Tove Fjeldvig
Institutt for privatretts avdeling for EDB-spørsmål
Niels Juels gate 16 - Oslo 2

UTVIKLING AV ENKLE METODER FOR TEKSTSØKING MED SØKEARGUMENTER I NATURLIG SPRÅK

1. Innledning

Det er spesielt 2 grunner til at det kan være ønskelig med spørsmål i naturlig språk i et fulltekstsøkesystem:

a) Søkeargumenter i naturlig språk stiller ingen eller få krav til forkunnskaper hos brukeren, og vil derfor gjøre det lettere for både nybegynnere og tilfeldige brukere å anvende systemet.

b) Enkelte søkeargumenter lar seg best formulere i naturlig språk. F.eks. at en jurist i en gitt sak blir presentert for en dom og ønsker å kontrollere hvorvidt det finnes andre dommer som angår samme spørsmål. I et slikt tilfelle vil søkeargumentet kunne bestå av f.eks. et sammendrag av dommen kombinert med med "FINN DOKUMENTER SOM LIGNER".

Nå vil imidlertid en uerfaren bruker også kunne anvende dagens tekstsøkesystemer uten all for mye veiledning, men da kun på det helt enkleste nivå. For å kunne anvende systemet effektivt og oppnå gode resultater, kreves lang erfaring og godt kjennskap til hvordan man kan utnytte systemets finesser.

De siste årene har det vært en økende interesse for søkesystemer med muligheten for spørsmål formulert i naturlig språk. Forskningen har primært vært rettet mot såkalte "spørsmål-svar" -systemer hvor man søker etter et konkret svar på et problem - og ikke sekundært informasjon i form av dokumenter. I mindre grad har forskningen vært rettet mot dokumentgjenfinningssystemer. Vi kjenner til bare noen få eksempler på dette området (som prosjektene CONDOR (tysk), SYSTEX (fransk), POLYTEXT (svensk) og RESPONSA (israelsk).

Vi fant det interessant å se nærmere på muligheten for spørsmål i naturlig språk i dokumentgjenfinningssystemer, og spesielt rettet oppmerksomheten mot muligheten for en slik strategi som et alternativ til eksisterende søkestrategier. Vi valgte derfor å konsentrere arbeidet om enkle og lite ressurskrevende metoder og legge vekten på brukerkrav som responstid, bruk av lagringsplass og - ikke minst - effektiv oppdatering av "databasen".

Utvikling av enkle metoder for teksktsøking med søkeargumenter i naturlig språk
Tove Fjeldvig
Proceedings of NODALIDA 1981, pages 175-190

Prosjektet ble initiert i begynnelsen av 1979 og er planlagt avsluttet ved utgangen av dette året. Arbeidet blir ledet og gjennomført av undertegnede, og prosjektet mottar økonomisk støtte fra Norges Teknisk- Naturvitenskaplige Forskningsråd.

2. Gjennomføring

Prosjektet har vært gjennomført som en serie med kontrollerte forsøk i tekstsøking

> Et kontrollert forsøk i tekstsøking går ut på at man for en gitt dokumentsamling definerer et sett med spørsmål som er aktuelle for samlingen. For hvert spørsmål går man gjennom alle dokumentene i hele samlingen og merker av de dokumenter som er relevant i forhold til spørsmålet (fasiten). Spørsmålene og fasiten defineres av de samme personer, og vi har lagt stor vekt på at vedkommende har godt kjennskap til dokumentsamlingen og dens fagområde.
>
> Spørsmålene danner grunnlaget for den maskinelle søkingen, og søkeresultatet blir sammenlignet med fasiten. Dette gir oss muligheten til å måle hvor mange av de relevante dokumenter som er funnet ved den maskinelle søkingen (recall), og hvor mange av de funne dokumenter som er relevante (presisjon).

Kontrollerte forsøk gjør det mulig til enhver tid å få feedback-informasjon om effekten av endringer i de valgte metoder. Dette bidrar til at vi kan styre forskningen i det vi tror er en riktig retning.

For å få spredning i dokumentmaterialet, har alle forsøk vært gjennomført på 3 ulike dokumentsamlinger:

a) 350 uttalelser fra Skattedirektøren, ca. 82 000 termer,

b) 1020 sammendrag av lagmannsrettsavgjørelser i familieskifte- og arverett, ca. 190 000 termer,

c) 1270 tinglysningsavgjørelser, ca. 218 000 termer.

Til hver dokumentsamling er det knyttet ca. 30 spørsmål med referanser til relevante dokumenter. Spørsmålene er definert av jurister med solid forankring innenfor dokumentsamlingens rettsområde. Referansene til de relevante dokumenter er stilt opp på grunnlag av en manuell gjennomlesing av alle dokumenter i hele samlingen.

3. Problembeskrivelse

Det ideelle resultat i et tekstsøkesystem (dokumentgjenfinningssystem) er å nå fram til alle og bare dokumenter som er relevant i forhold til brukerens problemstilling. Å konstruere et system som til enhver tid oppnår et slikt resultat, er ikke mulig. For det første fordi ordene i seg selv - eller språket - ikke er entydig, og for det andre fordi to forskjellige brukere ofte kan ha ulik forståelse av spørsmålet (søkeargumentet).

I de tradisjonelle tekstsøkesystemer, som f.eks. STAIRS, IMDOC, LEXIS og NOVA*STATUS, velger man bevisst søketermer som representerer idéene i problemstillingen, og som man forventer å finne i de relevante dokumenter. Dette vil ikke være tilfellet i et system basert på spørsmål i naturlig språk. Brukeren vil her velge ord og formuleringer med sikte på å bli forstått av et annet menneske. Man står med andre ord ovenfor ulike situasjoner i de to tilfellene, og dette er en av årsakene til at jeg tror at det er vanskelig - om ikke umulig - å utvikle et system basert på spørsmål i naturlig språk som gir like gode søkeresultater som et system av den første typen.

Av ressursmessige hensyn valgte vi å konsentrere arbeidet rundt selve spørsmålet og i så liten grad som mulig ta i bruk informasjon som krever nærmere analyse av dokumentmaterialet. Figur 1 på neste side gir et bilde av hvordan vi forestilte oss søkestrategien, og på denne bakgrunn ble prosjektarbeidet splittet i følgende hoveddeler:

a) Identifisering av søketermer og fraser i spørsmålet,
b) Utvidelse av søkeargumentet med synonymer,
c) Valg av regler for utvelging og rangering av dokumenter,
d) I hvilken grad tilbakeføring av informasjon til brukeren (feedback-informasjon) kan bidra til å øke søkeeffektiviteten til systemet.

Fram til i dag har oppmerksomheten vært festet til de 3 første punktene. I dette notatet vil jeg konsentrere meg om arbeidet under punkt a) og b).

4. Identifisering av søketermer og fraser i spørsmålet

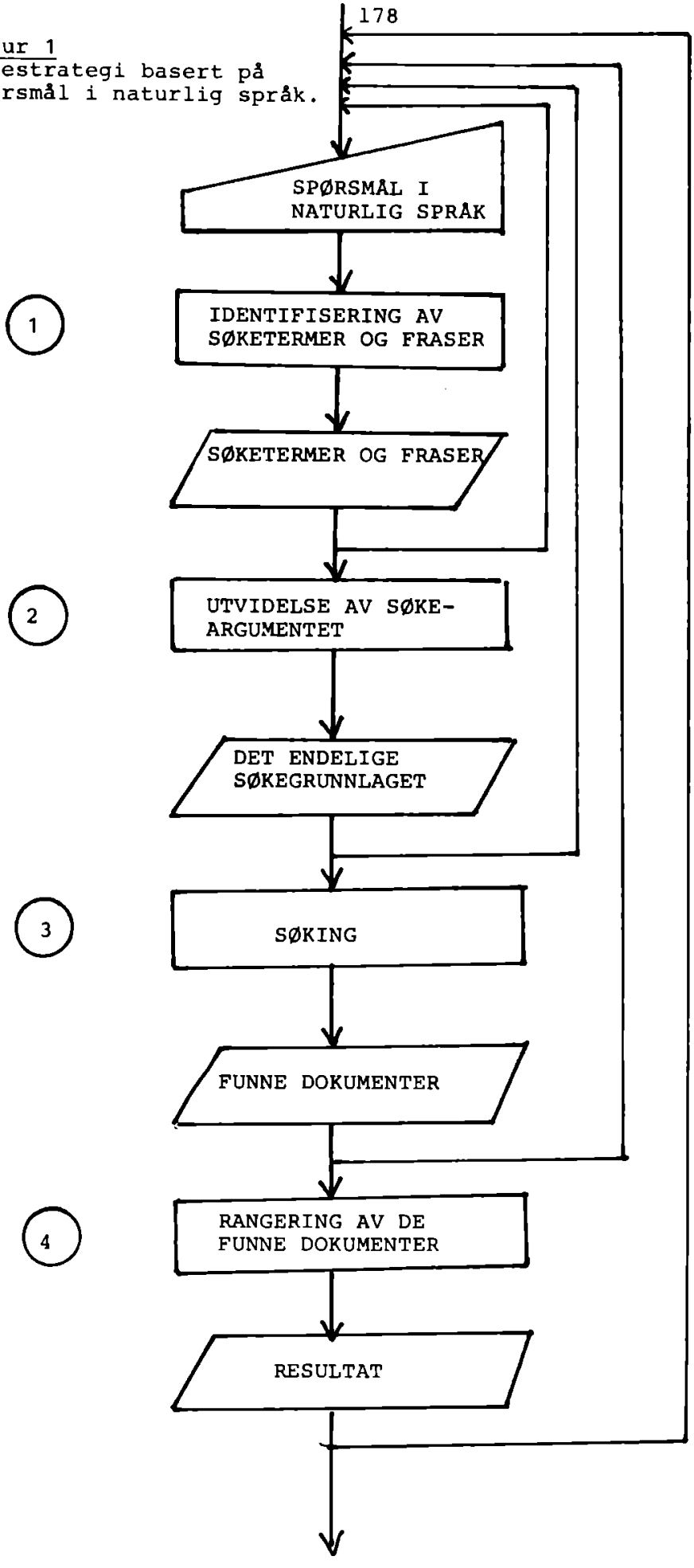
En søketerm skal ha den egenskapen at den kan bidra til å skille relevante dokumenter fra irrelevante. Betrakter man et spørsmål i naturlig språk, vil man finne at det er svært få ord som egentlig har denne egenskapen. De fleste ordene vil - betraktet isolert fra konteksten - være lite karakteristiske for selv meningsinnholdet i spørsmålet, men de vil ha en viktig funksjon i kommunikasjon med et menneske. La oss f.eks. betrakte spørsmålet:

"FINNES DET NOEN DOKUMENTER OM NORSKE TEKSTSØKESYSTEMER?"

Her vil kun ordene NORSKE og TEKSTSØKESYSTEMER ha betydning for søkeprosessen, mens de øvrige ordene vil bare skape støy i søkeprosessen. De har derfor fått betegnelsen støyord.

En søking kan også baseres på bestemte ordkombinasjoner som f.eks. et navn, et uttrykk eller en henvisning. For å kunne identifisere denne type ordstrenger i spørsmålet - eller fraser som vi har valgt å kalle dem i mangel på et bedre uttrykk - krever det enten kjennskap til disse på forhånd eller en nærmere analyse av spørsmålet.

Figur 1
Søkestrategi basert på spørsmål i naturlig språk.



Vi har foreløbig valgt å begrense arbeidet til de enkelte termer i spørsmålet og ikke ta hensyn til hvilke relasjoner som foreligger mellom dem. Et unntak er gjort når det gjelder tall, men dette vil jeg komme tilbake til nedenfor.

I steden for å forsøke å identifisere søketermene direkte, valgte vi å gå den motsatte vei å betrakte termer som ikke er egnet som søketermer - dvs. støyord.

Manuelt forsøk

For å få en nærmere innsikt i hva som karakteriserer støyord og omfanget av dem, ble det gjennomført et manuelt forsøk i å identifisere dem.

To personer fikk i oppdrag hver for seg å gå gjennom alle ulike termer i hele dokumentsamlingen og merke av de termer som ikke var egnet som søketermer. Begge hadde lang erfaring med tekstsøkesystemer og kjente godt til både prosjektet og den aktuelle dokumentsamling.

Forsøket var interessant på mange måter. For det første viste en sammenligning av resultatene fra de to, at knapt halvparten av de utvalgte termene var felles. I følge senere samarbeid kom det fram at dette i liten grad skyldes uenighet, men i større grad unøyaktigheter og mangel på fantasi. De hadde ofte oversett termer som helt opplagt var støyord eller glemt å få med alle grammatikalske varianter av ordene. Siden forsøket var basert på en skjønnsmessig vurdering av det enkelte ord, måtte man ofte forestille seg aktuelle problemstillinger før man kunne ta stilling til ordet. Mangel på fantasi var derfor en av årsakene til denne uoverensstemmelsen.

Resultatet viste bare hvor vanskelig denne type arbeid er - selv innenfor en nokså homogen dokumentsamling. En term kan gjerne være relevant i forhold til en problemstilling, men totalt uten interesse for en annen.

For det andre viste forsøket at en manuell utvelgelse av støyord på denne måten er svært tidkrevende - selv for små dokumentsamlinger. Det var faktisk nødvendig å gå gjennom hele ordlisten opp til flere ganger før den endelige støyordlisten forelå.

Eksperimentet ble gjennomført for 2 ulike dokumentsamlinger med henholdsvis ca. 7000 og 9000 ulike termer. Begge forsøkene ga de samme erfaringer.

Støyord i relasjon til ordklasser

Resultatene av det manuelle forsøket førte til at det ble rettet enda større oppmerksomheten om utviklingen av maskinelle metoder for identifisering av støyord. Det synes å være klart at hyppigheten (frekvensen) til en term og antall dokumenter termen forekommer i (spredningen), er av stor betydning for hvorvidt termen er egnet som søketerm eller ikke. Forekommer en term f.eks. i alle dokumentene, er det klart at den vil mangle evnen til å skille relevante dokumenter fra irrelevante.

Oversikten over støyord pekte også på interessante sammenhenger mellom ordklasser og støyord. Jeg skal nedenfor se litt nærmere på disse.

De mest typiske støyord finner vi blant ord som preposisjoner, pronomen, konjunksjoner, interjeksjoner og artikler. Dette er alle ord som forekommer svært hyppig og som i seg selv er lite meningsbærende. Fordelen med disse ordene er at de forekommer i et begrenset antall og kan derfor lett defineres én gang for alle.

På lik linje med disse ordene finner vi også en rekke adverb som SÅ, OFTE, DA, NÅR, NETTOPP, GANSKE, VISST osv. Dette er en langt større ordgruppe og som vanskelig lar seg gjenkjenne uten en gjennomlesing av alle ordene i hele samlingen. Litt hjelp vil man imidlertid også kunne ha av en frekvensordliste

En annen type adverb er de som står direkte knyttet til verbet og forteller noe om hvordan ting skjer. På samme måte vil adjektiv stå direkte knyttet til substantivet og forteller noe om hvilke egenskaper det har. Felles for disse ordene er at de kan bidra til å karakterisere en problemstilling hvis de betraktes i sammenheng med de ordene som de står knyttet til, men vil de egne seg som søketermer alene?

Forsøksresultatene viste imidlertid at de spilte en positiv rolle i søkeprosessen. De bidrog til at søkeordfrekvensen i de relevante dokumenter økte mer enn i de irrelevante. Dermed ble det lettere å skille ut de relevante dokumenter. På den annen side bidrog de også til at flere irrelevante dokumenter ble funnet, men denne ulempen kan man til dels unngå ved å se bort fra dokumenter som er funnet bare på grunnlag av adjektiv eller adverb.

Adjektivene vil i mange tilfeller kunne la seg identifisere automatisk enten ut fra særegne suffikser eller på grunnlag av deres posisjon mellom en artikkel og et substantiv (forutsatt at disse er kjent). Den første metoden er blitt testet i prosjektet.

En interessant type fraser er de som inneholder tall eller tallord. Problemet med tall som søketermer, er at de forekommer svært hyppig og i en rekke ulike sammenhenger - ikke alle like interessant for tekstsøking. F.eks. ved søking i et juridisk dokumentmateriale vil tall spille en sentral rolle som del av en lovhenvisning eller dato, men mindre interessant som punktbenevnelse. Siden vi i dette prosjektet ikke har planer om å se nærmere på muligheten for å søke innen et gitt intervall (f.eks. at antall arvinger i et arveoppgjør skal være mindre enn 3), er det også mindre interessant å betrakte tall som spesifikasjon av et kvantum.

Vi valgte å konsentrere arbeidet om tilfeller hvor tallet forekommer som en del av identifikasjonen til en frase, f.eks. "SIDE 9", "PARAGRAF 4" eller "ÅRET 1918". Det er interessant å merke seg at også ordene i denne type fraser (som SIDE, PARAGRAF og ÅRET) forekommer så hyppig, at de i enkelte tilfeller kan ha en negativ effekt på søkeresultatene hvis de behandles som individuelle søketermer.

I ett av våre forsøk med tall-fraser viste det seg at bare ved å knytte ordet PARAGRAF til paragrafnummeret, fikk vi redusert antall funne irrelevante dokumenter med 14% uten å miste noen av de relevante.

I arbeidet med å utvikle en metode for automatisk gjenkjenning av denne type tall-fraser, kjørte vi ut en oversikt over alle tallforekomster i hele dokumentmaterialet. I følge denne oversikten forekom tall i ca. 90% av tilfellene enten rett foran eller rett bak ordet (eller ordene) som det var tilknyttet. Så vi bort fra alle ord som forekom sammen med tall 4 ganger eller mindre, sto vi igjen med en liten gruppe ord av typen ÅR, PARAGRAF, LEDD, PROSENT, JANUAR etc. Dette er alle ord som betraktet isolert sett - i allefall i vårt datamateriale - er typiske støyord, men som i sammenheng med et tall vil kunne utgjøre en viktig del av søkeargumentet. Forsøket viste med andre ord at det var mulig å gjenkjenne automatisk de mest vanlige tall-fraser av denne type med enkle metoder.

De fleste søketermer finner vi blant substantiv og i langt mindre grad blant verb. Dette henger kanskje sammen med at substantiv ofte beskriver ting og idéer, mens verb sier noe om handlingen.

Blant verbene finner man i første rekke hjelpeverb og uselvstendige verb som typiske støyord. Også verb som ANSE, UTTALE, ANTA, NEVNE etc. er å betrakte som støyord da de sier ingenting om hva som omtales, men på hvilken måte det omtales. Disse verbene vil i de fleste tilfeller forekomme såpass ofte, at de lett lar seg skille ut i en spredning- eller frekvens-ordliste.

Enkelte andre verb vil ha like stor evne til å karakterisere idéer som et substantiv. Dette er ord som ofte er avledet av et substantiv, og som forekommer i et langt mer begrenset omfang enn de øvrige.

Substantivene vil nesten alltid ha evnen til å karakterisere ting eller idéer, og de er derfor godt egnet som søketermer. I svært homogene dokumentsamlinger hender det imidlertid at man står ovenfor substantiv som er så typiske for temaet som behandles, at de ikke lenger kan bidra til å skille relevante dokumenter fra irrelevante. Eksempel på denne type substantiv er BARN og FARSKAP i en samling med bare farskapssaker, og SKATT i en samling med bare skatteavgjørelser. Disse ordene lar seg lett skille ut i en frekvensordliste, men man skal ikke alltid stole helt på frekvensen eller spredningen i bedømmelsen om en term er egnet som søketerm eller ikke. F.eks. i en samling med familierettsavgjørelser, hvor farskapssaker kanskje utgjør 80-90%, vil ord som BARN og FARSKAP spille en sentral rolle i karakteristikken av et farskapsproblem. De vil kunne bidra til å avgrense søkingen til nettopp denne delen av basen, og av den grunn heller bli tildelt større vekt enn de øvrige søketermer.

De fleste verb og substantiv vil kunne gjenkjennes på grunnlag av særegne suffikser. Dette er også forsøkt implemenert i vårt forsøkssystem, og resultatet vil vi komme tilbake til nedenfor.

Studiet av støyord i relasjon til ordklasser har pekt på enkelte sammenhenger som kan være nyttig ved automatisk identifisering av støyord. Nå er imidlertid disse sammenhengene ikke like entydig for alle ordklasser, og av den grunn har vi foreløbig valgt å ikke legge for stor vekt på ordklasser ved generering av støyordlister.

Informasjon om hvilken ordklasse et ord tilhører, vil i mange tilfeller kunne bestemmes ut fra suffiksen til ordet. Vi gjennomførte et forsøk, hvor vi ut fra en liste med preposisjoner, konjunksjoner, interjeksjoner etc. (jfr. ovenfor) og de mest karakteristiske suffikser for hver ordklasse, forsøkte å bestemme automatisk ordklassen til ordene. Resultatet viste en presisjon på ca. 90%.

Maskinell metode for identifisering av støyord

Ved utvikling av en maskinell metode for identifisering av støyord tok vi primært utgangspunkt i frekvensen og spredningen, og bare i enkelte tilfeller ordklassen. Det falt seg naturlig å betrakte deskriptorer - og ikke det enkelte ord - da ordenes form (bøyning) er uten interesse i denne sammenheng.

Med en deskriptor mener vi her en gruppe av ord som alle er avledet av samme grunnform. En deskriptor kan derfor gjerne omfatte både substantiv, verb, adjektiv etc. - f.eks. deskriptoren "ARV, ARVEN, ARVENE, ARVE, ARVER, ARVENE, ARVING". Grupperingen av deskriptorer foregikk maskinelt, og resultatet vil jeg komme tilbake til under avsnitt 5.

Deskriptorene ble tildelt en vekt som var beregnet på grunnlag av deskriptorens frekvens og spredning i dokumentsamlingen. Vekten ble beregnet ut fra følgende spredningsmål (jfr. Rosengren 1970):

$$K = n \left(\frac{\sum_{i=1}^{n} \sqrt{x_i}}{n} \right)^2$$

n: antall dokumenter
x_i: frekvensen til deskriptoren i dokument i

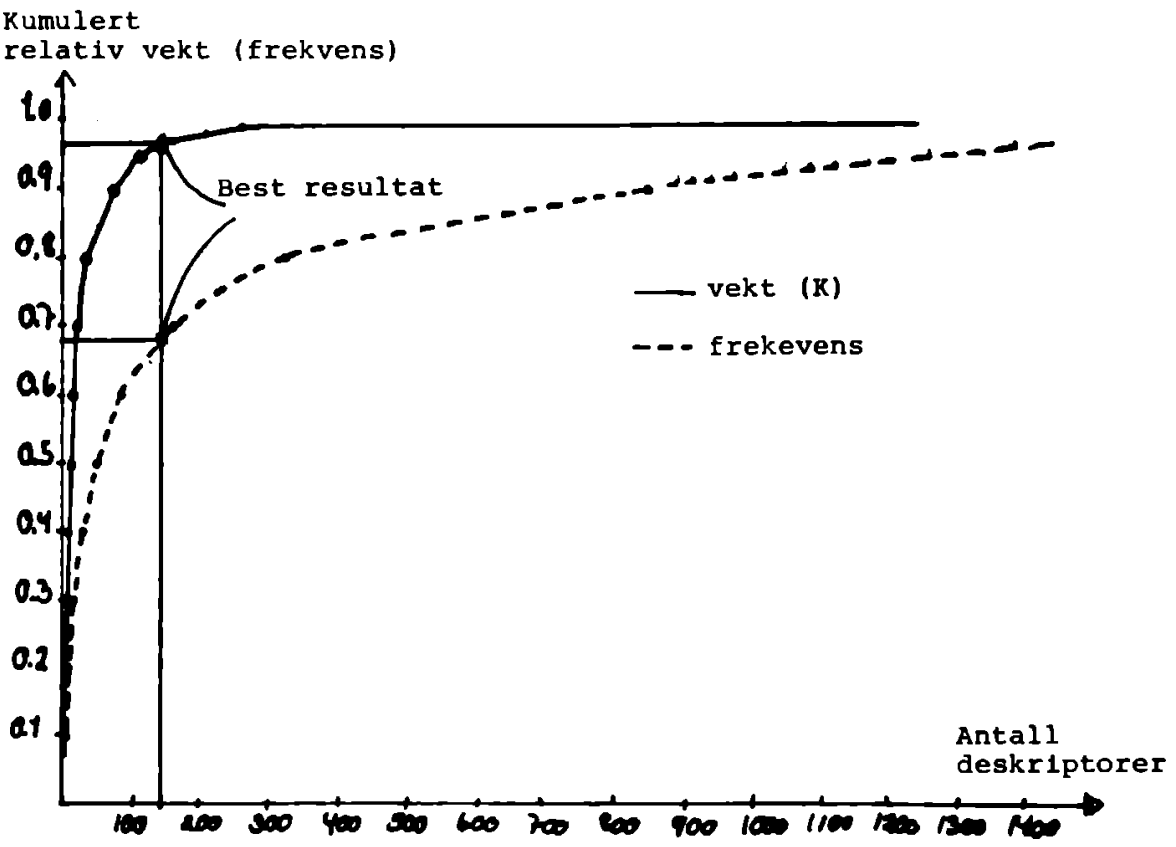
og øker med deskriptorens hyppighet og antall dokumenter den forekommer i. Vi forsøkte også å korrigere deskriptorens frekvens i et dokument (x_i) med lengden på dokumentet, men dette hadde liten innvirkning på resultatet i vårt tilfelle.

Vektene ble anvendt til å rangere deskriptorene, slik at de med høyst vekt kom øverst på listen. Denne listen viste seg på mange måter å gi et godt utgangspunkt for gjenkjenning av støyord - og langt bedre enn f.eks. en rangering basert på frekvensen alene.

For det første ga den en ganske bra rangering av deskriptorene i forhold til hvor godt de var egnet som søkeord. I figur 3 har vi gjengitt deler fra en slik liste basert på samlingen av tinglysningsavgjørelser. Øverst på listen finner vi de mest typiske støyord som preposisjoner, konjunksjoner, hjelpeverb, pronomen etc. Lenger nede framkommer tall, vanlige verb og substantiv som er svært karakteristiske for temaet i dokument-samlingen. Beveger vi oss langt nok ned, kommer vi til de mer karakteristiske ord.

For det andre fikk vi en meget gunstig fordeling av verdien på vektene med tanke på å kunne automatisk gjenkjenne hvor grensen går mellom støyord og ord som er bedre egnet som søkeord. Det viste seg nemlig at det var en relativt liten gruppe med deskriptorer som oppnådd en høy vekt, mens de fleste andre fikk en liten vekt. F.eks. for en av dokumentsamlingene utgjorde de 200 deskriptorer øverst på listen hele 98% av den totale summen av alle vektene (totalt ca. 11 200 deskriptorer). Målt i frekvens, utgjorde disse bare 73% av det totale antall termer i hele dokumentsamlingen. Dette kommer også til uttrykk i figuren nedenfor, hvor det nærmest er en knekk på kurven ved overgangen fra de hyppige deskriptorer til de mer skjeldene. Figuren viser antall deskriptorer i relasjon til den kumulerte relative vekten (avt. frekvensen).

Figur 2



Figur 3 Liste over deskriptorer sortert på vekten (K)
(basert på en samling med tinglysningsavgjørelser)

		VEKT
1.	I	5206220
2.	VÆRE, VAR, ER ..	5169546
3.	AV	4541176
4.	AT	44523941
5.	til	.
6.	OG	.
7.	PRG, PRG'S, ..	.
8.	SOM	.
	.	
	.	
	.	
	.	
14.	TINGLYSE, TINGLYSING, ..	.
15.	PÅ	.
16.	BLI, BLIR, ..	
	.	
	.	
24.	EIENDOM, EIENDOMMER, ..	914649
	.	
	.	
37.	GI, GIR, ..	.
	.	
43.	ANTA, ANTAR, ..	313116
44.	SIDE	.
45.	3	.
46.	BESTEMMELSE, ..	.
	.	
	.	
	.	
73.	TA	.
74.	BURDE, ..	.
75.	GJØRE, ..	.
	.	
	.	
	.	
131.	KREVE	.
	.	
	.	
	.	
199.	BOSETTE, ..	.
200.	BETALE, ..	13846

Det ble gjennomført forsøk med alternative støyordlister som omfattet et ulikt antall deskriptorer fra toppen av listen - f.eks. én støyordliste kunne omfatte de 80 øverste deskriptorene og én annen 100.

Det beste resultatet ble oppnådd ved å sette grensen akkurat i det området hvor kurven knekker, jfr. figur 2. Deskriptorene vi da fikk med utgjorde ca. 68% av det totale antall termer i hele dokumentsamlingen, og summen av vektene viste ca. 97% av den totale sum for alle vektene. Disse prosentandelene var omtrent de samme for alle våre 3 dokumentsamlinger (jfr. avsnitt 2).

Ved å eliminere tall og deskriptorer som bare omfattet substantiv, sto vi igjen med en nokså tilfredstillende støyordliste. I tillegg supplerte vi den med utelatte preposisjoner, konjunksjoner og pronomen. Også utelatte adverb, som lot seg gjenkjenne på grunnlag av suffiksen, ble lagt til listen.

Hele prosessen med å få identifisert støyord foregikk maskinelt, og det endelige resultatet viste at det var ytterst få ord man kunne reise tvil om hørte hjemme på en støyordliste. Sammenliknet med den manuelle støyordlisten omfattet den maskinelle langt færre ord. Allikevel viste søkeresultatene at bruk av den maskinelle listen ga vel så gode resultater som bruk av den manuelle.

5. Utvidelse av søkeargumentet

Termene i et spørsmål vil ofte være tilfeldig valgt og bestemt ut fra den formulering de er en del av. For å kunne komme så nær en fullstendig beskrivelse av meningsinnholdet i spørsmålet som mulig, burde hver av de utvalgte søketermer ha blitt supplert med synonymer. Dette forutsetter imidlertid informasjon om hvilke termer som er synonymer, og vi er her inne på et problem som ikke bare gjelder naturspråkbasert søkestrategier, men også tekstsøkesystemer generelt.

Enkelte systemer løser synonymproblemet ved hjelp av en synonymtesaurus, andre ved at brukeren selv definerer sine synonymer. Det arbeides i dag også med metoder basert på at systemet selv skal bygge opp et begrepsnettverk som gjenspeiler den semantiske strukturen i dokumentsamlingen (jfr. CONDOR-prosjektet, f.eks. Banerjee 1977).

Foreløbig har vi begrenset prosjektarbeidet til den enkleste form for synonymer, nemlig ord som er grammatikalske varianter av samme grunnform - dvs. deskriptorer. I tillegg har vi sett litt nærmere på effekten av å splitte sammensatte ord i spørsmålet. I neste fase ønsker vi å gå et skritt videre, å studere i hvilken grad vi kan få systemet selv til å generere (og vedlikeholde) en synonymstruktur, f.eks. ved å trekke på erfaringer fra tidligere søk og utnytte den respons det vil kunne få ved å føre informasjon tilbake til brukeren (f.eks. gjennom en dialog).

Gruppering av deskriptorer

Grupperingen av deskriptorer foregår maskinelt på grunnlag av en liste med

a) vanlige norske suffikser,
b) sterke verb,
c) uregelmessige bøyninger for substantiv,
d) preposisjoner, konjunksjoner, artikler og pronomen,
e) tall og tallord.

Til hver av disse elementene er det knyttet informasjon om hvilken ordklasse de tilhører. Listen er framkommet primært gjennom prøving og feiling med de tilgjengelige data. Til å begynne med, undersøkte vi muligheten for å anskaffe en slik liste fra miljøer som arbeider med språklig databehandling i Norge, men det eneste vi oppnådde var en liste med 67 suffikser basert på Ibsens verker. Listen var utviklet av NAVF's EDB-senter for humanistisk forskning, men var dessverre noe spesiell for vårt formål.

I dag omfatter listen ca. 1000 elementer. Den kan selvfølgelig gjøres enda bedre ved å supplere den med ytterligere elementer, med det viser seg at over et visst nivå vil elementene etterhvert bli så spesielle, at det skal mange til før men gjennomsnittlig seg merker positive utslag på søkeresultatene. Det vil derfor være et kostnad/nytte spørsmål hvor langt man skal gå i denne utvidelsen.

Deskriptoren til et ord dannes ved at man først finner fram til grunnformen på ordet, og deretter søker i den inverterte filen etter ord med samme grunnform.

Grunnformen til et ord genereres ved hjelp av listen ovenfor, og i følge våre forsøksresultater lykkes dette i gjennomsnittlig 90% av tilfellene. Det bør da tilføyes at suffikslisten også omfatter genitivs s, ing-form og lignende suffikser som ikke alltid er like lett å hanskes med. En stor del av feilprosenten kan defor føres tilbake til disse.

Den siste prosessen - søking etter ord med samme grunnform - kan unngås hvis man i den inverterte filen bare har grunnformen til ordene. Dette vil imidlertid føre til at man mister informasjonen om ordenes opprinnelige form, og dermed også muligheten til å søke etter disse i teksten (f.eks. som del av fraser). I dag arbeider vi med å finne metoder som ut fra grunnformen på ordet automatisk kan generere alle grammatikalske bøyninger av det uten ordbok. Dette byr ikke på vesentlige problemer innenfor én og samme ordklasse, men gjør det vanskelig når deskriptoren omfatter ord fra flere ordklasser.

Automatisk trunkering

De fleste tekstsøkesystemer i dag gir muligheten til trunkering av søketermer. Den mest vanlige form for trunkering er høyre-trunkering hvor brukeren kan søke på alle termer som begynner med en gitt tegnstreng. Tidligere forsøk med trunkering har

vist man gjennom høyretrunkering får med gjennomsnittlig 75% av alle kontekstuavhengige synonymer til de trunkerte ordene (jfr. Harvold 1974). Man vil i første rekke få med alle grammatikalske varianter av ordet, men også sammensatte ord og i enkelte tilfeller ord som er irrelevant for problemstillingen.

Vi fant det interessant å studere nærmere automatisk trunkering som et alternativ til utvidelse med bare deskriptorer. Foreløbig har vi bare sett på effekten av å trunkere stammen på ordet, og resultatet har gitt et noe blandet inntrykk.

Trunkeringen førte på den ene siden til at flere relevante dokumenter ble funnet og til at søkeordfrekvensen i de relevante dokumenter økte. På den annen side førte den også til at flere irrelevante ord ble fanget opp, og dette økte antall funne irrelevante dokumenter. De gjennomsnittlige søkeresultatene viste allikevel minst like gode som en utvidelse med bare deskriptorer. Man skal heller ikke se bort fra at metoden for automatisk trunkering kan gjøres enda bedre ved å ta hensyn til f.eks. ordklasser og uregelmessige bøyninger.

Splitting av sammensatte ord

På norsk er det nokså vanlig å anvende sammensatte ord, og i ett av våre spørsmålssett besto faktisk 38% av de ulike søketermene av sammensatte ord. Vi fant det derfor interessant å undersøke i hvilken grad en splittelse av disse ordene vil påvirke søkeresultatet.

Et sammensatt ord vil for det første bestå av flere ord, og en splittelse vil derfor kunne bidra til en bedre representasjon av meningsinnholdet i spørsmålet. For det andre vil det i mange tilfeller være en mindre sjanse for å finne et sammensatt ord enn de individuelle ordene - blant annet fordi at et sammensatt ord ofte kan skrives om til en frase.

På den annen side vil en splittelse også kunne ha en negativ effekt. Selv om ett av de individuelle ordene er til stede i dokumentet, er det slett ikke sikkert at idéen bak det sammensatte ordet er til stede. Dette kan føre til at flere irrelevante dokumenter blir funnet, men ved å prioritere de dokumenter som inneholder alle de individuelle ordene, vil denne effekten kunne reduseres.

Forsøket med splitting av sammensatte ord pågår fremdeles, og for tiden venter vi på å få tilgang til en rutine som gjennomfører splittingen automatisk uten ordbok. I følge de forsøk vi har gjennomført fra til nå, synes det som om en splittelse vil gi bedre søkeresultater.

6. Valg av regler for utvelging/rangering av dokumenter

I dette notatet vil jeg ikke gå nærmere inn på de forsøk som er gjennomført med rangering av de funne dokumenter, da dette faller utenfor temaet for denne konferansen. Allikvel vil jeg kort understreke hvor viktig denne delen av søkestrategien er.

Ut fra en antagelse om at ethvert dokument som inneholder minst én søketerm har en sannsynlighet for relevans, vil det alltid være en tendens til at for mange dokumenter blir funnet. Dette kommer spesielt til uttrykk i et tekstsøkesystem basert på argumenter i naturlig språk, på grunn av usikkerheten omkring valg av søketermer.

Formålet med en rangering av de funne dokumenter er derfor å få plassert dokumentene i en slik rekkefølge for brukeren, at de dokumenter som har størst sannsynlighet for relevans, blir plasert øverst på resultatlisten.

En viktig del av prosjektetarbeidet har derfor vært - og vil fortsatt være - å utvikle effektive metoder for rangering. Resultatene i dag viser at i ca. 64% av tilfellene får vi plassert relevante dokumenter øverst på resultatlisten, men fremdeles gjennstår mange uprøvde idéer - f.eks. å utnytte informasjon om ordklasser.

7. Oppsummering, konklusjon og framtidsutsikter

Hovedformålet med dette forskningsprosjektet er ikke primært å utvikle så effektive strategier for tekstsøking som mulig, men å undersøke muligheten for enkle og lite ressurskrevende søkestrategier basert på argumenter (spørsmål) i naturlig språk.

Prosjektet har i første rekke gitt oss en god innsikt i hvilke problemer som oppstår ved bruk av søkeargumenter i naturlig språk. Erfaringene viser - nokså naturlig - at det største problemet er overgangen fra et spørsmål i naturlig språk til et søkeargument som kan danne grunnlaget for søkingen. I denne forbindelse er det lagt mye arbeid i studiet av metoder for

- identifisering av søketermer og fraser i spørsmålet,
- utvidelse av søkeargumentet med termer som er avledet av søketermene spesifisert i spørsmålet (f.eks. ord med samme grunnform).

Det er grunn til å påpeke at metodene som er utviklet, er svært enkle og krever nesten ingen andre data enn de som vanligvis eksisterer i et tradisjonelt tekstsøkesystem. (Et unntak er listen med grammatikalske bøyningsregler og ord som preposisjoner, konjunksjoner, pronomen etc.)

En sentral del av prosjektarbeidet har også vært studiet av effektive metoder for rangering av de funne dokumenter. Fram til i dag har disse metodene bare tatt utgangspunkt i statistiske data, men vi ser ikke bort fra at de kan gjøres enda bedre ved trekke inn lingvistiske data.

Som en konklusjon på vårt arbeide vil vi si at forsøksresultatene har gitt oss tro på at det er mulig å oppnå er akseptabelt søkeresultat med så enkle søkestrategier som her beskrevet. Hva som er et akseptabelt søkeresultat vil imidlertid avhenge av den enkelte bruker og den aktuelle søkesituasjon.

Prosjektet vil bli avsluttet ved utgangen av dette året. Mens arbeidet så langt har vært rettet mot de metodiske spørsmål, vil vi i tiden framover konsentrere oss om spesifikasjonen av en modul for naturspråkbasert søking.

Det er fra neste år av søkt om økonomiske midler til et nytt prosjekt hvor vi ønsker å gå et skritt videre i arbeidet med tekstsøkesystemer basert på spørsmål i naturlig språk. Formålet med dette prosjektet vil være å utvikle en "intelligent" preprosessor (forsats) til et tekstsøkesystem, som gjennom en dialog med brukeren kan bidra til et rikere søkeargument enn det en typisk bruker vil kunne spesifisere uten hjelp. Prosjektet har fått navnet FORT (FORsats til Tekstsøkesystem) og er estimert til 3 år.

REFERANSER

Allén, Sture/Thavenius, Jan (1970)
Språklig databehandling; Studentlitteratur, Lund.

Banjerjee, N. (1967) "CONDOR - Communication in Natural language with dialogue oriented retrieval systems"; Schneider/Hein 1977: 163-172.

Bing, Jon/Fjeldvig, Tove/Flataker, Ole Bjørn/Harvold, Trygve (1980) Lovspråk og juristspråk;
Skriftserien Jus og EDB nr. 41, Oslo.

Fjeldvig, Tove (1976) Kontrollert forsøk i tekst-søking på uttalelser fra Skattedirektøren;
NORIS (8) III, Skriftserien Jus og EDB nr. 16, Oslo.

Harvold, Trygve (1976) "Belysning av synonymproblemet i norske, formuerettslige lover";
Bing/Fjeldvig/Flataker/Harvold 1980: 127-144.

Prestel, B.M. (1971) "Datenverarbeitung im Dienste juristischer Dokumentation", EDV und Recht, vol 3.

Rosengren, I. (1970) "Prosjektet Modernt tysk tidnings-språk"; Allén/-Thavenlus 1970:61-76.

Schneider, W./Hein, A.-I. Sågvall (1977)
Computational linguistics in Medicine; North Holland, Amsterdam.