samme innholdet (grunnidéen).

> I et tekstsøkesystem vil i prinsippet alle ordene i dokumentene være søkbare. Det vil bl.a. si at en bruker må selv definere alle mulige bøyninger og avledninger av aktuelle søkeord. Hvis man for eksempel bare angir søkeordet "BIL", vil man ikke finne de dokumenter som inneholder ordene "BILEN", "BILER" eller "BILENE".

Man fant det også interessant å undersøke om en slik rutine representerte et alternativ til manuell høyre-trunkering, eller om den kunne inngå som et ledd i en rutine for automatisk trunkering.

> Trunkering er en måte å spesifisere søkeord på ved å definere en viss følge av tegn som søkeordet skal inneholde. Alle ord som inneholder den definerte tegnstrengen anses kvalifisert som søkeord. Den mest vanlige form for trunkering er høyre-trunkering, hvor tegnstrengens høyre del er uspesifisert. Søkeargumentet BIL* (hvor * er brukt som trunkeringstegn) vil omfatte alle ord som begynner med bokstavene "bil", f.eks. BILER, BILEN, BILENE, BILHOLD, BILDE, BILLION. Ulempen med trunkering er at den medfører en del støy og ikke inkluderer vokalvekslinger (f.eks. sterke verb og uregelmessige substantiv).

Dessuten ville en slik rutine gjøre analyse av et søkeargument i naturlig språk lettere, og dermed også øke muligheten for et bedre søkegrunnlag.

I NORIS (34) ble det etterhvert et spørsmål om automatisk rotlemmatisering. Ved prosjektets opphør i 1981 fant vi resultatene såpass interessante, at vi ønsket å fortsette studiene.

Samtidig med prosjekt NORIS (34) pågikk det også et prosjekt LÆREBOKSPRÅK ved Nordisk institutt (UIO) hvor man var opptatt av lemmatiseringsproblematikken. Prosjektet var ledet av amanuensis Anne Hvenekilde og cand. philol. Anne Golden og finansiert av Kirke- og undervisningsdepartementet. Formålet med prosjektet var å kartlegge de høyfrekvente ordene i en del fagbøker for grunnskolen, slik at det kunne lages støttemateriell i norsk for innvandrerelever. Støttemateriellet skulle i første omgang konsentrere seg om vokabularet i fagbøkene, og man ønsket derfor å finne fram til de ordene som de fremmedspråklige elevene fikk mest nytte av å ha lært ved lesing av fagbøker i skolen. I dette arbeidet var det ikke tilstrekkelig å ta utgangspunkt i grafordenes frekvens - og heller ikke lemmaets frekvens (dvs. den samlede frekvens for de ulike bøyningsformer av samme grunnord). Det riktigste bilde av ordforådet fikk man ved å beregne frekvensen til et grunnord med alle dets avledninger. Med andre ord: det var nødvendig å rotlemmatisere ordene.

> F.eks. hvis en fremmedspråklig elev har lært ordet "ANVENDE", vil hun (evt. han) også forstå ordene "ANVENDELSE" og "ANVENDELIG" så snart vedkommende også har lært noen enkle regler for ordlaging i norsk.

I dette prosjektet ble grupperingen av ordene foretatt manuelt, og det utviklet seg mange diskusjoner omkring hvilke ord som tilhørte samme semantiske rotlemma.

Ved de Nordiske datalingvistdager 1981 ble vi oppmerksom på vår felles interesse for automatisk rotlemmatisering, og et samarbeid ble etablert. Fordi den ene hadde kompetanse i edb og den andre i lingvistikk, kunne vi nå ta fatt på en rekke av de uløste problemer som vi hver for oss hadde stått ovenfor.

Arbeidet med automatisk rotlemmatisering ble derfor intensivert, og i dag utgjør det et eget delprosjekt under NORIS (58). Dette prosjektet er finansiert av NTNF og tar sikte på å utvikle en "intelligent forsats" til tekstsøkesystemer.

## 3. Målsetning

### 3.1 Programsystem

Målsetningen for arbeidet med den automatiske rotlemmatiseringen var å utvikle et programsystem for rotlemmatisering som ikke var for ressurskrevende. Dette var spesielt viktig med tanke på implementering av et slikt program i tekstsøkesystemer, da responstiden i slike systemer spiller en meget viktig rolle.

Det var derfor utelukket å basere programsystemet på et manuelt utviklet leksikon. I steden valgte vi å basere oss på et sett med generelle regler for rotlemmatisering som var uavhengig av datamaterialet. Dette regelsettet kunne forsåvidt også inneholde ord, men disse måtte i tilfelle tilhøre lukkede ordklasser (f.eks. funksjonsord, sterke verb etc.) slik at det ikke oppsto behov for endring av regelsettet ved oppdatering av datamaterialet.

### 3.2 Rotlemmatisering

Rotlemmatiseringen skulle bidra til at ord med felles grunnform ble gruppert. Dette gjelder ikke ord som i vesentlig grad har fått endret sin betydning ved at de har fått lagt til endelser eller avledninger (f.eks. BEHOLDE og BEHOLDNING, KOMMUNE og KOMMUNIST, OPPDRAG og OPPDRAGELSE, STAT og STATISK.) Eksempel på ord som kan sies å tilhøre samme rotlemma er:

| | |
|---|---|
| AMERIKA | ANTA |
| AMERIKAS | ANTAS |
| AMERIKANSK | ANTOK |
| AMERIKANSKE | ANTATT |
| AMERIKANER | ANTATTE |
| AMERIKANERE | ANTAKELSE |
| AMERIKANERNE | ANTAGELSE |
| AMERIKANISERE | ANTAGELSEN |
| AMERIKANISERT | ANTAKELIG |
| : | : |
| : | : |

### 3.3 Homografseparering

Målsettingen omfattet ikke kartlegging av homografer. Dette problemområdet ble ansett for å være for omfattende innenfor rammen av prosjektet. Imidlertid vil en langt større del av homografene være interne homografer (dvs. homografer som har samme rotlemmatilhørighet) enn tilfellet er ved vanlig lemmatisering.

> ARBEIDER (nomen agentis) og ARBEIDER (verb, presens) er eksempler på interne homografer i en rotlemmatiserings-prosess. De skal grupperes sammen og skaper derfor ikke noe problem. I en vanlig lemmatiseringsprosess ville disse regnes som eksterne homografer fordi de tilhører hvert sitt lemma.

Derimot vil eksterne homografer virke forstyrrende (f.eks. ordet HELT som både kan være et substantiv, et adverb og et verb i perfektum partisipp). Retningslinjen for grupperingen var at vi skulle forsøke å plassere ordet i det rotlemmaet som vi regnet med hadde høyest frekvens, men generelt skulle rotlemmatiseringen aksepteres så sant ordet ble plassert i ett av de riktige rotlemmaene.

## 4. Gjennomføring

Prosjektet har følgende hovedaktiviteter:

1) Tilretteleggelse av datamateriale
2) Utvikling av et regelsett
3) Utvikling av ett programsystem
4) Testing

I utviklingen av regelsettet var det nødvendig med et eksperimentmateriale. I vårt tilfelle var det naturlig å ta utgangspunkt i det materialet som allerede var tilgjengelig, og eksperimentmaterialet ble derfor sammensatt av ulike fagbøker for grunnskolen (geografi, fysikk og historie) og 2 juridiske dokumentsamlinger (tinglysingsavgjørelser og sammendrag av lagmannsrettsavgjørelser i familie-, skifte- og arverett). Tilsammen besto korpuset av ca. 1/2 mill. løpende ord, hvorav de juridiske samlinger utgjorde ca. 2/3. Vi fant det hensiktsmessig å fjerne skrivefeil, utenlandske ord, nynorske ord, forkortelser og noen ord med gammel skriveform. Når det gjaldt navn, beholdt vi bare egennavn som hadde substantiv som siste ledd og navn på land og verdensdeler. Antall ulike ord i korpuset ble som følge av dette redusert fra ca. 29.000 til ca. 25.000.

Tyngden i prosjektet har helt opplagt ligget i spesifiseringen av regelsettet. Gjennomføringen kan deles i tre stadier:

a) Forslag til hovedregler ble satt opp på bakgrunn av en systematisering av formverket i norsk.
b) Hovedregler ble testet og spesialregler ble innført.
c) Reglene ble vurdert ut fra deres hyppighet.

Arbeidet har nærmest tatt form av en "feedback-prosess". Vi startet med et sett med hovedregler. Disse ble så testet på

eksperimentmaterialet, og ut fra en vurdering av feilene ble spesialregler satt opp. Vi gjentok så prosessen med det nye regelsettet, og fortsatte inntil vi sto igjen med en fullstendig systematisering og kategorisering av alle ordene i eksperiment-materialet. Før det endelige regelsettet ble fastsatt, ble det foretatt en vurdering av de enkelte reglers "eksistensberet-tigelse" på grunnlag av hvor hyppig de ble brukt - dvs. hvor mange ulike og løpende ord de dekket. Særlig gjaldt dette spesialreglene.

I overensstemmelse med målsettingen ble det lagt vekt på at regelsettet skulle være så uavhengig av korpuset som mulig. Det var allikevel vanskelig å unngå at enkelte regler ble noe "preget" av vårt materiale, f.eks. spesialreglene. For å få en generell bedømmelse av regelsettet til slutt, ble det testet mot et helt tilfeldig valgt materiale (barneboken "Ole Brumm").

## 5. Regelsettet

Den ordbehandlingen som regelsettet dekker, kan deles i 7 kategorier:

1) fjerning av bøyningsendelser
2) fjerning av avledningsendelser
3) nøytralisering av vokalvekslinger
4) nøytralisering av konsonantforenklinger og -fordoblinger
5) nøytralisering av stavelsessammentrekninger
6) nøytralisering av skriftvarianter av samme leksikalske ord
7) markering av ord som får felles oppslagsform som følge av vår behandling uten at de tilhører samme rotlemma.

En del ord må behandles med hensyn til flere av disse kategoriene samtidig. Det finnes derfor forskjellige typer regler. En type er "enkel" og tar bare for seg en kategori av gangen. Disse reglene er i noen tilfeller temporære, dvs. de er

kodet slik at de sender ordet til viderebehandling i motsetning til de endelige reglene som avslutter behandlingen av ordet. Andre regler er sammensatte, de dekker flere av kategoriene på ren gang.

Etter en systematisk gjennomgåelse av de forskjellige ordklassers paradigmer, satte vi opp et forslag til fjerning av bøyningsendelsene (pkt. 1). Likeledes valgte vi ut en "normalform" når det gjaldt paradigmer som inneholdt vokalvekslinger (pkt. 3), konsonantforenklinger/-fordoblinger (pkt. 4) og stavelsessammentrekninger (pkt. 5). Ord som forekom hyppig i forskjellige skriftvarianter (f.eks. fram/frem, nå/nu) ble lagt inn spesielt (pkt. 6).

> De sterke verbene FINNE og VINNE har vokalvekslingene i-a-u. Vi regner infinitivs-/presensformen som normal form og erstatter preteritums- og perfektumsformen med denne, dvs. -ANT og -UNNET strykes og -INN settes i stedet. (For å forenkle reglene strykes alltid e'en når den er utlyd).

Avledningsendelsene ble vurdert i forhold til hvor stor grad de endret det semantiske innholdet av ordet, og i første omgang ble de aller fleste lagt inn i regelsettet med beskjed om at de skulle fjernes (pkt.2). Vi undersøkte også hvilke prefikser (preposisjoner og adverb) som forekom hyppig i sammensetninger med sterke verb (f.eks. INNGÅ) og funksjonsord (f.eks. DERPÅ). Disse ble samlet i 2 forskjellige grupper og skulle hjelpe til å bestemme hvorvidt et ord var et sterkt verb (som måtte nøytralisere vokalvekslingen) eller et funksjonsord (som i de fleste tilfeller skulle beholde den formen det hadde).

De reglene vi så kom fram til, kjørte vi ut på vårt eksperimentmateriale for å kartlegge omfanget av følgende problemer:

1) ord med "falske" endelser
2) ord med større uregelmessighet i rotlemmakomponentene enn hovedreglene dekket
3) uheldige grupperinger

## 1) En falsk endelse

En falsk endelse er en bokstav eller en bokstavkombinasjon som er en del av stammen, men som har en form som er identisk med en bøynings- eller avledningsendelse. I prinsippet kan alle bøynings- og avledningsendelser ha en "tvilling" som er falsk, men det er stor forskjell på hyppigheten av forekomstene av disse falske endelsene. Eksempler på falske endelser er -EN i LAKEN, -ER i METER, -S i PRIS og -A i KOLLEGA. Hvis målsettingen hadde vært å finne fram til stammen i ordene, måtte disse falske endelsene beholdes. Men med vår målsetting kan vi tillate oss å la disse bli fjernet så lenge oppslagsordet for rotlemmaet (dvs. det som blir igjen av stammen) blir entydig.

> -A'en i utlyd kan være en bøyningsendelse (bestemt form entall hunkjønn eller bestemt form flertall intetkjønn) og skal da fjernes. Men den kan være en del av stammen (eks. KOLLEGA) eller den kan tilhøre et sterkt verb i preteritum (INNLA). Når A'en er del av stammen, burde den beholdes, når den er utlyd i forbindelse med et sterkt verb, burde det sterke verbet forandres til "normalformen" (se over). Dette har vi løst på følgende måte: Hovedregelen er at A'en fjernes i utlyd. At den dermed blir fjernet i ord som KOLLEGA, løser vi ved også å fjerne den i de andre bøyningsformene i ord med -A som siste bokstav i stammen (-AEN, -AER, -AENE i eksemplet med kollega). Så lenge oppslagsordet er entydig, er vårt krav oppfylt. Ved endelser som har en bokstavkombinasjon som kunne tilsi at de er sterke verb, kaller vi opp prefikslisten for sterke verb og sjekker ordets begynnelse mot denne. Det viser seg nemlig at svært mange av de sammensatte sterke verbene nettopp består av en prefiks + verbet (det viktigste unntaket er LEGGE som også danner forholdsvis mange sammensetninger med substantiv). Hvis vi får tilslag på den aktuelle prefikslisten, blir ordet oppfattet som et sammensatt sterkt verb og behandlet deretter. I eksempelet med INNLA ville INN være å finne blant prefiksene for verb, og ordet ville forandres til INNLEGG, mens f.eks. KULA ikke ville få tilslag på prefikslisten og ville følge hovedregelen.

Hvis derimot oppslagsordet for et rotlemma faller sammen med oppslagsord for andre rotlemma, kan ikke de falske endelsene fjernes. I disse tilfellene må vi legge inn spesialregler for å kunne skille mellom de ekte og de falske endelsene.

I noen tilfeller sløyfet vi hovedregelen, og de ordene som hadde denne bokstavkombinasjonen som en virkelig avlednings- eller bøyningsendelse, fikk spesialregler. I andre tilfeller var det ordene med falske endelser som fikk spesialreglene. Metoden vi valgte var alltid den som gav færrest regler totalt sett.

2) **Ord med større uregelmessighet i rotlemmakomponentene enn hovedreglene dekket**

En del ord som klart tilhører samme rotlemma, viser større uregelmessigehet enn våre hovedregler tilsier. Dette gjelder i første omgang fremmedord, lånt fra latin og gresk. I slike tilfeller måtte vi legge inn noen spesialregler som gikk forbi suffiksgrensen og behandlet stammen i ordet.

PRODUSERE og PRODUKSJON tilhører samme rotlemma og skal derfor grupperes sammen. Ved å fjerne -ERE og -SJON ville vi stå igjen med oppslagsordene PRODUS og PRODUK som må tillempes hverandre. I dette tilfellet vil PRODU være en entydig oppslagsform, og vi la inn regler som fjernet S'en og K'en i forbindelse med disse suffiksene.

3) **Uheldige grupperinger**

Uheldige grupperinger er ord som tilhører forskjellige rotlemmaer, men som får felles oppslagsform uten at de i utgangspunktet er homografer. I mange tilfeller dreier det seg om innholdsord som får samme form som et funksjonsord. I slike tilfeller har vi lagt inn og markert de aktuelle funksjonsordene (de tilhører jo en del av ordforrådet som ikke ekspanderer), slik at vi kan skille gruppene fra hverandre.

- E'en i utlyd blir alltid fjernet, og ordet MENE vil derfor få rotlemmaet MEN. For at det ikke skal bli gruppert sammen med konjuksjonen MEN, har vi lagt inn konjunksjonen og markert denne. Dermed får vi skilt de to rotlemmaene fra hverandre.

## 6. Programsystem for automatisk rotlemmatisering

### 6.1 Oversikt

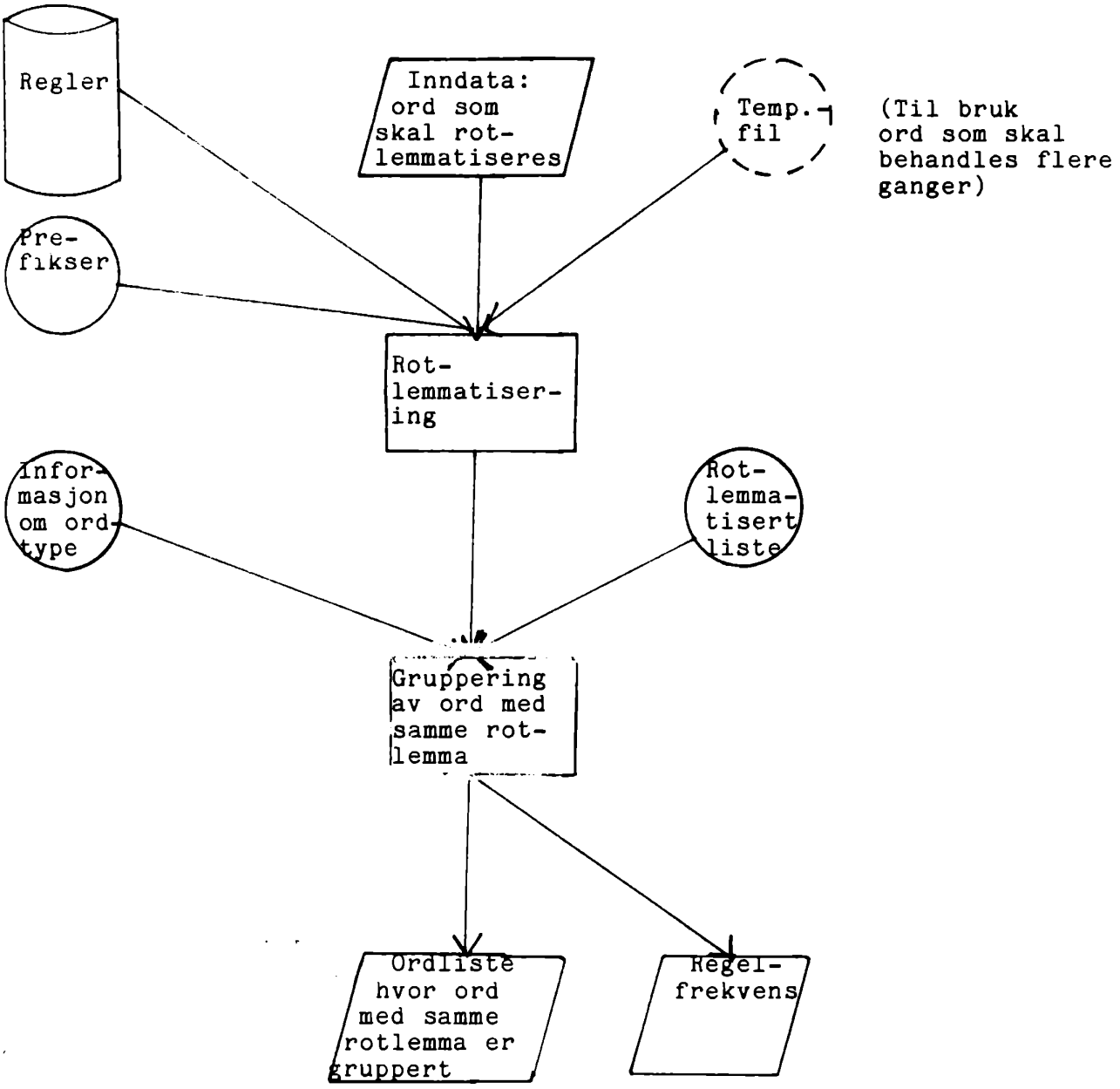
Figuren på neste side gir en oversikt over programsystemet for automatisk rotlemmatisering.

Inndata til programmet er ett eller flere ord, f.eks. en frekvensordliste som i vårt tilfelle. Resultat er en rotlemmatisert liste hvor hvert ord er angitt med rotlemma og en del tilleggsinformasjon, f.eks. hvilke regler som er brukt og om ordet er et funksjonsord. Det siste er nyttig informasjon ved gruppering av ordene fordi vi ikke ønsker å gruppere funksjonsord sammen med andre ord, (jfr. eksemplet med MEN og MENE).

Som del av resultatet får man også en oversikt over hvor hyppig den enkelte regel er brukt. F.eks. endelsen -EN er brukt 3399 ganger, mens -ENE er brukt bare 1681 ganger. Denne informasjon var til stor nytte for oss ved spesifisering av regellisten, og den kan også være interessant for dem som ønsker å studere suffiksene i et datamateriale.

Den rotlemmatiserte listen blir til slutt gitt som "input" til et program som grupperer alle ord med samme rotlemma. Dette programmet beregner også den samlede frekvens for rotlemmaeene.

Oversikt over programsystemet for automatisk rotlemmatisering.



## 6.2. Spesifisering av reglene

Hver regel inneholder en tegnstreng, en typebetegnelse, en ordre og et krav.

Tegnstrengen består av ett eller flere tegn som skal sjekkes mot ordets høyre del. I enkelte tilfeller kan eller må tegnstrengen utgjøre hele ordet.

Typebetegnelsen angir om ordet f.eks. er et funksjonsord eller et sterkt verb. Denne informasjonen hindrer at bestemte typer ord (f.eks. funksjonsord) blir gruppert sammen med andre ord.

Ordren gir informasjon om hvor stor del av ordet som eventuelt skal fjernes og hvilken tegnstreng som eventuelt skal legges til. Dessuten gir ordren beskjed om ordet skal behandles på nytt etter at ordren er utført (f.eks. ord med genitivs s).

Et generelt krav til rotlemmaet er at det minst må bestå av to tegn, hvorav det ene må være en vokal. I tillegg kan hver regel stille krav til

a) hvor stor del av ordet tegnstrengen skal utgjøre (f.eks. hele ordet eller bare en del av det),
b) ordets begynnelse (f.eks. at det skal være en prefiks),
c) at tegnstrengen ikke er etterfulgt av andre teng (dvs. at ordet ikke har vært behandlet tidligere). I slike tilfeller kaller vi regelen "lukket".

Eksempler på regler og bruken av dem er gitt på neste side.

## 6.3 Algoritme

Programmet behandler ordet bakfra. Etterhvert som det beveger seg et tegn mot venstre, sjekkes ordets høyre del mot tegnstrengen i regelen. Dette gjentas for hver regel inntil

ordets høyre del finner en regel som passer, eller at alle reglene er sjekket. I det siste tilfellet får ordet et rotlemma som er lik ordet.

Hvis regelen passer - dvs. at tegnstrengen er lik ordets høyre del eller hele ordet - sjekkes først regelens krav. Er disse også tilfredsstilt, utføres ordren. Hvis ikke, fortsetter søkeprosessen nedover regellisten.

I de tilfeller et ord skal behandles på nytt igjen, merkes ordet og legges ut på en temporær fil. Denne filen behandles til slutt som om den var en vanlig inputfil. Det er ingen grenser for hvor mange ganger et ord kan behandles før rotlemmaet er bestemt.

Regellisten er organisert som en kjedet fil for å gjøre søkingen mer effektiv. Dessuten er både inndata og regellisten (tegnstrengene) sortert bakfra i samme rekkefølge, slik at programmet ikke starter øverst på resultatlisten for hvert nytt ord. Det finnes mange alternative søkeprosesser som er mer effektive, og programsystemet BETA ville f.eks. ha vært velegnet i dette tilfellet (jfr. Benny Brodda 1982 "The BETA system", foredrag holdt på COLING i 1982).

## 6.4 Eksempel på automatisk rotlemmatisering

Vi ønsker å rotlemmatisere VARE og GÅRDEIERNES og forutsetter at reglene på neste side gjelder.

Behandling av VARE:

1. gang passer regel 6 og -E blir fjernet. Resultat: VAR.
   Ordren gir beskjed om at ordet skal behandles på nytt.
2. gang stemmer ordet overens med tegnstrengen i regel 3, men kravene til regelen er ikke tilfredsstilt (jfr. krav c). Søkingen fortsetter og regel 4 passer. Ordet tilfredsstiller kravene og får rotlemmaet VAR.

| Regel nr. | Tegn-streng | Type-betegn. | Ordre | | | Krav a): Tegn-strengen må ut-gjøre ... | Krav b): Sjekk prefiks-liste | Krav c Lukket |
|---|---|---|---|---|---|---|---|---|
| | | | Slett | Legg | Behandl. | | | |
| 1 | S | | 1 | | Ja | høyre del av ordet | | |
| 2 | ER | | 2 | | Nei | høyre del av ordet | | |
| 3 | VAR | sterkt verb | 2 | ÆR | Nei | hele ordet | | Ja |
| 4 | R | | 0 | | Nei | høyre del av ordet | | |
| 5 | ERNE | subst. eller verb | 2 | | Ja | høyre del av ordet | | |
| 6 | E | | 1 | | Ja | høyre del av ordet | | |
| 7 | LA | sterkt verb | 1 | EGG | Nei | høyre del av ordet el. hele ordet | Ja | Ja |
| 8 | A | subst. eller verb | 1 | | Nei | høyre del av ordet | | |

Behandling av GÅRDEIERNES:

1. gang passer regel 1 og endelsen -S blir fjernet.
   Resultat:GÅRDEIERNE.
   Regelen inneholder ikke typebetegnelse, men gir beskjed om at ordet skal behandles på nytt.
2. gang stopper prosessen ved regel 5 og endelsen -NE fjernes.
   Resultat:GÅRDEIER og typebetegnelse "verb eller substantiv".
   Også denne regelen gir beskjed om at ordet skal behandles på nytt.
3. gang passer regel 2 og -ER fjernes. Ordet får rotlemmaet GÅRDEI og typebetegnelsen beholdes.

## 7. Resultat

### 7.1 Oversikt

Eksperimentmaterialet besto av 489.382 løpende ord og 23.890 ulike graford. Av disse ble 685 (2,8 %) ulike graford gruppert feil. Tilsammen utgjorde disse 6.740 løpende ord - dvs. 1,4 % av det totale antall ord i korpuset.

### 7.2 Typer feil.

Feilene er inndelt i 3 hovedkategorier.

a) "Tunge" feil - (utgjør 2,3 %)
b) "Lette" feil - ( " 0,3 %)
c) "Vriene" feil- ( " 0,3 %)

Tunge feil har vi valgt som betegnelse på ord som ikke er behandlet i samsvar med de 7 kategoriene i avsnitt 5. Disse kan inndeles i 3 undergrupper.

a1) Rotlemmaet inneholder endelser som burde ha vært fjernet - eller at rotlemmaet ikke er fullstendig. Det siste tilfellet gjelder spesielt avledningsendelser som fører til at ordet får endret sitt semantisk innholdet i forhold til ordstammen, for eksempel:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DEFINISJON | får | rotlemma | DEFINISJON, | mens | ønsket | rotlemma | er | DEFI |
| SYKT | " | " | SYKT | " | " | " | " | SYK |
| SMERTE | " | " | SMER | " | " | " | " | SMERT |

a2) Rotlemmaet blir for lite og derfor tvetydig, for eksempel:

FETTER får rotlemma FETT, mens ønsket rotlemma er FETTER

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SETET | " | " | SE | " | " | " | " | SET |
| LEVERE | " | " | LEV | " | " | " | " | LEVER |

a3) Ordet inneholder stavelsessammensetninger som ikke blir nøytralisert, for eksempel:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| USSEL | får | rotlemma | USSL, | mens | ønsket | rotlemma | er | USL |
| SYKKEL | " | " | SYKKL | " | " | " | " | SYKL |

Lette feil omfatter ord hvor avledningsendelsen ikke er fjernet, men hvor endelsen i en viss grad endrer det semantiske innholdet til ordet i forhold til ordstammen. Endringen er allikevel ikke så stor at vi vil beholde endelsen, eksempelvis:

| | | | | | |
|---|---|---|---|---|---|
| ADRESSAT | er | ikke | gruppert | sammen | med ADRESSE |
| BILIST | " | " | " | " | " BIL |
| GARANTIST | " | " | " | " | " GARANTI, GARANTERE |

Vriene feil gjelder ord som har så avvikende skrivemåte i de ulike former, at det kreves spesialbehandling i hvert enkelt tilfelle, eksempelvis:

KONTO og KONTI
EPILEPSI og EPILEPTISKE
FØDSELSDATO og FØDSELSDATUM

## 7.3 Homografer

Målsetningen for den automatiske rotlemmatiseringen omfattet ikke homografseparering (jfr. også pkt. 3.3). Ord med felles rot er derfor plassert i samme gruppe, selv om de har avvikende innhold. Dette har vi ikke regnet som feil. Det samme gjelder selv om ordet i sin bøyde form er entydig, f.eks.:

RETTSLIG som har fått rotlemmaet RETT (homograf en (varm)rett)

IRI

MUNNING " " " " MUNN ( " " munn)

### 7.4 Rotlemmatisering av et tilfeldig valgt materiale

For å få en så generell bedømmelse av den automatiske rotlemmatiseringen som mulig, ble regelsettet helt til slutt utprøvd på et tilfeldig valgt materiale. Vi valgte barneboken Ole Brumm som består av 19725 løpende ord og 2.369 ulike graford.

Resultatet viste at ca. 97,6 % av de ulike ordene ble plassert i riktig gruppe.

## 8. Videreføring

Språkbruken i tekster fra forskjellige fagområder kan variere sterkt, og et regelsett som har som mål å være tekstuavhengig må nødvendigvis bli ganske omfattende. Dessuten vil kravet til korrekthet variere alt etter hva rotlemmatiseringen skal brukes til. I de to konkrete problemstilligene vi tok utgangspunkt i, var korrekthetskravet forskjellig. I det førstnevnte - tekstsøkingsproblemet - er det viktig at de ordene som blir vurdert som søkeord blir gruppert riktig, mens det ikke gjør noe om det forekommer feilgrupperinger blant støyordene. Den andre problemstillingen - utskillingen av høyfrekvente ord - krever derimot at alle typer ord blir gruppert riktig. Hvis ikke ville frekvensen forskyve seg, og man vil ikke finne fram til de gruppene man ønsket. Man kunne derfor tenke seg at et regelsett ble delt inn i forskjellige lag eller pakker. Utgangspunktet kunne være en bunke med basisregler, og man kunne tilføre lag med spesialregler som gav større korrekthet. Dessuten kunne man tenke seg spesielle fagpakker eller genre-pakker, f.eks. en juss-pakke, en fysikk-pakke, en medisin-pakke eller en språkkonservativ-pakke og en språkradikal-pakke. På denne måten kunne man få et regelsett som var tilpasset både teksten og bruken.

En del av reglene er markert med hensyn til ordklassetilhørighet. Alle reglene er markert enten som funksjonsord eller som innholdsord. Denne markeringen er først og fremst lagt inn for å skille mellom ord som ellers ville bli gruppert sammen. Dessuten har det vært en hjelp under arbeidet å vite hvilke ord en regel er ment å dekke. Dette arbeidet kunne lett utvides, og en automatisk markering av ordklassetilhørighet skulle være innen rekkevidde.

# ALLIGATORISK ROT-LEMMATISERING