# Hvor kort er godt?

## En evaluering av NorSum

### - en automatisk tekstsammenfatter for norsk

**Anja Therese Liseth**

**Hovedfagsoppgave i datalingvistikk og språkteknologi**

**Seksjon for lingvistiske fag**

**Universitetet i Bergen**

**September 2004**

# Innhold

## *Abstract*

This thesis has been carried out in collaboration with the Scandinavian science nettwork ScandSum, and it presents an evaluation of NorSum, an automatic text summarizer for Norwegian. The evaluation is an intrinsic one, which compares the automatic summaries against a gold standard. The gold standard is built from manually made summaries. The evaluation had two goals; the most important one was to do a quantitative evaluation of NorSum to investigate its performance. It was interesting to see if a quantitative evaluation could provide any information about the quality of the summarizer. The second goal was to develop a program that could automatically generate a gold standard, which the automatic summary could be compared against. The results show that there is a relative good overlap between the gold standard and the automatic summary, and regarding the deviation between the automatic summaries and the manual summaries; it was less than expected.

## *Sammendrag*

I samarbeid med forskningsnettverket ScandSum har det i denne hovedoppgaven blitt utført en evaluering av NorSum, som er den norske versjonen av den automatiske sammenfatteren SweSum. Hovedoppgaven presenterer en evaluering av NorSum, et system for automatisk sammenfatning av tekster på norsk. Oppgaven hadde to mål, hvor det viktigste var å utføre en kvantitativ evaluering, for å undersøke om en kvantitativ metode kan gi nyttig informasjon om kvaliteten på sammendragene. For å sammenligne de automatiske sammendragene med manuelle sammendrag, ble det automatisk generert et referansesammendrag, en gullstandard, ut fra de manuelle sammendragene. Resultatene av den interne evalueringen viser at det er en klar, men relativt liten overlapping mellom de automatiske sammendragene og referansesammendraget. Avviket mellom sammendrag og referansesammendrag er større for de automatiske enn de manuelle sammendragene, men mindre enn på forhånd antatt, tatt i betraktning at referansesammendraget er laget ut fra de manuelle sammendragene. I tillegg viste det seg nyttig å generere referansesammendragene automatisk, da dette både sparte tid og arbeid, siden evalueringen i seg selv var en tidkrevende prosess.

## *Forord*

Denne hovedfagsoppgaven i datalingvistikk og språkteknologi er utført og avlagt ved Seksjon for lingvistiske fag, Institutt for lingvistikk og litteraturvitenskap, Universitetet i Bergen. Arbeidet ble påbegynt i august 2003 og avsluttet i september 2004.

Oppgaven er utført i samarbeid med det skandinaviske forskningsnettverket ScandSum, som var finansiert av Nordisk Ministerråd. Prosjektet ble avsluttet våren 2004. Jeg har deltatt på noen nettverksmøter i løpet av den tiden jeg har arbeidet med oppgaven, og det har bidratt til å inspirere og motivere i arbeidet med oppgaven. I den sammenheng vil jeg takke Hercules Dalianis og Martin Hassel ved Kungliga Tekniska Högskolan – KTH – som har utviklet den automatiske sammenfatteren jeg har evaluert.

Jeg vil få takke min veileder prof. Koenraad de Smedt ved seksjonen og Aleksander Krzywinski som utviklet databasen hvor sammendragsmaterialet er lagret. I tillegg vil jeg takke alle som har bidratt med å lage manuelle sammendrag til databasen.

# 1 Innledning

Automatisk sammenfatning (*automatic summarization*) hadde sitt utspring ved de store forskningsbibliotekene i USA i 60-årene. Det var på denne tiden ønskelig å lagre vitenskapelige artikler og bøker digitalt og gjøre dem søkbare. Men på grunn av begrenset lagringskapasitet var det ikke mulig å lagre dem i sin fulle form i databasen, og derfor ble det lagret sammendrag som ble indeksert og gjort søkbare. I dagens samfunn kan automatisk sammenfatning være nyttig for flere enn store institusjoner med omfattende dokumentmengder. For eksempel kan et søk på internett i dag gi oss tilgang på veldig store informasjonsmengder, og automatisk sammenfatning kan bidra til å gi oss den oversikten vi trenger. I litteraturen blir automatisk sammenfatning gjerne definert på denne måten: *In automatic text summarization, the most relevant parts of a document are extracted and put together in a non-redundant summary that is shorter than the original document* (Dalianis et al. 2003).

I 2000 ble forskningsnettverket ScandSum startet, finansiert av Nordisk ministerråds språkteknologiske forskningsprogram. Målet var å øke fokuset på automatisk sammenfatning for de skandinaviske språkene. Ved KTH – Kungliga Tekniska Högskolan i Stockholm – ble det utviklet en sammenfatter, SweSum, som er vellykket overført til norsk, dansk og noen andre språk. I tillegg finnes det også en versjon som ikke er koblet til et leksikon og dermed ikke tar hensyn til hvilket språk teksten er skrevet på.

I enhver utvikling av et dataverktøy er det nødvendig med evalueringer underveis fordi dette kan si noe om kvaliteten på verktøyet, i dette tilfellet en automatisk sammenfatter. En evaluering kan si noe om prestasjonen til sammenfatteren, og kan gi nyttig tilbakemelding til utviklerne. Det har allerede blitt utført en rekke evalueringer av SweSum, og disse har bidratt med viktig informasjon til utviklerne, i tillegg til at evalueringsmetodene i seg selv også har blitt vurdert, for å kunne finne frem til den som fungerer best.

Denne oppgaven tar for seg en evaluering av NorSum, som er den versjonen av SweSum som er koblet til norske språkresurser. Evalueringen er utført ved hjelp av kvantitative metoder i et håp om å dempe den subjektiviteten som ofte preger evalueringer av automatiske sammenfattere. I den forbindelse ble det utviklet et program som automatisk

genererte et referansesammendrag (RS) som de automatiske sammendragene ble sammenlignet mot. Dermed har ikke de kvalitative egenskapene ved de automatiske sammendragene blitt evaluert, med unntak av det som kan tolkes ut av tallmaterialet. Den språkuavhengige versjonen er også evaluert, for på den måten å indirekte teste om leksikonet i NorSum bidrar til en vesentlig forbedring av sammendragene.

Kapittel 2 går gjennom forskjellige metoder for å generere automatiske sammendrag og anvendelsesområder for automatiske sammendrag, samt den metodiske bakgrunnen for SweSum. Kapittel 3 tar for seg forskjellige evalueringsmetoder og de evalueringene som har blitt gjort av SweSum. Så følger kapittel 4 med beskrivelse av materialet jeg har samlet inn og som danner grunnlaget for evalueringen av NorSum, i kapittel 5 beskrives selve utførelsen av evalueringen. Til slutt i kapittel 6 følger tolkning av resultater og konklusjon, samt fremtidsperspektiver for forskningsfeltet.

# 2 Metoder for automatisk sammendrag

Automatiske sammenfatningsverktøy kan ha mange bruksområder og bruksområdene kan ha avgjørende betydning for valg av metode som skal generere sammendragene. Noen av applikasjonsområdene blir her presentert sammen med en oversikt over de forskjellige typer av sammendrag. Det blir også gitt en oversikt over de ulike metodene som benyttes for å generere sammendrag, inkludert den metodiske bakgrunnen for utviklingen av SweSum / NorSum.

## *2.1 Anvendelsesområder*

Både Mani (2001b) og Dalianis et al. (2003) gir gode oversikter over de mange områder hvor automatisk sammenfatning kan anvendes. De fleste produktene innenfor hvert område er utviklet for engelsk. Det finnes et utall kommersielle produkter på markedet, som AutoSummarize i Microsoft Office, InXight-sammenfatter i Alta Vista Discovery søkemotor, IBM sin Intelligent Miner for tekst, DimSum sammenfatter fra SRA Corporation, en sammenfatter fra General Electric R&D Labs og mange flere. I Norge har CognIT as (CognIT 2004) utviklet en sammenfatter som lager sammendrag av tekster på norsk, svensk, tysk og engelsk. Sammenfatteren er en inkorporert del av et større system for dokumentanalyse; SLATE!! - the CORPORUM™ Desktop Navigator. I tillegg finnes det også mange systemer som blir utviklet innenfor forskjellige forskningsinstitusjoner rundt om i verden, uten at det nødvendigvis blir kommersielle produkter ut av disse.

Isteden for å liste alle kommersielle produkter, er det mer hensiktsmessig å komme inn på spesifikke anvendelsesområder for automatisk sammenfatning som Mani (2001b) gjør. Et relativt nytt og spennende område er nyhetssammendrag for multimedia. Denne teknologien vil tillate en viss grad av omstrukturering av nyhetsformidling via multimedia ("se på nyhetene og fortell meg hva som har skjedd mens jeg var borte"). Et annet interessant område er sammenfattere som kan være til hjelp for leger. Et forskningsprosjekt ved Columbia University (McKeown et al. 1998) tok sikte på å tilby leger sammendrag av medisinsk litteratur som var tilgjengelig på internett, relatert til en pasients medisinske journal ("sammenfatt og sammenlign den anbefalte behandlingen for denne pasienten"). Sammendrag av møter basert på talegjenkjenning er også en mulighet som ikke ligger langt unna i tid. Her kan f.eks en bruker skumme gjennom og få oversikt over innholdet i ett eller flere møter som personen har gått glipp av. Ved å benytte en

metode som er uavhengig av domene, så kan det være mulig å generere sammendrag av f.eks telefonkonferanser uansett tema. Shiffman et al. (2000) beskriver et annet område hvor det kan være interessant å benytte sammenfatning, og det er i etterforskning av kriminalsaker. Her kan en sammenfatter ta som innputt forskjellige dokumenter og generere et dossier om en person som er omtalt i dokumentene ("Lag en 500 ords biografi over G. Bush"). Dette kan benyttes av de taktiske etterforskerne i deres analyser av en forbrytelse. Det området som kanskje er mest interessant, og hvor det ikke eksisterer noe tilbud til dags dato, er søketrefflister fra søkemotorer. Når en spørring er skrevet inn i søkefeltet til en søkemotor på internett, får man som oftest opp en svært lang liste med treff, hvor det kan være vanskelig å orientere seg om hvilke dokumenter som har relevans utifra søkeordene. Goldstein et al. (2000) og Radev og Fan (2000) beskriver forskjellige metoder for å sammenfatte informasjon i trefflister returnert av søkemotorer.

Det er ikke bare i forbindelse med informasjonssøk over internett at en sammenfatter er nyttig. I større grad enn før kan nyheter og annen informasjon også leses ved hjelp av mobiltelefonen, både via WAP og SMS, eller en personlig digital hjelper (PDA), og her skapes det nye behov i takt med den teknologiske utviklingen. Ofte er det et ønske fra nyhetsformidlerne at for eksempel nyheter presenteres med samme innhold, men i forskjellig format, i forskjellige medier, så som nettaviser, papiraviser, WAP-tjenester o.l. Men skal dette editeringsarbeidet gjøres manuelt, er det både tid- og resurskrevende. Her kan en automatisk sammenfatter automatisere dette arbeidet enten fullstendig eller assistere i prosessen.

Ikke all informasjon man er interessert i finnes på et språk man kjenner til. Ofte kan man få treff på artikler som synes interessante, men som er skrevet på et annet språk enn det man har som morsmål. Siden oversettelse også krever tid og resurser, kan teksten med fordel bli kortet ned først og så kan sammendraget bli oversatt. På den måten vil det være mulig å kontrollere at dokumentet man har funnet er så interessant som man tror det er. En mulig løsning vil være at en sammenfatter er knyttet til programmet som automatisk oversetter teksten, slik at man først kan sammenfatte teksten og deretter oversette den.

I tillegg til de bruksområdene jeg har skissert så langt, er det også viktig å nevne nytteverdien sammenfattere kan ha for mennesker med forskjellige handikap. I forbindelse med oppleste avistjenester for synshemmede kan en sammenfatter være

nyttig. Da kan det først bli lest opp et sammendrag av avisartikkelen og så kan brukeren selv bestemme etterpå om det er interessant å få lest opp artikkelen i sin fulle form. Det kan også være interessant å se på sammenfatning i forbindelse med hørselshemmede, hvor f.eks nyhetssendinger kan bli sammenfattet og senere presentert i sammenfattet form. Slike hjelpemidler ligger enda på utprøvingsstadiet.

Innenfor forskjellige anvendelsesområder kan man se at sammenfatning kan være et viktig verktøy for å finne frem til relevant informasjon på en mer effektiv måte. Man kan få raskere oversikt over store dokumentmengder og det kan være lettere å sile bort informasjon som ikke er relevant.

## *2.2 De ulike typer sammendrag*

Radev et al. (2002) deler de ulike sammendragstypene inn i fire kategorier. Den første er kalt *indikative* sammendrag. Dette er en type sammendrag som gir en idé om hva kildeteksten handler om, uten å gjøre kjent noe spesielt innhold. Det vil si at ingen av temaene i kildeteksten blir utdypet i noen grad, men sveipes i overflaten. Denne typen sammendrag kan sies å ha en referansefunksjon for å kunne velge ut dokumenter til mer dybdelesning. Den neste kategorien er *informative* sammendrag som gir en forkortet versjon av innholdet og dekker all viktig informasjon i kilden på et visst detaljnivå. Her er alle de viktigste begrepene fra kildeteksten tatt med. Denne type sammendrag kan erstatte kildeteksten og det er ikke nødvendig å lese kildeteksten for å ha en forståelse av hva den omtaler og inneholder av emner. Den tredje kategorien er *emneorienterte* sammendrag og her fokuseres det på brukerens ønskede emner. Det vil si at brukeren kan legge inn de nøkkelordene som skal vektlegges i sammendraget og få returnert et sammendrag som er fokusert på disse nøkkelordene. Denne typen kalles også styrte sammendrag. Den siste kategorien er *generelle* sammendrag som reflekterer forfatterens synsvinkel. Man kan også si at de to siste kategoriene er overkategorier av de to første typene. Et sammendrag kan altså både være indikativt og emneorientert.

Firmin og Chrzanowski (1999) fokuserer på tre aspekter ved automatiske sammendrag: *Hensikt*, *fokus* og *dekningsområde*. *Hensikt* beskriver den potensielle bruken av sammendraget, som kan være indikativt, informativt eller evaluerende. Indikative sammendrag gir akkurat nok informasjon til å kunne avgjøre relevansen til kildeteksten, eller gi en kort oversikt over temaet i teksten. Informative sammendrag kan tjene som

erstatning for kildeteksten og beholder de viktige detaljene, men reduserer mengden informasjon som blir gitt til brukeren. Evaluerende sammendrag fanger opp forfatterens synsvinkel på et gitt tema. *Fokus* refererer til rekkevidden (scope) av sammendraget og det kan være enten generelt eller spørringsrelevant. Et generelt sammendrag er basert på hovedtemaet/-ene for et dokument, mens et spørringsrelevant sammendrag er generert med fokus på det emnet som brukeren ønsker. *Dekning* viser til om sammendraget er basert på ett enkelt dokument eller flere dokumenter relatert til det samme emnet. Mye av det tidligere arbeidet innenfor automatiske sammenfatningssystemer har vært rettet mot genereringen av indikative, generelle sammendrag av enkeltdokumenter, og Firmin og Chrzanowski (1999) mener at både Luhn (1958), Edmundson (1969), Johnson et al. (1993) og Brandow, Mitze og Rau (1995) fokuserte på denne typen sammendrag, selv om deres fremgangsmåter inkluderte ulike kombinasjoner av statistiske og lingvistiske teknikker. De fleste av disse arbeidene hevder å ha en viss grad av uavhengighet til domene, men metodene har kun blitt prøvd ut på en spesifikk type data, nemlig avisartikler eller tekniske artikler.

Mani (2001b) mener at skillet mellom de forskjellig typer sammendrag må sees på som mer pragmatisk enn teoretisk, og at det egentlig er mer et utgangspunkt for retningslinjer til bruk for profesjonelle sammenfattere. Her refererer han til ANSI (American National Standards Institute) sine retningslinjer for profesjonelle sammenfattere, hvor skillet mellom indikative og informative sammendrag blir beskrevet på denne måten: For en vitenskapelig artikkel som beskriver et eksperiment eller en undersøkelse som er utført, bør et indikativt sammendrag inneholde informasjon om artikkelens *formål*, *rekkevidde* (scope) og *tilnærming*. Et informativt sammendrag derimot, bør inneholde denne informasjonen, men i tillegg også *resultat*, *konklusjon* og *anbefalinger*. Andre forskere (f.eks Spärck Jones 1999) påpeker at det er nyttig med forskning på hvordan profesjonelle sammenfattere arbeider, nettopp for å kunne overføre noen erfaringer fra dette til arbeidet med utvikling av automatiske sammenfattere.

Hassel (2004) har satt opp en matrise som viser de forskjellige aspektene ved sammendrag på en oversiktlig måte og som gir en nyttig oversikt over emnet:
Kildetekst (innputt):

- Kilde: enkeltdokument vs. multidokument
- Språk: monolingual vs. multilingual

- Sjanger: avisartikler vs. teknisk artikkel
- Spesifikasjon: domenespesifikk vs. generell
- Lengde: kort (1-2 sider) vs. lang (> 50 sider)
- Media: tekst, grafikk, audio, video, multimedia

Formål:

- Bruk: generell vs. spørringsspesifikk
- Formål: Hva skal sammendraget brukes til?
- Publikum: ikke-målrettet vs. målrettet

Sammendrag (utputt):

- Avledning: ekstrakt vs. abstrakt
- Format: løpende tekst, tabell, geografiske oversikter, tidslinjer, diagrammer osv
- Partiskhet: nøytral vs. evaluerende

En del av disse kategoriene faller inn under det Spärck Jones (1999) kaller *formålsfaktorer* (*purpose factors*), hvor *bruk* av sammendrag ansees som den viktigste (ovenfor: *formål*). Disse faktorene burde være mest vektlagt i utviklingen av sammenfatningsstrategier, men hun påpeker at i praktisk utførelse er de som oftest oversett. Selv om *formål* og *bruk* av sammendragene er implisitt, så er det ofte nyttig å tydeliggjøre disse aspektene. Publikum kan være målrettet, som f.eks forskere som leser vitenskapelige artikler, eller ikke-målrettet, som f.eks lesere av nyheter i en avis. Den kanskje viktigste faktoren er *bruk* av sammendraget. Hva sammendraget skal brukes til henger ofte sammen med hvilken type sammendrag som bør velges. Eksempler på bruk av sammendrag som også henger sammen med valg av type sammendrag, inkluderer å bruke en sammenfatter som verktøy for å lokalisere kildedokumenter av interesse og som hjelp til å få oversikt over et dokument før man leser det. Her ville det være naturlig å velge et indikativt sammendrag. En sammenfatter kan også generere et sammendrag som kan erstatte et gitt dokument hvis det ikke er nødvendig å lese hele dokumentet. Eller sammendraget kan fungere som en oppfrisker av et dokument man har lest tidligere, men hvor man trenger å bli minnet på hovedpunktene i teksten. I disse tilfellene ville det være naturlig å velge et informativt sammendrag. Men det finnes også situasjoner hvor ett og samme sammendrag kan ha forskjellig bruksområder, som f.eks i en forelesningssituasjon, hvor sammendraget først kan fungere som en oversikt over

hovedtemaene før forelesningen og som en oppfriskning av emnene som ble gjennomgått i tiden etter forelesningen.

## 2.3 Metoder for generering av sammendrag

Før de mer spesifikke metodene for generering av automatiske sammendrag omtales, er det nødvendig å si noe generelt om de to hovedtilnærmingene til sammenfatning. Fra et lingvistisk ståsted kan man enten ta utgangspunkt i en *grunn analyse* (*shallow analysis*) av en tekst, eller en *dypere analyse* (*deep analysis*). En grunn analyse vil si at representasjonen av teksten ikke går dypere enn til det syntaktiske nivået, og som oftest blir den bare analysert morfologisk. Ord kan bli analysert på et semantisk nivå, men som oftest skjer ikke dette. Denne fremgangsmåten benyttes hovedsaklig for å produsere ekstrakter; uttrekk av hele setninger. Den dypere analysen foretar en grundigere parsing av teksten og setningene kan bli analysert på semantisk nivå. Sammendraget blir som oftest generert i omskrevet form, og dette innebærer en naturlig språkgenerering fra en semantisk modell eller diskursmodell. Sammenfatningsprosessen kan også deles opp i tre stadier, uavhengig av praktisk fremgangsmåte; *analyse*, *omforming* (*transformation*) og *syntese*. I *analysefasen* analyseres innputt og det bygges en intern representasjon av den. *Omformingfasen* oversetter den interne representasjonen til en representasjon av sammendraget, og *syntesefasen* gjengir sammendragsrepresentasjonen i naturlig språk. Den grunne analysen som hovedsaklig produserer ekstrakter, vil nødvendigvis ikke ha en omfattende omformingsfase i og med at kildeteksten ikke forandres. I syntesefasen vil sammendraget bli presentert med setninger som er å finne i kildeteksten. For den dypere analysemetoden vil nødvendigvis omformingsfasen være av en mer omfattende karakter enn for den grunne analysemetoden, siden det her som regel blir generert ny tekst og ikke bare et utdrag av den opprinnelige kildeteksten. Den interne representasjonen kan også gjennomgå flere komprimeringsoperasjoner, for å trekke sammen begreper og tilpasse spesialiserte begreper til mer generelle med det formål å korte ned teksten.

Uavhengig av hvilken tilnærmingsmåte som velges, er det viktig å ha fokus på et av hovedproblemene innenfor sammenfatning, nemlig problemet med *koherens*, dvs sammenheng og flyt i teksten. En strategi som har som mål å ordne dette etter at sammendraget har blitt generert er ikke en optimal strategi. Sammendraget vil være av bedre kvalitet om det fokuseres på dette problemet allerede i prosessen når metoder for automatiske sammendrag blir utviklet. Det kan for eksempel være nyttig å utvikle en

metode som tar høyde for de sjangerspesifikke trekkene i teksten; hvor avisartikler kan være et eksempel på en slik sjanger. Dermed kan problemet løses i utgangspunktet ved at det utvikles en metode som er spesifikk for en sjanger og legger vekt på dens særtrekk.

Mani (2001b) lister tre hovedtyper av problemer med mangel på koherens:

- Hengende / løse anaforer: Dette oppstår hvis en anafor (f.eks et pronomen som "de") er inkludert i sammendraget, mens beskrivelsen av hvem "de" er, altså referenten, er utelatt.

- Hull: Oppstår når emnene i teksten ikke er bundet sammen og hvis ikke alle emnene er tatt med. Problemet oppstår først og fremst hvis overgangene mellom emnene blir utelatt. Her illustrert med et eksempel: "Spesielt heftig var ordbruken i Indonesia og Malaysia. Indonesia er verdens mest folkerike muslimske land."

- Strukturerte oversikter: Lister, tabeller eller logiske argumenter kan ikke vilkårlig bli delt opp. Hvis kildeteksten inneholder f.eks "Opprørerne hadde tre krav..."og alle tre blir listet opp, enten som en punktliste eller i vanlig tekst i kildeteksten, mens sammendraget bare har med to av dem, så vil dette fremstå som villedende for leseren.

Luhn (1958) var trolig den første som presenterte en algoritme som var en sjangerspesifikk teknikk for automatisk generering av sammendrag. For å generere sammendraget ble de viktigste setningene i teksten plukket ut etter bestemte kriterier. Først ble nøkkelordene lokalisert på bakgrunn av frekvens og lengde på ordene. Hvor ofte disse nøkkelordene forekom i en setning, altså tettheten av nøkkelordene, definerte vektingen av hver setning. Setningene med høyest vekting ble dermed identifisert som de viktigste setningene i teksten, og det var disse som tilslutt utgjorde det automatiske sammendraget. Algoritmen brukte først en stoppliste for å sortere ut lukkede ordklasser som pronomen, preposisjoner og artikler. En stoppliste er en liste over ord som et program skal se bort i fra og ikke gi vekting, og disse ordene tilhører vanligvis lukkede ordklasser. Deretter ble resten av ordene normalisert, det vil si at ord som er ortografisk like, men har ulik bøyning, ble slått sammen; som f.eks "similar" og "similarity". Deretter ble frekvensen for disse sammenslåtte termene talt opp og de med lav frekvens ble forkastet. Setningene blir ikke bare vektet ut ifra hvor mange signifikante ord (nøkkelord) de inneholder, men også ut ifra tettheten av disse signifikante ordene. Luhn (1958)

beskriver også utvidelsesmuligheter for algoritmen, så som varierende lengde på sammendraget og ekstra vekting til ord som finnes i en domenespesifikk ordliste, såkalte bonusord. Han kommer også inn på bruksområder for automatiske sammendrag og nevner i den sammenheng oversettelse og innhenting av informasjon (Information Retrieval – IR).

Edmundson (1969) utvidet denne algoritmen til å inkludere tre andre komponenter i tillegg til ordfrekvens; stikkordsfraser (cue phrase) – f.eks "significant", "in conclusion", "hardly" – tittel/overskrift og setningsplassering. Han antyder at metoden fra Luhn (1958) var av tilstrekkelig kvalitet til å oppmuntre til videre forskning, men at en rent statistisk metode ville være mangelfull i forhold til å generere sammendrag, og at man derfor måtte søke etter nye metoder. Likevel har Luhns metode vært med på å legge grunnlaget for det som i dag blir kalt ekstrahering, og det er fremdeles de statistiske metodene som danner basisen.

### *2.3.1 Ekstrahering og abstrahering*

Metoder for generering av automatiske tekstsammendrag deles gjerne opp i to hovedretninger, den ene er *abstrahering* (*abstracting*) og den andre er *ekstrahering* (*extracting*). Abstrahering utfører til en viss grad en semantisk analyse av kildeteksten, noe som ikke gjennomføres innenfor ekstrahering. I tillegg genereres det ny tekst i abstraktet, mens ekstrahering kun presenterer setninger som bokstavelig finnes i kildeteksten. Et abstrakt er altså et sammendrag hvor innholdet blir omformulert og de samme setningene ikke nødvendigvis var inkludert i kildeteksten. Innenfor abstrahering eksisterer det fremdeles ikke en fullgod metode som har latt seg utvikle til et velfungerende produkt. Det har vært prøvd ut forskjellige strategier som hver for seg har sine kvaliteter og mangler.

Selv om det finnes forskjellige abstraheringsmetoder, har de likevel en overordnet struktur til felles, hvor abstraheringen foregår i tre trinn (grovt inndelt):

1. Det blir utført en semantisk analyse av innputteksten og en intern representasjon av setningene i teksten blir konstruert
2. Det blir utført forskjellige operasjoner på den interne representasjonen, først en filtrering av elementer (*selection*), ofte tokenisering. Deretter blir aktuelle begreper slått sammen (*aggregation*), f.eks "spurv" og "ugle" blir slått sammen til "fugler". Så

blir begreper byttet ut med mer generaliserende eller abstraherende termer (*generalization*) der det er nødvendig, f.eks "mannen som plantet blomstene..." kan bli byttet ut med "gartneren...". Dette blir utført for å danne nye semantiske representasjoner. I løpet av denne prosessen kan en representasjon på diskursnivå bli dannet. En kunnskapsbase som inneholder verdenskunnskap kan også bli benyttet.

3. Sammendraget blir generert i naturlig språk, ut fra den semantiske representasjonen

Det har blitt prøvd ut forskjellige metoder innenfor abstrahering og nedenfor nevnes de hovedretningene som har vært mest fremtredende i den seneste tiden:

- Abstrahering fra templater: Informasjonen som er forhåndsdefinert av templater blir sammenfattet og bakgrunnsinformasjonen som kreves er gitt av templatplasser som skal utfylles. Måltemplatene er forhåndsdefinert av enten domene eller sjanger. Denne teknikken er som oftest sjangerspesifikk, men ikke nødvendigvis domenespesifikk. En forhåndsdefinert sjanger som f.eks tekniske artikler dekke forskjellige domener som f.eks kjemi, fysikk, medisin.

  Fordeler: Denne metoden kan sørge for en høy grad av kompresjon og kan dermed være nyttig ved mulitidokumentsammenfatning hvor en høy kompresjonsgrad er nødvendig. I tillegg kan korpusbaserte metoder forfines i automatisk templatutfylling. Utfylling av templater kan også være basert på morfologisk analyse i noen tilfeller.

  Svakheter: Templater er kostbare å utvikle og knyttet til en begrenset klasse av domener hvor de kan benyttes. Templater kan produsere ukorrekte sammendrag på grunn av ukorrekt templatutfylling og metoden har ingen *generalisering*.

- Abstrahering ved omskriving av termer: Logiske termer i semantiske representasjoner blir valgt ut, samlet (*aggregate*) og slått sammen (*merge*), f.eks "Kari så en spurv og en ugle" blir gjort om til "Kari så fugler". *Viktighet* (*salience*) baseres på telling av emner i teksten.

  Fordeler: Håndterer *generalisering* og sammenfatning blir sett på som en lingvistisk prosess med omskriving av symbolstrenger. Metoden kombinerer også domenekunnskap med referansefrekvens

  Svakheter: Metoden krever spesifikke verktøy for å konstruere semantiske representasjoner. Regler for omskrivning er ofte nært knyttet til syntaktiske regler og de må derfor inkluderes. Dessuten krever *generalisering* mye verdenskunnskap og trenger avgrensninger.

- Abstrahering ved å benytte relasjoner mellom hendelser: *Viktighet* baseres på telling av hendelsesrelasjoner eller forbindelser i et diagram over semantisk slektskap, og bakgrunnsinformasjon brukes for å relatere hendelsene.
  Fordeler: Kombinerer domenekunnskap med referansefrekvens, *generalisering* og bakgrunnsinformasjon kan bli introdusert. Vellykket brukt i sammenfatning av loggen til en kampsimulator.
  Svakheter: Bundet til spesifikke domener hvor hendelsesstrukturene er kjent.

- Abstrahering ved å benytte et emnehierarki: Metoden utfører en grunn parsing av nomen og er avhengig av et leksikon som er lenket til en kunnskapsbase med emnedomener. *Viktighet* baseres på emnetelling og *generalisering* baseres på hierarkier (generelle eller domenespesifikke) over disse emneene.
  Fordeler: Bakgrunnsinformasjonen kommer fra hierarkiet og allerede eksisterende synonymordbøker kan benyttes. I tillegg kan man kontrollere hvor detaljert nivået på *generalisering* skal være.
  Svakheter: Et hierarki må være tilgjengelig og inneholde betydninger av ord innenfor et domene. Resultatene fra *generalisering* er ikke alltid leselig siden det ikke produseres et abstrakt, men en oversikt over emnene fra teksten.

Ekstrahering gjør bruk av en mye grunnere analyse enn abstrahering. Her skjer analysene på ord-, eller morfemplan (subord) og tar svært sjelden hensyn til f.eks. setningssemantikk. Noen metoder utfører en viss analyse for å lokalisere egennavn og slik navnegjenkjenning er implementert i SweSum for svensk. En grunn analyse kan være fordelaktig ved at det er lettere å implementere i et dataprogram fordi det ikke skal genereres ny tekst i naturlig språk.

Ekstrahering tar utgangspunkt i en kildetekst og ved hjelp av både statistiske og lingvistiske metoder, men også en del heuristikker, rangeres setningene etter visse kriterier og de høyest rangerte plukkes ut til å være med i det automatiske sammendraget. De statistiske metodene som benyttes er i stor grad de samme som Luhn (1958) benyttet seg av; dvs at nøkkelord blant annet blir plukket ut på bakgrunn av frekvens. I tillegg benyttes også lingvistiske metoder i analysen av teksten, ved at det blir gjort en morfologisk analyse som f.eks avdekker stammen av et ord for å kunne identifisere like

ord med ulik bøyning. Utover dette brukes det ofte en del heuristikker også, som er en type parametre som angir hvordan setninger skal vektes, f.eks uthevet skrift, overskriftstagging og lignende. Sammendraget inneholder nøyaktig de samme setningene som kildeteksten og de opptrer også i samme rekkefølge. Den opprinnelige teksten blir altså ikke skrevet om eller forandret på noen annen måte. Ikke alle metoder trekker ut hele setninger, noen trekker ut hele avsnitt eller fraser, men det er setninger som er mest hensiktsmessig å bruke, noe Mani (2001b) begrunner i at setninger er et lingvistisk element i motsetning til avsnitt som er en formateringsenhet. Man kan selvsagt argumentere med at elementer under setningsnivå kunne blitt trukket ut – så som ord, fraser eller uttrykk (clauses) – siden disse jo også er lingvistiske enheter. Men sammendrag på dette nivået ville sannsynligvis mangle flyt og sammenheng i teksten. Dessuten er det fullt mulig å trekke ut disse enhetene etterpå, hvis man skulle trenge navn, stedsnavn o.l., når man allerede har setningsenhetene.

Mani (2001b) hevder at automatisk sammenfatning på mange måter er en praktisk disiplin og at det ikke er en dypere teori bak sammenfatning, selv om det selvfølgelig er teoretiske rammeverk som blir undersøkt. Videre sier han at mye av forskningen på området kommer ofte fra en slags smart 'fikling'; utprøving av forskjellige hypoteser og metoder og utvikling av forskjellig programvareprototyper for eksperimenter. Spärck Jones (1999) mener at denne fremgangsmåten vil komme til kort på lang sikt. Det er mulig å utvikle gode verktøy for en applikasjon ved å simpelthen prøve ut en strategi man har for hånden og se om det gir en tilfredsstillende utputt. Men hun mener at dette er en uferdig metode, og at dette henger sammen med en generell villfarelse om bruk av sammendrag: *"It is important to recognize the role of context factors because the idea of a general-purpose summary is manifestly an ignis fatuus."*[1] Med *kontekstfaktorer* (*context factors*) mener Spärck Jones (1999) *innputt*, *formål* og *utputt*, og at det kun ved en grundigere analyse av disse aspektene er mulig å utvikle generelle sammenfatningsstrategier. Det er ingen grunn til å tro at ett enkelt sammendrag, selv ikke et bra sammendrag, skal kunne imøtekomme de forskjellige begrensningene i kontekst, så som kildetekst, formål og lesere av utputt (sammendraget). Det pekes spesielt på *formålsfaktorene*, som Spärck Jones mener er de viktigste. Innenfor formålet til et sammendrag bør det være en analyse av situasjon, publikum og bruk, men at disse aspektene ofte blir neglisjert i utviklingen av en metodologi, nettopp fordi mange av systemene som blir utviklet oppstår fra en type smart

1 Medieval Latin: ignis - fire, fatuus - foolish. "Something that misleads or deludes; an illusion". Fra: http://www.dictionary.com

'fikling'.

Hittil har det meste av forskningen dreid seg om ekstraheringsmetoder, da det er lettere å implementere disse og oppnå brukbare produkter. Men selv om denne metoden i seg selv kan generere akseptable sammendrag, mener Spärck Jones (1999) at det enda er mange spørsmål som ikke er besvart, og at det må en grundigere undersøkelse og videre forskning til for å kunne løse disse. Et eksempel på spørsmål det bør fokuseres på er diskurs, hvor en mulig løsning er å foreta en dypere analyse enn overfladisk setningsekstrahering. Hun foreslår en strategi hvor setninger får en midlertidig parsing til logisk form, med lokal anaforoppløsning. Dermed blir setninger analysert lingvistisk i den utstrekning det er mulig uten å inkludere en referansemodell med verdenskunnskap.

## *2.4 Arkitekturen til SweSum*

I denne oppgaven vil både SweSum og NorSum bli omtalt og skillet mellom de to sammenfatterne kan av og til synes litt uklart. Forskjellene mellom SweSum og NorSum blir presisert her. Først av alt er det viktig å påpeke at SweSum både er navnet på den bakenforliggende algoritmen / arkitekturen til sammenfatteren og navnet på den svenske applikasjonen. Derfor kommer jeg til å omtale arkitekturen som SweSum-arkitekturen, mens SweSum tilsvarer applikasjonen som er tilknyttet de svenske språkresursene. Dermed kan man si at både SweSum og NorSum bygger på SweSum-arkitekturen. Det grafiske grensesnittet for SweSum og NorSum er også identisk, og det som skiller mellom de to applikasjonene er en meny hvor det er mulig å velge hvilket språk man ønsker å sammenfatte på. I motsetning til DanSum har det ikke vært midler til å utvikle et eget grafisk grensesnitt for norsk. Dermed blir den eneste praktiske forskjellen mellom applikasjonene SweSum og NorSum at NorSum er koblet til norske språkresurser, mens SweSum er koblet til svenske språkresurser. Så i praksis kunne man ha kalt sammenfatteren SweSum med norsk leksikon, men jeg synes det er lettere å skille sammenfatterne fra hverandre om man opprettholder forskjellige navn.

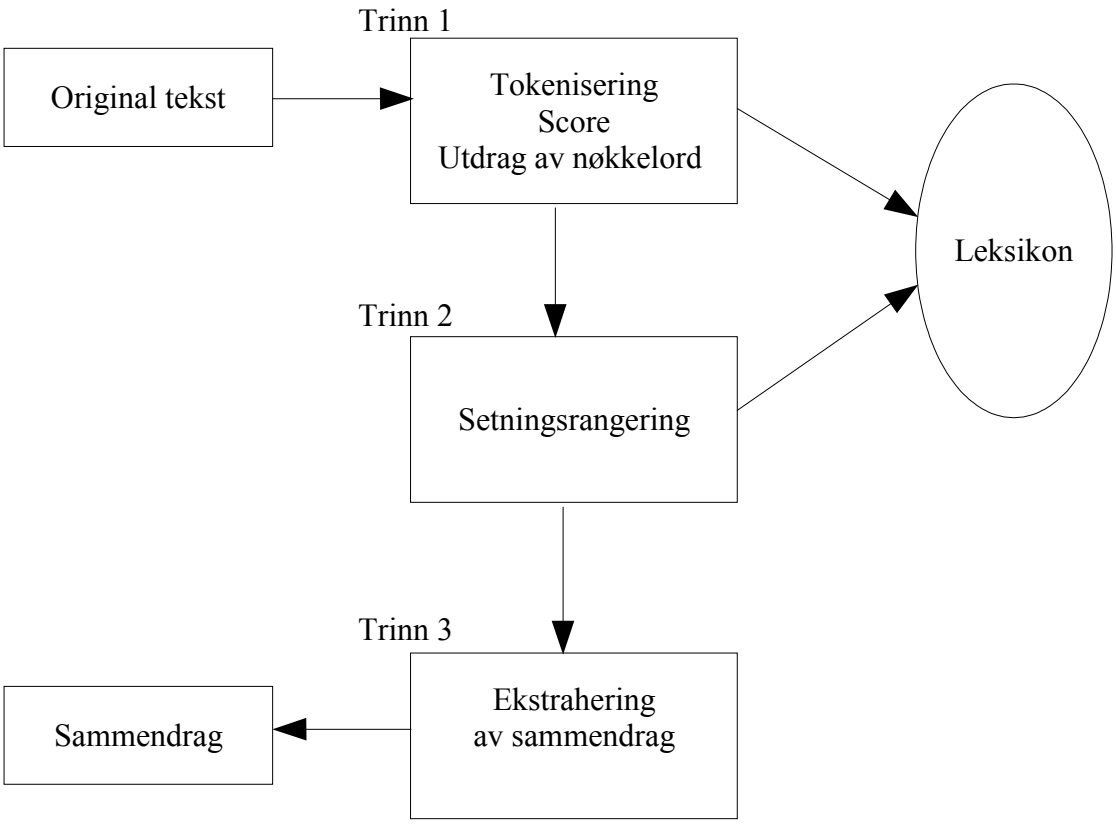



Fig. 1. Skjematisk oversikt over SweSum-arkitekturen (Fra Mazdak, 2004)

Denne stegvise tegningen viser hvordan arkitekturen til SweSum er bygget opp. Sammenfatteren arbeider i tre forskjellige trinn: I den første fasen foregår identifiseringen av hovedemnet og identifiseringen av nøkkelord. Dette gjøres ved hjelp av tokenisering, som fjerner mellomrom og all tegnsetting og skiller ut de enkelte ord. I tillegg kobles søk av nøkkelord til leksikonet, og på bakgrunn av dette blir nøkkelordene trukket ut. Setningene blir også tildelt en score. I andre fase utføres rangering av setningene utifra score i henhold til heuristikkene, og i tredje og siste fase blir selve sammendraget generert.

Sammenfatteren genererer automatiske sammendrag ved å trekke ut hele setninger fra en avisartikkel, vektet etter blant annet innhold av nøkkelord. Setningene settes sammen i samme rekkefølge som i kildeteksten og på denne måten lages et sammendragsekstrakt. Teksten skrives ikke om på noen måte og setningene opptrer i den rekkefølgen de hadde i kildeteksten. Sammenfatteren ignorerer forøvrig tagger som formaterer siden, men nyttiggjør taggene som formaterer teksten, som for eksempel at fet skrift som markerer avsnittsoverskrifter får tyngre vektlegging. SweSum-arkitekturen er programmert i Perl,

og baserer seg på både statistiske og lingvistiske metoder, i tillegg til en del innebygde parametre (heuristikker).

Sammenfatteren plukker ut setninger i artikkelen som har fått en høy score, setningene med lav score blir utelatt, og genererer sammendraget ut ifra disse. Setninger som inneholder nøkkelord knyttet til ordlisten, får en høyere score enn de som ikke inneholder nøkkelord. Ordlisten består av alle åpne ordklasser, det vil si innholdsord og funksjonsord. I tillegg så kan også nøkkelord plukkes ut fra teksten på bakgrunn av lengden på ordet og stopplisten. Nøkkelord med høy frekvens får også en høyere score enn nøkkelord med lav frekvens. Kriteriene for hvilke setninger som får høy score er listet opp nedenfor, men siden domenet for sammenfatteren er avistekster på mellom en og to sider, er det opplagt at det også blir tatt høyde for sjangerspesifikke trekk. For eksempel så viser det seg at for denne sjangeren av tekst så forekommer de viktigste termene innenfor de fire første avsnittene (Dalianis 2000).

SweSum-arkitekturen utfører tekstsammenfatningen hovedsaklig i tre trinn; det første er identifisering av hovedemnet i teksten, det neste er å trekke ut viktige deler av teksten i henhold til det identifiserte hovedemnet og til slutt generering av sammendraget. Emneidentifiseringen, eller identifiseringen av nøkkelord og viktige deler av teksten, gjøres ut ifra et sett av parametre (heuristikker), som presenteres i listen nedenfor (Dalianis et al. 2003).

- *Baseline*: Rekkefølgen av setninger angir viktigheten av setningen. Første setning får høyest rangering – siste setning får lavest rangering.
- Tittel: Ord i tittelen, og i setningene som følger umiddelbart etter, får høy score.
- Ordfrekvens (tf): Ord fra åpne ordklasser som er frekvente i teksten, er viktigere enn ord med lav frekvens i teksten.
- *Position score*: Forskjellige tekstsjangre har forskjellige trekk og ett trekk er at viktige setninger står i spesielle posisjoner, derfor får setninger som er plassert tidlig (innenfor de fire første avsnittene) i en avisartikkel høyere score enn setninger som kommer til slutt.
- Setningslengde: Lengden på setninger impliserer hvilke setninger som er viktige, dvs at lange setninger er mer viktig enn korte
- *Average lexical connectivity*: Setninger med numeriske data som deles med andre

setninger. Det viser seg at setninger som deler flere termer med andre setninger er mer viktig.

- Numeriske data: Setninger med numeriske data får en høyere score enn setninger uten numeriske verdier.
- Spørringssignatur: Brukerens spørring kan bli brukt til å påvirke hvilke nøkkelord som blir plukket ut, og sammendraget vil da inneholde disse nøkkelordene. Dette vil da bli et styrt sammendrag.

Alle de ovennevnte parametrene normaliseres og puttes i en naiv kombinasjonsfunksjon med en modifisert vekting for å fremskaffe den totale scoren for hver setning. I tillegg til disse heuristikkene tildeles setninger også ekstra vekting ut ifra enkelte HTML-tagger. Dette gjelder for de som markerer fet skrift, som ofte indikerer (avsnitts-) overskrifter, og avsnittsmarkering, fordi setninger som står først i et avsnitt får tildelt mer vekting enn de som er i slutten av et avsnitt. Det er muligheter for at evalueringen av applikasjonen NorSum avdekker om disse heuristikkene er de mest hensiktsmessige, eller om de må opp til vurdering på bakgrunn av resultatene som fremkommer. Det er likevel sannsynlig at det må en kvalitativ analyse til for å kunne svare på dette.

Det grafiske grensesnittet som er lagt til SweSum-arkitekturen (se appendiks E) er ganske enkelt å forholde seg til, man skriver inn en URL i et lite tekstfelt, og i feltet under kan man skrive inn nøkkelord som man synes er viktig å vektlegge i teksten. Hvis det blir skrevet inn noen nøkkelord, vil det i så fall bli generert et brukerstyrt sammendrag. I tillegg er det noen andre tekstbokser hvor man kan velge kompresjonsgraden og klikke av for det språket kildeteksten er skrevet på. Man kan også oppgi om det er en avisartikkel eller en akademisk tekst. Utover dette enkle grensesnittet er det en lenke på siden hvor man kan få flere valgmuligheter. Hvis man klikker på denne lenken får man opp en ny side hvor man får flere valgmuligheter for hvordan innputt kan gies, men dette er ikke det viktigste. Det som er interessant å gjøre på denne siden er at man selv kan justere parametrene som styrer deler av vektingen av setningene. Her kan vektingen av første setning, fet skrift, numeriske verdier, nøkkelord og anvendte nøkkelord fritt forandres ut ifra de behov man måtte ha. Dette forutsetter riktignok en del kunnskaper, men det er ingenting i veien for å eksperimentere med forskjellige kombinasjoner av parametrene.

Som nevnt ovenfor så styrer brukeren i hvilken grad teksten skal komprimeres. Vanligvis

så regnes 30% som den ideelle kompresjonsraten når det gjelder avisartikler. Det vil si at sammendraget består av 30% av den opprinnelige teksten, og det er også det som står som standard hvis man ikke selv endrer dette.

Det er viktig å påpeke at det ikke har vært lagt noe større vekt på utformingen av det grafiske grensesnittet av sammenfatteren. Det er meningen at sammenfatteren skal fungere som en forskningsprototyp og som en demo for hva den kan prestere så langt, og ikke et kommersielt produkt. Det som var tanken bak arkitekturen var å utvikle en sammenfattningsmotor, eller kjerne, heller enn å utvikle et grensesnitt rettet mot en bestemt funksjon. Det har dessuten blitt påpekt i en av evalueringene (Fallahi 2003), at hvis sammenfatteren skal ha noen praktisk nytteverdi er det nødvendig å integrere den sømløst i et annet program eller verktøy, som for eksempel et skriveprogram eller en søkemotor.

## *2.5 Scandsum og NorSum*

Det skandinaviske forskningsprosjektet ScandSum (ScandSum 2003) har hjulpet til med å koordinere forskningsarbeidet innenfor automatisk tekstsammenfatning i skandinavia. Dette kom som et svar på behovet for mer forskning på sammenfatning for de skandinaviske språkene, ettersom det frem til da ikke fantes noe brukbart verktøy for disse språkene (Hassel 2004). Ved KTH i Stockholm ble det i 1999 utviklet en sammenfatter som genererte automatiske sammendrag av avistekster, og dette var den første versjonen av SweSum. Det ble først utviklet for svensk, men i samarbeid med Universitetet i Bergen, ble den norske versjonen ferdig våren 2003, og ble kalt NorSum. Center for sprogteknologi (CST) i København, Danmark, deltok i utviklingen av den danske versjonen, DanSum, høsten 2002. I samarbeid med UPS i Barcelona, Spania, og med ENST i Paris, Frankrike, ble det også lagt til moduler for henholdsvis spansk og fransk. Arbeidet med å tilrettelegge for disse språkmodulene, ble avsluttet høsten 2001. I tillegg så har også tysk og farsi blitt lagt til ved hjelp av andre samarbeidspartnere. På grunn av SweSum sin arkitektur er det mulig å tilknytte forskjellige språkspesifikke resurser på en relativt enkel måte og dermed forenkles arbeidet med å utvide bruksområdet som sammenfatteren. Det som er viktig å presisere i denne sammenheng er at alle de språkspesifikke versjonene består av den samme arkitekturen, men er tilknyttet språkresurser for de forskjellige språkene. Språkresursene omfatter som oftest leksikon og lister over vanlige forkortelser.

NorSum bygger på samme programstruktur som SweSum, men er som nevnt tilknyttet ulike språkresurser. Domenet av tekster som NorSum tar seg av er html-taggete avisartikler på norsk (bokmål), og i utarbeidelsen av testmaterialet til denne oppgaven er det brukt artikler hentet fra Bergens Tidende sitt nyhetsarkiv. Det benyttes også frekvenslister som er utviklet på bakgrunn av et aviskorpus i Bergen.

Et av problemene som det måtte taes høyde for i NorSum var den store variasjonen av skriftlige normer i bokmål. Det er for eksempel like riktig å skrive *høyesterett* som *høgsterett*. Og selv om det ikke er sannsynlig at en tekst vil inneholde forskjellige skriftnormer, krever dette likevel en spesiell håndtering for å oppnå en sikker og pålitelig identifisering av nøkkelord. Dermed ble det naturlig å gjenbruke ordformsleksikonet fra forskningsprosjektet SCARRIE, som nå er avsluttet (Rosén og de Smedt 1999). Ordformsleksikonet inneholder ordboksoppslag med eksplisitte relasjoner mellom de forskjellige ordformene, med unntak av genitiv og avledninger. Dermed vil f.eks *høyesterett* og *høgsterett* regnes som ett nøkkelord og ikke to. Dette er nødvendig for å unngå overlapping i utvelgelsen av nøkkelord. I tillegg er NorSum tilknyttet en liste av norske forkortelser som er nødvendig for korrekt å kunne identifisere setningsgrenser. I tillegg til disse språkresursene var det også tanker om å prøve ut bruk av en part-of-speech-tagger som tidligere har blitt utviklet i et samarbeid mellom Senter for Humanistisk Informasjonsteknologi[2] og Tekstlaboratoriet ved Universitetet i Oslo[3]. Men siden dette både er et tid- og resurskrevende arbeid, har det ikke vært mulig å prøve ut dette pr dags dato.

I dette kapittelet har forskjellige anvendelsesområder og metoder for generering av automatiske sammendrag blitt gjennomgått. Flere forskere har påpekt at disse to områdene henger tett sammen og at valg av metode ofte henger sammen med hva sammenfatteren skal brukes til og innenfor hvilket domene. Men det er ikke bare i forbindelse med genereringen av sammendrag det er viktig å se på hva sammendraget skal brukes til. Også når det gjelder evaluering er dette et aspekt som det må taes hensyn til. En del forskere, bl.a. Spärck Jones (1999), mener at bruken av sammendragene spiller en viktig rolle når de skal evalueres, og at selv om dette nevnes implisitt, så er det en side ved sammendrag det må fokuseres på i større grad enn tidligere.

2 http://www.aksis.uib.no/projects
3 http://www.hf.uio.no/tekstlab/

# 3 Metoder for evaluering av automatiske sammendrag

Det er nødvendig å kunne evaluere sammenfattere på samme måte som det er nødvendig å kunne evaluere f.eks automatiske oversettelsesprogrammer. En evaluering vil kunne si noe om kvalitetene til et program, og også noe om produktet som blir produsert av programmet, det være et sammendrag eller en oversettelse. Et annet aspekt ved evaluering er at det kan gjøre det mulig å sammenligne forskjellige sammenfattere med hverandre, forutsatt at metodene som brukes gir sammenlignbare resultater. Og her kommer man inn på kjerneproblemet ved evaluering; fordi det ikke finnes et fasitsammendrag for en tekst, et sammendrag som er det eneste sanne og som alle andre sammendrag kan sammenlignes mot, blir det også vanskelig å sammenligne resultater mellom forskjellige sammenfattere. Dermed eksisterer det også lite konsensus omkring metoder for evaluering og det er vanskelig å sammenligne de forskjellige resultatene. Ofte er det vanskelig å unngå at evalueringen blir subjektiv, i og med at hva som er et godt sammendrag er en subjektiv vurdering.

## *3.1 Generell oversikt*

Evaluering har lenge vært av interesse for automatisk sammenfatning. Til og med i den tidligste forskningen, som ved det klassiske arbeidet til Edmundson (1969), ble det lagt stor vekt på evalueringsspørsmål. Med sitt evalueringsgrunnlag på 200 dokumenter innenfor emnet kjemi, og sammenligningen mellom sammendrag laget av profesjonelle sammenfattere og fire ekstraheringsmetoder, er dette fremdeles en av de største evalueringene som er utført. Selv om mange forskningsprosjekter presenteres sammen med en evaluering av en eller annen type, kan det synes overraskende at det ikke finnes noen konsensus i disse evalueringsspørsmålene (Mani, 2001b). Noe av årsaken til denne mangelen finner man i at det som skal evalueres, altså sammenfatteren, i seg selv ikke produserer en standard som lett lar seg måle. Som Jing et al. (1998) påpeker; de fleste sammenfatterne som utvikles har også med en evalueringsdel, men der blir ofte problemet med evaluering stadfestet først, og så blir det anvendt en evalueringsmetode som synes passende. Problemet med disse individuelle evalueringsmetodene er at det er umulig for en sluttbruker å sammenligne de forskjellige sammenfatterne, nettopp fordi de er individuelt basert og ikke standardisert. Så problemene omkring konsensus innenfor evaluering har ikke nødvendigvis sitt utspring i interne stridigheter, men heller med bakgrunn i det som skal evalueres. For uansett hvilken metode som velges så er det en

subjektiv avgjørelse hva som er et godt og brukbart sammendrag. Det finnes ikke bare ett sammendrag som er det eneste riktige, dermed blir det også til syvende og sist vanskelig å avgjøre på et objektiv grunnlag om ett sammendrag er bedre enn et annet. Til og med for relativt "enkle" avisartikler viser det seg at de som lager manuelle sammendrag stort sett er enig i kun 60% av tilfellene, når man måler overlapping av setningsinnhold (Radev et al. 2002).

På tidlige stadier i utviklingen av et sammenfatningsverktøy er det nødvendig med interne evalueringer for å kunne lokalisere problemer med programvaren som må løses. På senere stadier blir det viktigere med brukerevaluering, i og med at programvaren til slutt skal benyttes av en bruker. Brukerevalueringer er både resurs- og tidkrevende, og behovet for helt eller delvis å automatisere denne prosessen er innlysende.

Det finnes ikke noen entydig evalueringsmetode for automatiske sammenfattere. Årsaken til at metodene ofte blir individuelle og vanskelig å sammenligne på tvers av systemene, er at det ikke finnes noe fasitsvar, det finnes ikke ett sammendrag som er det eneste riktige, og det kan være vanskelig å velge ett av to sammendrag, fordi de uttrykker samme innhold på to litt forskjellige måter. Firmin og Chrzanowski (1999) påpeker at det eksisterer nesten en uniform enighet om at det ikke finnes noe perfekt sammendrag, en gullstandard, som alle andre sammendrag kan måles opp mot. Det finnes som regel flere gode sammendrag som alle gjengir innholdet i et gitt dokument på en tilfredsstillende måte. Og det er dette som er kjerneproblemet i evalueringsarbeidet som utføres innenfor forskningsfeltet, det er ikke mulig å generere en standard som kan gjelde for alle sammenfatningssystemer. Dermed må man hele tiden ta høyde for at det er flere sammendrag som kan være riktige, men Firmin og Chrzanowski (1999) sier også at enigheten om hvilke setninger som skal være med i et sammendrag øker når også kompresjonsgraden øker. Dermed er det mulig å oppnå en tilnærmet enighet om hvordan et referansesammendrag skal utformes innenfor et system, men også bare innenfor en relativt høy kompresjonsgrad, jfr Jing et al. (1998). I lys av disse aspektene blir det dermed ikke helt korrekt å kalle et referansesammendrag for en gullstandard, nettopp fordi dette kan gi assosiasjoner til at den representerer det eneste sanne sammendraget for en kildetekst.

Evalueringer som tar for seg sammenfattere laget for engelskspråklige tekster har en stor

fordel i motsetning til programmer laget for tekster på svensk eller norsk. I USA blir det for eksempel avholdt store evalueringskonferanser, som Text REtrieval Conference - TREC[4], som er sponset av National Institute of Standards and Technology - NIST og Forsvarsdepartementet. I tillegg til denne finnes også en annen konferanse, som også er sponset av NIST, nemlig Document Understanding Conference – DUC[5]. Disse konferansene gjør tilgjengelig store mengder testdata, som for eksempel taggete korpus som inneholder både kildetekst og sammendrag, og dette bidrar til at det kan utføres evalueringer i en annen skala enn for de skandinaviske språkene.

Det som kan synes spesielt for de aller fleste evalueringsmetoder som blir benyttet, er at de evaluerer ikke selve sammenfatteren og dens bestanddeler, men det som blir produsert; sammendraget. Men dette har sine naturlige grunner. En sammenfatter består av et helhetlig system som det kan være vanskelig å plukke fra hverandre. De statistiske metodene kan justeres og endres ut fra gitte spesifikasjoner, men det kan være vanskelig å isolere denne komponenten for å analysere den. Likeledes kan heller ikke de lingvistiske komponentene plukkes ut og analyseres for seg, fordi en sammenfatter inneholder for eksempel ikke en parser eller en grammatikk som kan hentes ut og testes. Den eneste komponenten som kan vurderes for seg er et eventuelt leksikon, men også her kan man støte på problemer, for hva skal det vurderes mot? Det som er det interessante aspektet ved et leksikon er om det tilfører noe ekstra til sammendraget, som for eksempel at sammendraget blir mer koherent og lesbart. I evalueringen av NorSum vil leksikonet bli evaluert indirekte. Med indirekte menes her at leksikonet ikke blir analysert isolert, men at et sammendrag generert ved bruk av leksikon og et sammendrag uten bruk av leksikon blir begge sammenlignet mot referansesammendraget.

Når automatiske sammendrag og sammendragsystemer skal evalueres er det generelt sett to egenskaper ved sammendrag som må måles og vurderes, og dette gjelder uansett hvilken praktisk metode som brukes. Det første er *kompresjonsraten* (CR), det vil si hvor mye kortere sammendraget er enn kildeteksten.

CR = lengden på sammendrag / lengden på full tekst

Det andre er *bevaringsgraden* (*Retention Ratio* – RR), det vil si hvor mye informasjon som er bevart fra kildeteksten.

RR = informasjon i sammendrag / informasjon i kildetekst

4 http://trec.nist.gov/
5 http://duc.nist.gov/

En evaluering av et system for automatiske sammendrag må i det minste takle begge disse egenskapene på en eller annen måte. Men i tillegg så må også sammendragenes kvalitative egenskaper vurderes, som for eksempel hvor sammenhengende og forståelig teksten er (Hassel, 2004).

*Precision* er måleenheten for det antall setninger i det automatiske sammendraget som også finnes i referansesammendraget. *Recall* blir definert som det antall setninger i referansesammendraget som finnes i det automatiske sammendraget. Ofte blir disse måleenhetene brukt til å presentere kvantitative resultater fra evalueringer, men det hender like ofte at disse angivelsene er mer villedende enn veiledende. Som man kan se av definisjonene så kan man hente ut den samme informasjonen fra *precision* og *recall*, derfor blir vanligvis bare en av dem oppgitt som en del av resultatene. I følge eksperimentene til Jing et al. (1998) viser resultatene deres at *precision* og *recall* lett blir påvirket av lengden på sammendragene. Det gjelder både lengden på sammendragene som sammenlignes og sammendragene som referansesammendraget er basert på. Og i og med at forskjellige evalueringer har forskjellige lengde på sammendragene, og at ikke alle evalueringer har en fastsatt lengde i sine eksperimenter, fører dette til at precision og recall på mange måter er måleenheter som stiller store krav ved tolkning fordi de ikke bidrar til at resultater kan sammenlignes på tvers av de ulike systemene. Selv innenfor sitt eget eksperiment var det vanskelig å sammenligne de ulike resultatene og resultatene variert ut ifra lengden på sammendragene. F.eks kunne *precision* være på 61% for et sammendrag på 10%, mens hvis sammendraget økte til 20% ble *precision* på 47%. Og sammenligning mellom de ulike systemene hadde begrenset verdi fordi sammenfatterne beregnet lengde ut ifra forskjellige metoder (Jing et al. 1998). Mye av årsaken til disse problemene kan spores til den binære strukturen til måleenhetene; svarene som returneres er enten rett eller gal, og dette passer dårlig sammen med den subjektive strukturen til sammendrag, hvor forskjellige setninger kan presentere de sammen nøkkelemnene i en tekst.

Et annet eksempel er hentet fra Firmin og Chrzanowski (1999). I deres studie skulle informanter avgjøre om et sammendrag var relevant i forhold til et forhåndsdefinert emne. I denne evalueringen så var det knyttet både relevante og ikke-relevante dokumenter til de forhåndsdefinerte emnene, og informantene skulle ved lesning av sammendragene avgjøre om kildedokumentet var relevant i forhold til emnet eller ikke. Precision og recall ble i

evalueringen definert slik:

Precision = True Positive / (True Positive + False Positive)

Dette betyr at ut fra den totale mengden av sammendrag som er vurdert av informantene til å være relevante, hvor mange er virkelig det? Siden kildedokumentene allerede var definert som enten relevante eller ikke, hadde man en slags fasit å forholde seg til.

Recall = True Positive / (True Positive + False Negative)

Dette betyr at ut fra den totale mengden av kildetekster som var klassifisert til å være relevante til emnet, hvor mange av de korresponderende sammendragene ble vurdert likt?

Selv om relevans til et emne allerede er forhåndsdefinert og at det dermed kan synes passende å bruke disse måleenhetene, så viser forfatterne til en del svakheter ved å presentere resultater på denne måten. På forhånd var altså alle kildetekstene definert som enten tilhørende eller ikke tilhørende til et emne. Innenfor eksperimentene som ble utført var det definert fem emner og innenfor hvert emne var fordelingen omtrent 50/50 for dokumenter med og uten relevans til emnet. Problemene oppsto hvis et dokument som er klassifisert som ikke relevant til emnet likevel inneholder mindre tekstsegmenter som er relevante. Hvordan skal da det tilhørende sammendraget klassifiseres hvis det inneholder disse tekstsegmentene? Da vil det tilsynelatende være relevant. Og kan man da kategorisk si at kildedokumentet ikke er relevant til emnet når det inneholder deler som er relevante? I vanlig dagligtale ville et slikt dokument bli klassifisert som litt relevant. Og dermed er man tilbake til utgangspunktet; den binære egenskapen ved måleenhetene *precision* og *recall* vs. den nyanserte / graderte egenskapen ved sammendragene og kildetekster.

### *3.1.1 Interne og eksterne metoder*

Evalueringsmetoder kan generelt deles inn i to grupper; *eksterne* (extrinsic) og *interne* (intrinsic) metoder (Spärk Jones, 1999 og Mani, 2001a). Og sagt på en litt forenklet måte, så benytter man en ekstern metode hvis man vil undersøke hvordan sammendraget kan brukes, mens ved bruk av en intern metode så undersøker man de interne kvalitetene ved sammendraget, enten kvantitativt eller kvalitativt. Ved bruk av eksterne metoder blir kvaliteten på sammendraget vurdert ut ifra hvordan det tjener et formål, f.eks i forbindelse med innhenting av informasjon (IR). Dette kan være at en bruker avgjør relevansen av et dokument i forhold til et gitt emne, eller at en bruker svarer på spørsmål basert på lesing av sammendrag. Ved interne evalueringsmetoder vurderer man kvaliteten på

sammendrag basert på en direkte analyse av sammendraget. Dette kan for eksempel innbefatte en vurdering av flyten i teksten, dekningsgraden av antatte nøkkelemner eller likhet med et såkalt ideelt sammendrag. En av forskjellene mellom den eksterne og interne metoden er at den interne metoden kan helt eller delvis automatiseres. Hvis det automatiske sammendraget f.eks skal sammenlignes mot et referansesammendrag, så er dette noe som kan gjøres automatisk, ved at sammendragene blir sammenlignet setning for setning og man kan få resultatene presentert i f.eks antall overlappende setninger. Vurdering av et sammendrags relevans til et gitt emne kan derimot vanskelig la seg automatiseres.

Hassel (2004) påpeker at det ved interne metoder i hovedsak fokuseres på koherens og bevaring av informasjon fordi det er her det er antatt at de største problemene forekommer. Sammendrag som blir generert ved bruk av ekstraheringsmetoder, det vil si en slags klipp-og-lim-metode hvor fraser, setninger eller avsnitt blir trukket ut av teksten og satt sammen til et sammendrag, lider ofte av problemer med koherens. Det kan være løse anaforer eller hull i den retoriske strukturen av teksten som oppstår fordi delene ofte blir ekstrahert ut av kontekst. Graden av koherens i sammendraget kan så bli vurdert av testpersoner som sammenligner dette med et referansesammendrag eller med kildeteksten.

Problemet med bevaring av informasjon (Retention Ratio – RR), det vil si i hvor stor grad informasjonen (nøkkelemnene) er bevart i sammendraget, er det andre aspektet det må taes høyde for i evalueringen. For å undersøke dette kan testpersonene enten sammenligne sammendraget mot kildeteksten for å se hvor mye informasjon som er ivaretatt, eller mot referansesammendraget for å se om det genererte sammendraget inneholder de samme nøkkelemnene som referansesammendraget gjør. Mani (2001b) diskuterer de samme aspektene under punktene *Quality* og *Informativeness*, hvor han påpeker at en vurdering av *kvalitet* nødvendigvis alltid vil utføres av mennesker. Konsekvensen av det er at det må taes høyde for den subjektiviteten som da vil spille inn. For selv om personene som skal utføre evalueringen har et sett av kriterier å forholde seg til, er det likevel mulig at disse kan tolkes forskjellig. Denne typen vurdering av kvaliteten på et sammendrag og resultatene, kan være vanskelig å kvantifisere, for hvordan kan man gjøre om "svært enig" til tall? En av mulighetene som Mani (2001b) skisserer er måleenheten *Kappa*, som er relativt mye brukt innenfor datalingvistiske eksperimenter. *Kappa* blir regnet ut på følgende måte:

$$K = P(A) – P(E) / 1 – P(E)$$

Her er P(A) antall ganger testpersonene er enig og P(E) er antall ganger man forventer at testpersonene skal være enig, justert for tilfeldig enighet. Dette vil da gi K = 1 hvis testpersonene er fullstendig enig og K = 0 hvis det ikke er annen enighet mellom testpersonene enn det som kan forventes av tilfeldig enighet.

På den annen side, så er ikke kvaliteten på et sammendrag, på den måten Mani (2001b) beskriver det under *Quality*, det eneste kriteriet man må evaluere. Det er fullt mulig å ha et sammendrag som er godt skrevet, har god flyt og ingen løse anaforer, men som likevel kan være ukorrekt, irrelevant eller på andre måter ubrukelig i forhold til hva sammendraget var ment å brukes til. Dermed viser det seg at *bruk* av sammendrag også er et viktig aspekt ved en evaluering nettopp fordi det sier noe om kravene som må stilles til sammendraget, i tillegg til kvaliteten.

En tidlig undersøkelse som bidrar til å understreke problemet med subjektive forskjeller i genereringen av sammendrag og problemet med at det ikke finnes et fasitsammendrag å sammenligne mot, er evalueringen som ble utført av Rath et al. (1961). Denne evalueringen kan også tjene som et godt eksempel på bruk av en intern evalueringsmetode. Her var formålet med evalueringen å avdekke de intersubjektive forskjellene, forskjellene mellom maskinprogrammene og menneske-maskin forskjeller med hensyn til utvelgelse av setninger. Evalueringen ble gjennomført på følgende måte: Seks informanter fikk ti Scientific American-artikler hver (de samme artiklene til hver informant) hvor de skulle plukke ut de tyve mest representative setningene og rangere disse setningene etter hvor representative de var for innholdet i artikkelen. I tillegg ble det generert fem automatiske sammendrag etter fem forskjellige metoder, se Rath et al. (1961) for nærmere spesifikasjoner. Resultatene som ble presentert viste at det var stor forskjell mellom informantene, alle seks var enig om gjennomsnittlig 1.6 setninger pr artikkel, dvs 8%. Men fem av seks informanter var enig om 6.4 setninger, altså 32% av setningene. I kontrast til dette så var de fem maskinmetodene enig om et gjennomsnitt på 9.2 setninger pr artikkel, dvs 46%. Men fire av fem maskinmetoder var enig om 17 av 20 setninger. Sammenligningen mellom informantene og maskinmetodene viste at det kun var enighet om 0.2 setninger i gjennomsnitt. Så konklusjonen her ble at det var størst forskjell mellom maskinmetodene og informantene. I tillegg til denne evalueringen hadde Rath et al. (1961) også et annet eksperiment hvor 5 informanter ble presentert for seks

Scientific American-artikler, var det bare et sammenfall på 55% når informantene ble presentert for de samme artiklene åtte uker senere. Hvorfor blir sammendrag subjektive? Personen som lager sammendrag er influert av sin bakgrunn, holdninger og disposisjon, og personens egen mening eller nåtidige interesser kan noen ganger påvirke dens tolkning av forfatterens ideer (Luhn 1958).

Spärck Jones (1999) tar opp igjen problemstillingene rundt *kontekstfaktorer* når det kommer til evaluering, og mener det er umulig å evaluere sammendrag skikkelig uten å vite hva de skal brukes til og i hvilke situasjoner. Hun peker på at eksterne metoder som gjør bruk av spørsmål-og-svar-skjemaer på forskjellig måte, er mangelfulle. Og at interne metoder som sammenligner manuelle sammendrag med automatiske, er det beste alternativet, men at det ikke er en triviell oppgave. Problemet med denne evalueringsformen er at den antar at det manuelle sammendraget er den beste referansestandarden, noe det ikke nødvendigvis er. Og da kommer man igjen inn på kjerneproblemet innenfor automatisk tekstsammenfatning; at ikke finnes noen standard som sammendrag kan måles mot, nettopp fordi det finnes flere sammendrag av en kildetekst som er sanne, og kan være brukbare, og at denne avgjørelsen er en subjektiv avgjørelse.

## *3.2 Evaluering av SweSum*

I forbindelse med ScandSum-nettverket har det blitt utført en analyse av hvordan manuelle sammendrag blir laget og hvordan SweSum kan bli brukt i en svensk avisredaksjon (Fallahi 2003). Formålet med undersøkelsen var å se på hvordan en sammenfatter kunne ha vært integrert i en avisredaksjon. De mulige behovene som ble kartlagt var at journalister kunne trenge et verktøy som effektivt kunne korte ned artiklene de skrev, og redaksjonen kunne benytte en sammenfatter for å korte ned artiklene slik at de passer inn i den tilmålte plassen på en avisside. I tillegg kunne det vært interessant å ha et verktøy som lett kunne forkortet en tekst ned til et format som passer til SMS eller WAP. 308 artikler ble undersøkt, men av disse ble bare 39 prosent, dvs 119 artikler, sammenfattet i redaksjonen. De samme artiklene ble deretter kjørt gjennom SweSum for å kunne sammenligne de manuelle sammendragene med de automatiske. De kvantitative resultatene viste at av 106 artikler er det for 17 av de et avvik på mer enn 10 prosent mellom de manuelle og de automatiske sammendragene. Hvis man så ser bort ifra disse 17 avviker de manuelle fra de automatiske med bare to tegn. Da kan det

tilsynelatende se ut som at de manuelle og de automatiske sammendragen nærmest er identiske. Men i tillegg til dette ble det også gjort en kvalitativ undersøkelse hvor åtte personer skulle vurdere innhold, grammatikk, koherens og innhold med hensyn til originalartikkelen. Resultatene viser at her kom de automatiske sammendragene dårligst ut. De manuelle scoret gjennomgående mye høyere enn de automatiske (over 50 prosent bedre). I konklusjonen peker Fallahi på at mange små detaljer senker kvaliteten og at det kan være nyttig å revurdere heuristikkene, som f.eks at det skal kunne være mulig å dele opp lange setninger og hente ut deler av en setning. Av de mer positive resultatene er at 72% av artiklene som ble kortet ned til SMS-format (160 tegn) ble oppfattet som like gode som de som ble manuelt forkortet. Det kan antas at de gode resultatene skyldes at 160 tegn tilsvarer 2-3 setninger, og at for sammendrag av såpass kort lengde er både informanter og automatiske sammenfattere relativt ofte enige, jfr diskusjon i kap. 4.1.3.

Det har blitt utført en rekke evalueringer av SweSum i løpet av utviklingen av programmet, ved å benytte både eksterne og interne metoder. Dette ble gjort for å lettere kunne lokalisere svakheter ved programmet, men også for å få nyttig tilbakemelding fra brukere. Et eksempel på bruk av en ekstern metode finnes i Dalianis (2000). Her ble ni studenter gitt ti tekster hver, avisnyheter og filmanmeldelser, som de skulle lage automatiske sammendrag av. Først leste de gjennom hele teksten og deretter skulle de generere sammendrag ved hjelp av SweSum hvor de gradvis senket prosentsatsen på hvor langt sammendraget skulle være i forhold til kildeteksten. Samtidig skulle de notere på et skjema når sammenhengen ble brutt og når viktig informasjon gikk tapt. På grunn av mangelfullt materiale ble det valgt å regne ut medianen istedenfor gjennomsnittet, og resultatene ble at informasjonen var beholdt inntil en kompresjonsgrad på 30%, mens koherens ble bevart ned til en kompresjonsgrad på 24%.

En annen type evaluering benyttet seg av spørsmål og svar-skjema, hvor ti studenter skulle vurdere tre sammendrag med forskjellig kompresjonsgrad, dvs 10, 20 og 30 prosent, og krysse av for ved hvilken kompresjonsgrad det ikke var mulig å få svar på spørsmålene. Begge disse evalueringene ble gjort i henhold til metodene til Firmin og Chrzanowski (1999), hvor det ble lagt vekt på koherens og innhold, dvs bevaring av informasjon. Det viste seg at koherens var intakt ved sammendrag komprimert til 30 prosent, mens innholdet var intakt ved 25 prosent. Denne typen evaluering er kvalitativ og subjektiv, og det ble etterhvert et ønske å kunne foreta en mer objektiv evaluering med

muligheter for å kunne automatisere deler av prosessen, men til dette var det behov for et annotert korpus, og noe slikt var ikke tilgjengelig på svensk (Dalianis og Hassel 2001).

Hassel (2003) utviklet senere et slikt korpus, som inneholder et antall avisartikler og tilhørende manuelle sammendrag (ekstrakter). Det ble også utviklet et grafisk grensesnitt til korpuset som informantene skulle benytte seg av når de lagde sammendrag. Evalueringen som ble utført med dette materialet er et eksempel på en intern metode, og ble brukt i evalueringen av SweNam[6]; SweSum med navnegjenkjenner. Her ble det samlet inn manuelle sammendrag som ble laget av informanter, til sammen 96 sammendrag av 10 nyhetsartikler, og disse ble lagret i en database, sammen med de tilhørende avisartiklene. Deretter ble det generert automatiske sammendrag av alle avisartiklene både med SweSum og med SweNam.

Ut ifra de manuelle sammendragene ble det laget et majoritetssammendrag, som inneholdt de mest frekvente setningene tilsvarende gjennomsnittslengden for alle sammendragene. Dette majoritetssammendraget ble altså laget ut ifra de setningene som ble plukket ut av flest informanter. Når det ble regnet ut gjennomsnitt for alle artiklene viste det seg at enigheten om hvilke setninger som skulle plukkes ut til et sammendrag, lå på bare 39,6%. Men når majoritetssammendragene ble generert, og den gjennomsnittlige enigheten kun ble regnet ut på disse, så steg den til 68,9%. Veldig få av setningene i majoritetssammendraget ble valgt av så få som en tredjedel eller mindre av informantene, men enda færre setninger ble plukket ut av alle informantene. Majoritetssammendraget ble sammenlignet med begge de automatiske sammendragene på setningsnivå og dette ble gjort for hver av artiklene. Resultatene ble at majoritetssammendraget og det automatiske som ble generert av SweNam, hadde bare 33,9% av setningene til felles, mens sammenlignet med sammendraget generert av SweSum hadde de 57,2 % av setningene til felles. Dette sier selvsagt ikke noe om hvor bra de automatiske sammendragene er, bare hvor mye de overlapper med det som informantene synes var de mest informative setningene.

Et av hovedproblemene som ble avdekket i denne evalueringen var at navnegjenkjenningsmodulen hadde en tendens til å favorisere setninger som hadde en utdypende funksjon heller enn setninger som introduserte ny informasjon. I avisartikler vil

6 http://www.nada.kth.se/iplab/hlt/swenam/index.html

dette i praksis si setninger som kommer et stykke nedover i avsnittene, istedenfor setningene som kommer først i et avsnitt. Dermed går en del bakgrunnsinformasjon tapt i og med at dette ofte blir introdusert først i et avsnitt. En problemstilling det dermed må taes hensyn til er hvordan navnegjenkjenneren skal vektes i forhold til andre parametre, f.eks de som vekter setningsplassering. I tillegg var det også et problem at sammendragene fikk et noe hakkete uttrykk fordi setninger som inneholdt navn som oftest ble plukket ut, med det resultat at sammenhengen ble veldig repeterende. Dermed vil det være nødvendig å løse dette ved å bytte ut noen av navnene med tilsvarende pronomen.

Selv om det bør finnes løsninger på de nevnte problemene, så kan likevel sammendrag med navnegjenkjenning ha noe for seg. I nyhetssammenhenger hvor bakgrunnsinformasjon kan være mer eller mindre kjent fra før kan det være tilstrekkelig å lese denne typen av sammendrag som fokuserer på f.eks personnavn og stedsnavn. Da vil man kunne få en kjapp oppdatering på de siste hendelsene og hvem som eventuelt var involvert. For flere og utdypende detaljer omkring bruk av navnegjenkjenner i forbindelse med sammenfatteren SweSum, se Hassel (2003).

# 4 Evaluering av NorSum

I evalueringen av NorSum var det to hovedmål det var interessant å undersøke, hvor det ene var et delmål av det andre. En evaluering av en sammenfatter kan utføres på flere måter, men felles for dem alle er at de ofte er subjektive og tidkrevende. Jeg ønsket å undersøke om det var mulig å benytte en intern kvantitativ metode, for å dempe den subjektive påvirkningen, men likevel få tilbake informasjon om kvaliteten på de automatiske sammendragene. Metoden jeg benyttet gikk ut på å sammenligne de automatiske sammendragene mot referansesammendrag (RS) og undersøke både overlapping og avvik mellom disse. Delmålet i dette spørsmålet var om det ville la seg gjøre å generere referansesammendraget automatisk på bakgrunn av manuelle sammendrag og hvorvidt det var tidsbesparende å gjøre dette.

Valget av evalueringsmetode, tok utgangspunkt i det evalueringsarbeidet som ble gjort ved KTH for SweSum. Det ble dermed valgt en intern evalueringsmetode som fokuserer på en intern vurdering av sammendragene som blir produsert av sammenfatteren. Disse sammendragene blir sammenlignet med et referansesammendrag (RS) som blir generert av et program skrevet i Perl, som benytter manuelle sammendrag som innputt. I evalueringsfasene vil RS bli målt mot automatiske sammendrag generert av NorSum tilknyttet de norske språkresursene, og mot automatiske sammendrag generert av NorSum uten noen spesifikke språkresurser.

Det var også praktiske hensyn som talte for valg av evalueringsmetode; det var ikke mulig å få tilgang på et stort utvalg av testpersoner eller profesjonelle sammenfattere, og dermed ville det bli vanskelig å utføre en evaluering av ekstern type. Men for å kunne hente inn nok manuelle sammendrag, var jeg likevel avhengig av å ha en del personer til å lage disse. Så selv om det var mulig å lage sammendragene via en side på internett viste det seg at det ble en lang og tidkrevende prosess å hente inn selv et minimum av manuelle sammendrag.

## *4.1 Utvikling av testgrunnlaget*

Valg av evalueringsmetode for NorSum baserer seg på erfaringene fra evalueringen av SweSum, og etter flere evalueringsrunder ved KTH var konklusjonen derfra at den mest hensiktsmessige måten for evaluering for dette systemet, var å lage et

referansesammendrag og så sammenligne det automatiske sammendraget med dette. For å kunne bygge et referansesammendrag måtte det fremskaffes et lite korpus som besto av avisartikler og tilhørende manuelt lagde sammendrag. Jeg vil i dette kapitlet gå gjennom oppbyggingen av testgrunnlaget, det vil si innhentingen av avisartiklene og arbeidet med å hente inn de manuelle sammendragene.

### *4.1.1 Avisartiklene*

Ved oppstarten av oppgaven var intensjonen å hente ut både avisartikler og sammendrag fra Bergens Tidendes (BT)[7] nyhetsarkiv. Men det viser seg at BT ikke har rutiner for å lage sammendrag som ville tilfredsstilt mine krav, siden det i realiteten ikke blir laget noen sammendrag hos dem. Måten redaksjonen arbeider på er at journalisten får beskjed om hvilken type artikkel som skal skrives og lengden på artikkelen han eller hun skal skrive. Hvis det viser seg at artikkelen likevel blir for lang, så kortes den som oftest ned ved å kutte et avsnitt eller noen setninger på slutten. Dette kan gjøres på bakgrunn av avisartiklenes faste oppbygging og gjør det relativt enkelt for de ansatte på desken i en avis å tilpasse artiklene det formatet de skal inn i. Dermed måtte jeg vurdere andre metoder for å hente inn de manuelle sammendragene og valget mitt falt på bruk av frivillige personer som skulle lage sammendrag. (Se kap 4.1.3)

Det ble i første omgang hentet ut 30 tekster fra nyhetsarkivet til Bergens Tidende og det ble forventet at hver av disse tekstene skulle ha 10 sammendrag tilknyttet seg. Etterhvert ble dette kuttet ned til 20 artikler, med forhåpninger om å få inn minimum 15 sammendrag pr. artikkel (se kap.4.1.2). Tematisk er tekstene hentet fra forskjellige områder som sport, kultur, lokalnyheter, politikk og økonomi. Artiklene er nyhetsartikler og reportasjeartikler og noen av de inneholder også lengre dialog- (intervju) sekvenser. Dette er tekster som er refererende og ikke argumentative, som f.eks kronikker. Det er tilstrebet å ha ulik lengde på artiklene for å få et rimelig variert materiale. Så ut fra disse subjektive kriteriene om tema, sjanger og lengde på artiklene, ble det forsøkt å oppnå en viss variasjon av artikler, for på et senere tidspunkt å kunne vurdere om dette spiller inn på hvor enig informantene er angående valg av setninger til sammendragene.

### *4.1.2 Databasen*

Databasen som ble utviklet i forbindelse med denne oppgaven, ble utviklet i samarbeid

7 http://bt.no/

med Aleksander Krzywinski, student ved Institutt for informasjonsvitenskap, UiB. Det var behov for en slik database for å ha en oversiktlig organisering av avistekstene og deres tilhørende manuelle sammendrag. Det var nødvendig med en egen database for å lett kunne administrere og organisere tekstene her istedenfor å legge de inn i databasen som er utviklet ved KTH.

Databasen er bygget opp slik at hver avisartikkel får sin unike ID (tallnummerering). Hver artikkel er deretter delt opp i avsnitt som også får en unik ID, som er koblet til artikkel ID. I tillegg er teksten delt opp i setningsenheter som består av enkeltvise setninger som også har en unik ID. Dermed får man et hierarki som består av artikkel – avsnitt – setning (se appendiks C). Det er viktig at artiklene deles på denne måten med setninger som minste enhet, siden de manuelle sammendragene blir laget ved å trekke ut hele setninger. Avsnittinndelingen var viktig for at de manuelle sammendragene skulle inneholde de samme avsnittene som kildeartikkelen, og dermed se mer "riktig" ut for leseren.

Tekstene fra Bergens Tidende sitt nyhetsarkiv ble lagt inn i databasen etter en liten justering av formateringen. Det ble laget et grafisk grensesnitt, kodet i php, for å lette denne oppgaven. Informasjon som måtte legges til var avsnittinndeling, overskriftmarkering og uthevelser i fet skrift (gjaldt bare overskrifter og avsnittoverskrifter). Denne type informasjon var det viktig å ivareta av to grunner: For det første skulle artiklene senere leses av informanter som skulle lage de manuelle sammendragene, og det var viktig at artiklene tilnærmingsvis ble oppfattet som genuine avistekster og ikke som oppkonstruerte prøvekanin-tekster, for å forsøke å lage en tilnærmet autentisk situasjon. For det andre var det også viktig at tekstene som ble brukt i det manuelle arbeidet var identiske og så identisk ut, med de som ble brukt som innputt til NorSum. Dette var viktig for å senere kunne vise til at både de manuelle og de automatiske sammendragene var laget av det samme materialet. Og i denne sammenheng er det ikke nok at det tekstlige innholdet er det samme, den grafiske presentasjonen er også viktig, og særlig med hensyn til den formateringsinformasjonen som NorSum gjør nytte av (f.eks. avsnittsinndeling).

Forøvrig er databasen bygget opp slik at det er forholdsvis enkelt å legge til flere avisartikler om det senere skulle bli nødvendig. Jeg synes det var et poeng at databasen hadde en struktur slik at den senere har et potensiale for å danne grunnlag for et lite

korpus, og dermed være tilgjengelig for gjenbruk.

### *4.1.3 Manuelle sammendrag og informanter*

I forbindelse med utviklingen av testsettet til denne oppgaven var jeg avhengig av frivillige personer (heretter kalt informanter) som kunne lage manuelle sammendrag. I og med at valget av metode innbefattet et relativt stort antall manuelle sammendrag var det naturlig å inkludere informanter i utarbeidingen av sammendragene, og det ville uansett ikke være naturlig eller metodisk forsvarlig å lage disse selv. Etterhvert viste det seg at dette var en tid- og resurskrevende oppgave, og at innenfor rammene av en hovedfagsoppgave, tok det en større plass enn først forventet. Selv om det var viktig å tilstrebe en viss variasjon innenfor valg av testpersoner, er utvalget av frivillige langt fra tilfeldig nok, rent statistisk sett. Jeg har valgt personer ut ifra "rullende snøball"-prinsippet; jeg spurte så mange som jeg kjente og ba disse om å spørre sine bekjente igjen. Og i og med at oppgaven med å lage automatiske sammendrag kunne gjøres via en inernettside, så har jeg ikke hatt full kontroll med hvem som har laget sammendragene. Dermed må jeg ta en del for gitt angående de som har bidratt til å lage sammendrag; jeg må anta at de har et minimum av språksans og jeg må anta at de har tatt oppgaven seriøst og gjort sitt beste. En annen grunn til at det ikke var en absolutt nødvendighet å ha oversikt over informantene, var at det i denne oppgaven ikke blir studert hvordan mennesker lager sammendrag. Et studie av den prosessen ligger utenfor rammene av denne oppgaven.

Resultatet av prosessen med å hente inn manuelle sammendrag ble at det måtte settes en strek for innsamlingen av sammendragene på grunn av tidsmangel, slik at det var mulig å få gjennomført arbeidet i testing- og evalueringsfasen innenfor de rammene som var satt. Dermed er det muligheter for at testgrunnlaget er for smalt og at et større statistisk grunnlag vil kunne gi andre resultater. Jeg valgte likevel å gjennomføre testingen på denne måten, fordi utprøvingen av evalueringsmetoden og programmet som genererer referansesammendraget, kunne utføres innenfor rammene av det materialet jeg hadde tilgjengelig. Dermed må evalueringen av NorSum sees på mer som en pilotstudie for å prøve ut en metode hvor genereringen av referansesammendraget blir automatisert, heller enn en statistisk undersøkelse.

Det ble laget en side med grafisk grensesnitt på internett lenket til databasen. Dermed

kunne informantene gå inn på siden og få en liste med lenker til artiklene, klikke på en av dem og få artikkelen presentert i full tekst. Her skulle de så lage manuelle sammendrag ved å klikke på de setningene som de mente var de mest meningsbærende og interessante for et sammendrag. Når informantene holdt musemarkøren over en setning ble denne farget gult slik at det skulle være enkelt å identifisere den enkelte setning. Setningene ble lagt inn i sammendraget fortløpende etter hvert som de ble valgt og sammendraget ble presentert nederst på siden. Informantene fikk minimalt med instruksjoner, fordi det ikke var ønskelig å styre informantene på noen måte. Den eneste spesifikke instruksen de fikk var det maksimale antall setninger som kunne taes med. Den øvre grensen tilsvarte en kompresjonsgrad på ca 50%. Ved KTH hvor denne fremgangsmåten også ble brukt, valgte de en øvre og nedre grense. Det er nødvendig med en viss grad av styring angående lengden på sammendragene for å kunne oppnå et mest mulig ideelt sammenligningsgrunnlag når det skal benyttes sammen med NorSum. Den ideelle sammendragslengden av avistekster ved bruk av NorSum er 20-30 prosent, som også er det som generelt regnes som den optimale komprimeringsgraden (Dalianis 2000).

Det kan selvsagt debatteres i hvor stor grad man skal styre informantene som skal lage sammendrag. I flere studier viser det seg at i jo større grad informantene styres, jo oftere enes de om hvilke setninger som er de mest viktige for et sammendrag. I evalueringen som ble utført av Jing et al. (1998) fikk informantene spesifikke instrukser angående lengden på sammendragene. For hver av de 40 artiklene som var med i evalueringen, skulle informantene lage ett sammendrag på 10% og ett på 20% av originalteksten; regnet ut på setningsnivå. I tillegg til disse manuelle sammendragene ble det også automatisk generert sammendrag ved hjelp av tre forskjellige maskinelle metoder. Hver av disse metodene genererte ett sammendrag på 10% og ett sammendrag på 20% for hver artikkel. Og resultatene var ganske oppsiktsvekkende: Sammenfall av valg av setninger var på 96% i gjennomsnitt for sammendrag med en kompresjonsgrad på 10% av kildeteksten, og litt under 90% sammenfall for sammendrag som består av 20% av kildeteksten. Forfatterne mener årsaken til de gode resultatene kan tillegges de spesifikke instruksene, og også likheten mellom tekstene, som alle var avisartikler hentet fra TREC. Når man kikker på tabellen som viser sammendragene for én artikkel, ser man at informantene skulle plukke ut to setninger til et sammendrag på 10% og tre setninger for et sammendrag på 20%, og i den sammenheng blir kanskje ikke resultatene så oppsiktsvekkende likevel. Hvis man går ut ifra at antall manuelle sammendrag som det er

mulig å lage er tilnærmet ubegrenset, så viser det seg at jo mer spesifikke instrukser informantene får, jo mer begrenset blir antall sammendrag det er mulig å lage. Og når informantene blir informert om at for et sammendrag på 10% så skal de plukke ut to setninger, da sier det seg selv at antall sammendrag det er mulig å lage er færre enn om sammendragene kunne hatt ubegrenset lengde.

I forbindelse med resultatene fra Jing et al. (1998), vil jeg henvise til tall fra mitt eget materiale. Informantene som ble benyttet fikk som nevnt, få og generelle instrukser. De fikk oppgitt et maksimum antall setninger de kunne plukke ut, og utover det skulle de plukke ut setninger de selv synes var meningsbærende og informative. Likevel så viste det seg at det var stor grad av enighet om hvilke to setninger som var de viktigste i artikkelen. Det var en enighet på omtrent 86% for de to mest frekvente setningene, mens uenigheten var mye større for hvor mange setninger som skulle inkluderes i sammendraget og hvilke setninger som skulle taes med utover de to mest frekvente. Dette kan selvsagt skyldes tilfeldigheter, men det at det var avisartikler med en ganske fast oppbygging kan selvsagt også spille inn. Det må jo også sies at Jing et al. (1998) tillegger artikkelsjangeren en viss påvirkning av sine resultater. Dermed kan det synes som at det ikke bare er spesifikke instrukser som påvirker enigheten mellom informantene, men like mye den fastsatte sjangeren artiklene befinner seg innenfor. Så om tiden hadde tillatt det hadde det vært interessant å sett på om informantene hadde vært like samkjørte om de hadde fått spesifikk instruks om å lage et sammendrag på 10%, og om de da også hadde plukket ut de samme setningene.

Som nevnt ovenfor var det viktig å opprettholde avisartiklenes opprinnelige utseende slik at informantene fikk et tilnærmet autentisk inntrykk av å lese avisartikler og ikke oppkonstruerte liksom-artikler. Den eneste vesentlige "mangelen" i avisartiklene er ingressen, men denne var heller ikke lagret i nyhetsarkivet til BT, mest sannsynlig fordi denne blir skrevet i etterkant. Jeg vil også påstå at en ingress vil være overflødig i denne sammenhengen i og med at den opptrer nettopp som et lite sammendrag av artikkelen og dermed ville kunne være med å styre informantenes tankegang og valg av setninger som skulle velges ut.

Informantene som benyttes i denne oppgaven er anonyme. Det var ikke interessant å kikke på hvordan forskjellige personer lagde sammendrag og sammenligne disse med

hverandre fordi som nevnt, lå dette utenfor oppgavens formål. Det som var interessant var å tilstrebe et statisk materiale som kunne gjenbrukes senere og dermed ble det også uvesentlig hvem som hadde laget hvilket sammendrag. Ved evalueringen av SweSum utført ble det en stund logget IP-adresser for å kunne følge med hvilke som lagde hvilke sammendrag, men dette ble mest en kuriositet, heller enn noe statistisk materiale i og med at en datamaskin potensielt kan brukes av flere personer.

## 4.2 Utvikling av testsettet og referansesammendrag

I dette kapitlet vil jeg gå nærmere inn på arbeidet med å utvikle referansesammendragene som skal danne grunnlaget for sammenligningen med det automatiske sammendraget. Først gir jeg en definisjon av hva begrepet referansesammendrag uttrykker i denne sammenhengen. Deretter tar jeg opp problemer med utviklingen av mine referansesammendrag og formulerer noen spørsmål som det ble arbeidet ut fra. Til sist presenteres programmet som ble utviklet for å generere referansesammendragene automatisk.

### 4.2.1 Definisjon av et referansesammendrag (RS)

Det er vanskelig å gi en entydig definisjon av ordet gullstandard. Ordet i seg selv henspeiler på en standard eller norm som det er ønskelig å oppnå, eller et ideal man streber etter, men som kan være uoppnåelig. I tillegg har denne definisjonen mange navn; gullstandard, majoritetssammendrag, "ideelt" sammendrag og referansesammendrag, som er den ternen jeg har valgt å bruke. Spärck Jones (1999) kritiserer bruken av gullstandarder, fordi denne strategien antar at de manuelt lagde sammendragene utgjør det beste sammenligningsgrunnlaget, noe som ikke nødvendigvis er tilfellet. Hassel (2003) definerer et majoritetssammendrag som et sammendrag som består av de mest frekvente setningene. Lengden på dette sammendraget tilsvarer gjennomsnittslengden av de manuelle sammendragene. Men for å kunne definere en klar majoritet vil det i denne sammenhengen være nødvendig med flere enn ti sammendrag pr avisartikkel. Derfor har jeg valgt å kalle dette sammendraget for et referansesammendrag, fordi å kalle det for majoritetssammendrag vil være litt misvisende i denne oppgaven, og fordi en gullstandard nettopp gir feile assosiasjoner til hva dette sammendraget er. Men uansett hvilket navn man velger å ha på et referansesammendrag, så er det stort sett bygget opp på samme måte, og dermed blir det vel også en smakssak hva man ønsker å kalle det.

Det som er interessant å kikke på når man benytter et refereansesammendrag som sammenligningsgrunnlag, er graden av overlapping det har med det automatisk genererte sammendraget. I tillegg kan man undersøke i hvor stor grad ett enkelt av de manuelle sammendragene overlapper med det automatiske sammendraget, og i hvor stor grad det overlapper med RS. Selv om det må antas at RS i tilstrekkelig grad representerer de manuelle sammendragene i og med at det er generert statistisk fra alle de manuelle sammendragene, er dette noe som også bør bli undersøkt kvantitativt. Det er også interessant å kikke på i hvor stor grad de manuelle sammendragene avviker fra RS, og ikke bare på graden av overlappende setninger.

I siste evalueringsrunde ved KTH (Hassel 2003) ble det laget mellom ti og femten manuelle sammendrag av forskjellige informanter, for hver artikkel. Alle setningene i hvert sammendrag blir satt opp i en frekvensliste hvor den setningen som flest informanter har plukket ut til sitt sammendrag, altså setningen med høyest frekvens, står øverst. Fra denne listen blir det deretter plukket ut det antall frekvente setninger som svarer til gjennomsnittslengden av sammendragene.

Dermed får man et referansesammendrag som inneholder de mest frekvente setningene for en artikkel. Lengden på referansesammendraget tilsvarer gjennomsnittslengden til de manuelle sammendragene. Problemet med å benytte en såpass enkel metode er at sannsynligheten for at flere setninger har samme frekvens er relativt høy. Dermed kan det oppstå et dilemma omkring hvilke setninger som skal inkluderes i referansesammendraget, som jo på forhånd har en bestemt lengde. Den løsningen som har vært benyttet tidligere har vært å plukke ut de setningene som forekommer tidligst i avisartikkelen, noe som kan være en gyldig metode på bakgrunn av diskursstrukturen i avisartikler. Det som det derimot også kan være interessant å kikke på i denne sammenheng, er hvilke setninger som oftest opptrer sammen i de manuelle sammendragene. Ved å fokusere på hvilke setninger som opptrer sammen, og ikke bare frekvens og plassering, kan det være mulig å utvikle et litt mer avansert referansesammendrag. Utvikling av metoden og programmet som genererer dette referansesammendraget blir forklart nærmere i kommende kapitler.

Ved å basere seg på en statistisk modell i utformingen av et referansesammendrag,

reduseres subjektiviteten til en viss grad, og det regnes dermed som en bedre løsning enn å ta utgangspunkt i kun ett manuelt laget sammendrag. Det kan likevel være interessant å undersøke i etterkant om referansesammendraget likevel er identisk med et av de manuelle sammendragene.

Mani (2001b) problematiserer begrepet gullstandard, ved å peke på at det også ofte blir kalt et majoritetssammendrag. Hvis en gullstandard skal velges ut på bakgrunn av hvilke setninger som har høyest frekvens, hva er da et majoritetsvalg? Hvis 4 av 5 informanter har valgt en setning, er denne kvalifisert til en gullstandard. Men hvis 3 av 5 informanter har valgt setningen, er det da en tilstrekkelig majoritet? I mitt materiale er referansesammendraget hovedsaklig generert ut ifra frekvens og dermed blir dette i realiteten majoritetssammendrag. Og hvis man antar at en majoritet er en majoritet så lenge den inneholder flere enn halvparten av stemmene, så inneholder mitt materiale likevel følgende fakta: For seks av sammendragene er minst halvparten av setningene "minoritetssetninger" dvs de er valgt ut av halvparten eller færre informanter. Et konkret eksempel: Artikkel nr.13 har et referansesammendrag som består av 12 setninger, av disse er 5 plukket ut av 5 informanter, mens 2 av de er plukket ut av 4 informanter. Dermed kan det være vanskelig å kalle dette et majoritetssammendrag, og jeg velger heller å kalle det for et referansesammendrag. Med et større statistisk materiale må man anta at dette kanskje er en problemstilling som løser seg selv.

Jing et al. (1998) foreslår en løsning på nevnte problemstilling hvor setninger ikke får en binær verdi, men en gradert vekting. I eksempelet over ville den første setningen fått verdien 4/5 og den neste setningen ville fått 3/5. På denne måten kan det være lettere å selektere de "riktige" setningene. Men for min del må jeg akseptere at dette er en begrensning i mitt testmateriale, og at de statistiske resultatene vil få en begrenset betydning.

### *4.2.2 Metoden bak referansesammendraget i oppgaven*

I dette kapitlet har jeg formulert to metoder som danner den teoretiske bakgrunnen for programmet som ble laget for å generere referansesammendragene automatisk. Utover å danne et generelt referansesammendrag kun basert på frekvens, plasseringen til setningene og gjennomsnittslengde av de manuelle sammendragene, så var det også interessant å kikke på hvilke setninger som opptrådte sammen i sammendragene, dette

gjaldt spesielt for de setningene som hadde lik frekvens. Når setningene som skulle utgjøre referansesammendraget ble plukket ut så ble disse valgt fra en frekvensliste, men i noen tilfeller hadde flere setninger samme frekvens, og hvis alle setningene innenfor en frekvensgruppe ble tatt med ville referansesammendraget bli for langt, ut fra gjennomsnittslengden til de manuelle sammendragene. Denne konflikten omkring frekvensgrupper er bare aktuell når de siste setningene til referansesammendraget skal plukkes ut. Databasen innhentet frekvenslistene automatisk, etterhvert som de manuelle sammendragene ble laget. Så på hvilket grunnlag skal så de siste setningene velges ut? Her har det blitt lagt vekt på to fremgangsmåter som blir illustrert med et tenkt eksempel lenger nede.

Metode I. Den første fremgangsmåten tar utgangspunkt i gruppen av setninger som har samme frekvens (den "usikre" gruppen), her kan det f.eks. være fem setninger, hvor bare tre av dem skal med i referansesammendraget. I det som da ble kalt metode I, ble to og to setninger fra den "usikre" gruppen sammenlignet med ett og ett manuelt sammendrag for å få en frekvens på hvor mange ganger disse to setningene opptrådte sammen i samme sammendrag. De med høyest score ble til slutt valgt til referansesammendraget, men hvis det også her var for mange med samme frekvens, ble setningene valgt på bakgrunn av plassering, det vil si at de setningene som står tidligst i artikkelen ble valgt fremfor de som kommer senere. Bakgrunnen for at metode I kan være en gyldig metode er at det antas at setningene med høyere frekvens enn den "usikre" gruppen naturlig nok opptrer ofte sammen siden de har høy frekvens og som oftest opptrer i mer enn halvparten av sammendragene. Dermed ble det viktig at de siste setningene som skulle inkluderes også hadde høy samopptreden innad, og at det innenfor frekvensgruppen "usikre" var setningene med høyest samopptreden som ble plukket ut.

Metode II. Den andre fremgangsmåten tar utgangspunkt i de setningene som allerede er plukket ut til å være med i referansesammendraget, altså de "sikre" setningene. Her blir alle setningene innenfor den "usikre" frekvensgruppen sammenlignet med alle setningene i den "sikre" gruppen. Dette ble gjort på den måten at en setning fra den "usikre" gruppen og en setning fra den "sikre" gruppen ble sjekket mot ett og ett manuelt sammendrag for å undersøke hvor mange ganger disse setningene opptrådte sammen. Bakgrunnen for at dette også kan være en gyldig metode er at det kan være viktigere at de resterende setningene som skal med i referansesammendraget har høy samopptreden med de som

allerede er kvalifisert til å være med i referansesammendraget, enn at de har høy samopptreden innbyrdes.

Både metode I og II vil så bli illustrert med et tenkt eksempel for å vise at de kan produsere forskjellige resultater: La oss anta at tabellen viser seks sammendrag (1-6) som består av et utvalg av setninger (A-F). Som man kan se er gjennomsnittslengden på sammendragene på tre setninger, det vil si at referansesammendraget her også skal være på tre setninger. Hvis man først ser på setningsfrekvensen, ser man at setning D har fått høyest frekvens, 4, mens A, B og E har fått 3 i frekvens. Men siden referansesammendraget bare skal inneholde tre setninger til sammen, må en av setningene i den lavere frekvensgruppen forkastes. Hvis bare plassering (det vil si plassering i artikkelteksten; den setningen som opptrer tidligst, har høyest prioritet) hadde blitt lagt til grunn for å velge ut setninger, ville setning E blitt forkastet. Men hvis man igjen kikker på hvilke setninger som opptrer sammen, så forekommer ikke setning A og B i samme sammendrag i noen av tilfellene. Dermed er det en viss sannsynlighet for at man vil ende opp med et sammendrag som er redundant og dermed ikke optimalt i forhold til kildeartikkelen, og heller ikke gjenspeiler de manuelle sammendragene på en god måte.

| ***Sammendrag*** <br> ***Setninger*** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| A | I | I | | I | | |
| B | | | I | | I | I |
| C | | | | | | I |
| D | I | | I | | I | I |
| E | I | I | | I | | |
| F | | | I | | I | |

Tabell 1.

Ved å bruke metode I vil to og to av setningene innenfor gruppen med 3 i frekvens bli sammenlignet sammen med ett og ett sammendrag for å se hvor ofte de forskjellige setningene opptrer sammen. I dette eksempelet opptrer setningene A og E oftest sammen (3 ganger) og dermed ville disse to setningene bli plukket ut til referansesammendraget sammen med setning D. Men det som her kan synes litt merkelig er at den setningen som forekommer oftest i de samme sammendragene som setning D, nemlig B, blir forkastet.

Hvis derimot metode II ble brukt så ville denne vurdere hvilke av setningene innenfor gruppen med 3 i frekvens, også kalt den "usikre" gruppen, som opptrer oftest sammen med setningene som har høyere frekvens, altså de i den "sikre" gruppen. Innenfor denne metoden ville setning E blitt forkastet fordi setning B har høyest antall forekomster sammen med setning D, som er den eneste med høyere frekvens, og setning A blir beholdt på grunn av tidligst forekomst i kildeteksten. Kritikken mot metode II går på det som er nevnt ovenfor, nemlig at setningene A og B ikke opptrer i samme sammendrag og dermed kan referansesammendraget også være redundant. Men programmet benytter begge metodene i genereringen av referansesammendrag, og så vil det i etterkant bli undersøkt hvilken metode som har generert det beste referansesammendraget, målt i overlappingsgrad med de automatiske sammendragene. Dette vil riktignok bare bli utført kvantitativt ved å sjekke antall setninger som overlapper mellom referansesammendragene og de automatiske sammendragene. For å kunne vurdere hvilke referansesammendrag som f.eks har best flyt eller er minst redundant, må det en kvalitativ undersøkelse til, og det ligger utenfor denne oppgavens område.

Innenfor forskning om evaluering av automatiske sammenfattere har det lenge vært kommentert at det ikke finnes noen rene objektive metoder for evaluering og at graden av subjektivitet kan påvirke resultatet av evaluering. Ved å bruke en metode hvor man sammenligner de automatiske sammendragene mot et referansesammendrag forsøkes det å unngå at testpersoner skal bedømme kvaliteten på et sammendrag, eller hvor mye en tekst kan komprimeres før de ikke får svar på gitte spørsmål. Sannsynligheten for at variasjonen mellom sammendragene likevel vil være høy er absolutt til stede. Mani påpeker dette i sin artikkel om evalueringsmetoder (Mani, 2001a). Han viser til at i studier som er gjort så viser det seg at testpersonene enes kun om gjennomsnittlig 1,6 setninger pr sammendrag. I tillegg så viser studier at når testpersonene ble presentert for de samme tekstene åtte uker senere, så produserte de sammendrag som var forskjellig fra de som ble produsert i første omgang, disse samsvarte på litt over halvparten av setningene. Så det er nok naivt å anta at de forskjellige vurderingene hver informanter har gjort, både når det kommer til lengde på sammendragene og hvilke setninger som skal taes med, ikke skal ha noen innvirkning på materialet som blir benyttet. Det hadde vært nødvendig å gange antall manuelle sammendrag med minst ti om man skulle kunne si at subjektiviteten hadde blitt utvisket.

## 4.3 Programmet som genererer referansesammendraget

I henhold til beskrivelsen av den metodiske bakgrunnen beskrevet i 4.2.2, ble det utviklet et program skrevet i programmeringsspråket Perl. Det ble valgt å bruke Perl fordi dette er et gunstig språk når det gjelder behandling av forskjellig typer lister, som arrays og hashtabeller. I tillegg var det selvfølgelig praktiske hensyn som måtte taes, som for eksempel at dette er det programmeringsspråket jeg selv mener jeg behersker best. Programmet genererer automatisk et referansesammendrag på bakgrunn av frekvenslisten og de manuelle sammendragene. Programmet tar som nevnt, hensyn til samopptreden av setninger og deres plassering i artikkelen. Og alle referansesammendragene ble først generert ved bruk av metode I og deretter metode II, og så sammenlignet for å identifisere om det var noen forskjell og evt hva denne forskjellen besto i. Nedenfor vises en algoritmisk fremstilling av hvordan programmet arbeider.

Innputtfilene til programmet er én fil med setningsnummer og deres respektive frekvens, og én fil med alle de manuelle sammendragene og setningene de inneholder. Altså to innputtfiler for hver artikkel (se appendiks B).

1 Innputtfilene leses inn i tabeller (*arrays*)
2 Tabellen med frekvensdata behandles først og denne deles inn i en to-dimensjonal tabell (med x- og y-koordinater), setninger med null i frekvens blir ikke tatt med. Underveis blir frekvenstallene summert og lagret i en variabel som brukes senere når gjennomsnittslengden skal regnes ut.
3 Gjennomsnittslengden på sammendragene regnes ut. Dette gjøres på setningsnivå.
4 Plukker ut riktig antall setninger som skal med i referansesammendraget i henhold til gjennomsnittslengden. Sjekker så om frekvensen til den siste setningen er den samme som den første av de setningene som ikke kom med i referansesammendraget. Hvis de er ulike skrives referansesammendraget ut.
5 Hvis frekvensene er like deles setningene inn i to grupper, hvor grenseverdien er frekvensen fra forrige trinn. Dermed blir setningene med høyere frekvens satt i gruppen for “sikre”, de med frekvens lik grenseverdien blir satt i gruppen “usikre”, mens de med frekvens lavere enn grenseverdien blir forkastet.
6 Filen med sammendragsdata blir gjort om til en hash-tabell for å lettere få tilgang til dataene.

7 Deretter utføres det en sammenligning for å avgjøre hvilke av setningene i den "usikre" gruppen som skal velges ut til referansesammendraget.

7.1 "Intern metode". To og to setninger innenfor gruppen "usikre" sjekkes mot sammendragene for å finne hvor mange ganger disse to setningene opptrer sammen i sammendragene.

7.2 "Ekstern metode". En setning fra gruppen "usikre" og en setning fra gruppen "sikre" blir sjekket mot sammendragene for å finne ut hvor mange ganger disse setningene opptrer sammen i sammendragene.

8 Hver setning får en frekvens, uavhengig av hvilken metode som er brukt, etter hvor mange ganger de opptrer sammen med de andre setningene.

9 På bakgrunn av denne frekvensen blir setningene i gruppen "usikre" sortert og de med høyest frekvens blir med i referansesammendraget (i forhold til gjennomsnittslengden). Setninger som har lik frekvens, blir sortert etter lokasjon. Det vil si at setningene med lavest nummer kommer øverst på listen.

10 Referansesammendraget skrives ut.

# 5 Praktisk utførelse av evalueringen og resultater

I gjennomførelsen av selve evalueringen viste det seg å bli mye manuelt arbeid. Selv om referansesammendraget (RS) ble generert automatisk, så var det mye annet arbeid som måtte gjøres manuelt, og det ble en tidkrevende prosess.

## *5.1 Kompresjonsgrader*

Den første utregningen jeg gjorde var å regne ut gjennomsnittlig kompresjonsgrad for hver artikkel og så regne ut gjennomsnittet samlet for alle artiklene. Kompresjonsgraden for hvert RS blir regnet ut på ordnivå ved at antall ord i RS blir delt på antall ord i kildeartikkelen. Det er viktig at kompresjonsgraden regnes ut på ordnivå og ikke på setningsnivå, siden setningene kan ha ulik lengde og dermed føre til at sammendragene også får ulik lengde, selv om de skulle inneholde like mange setninger. Det er nemlig denne kompresjonsgraden som blir skrevet inn i NorSum når de automatiske sammendragene skal genereres, og for at sammenligningsgrunnlaget skal bli så likt som mulig er det viktig at både RS og det autmatiske sammendraget har lik kompresjonsgrad.

Det viste seg at informantene var ganske uenig om hvor langt et sammendrag skulle være. Den eneste instruksen som ble gitt angående lengde på sammendragene var som nevnt, en maksimumsgrense for hvor mange setninger som kunne taes med. I følge flere studier viser det seg at informanter er relativt enig om de to første setningene som skal taes med i et sammendrag, men at de utover det er svært uenig (kap. 3.1.1).

For de RSene som ble generert ved bruk av metode I, var gjennomsnittlig kompresjonsgrad på 35,5%. Den høyeste kompresjonsgraden, dvs det korteste sammendraget, var på 24,8%, og den laveste, dvs det lengste sammendraget, var på 50,1%. Den gjennomsnittlige kompresjonsgraden for RS generert ved bruk av metode II var på 35,5%. Den høyeste kompresjonsgraden, var på 24,8% og den laveste kompresjonsgraden var på 49,7%. Dermed kan man se at det var ikke så store variasjoner i kompresjonsgrad mellom de to metodene. Se forøvrig appendiks D for å se kompresjonsgraden for hver enkelt RS.

## *5.2 Referansesammendrag*

Fra de tyve artiklene som utgjør testmaterialet mitt, genererte perl-programmet RSene på

bakgrunn av de manuelle sammendragene. Det var mellom10 og 14 sammendrag pr artikkel og i gjennomsnitt ble det 11 sammendrag pr artikkel. RSene dannet basis for sammenligningene mellom manuelle og automatiske sammendrag. I henhold til de to metodene omtalt i kap. 4.2.2. kunne RS bli generert på to måter. Men det viste seg at dette kun fikk utslag for 8 av RSene. Hovedproblemet i genereringen av RS var om setningene som skulle inkluderes, måtte plukkes fra en frekvensgruppe som inneholdt flere setninger enn de som var nødvendig for å oppnå riktig lengde på RSene. Av de 20 RSene som skulle genereres hadde 8 av dem ikke dette problemet med frekvens. Her sammenfalt antall setninger som skulle plukkes ut med skillet mellom to frekvensgrupper. Det vil si at disse RSene også kunne ha blitt generert uten hensyn til andre aspekter enn frekvens. Av de resterende 12 hvor det måtte velges ut et gitt antall setninger fra én frekvensgruppe som inneholdt for mange setninger, valgte både metode I og II de samme setningene i 4 tilfeller. Det vil si at RSene ble identiske uavhengig av hvilken metode som ble benyttet. For de resterende 8 RSene ble det plukket ut ulike setninger avhengig av hvilken metode som ble benyttet. Ulikhetene var derimot ikke så store, stort sett var det bare en setning som var forskjellig. Dermed er det 28 RS med i denne evalueringen, og i de tilfellene hvor en artikkel opererer med to ulike RS, og hvor dette er av avgjørende betydning, vil det bli presisert.

I kap 4.2.1 ble forskjellige definisjoner av RS tatt opp. For eksempel så ble et majoritetssammendrag der definert som et sammendrag som består av de mest frekvente setningene hentet fra de manuelt lagde sammendragene. Når det gjelder mitt eget materiale, så viser det seg at i mange tilfeller så består RS av flere minoritetssetninger enn majoritetssetninger. Minoritetssetninger ble definert som setninger som har fått halvparten eller færre av stemmene. De reelle tallene viser at gjennomsnittlig så består RS av 33,92% minoritetssetninger. RSet med flest minoritetssetninger hadde 62,5%, dvs at for 62,5% av setningene var under halvparten av informantene enig om at disse skulle med i sammendraget. Det RSet som hadde færrest minoritetssetninger hadde bare 6,3% minoritetssetninger.

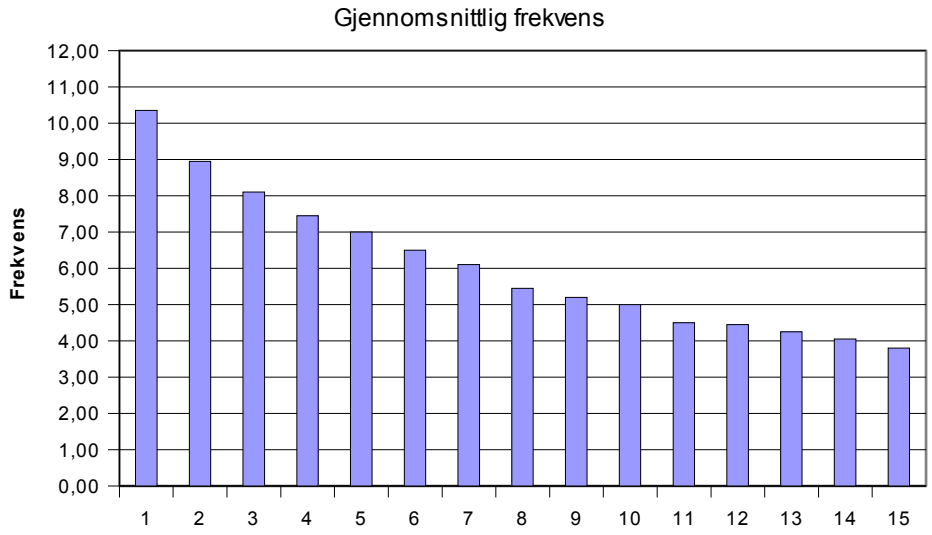



Diagram 1.

Diagrammet viser en oversikt over de mest frekvente setningene, regnet ut gjennomsnittlig over alle artiklene. Det vil altså si at den mest frekvente setningen har i gjennomsnitt en frekvens på 10 (10,35). Det betyr at 10 av informantene var enig om at dette var den viktigste setningen. Den neste setningen har en frekvens på 9 (8,95). Det betyr at 9 informanter var enig om at dette var en viktig setning å ha med i sammendraget. I gjennomsnitt var det 11 (11,15) sammendrag pr artikkel og det viser at det er relativt høy enighet blant informantene om hvilke to setninger er mest viktig for et sammendrag. Som man kan se av diagrammet så skrår søylene gradvis nedover og flater litt ut mot slutten. Dette tyder på at det er relativt enkelt å enes om de aller viktigste setningene, men at informantene varierer mer i avgjørelsene når det kommer til resten av setningene som skal med i sammendraget. Dette passer stemmer bra overens med hva andre studier også har kommet frem til, f.eks Jing et al. (1998), se kap. 4.1.3.

I forbindelse med diskusjonen omkring majoritets- og minoritetssetninger, så kan forskjellen mellom dem og deres innvirkning på RS tydeliggjøres ved å kikke på diagrammet. Siden det i gjennomsnitt var 11 sammendrag pr artikkel, så vil de setningene som har 5 eller færre stemmer i denne sammenhengen bli definert som minoritetssetninger. Av diagrammet kan vi se at 8 av setningene da blir å regne som minoritetssetninger. Det vil altså si at hvis diagrammet representerer et referansesammendrag med 15 setninger, så vil 8 av disse, over halvparten, være definert som minoritetssetninger. Kan dette RSet da sies å være en god gjennomsnittlig

fremstilling av de manuelle sammendragene? I og med at det er sannsynlig at det vil oppstå et problem i forbindelse med utvelgelse av setninger jfr. problemstillingen i kap 4.2.2, er det kanskje nødvendig å kikke på alternative metoder for å definere lengden på referansesammendragene. En mulig løsning kan være å avgjøre lengden på RS ved å bare ta med klare majoritetssetninger. Da vil man også være sikret at det er majoriteten som styrer hvilke setninger som er de viktigste for et sammendrag. Men dette kan nok se ganske annerledes ut ved et materiale som er ti ganger så stort. På den annen side så viser jo, som nevnt tidligere, flere studier at uenigheten mellom informantene øker i takt med lengden på sammendraget.

## *5.3 Sammenligningene*

I dette kapitlet vil jeg gå gjennom de forskjellige sammenligningene som har blitt utført i oppgaven. Den første sammenligningen som ble utført var mellom RS og NorSum tilknyttet leksikon. Her var det først og fremst antall overlappende setninger som var det mest interessante. Deretter ble RS sammenlignet med NorSum uten tilknytning til leksikon, dvs den språkuavhengige versjonen av SweSum-arkitekturen. Også her var det interessant å undersøke hvor mange setninger som overlappet, og sammenligne dette resultatet med det forrige. På den måten kan det undersøkes hvilken av de to versjonene av NorSum som hadde høyest overlapping, og på den måten kan også leksikonet evalueres på en indirekte måte. Den tredje sammenligningen som ble utført var mellom RS og de manuelle sammendragene. Dette ble gjort for å undersøke om RS virkelig representerte de manuelle sammendragene på en tilfredsstillende måte. Her ble det regnet ut både overlapping og avvik mellom RS og sammendragene. Alle sammenligningene ble utført manuelt på setningsnivå, og det viste seg at dette var en tidkrevende prosess. Dette har jo også blitt understreket i litteraturen og viser at det er et behov for å automatisere de prosessene som det er mulig å automatisere.

### *5.3.1 RS vs NorSum med leksikon*

Her ble RSet og det automatiske sammendraget generert av NorSum med leksikon sammenlignet. Den gjennomsnittlige overlappingen mellom RS generert av metode I og II og automatiske sammendrag var på 5,05 setninger. Det vil si at RSet og det automatiske sammendraget hadde 5,05 setninger til felles. Den høyeste overlappingen var på 8 setninger for RS generert av metode II, og 7 setninger for metode I. Den laveste overlappingen var 1 setning, uansett metode.

### *5.3.2 RS vs NorSum uten leksikon*

I denne sammenligningen ble RSet sammenlignet med det automatiske sammendraget som var generert av NorSum uten tilkobling til de norske språkresursene. Den gjennomsnittlige overlappingen var her 5,4 setninger for RS generert av både metode I og II. Den høyeste overlappingen var på 9 setninger og den laveste overlappingen var på 1 setning for både metode I og II.

I tillegg ble det regnet ut standardavviket både for denne sammenligningen og for den forrige. Når man sammenligner disse standardavvikene mellom NorSum med leksikon og NorSum uten leksikon, så er det en viss grad av variasjon her. For RSene som var generert etter metode I, ble disse sammenlignet med NorSum med leksikon og NorSum uten leksikon. Her var standardavviket på 1,54 med leksikon og 1,67 uten leksikon. For RSene generert etter metode II, var standardavviket på 1,70 ved sammenligning for NorSum med leksikon og 1,82 uten leksikon. Så selv om antall setninger som overlappet var likt for RS uansett hvilken metode de ble generert ut fra, viste det seg at variasjonen mellom artiklene var større for RS generert ved bruk av metode II enn metode I. For detaljert tallmateriale, se appendiks D (ark 2-3).

Resultatene fra overlapping av setninger er oppsummert i tabellen nedenfor. Her er det også tatt med overlapping mellom RS og manuelle sammendrag (MS) for sammenligningens skyld.

| | |
|---|---|
| Overlapp mellom RS og AS + leksikon | 5,05 med SD 1,54 (I) eller SD 1,70 (II) |
| Overlapp mellom RS og AS - leksikon | 5,40 med SD 1,67 (I) eller SD 1,82 (II) |
| Overlapp mellom RS og MS | 7,23 med SD 2,77 |

Tabell 2.

### *5.3.3 RS vs manuelle sammendrag (MS)*

Det neste steget i evalueringsprosessen var å sammenligne de manuelle sammendragene med RSene, for å undersøke i hvor stor grad hvert enkelt manuelt sammendrag avvek fra RSet. Det viste seg at ingen av de manuelle sammendragene sammenfalt fullstendig med et RS, noe som i seg selv er ganske interessant. Man kunne forventet at minst ett av de 223 sammendragene ville sammenfalle med et av RSene. Riktignok inneholdt noen av de manuelle sammendragene alle setningene som RSet besto av, men de inneholdt også flere setninger enn bare de i RSet. Sammenligningene ble utført delvis manuelt og delvis automatisk. Først ble det regnet ut hvor mange setninger som fantes i hvert sammendrag, men som samtidig ikke fantes i RSet. Deretter ble det automatisk regnet ut hvor mange setninger som fantes i RSet, men som samtidig ikke fantes i det manuelle sammendraget. For en nærmere studie av dette tallmaterialet, se appendiks D (ark 4-23).

Det ble regnet ut gjennomsnitt og standardavvik for de tallene som kom frem i disse sammenligningene. Dette ble gjort for å se på om avviket var større for setninger som var inkludert i manuelle sammendrag eller for setninger inkludert i RSene. Standardavviket ble regnet ut for å se på i hvilken gruppe variasjonen var størst. Det er usikkert om standardavviket egentlig tilfører noe informasjon i og med at materialet er relativt lite. Det ble også gjort en sammenligning hvor det samme avviket ble regnet ut for de automatiske sammendragene både med og uten leksikon. Det vil si hvilke setninger som fantes i de automatiske sammendragene, men ikke i RSene og hvilke som fantes i RSene, men ikke i de automatiske. Her ble det ikke regnet ut noe standardavvik siden det ikke ville bidratt med mer informasjon enn det man får av å se på differansen mellom de to tallene.

Resultatene fra sammenligningen mellom de manuelle sammendragene og RSene viser at de manuelle sammendragene avviker omtrent like mye fra RSene både når det gjelder setninger som er med i manuelle sammendrag, men ikke i RS og omvendt.

| | |
|---|---|
| I RS, men ikke i AS + leksikon | 7,00 |
| I RS, men ikke i AS - leksikon | 6,85 |
| I RS, men ikke i MS | 4,77 med SD 2,67 |
| I AS + leksikon, men ikke i RS | 11,10 |
| I AS - leksikon, men ikke i RS | 10,70 |
| I MS, men ikke i RS | 4,14 med SD 2,98 |

Tabell 3.

Det ble også regnet ut gjennomsnitt for disse sammenligningene, og det viser seg at forskjellen mellom de autmatiske og de manuelle sammendragene ikke var så stor som man kunne antatt. Antagelsen var at avviket ville være mindre for de manuelle sammendragene i og med at RSene er generert ut fra disse.

Det er viktig å påpeke her at disse tallene kan være mer villedende enn veiledende siden tallmaterialet for de to sammenligningene er ytterst forskjellig. De to øverste resultatene er beregnet på bakgrunn av tallgrupper med 10-14 tall, mens de to siste tallene er beregnet på bakgrunn av tallgrupper med 2 tall. Så man kan regne med at forskjellen kanskje ville blitt større ved et større utvalg.

# 6 Konklusjon

I denne oppgaven har jeg utført en evaluering av NorSum. Og i den forbindelse har jeg benyttet meg av en kvantitativ metode hvor jeg har sammenlignet de automatiske sammendragene med et referansesammendrag, som er generert fra manuelt lagde sammendrag. Jeg har utført en kvantitativ evaluering fordi jeg ønsket å undersøke om kvantitative resultater kan si noe om hvordan automatiske sammendrag forholder seg til manuelle sammendrag, uten å måtte utføre en subjektiv, kvalitativ undersøkelse. Resultatene viser at en kvantitativ undersøkelse også kan gi informasjon om hvordan en sammenfatter presterer i forhold til manuelle sammendrag.

## *6.1 Tolkning av resultatene*

Først vil jeg si noe om delmålet for oppgaven, nemlig automatisk generering av RS. Jeg ville undersøke om det var mulig og hensiktsmessig å generere RS automatisk, og i tillegg ta hensyn til andre aspekter enn bare frekvens i genereringen av disse. Det viste seg at forskjellene mellom metode I og II var minimale og at det var kun i 8 av 20 tilfeller at RSene var ulike, og i de fleste tilfellene var det bare en setning som var forskjellig. Dermed kan det se ut som at det ikke utgjør noen forskjell om man velger å generere RS ut fra metode I eller II. Det må i så fall en kvalitativ analyse til for å vurdere om det ene sammendraget er av bedre kvalitet enn det andre, og en vurdering av hvilke kriterier som eventuelt skal ligge til grunn for en slik analyse. Når det gjelder den gjennomsnittlige kompresjonsgraden for referansesammendragene er det også her minimale forskjeller mellom metodene. Dette henger selvsagt sammen med at kompresjonsgraden blir regnet ut på ordnivå, og når det var kun i 8 tilfeller at RSene varierte, og da som oftest bare med en setning, så kan det ikke forventes at det skal ha noen stor innvirkning på gjennomsnittslengden til RSene.

Det som forøvrig har vist seg å være nyttig er å generere referansesammendragene automatisk. Dette sparer både tid og arbeid, og det kan være en fordel å automatisere de deler av evalueringen som det er mulig å automatisere, siden evaluering har vist seg å være en tid- og resurskrevende prosess.

Når det gjelder resultatene fra evalueringen, synes jeg det er noen interessante punkter som er verdt å kikke på. Når det gjaldt antall overlappende setninger mellom det

automatiske sammendraget og RS, så var dette omtrent likt for NorSum både med og uten leksikon, selv om NorSum uten leksikon hadde en noe høyere grad av overlapping av setninger (5,05 mot 5,4 setninger). Man bør være forsiktig med å konkludere at et språkspesifikk leksikon ikke har noen hensikt, men det kan antydes at det ikke blir benyttet på en tilfredsstillende måte.

Når man vurderer avviket i antall setninger mellom manuelle sammendrag og RS, og mellom automatiske sammendrag og RS, er resultatene ganske interessante. For å referere til resultatene, så viser de at i gjennomsnitt så er det 4,1 setninger som finnes i de manuelle sammendragene som ikke finnes i RSet, og på samme måte så finnes det i gjennomsnitt 4,8 setninger i RSet som ikke finnes i de manuelle sammendragene. Kan man da konkludere med at RSene representerer de manuelle sammendragene på en tilfredsstillende måte?

De automatiske sammendragene hadde et større avvik fra RS, men noe annet kan heller ikke forventes siden RS er generert på bakgrunn av de manuelle sammendragene. Avviket fra RSet er altså større enn overlappingen. På den annen side så er det ikke sikkert at denne evalueringsformen er den beste for sammenfatteren. Hvis man skal følge oppfordringen fra flere forskere, så vil et større fokus på hva sammendragene skal brukes til og i hvilket miljø sammenfatteren benyttes, ha mye å si for hvilken evalueringsmetode som bør benyttes.

Det må kunne antas at med et større antall informanter vil resultatene fra denne evalueringen endre seg. Prosent betyr jo egentlig pr hundre, og med hundre sammendrag pr artikkel kunne resultatene blitt ganske annerledes. Et annet aspekt det også må taes høyde for er at det kan være stor forskjell på artiklene. Slik at det for noen artikler kan være stor uenighet blant informantene om hvilke setninger som skal med i sammendraget, mens det for andre artikler er relativt enkelt å avgjøre.

## *6.1 Perspektiver fremover*

På grunn av likhetene mellom de skandinaviske språkene har det vist seg nyttig å samkjøre forskningen innenfor automatisk sammenfatning (ScandSum 2003). Men det må også påpekes at den vellykkede overføringen av SweSum-arkitekturen til norsk og dansk

langt på vei var avhengig av gjenbruk av allerede eksisterende store leksikale resurser for de involverte språkene.

Det kan være vanskelig å spå om fremtiden, men Mani(2001b) har satt opp en ønskeliste for hvilke fokus forskningen bør ha fremover:

- Manuell sammenfatning: Økt studie av hvordan profesjonelle sammenfattere arbeider, slik at man kan trekke erfaringer til hvordan man kan gjøre det samme automatisk
- Abstrahering: Mer domeneuavhengig, korpusbasert forskning er nødvendig før man kan si at dette er en nyttig metode. Og det kan være nyttig å sammenligne abstrahering og ekstrahering for å se om det er mulig å trekke noen fordeler fra ekstrahering til bruk ved abstrahering
- Multidokumentsammenfatning: Dette er et område av interesse hvor det har vært lite empiriske studier på hvordan mennesker utfører denne typen sammenfatning. Interessante områder er biografiske sammendrag, oppdatert sammendrag av nyheter innenfor et domene og sammendrag av konkurrerende eller sammenfallende emner innenfor et domene. Sammenfatning av flere dokumenter krever ofte andre strategier enn sammendrag av enkeltdokumenter fordi kompresjonsgraden for hvert enkelt dokument ofte blir en tiendedel av hva som ofte er vanlig for enkeltdokumenter
- Multimediasammendrag: Forbedrete analyser av ikke-tekstlige media er nødvendig for å sette fart på dette forskningsområdet, som er et relativt nytt område i historisk sammenheng, og nye media krever nye metoder
- Evaluering: Nye evalueringsmetoder må undersøkes slik at man kan oppnå kostnadseffektive, brukersentrerte og repeterende evalueringer. I tillegg kan evaluering av spesielle oppgaver som f.eks sammenfatning for applikasjoner til mobiltelefoner, kaste lys over styrker og svakheter ved sammenfatning i konkurranse med annen teknologi. Det er også nødvendig at evalueringsmetoder på en adekvat måte modellerer virkelige situasjoner og informasjonsbehov, og konsentrerer seg om en variasjon av tekstsjangre og -lengder.

Utover dette vil fremskritt innenfor de forskjellige områdene være avhengig av at det gjøres tilgjengelig store annoterte korpus som består av manuelle sammendrag, det gjelder spesielt i forhold til evalueringsarbeidet. Dette er absolutt aktuelt for de skandinaviske språkene, som ikke har denne typen resurser tilgjengelig i like stor skala som f.eks engelsk.

Innenfor automatisk tekstsammenfatning kan man se at de generelle fremtidsperspektivene dreier seg mer og mer mot naturlig språkprosessering – NLP, og erfaringene som kan trekkes fra dette forskningsfeltet. Tekstsammenfatning vil helt klart kunne trekke fordeler av navnegjenkjenning, dvs identifisering og kategorisering av personer, steder, nasjonaliteter og land, anaforoppløsning og lenking av co-referanser. Navnegjenkjenning har allerede blitt prøvd ut for SweSum, med potensiale for gode resultater (Hassel 2003).

Innenfor forskningsområdet pågår det nå et norsk forskningsprosjekt innenfor KUNSTI-programmet til Norges forskningsråd; KunDoc – Kunnskapsbasert dokumentanalyse[8]. Dette er et samarbeidsprosjekt mellom CognIT as (CognIT 2004) og Universitetet i Bergen, og fokuserer på hvordan domenespesifikk semantisk kunnskap, lagret i ontologier, kan bli gjenbrukt i analyser av naturligspråklige tekster innenfor samme domene. Det vil spesielt bli lagt vekt på oppløsning av co-referanser ved å benytte verdenskunnskap lagret i ontologier.

For at SweSum skal få en reell nytteverdi i hverdagen er det nødvendig at programmet blir integrert sømløst i en nettleser eller et redigeringsverktøy, f.eks Illustrator eller QuarkExpress, på en måte hvor sammendragene kan bli generert ved hjelp av et enkelt museklikk (Fallahi 2003). Da kan det være til nytte i avisredaksjoner som har behov for å korte ned artikler, eller det kan benyttes av avislesere i deres jakt på oversikt og som en pekepinn på hvilke artikler som er interessante for videre lesning. Det må forøvrig påpekes at arkitekturen bak SweSum og utviklingen av en sammenfatter, var et forskningsstudium og ikke et prosjekt utviklet med tanke på å lage et kommersielt produkt.

Selv om prosjektet forsåvidt er avsluttet, er det likevel påtenkt forbedringer som forhåpentligvis vil bedre prestasjonene til SweSum / NorSum. Forbedringene listes kort opp her:

- Tagging av kildeteksten istedenfor et statisk leksikon
- Sammenfatning på frasenivå
- Forbedret navnegjenkjennelse

8 http://www.kundoc.net

- Forbedret pronomenresolusjon
- Leksikalske kjeder ved å bruke SIMPLE og/eller EuroWordNet
- Automatisk evalueringsmetode

Selv om det er et håp om at disse utvidelsene skal øke kvaliteten på sammendragene, er det en grundig evaluering som må avgjøre om det utgjør en reell forbedring. Derfor er det viktig at det i takt med forbedringene av SweSum også drives videre forskning på passende evalueringsmetoder.

# Referanser


Brandow, R., K. Mitze og L.F. Rau (1995): Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management*, 31(5): 675-685.

CognIT (2004): CognIT as, Halden, Norge. Tilgjengelig på http://www.cognit.com/home_multi/html/index.asp

Dalianis, H., M. Hassel, J. Wedekind, D. Haltrup, K. de Smedt og T. C. Lech (2003): From SweSum to ScandSum – Automatic text summarization for the Scandinavian languages. I: Holmboe, H. (red): *Nordisk Sprogteknologi 2002: Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Museum Tusculanums forlag.

Dalianis, H., M. Hassel, K. de Smedt, A. Liseth, T. C. Lech og J. Wedekind (2004): Porting and evaluation of automatic summarization. I: Holmboe, H. (red): *Nordisk Sprogteknologi 2003: Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Museum Tusculanums forlag.

Dalianis, Hercules (2000): "SweSum – A text summarizer for Swedish". Technical report, TRITA-NA-P0015, IPLab-174, NADA, KTH, Oktober 2000. Tilgjengelig på http://www.dsv.su.se/~hercules/Textsumsummary.html Sitert 10.09.03.

Dalianis, Hercules og Martin Hassel (2001): Development of a Swedish corpus for evaluating Summarizers and other IR-tools. Technical report, TRITA-NA-PO112, IPLab-188, NADA, KTH, Juni 2001. Tilgjengelig på http://www.dsv.su.se/~hercules/papers/TextsumEval.pdf Sitert 10.09.03.

Dalianis, Hercules (2004): What is automatic text summarization? Tilgjengelig på http://www.dsv.su.se/~hercules/textsammanfattningeng.html Sitert 22.06.04.

Edmundson, H. P. (1969): New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2):264-285. Også i: Mani, Inderjeet og Mark T. Maybury (red): *Advances in automatic text summarization*. MiT Press, 1999.

Fallahi, Sasan (2003): Computer aided text summarization. Using SweSum in a real newspaper environment. Slides tilgjengelig på: http://www.dsv.su.se/~hercules/scandsum/OHSasanFeforJan2003.pdf

Firmin,Thérèse og Michael J. Chrzanowski (1999): An evaluation of Automatic Text Summarization Systems. I: Mani, Inderjeet og Mark T. Maybury (red): *Advances in automatic text summarization*. MiT Press, 1999.

Goldstein, J., V. O. Mittal, J. G Carbonell og M. Kantrowitz (2000): Multi-Document Summarization by Sentence Extraction. I: *Proceedings of the Workshop on Automatic Summarization*, 40-48.

Hassel, Martin (2003): Exploitation of Named Entities in Automatic Text Summarization for Swedish. I: *Proceedings of NODALIDA '03 - 14th Nordic Conference on Computational Linguistics,* Reykjavik, Island.

Hassel, Martin (2004): Evaluation of Automatic Text Summarization. A practical implementation. Licentiate Thesis, KTH NADA, May 2004. ISBN 91-7283-753-5

Jing, H., R. Barzilay, R. McKeown og M. Elhadad (1998): Summarization Evaluation Methods: Experiments and Analysis. I: *Working notes of the Workshop on Intelligent Text Summarization*, 60-68. Menlo Park, California: American Association for Artificial Intelligence Spring Symposium Series.

Johnson, F. C., C. D. Paice, W. J. Black og A. P. Neal (1993): The Application of Linguistic Processing to Automatic Abstract Generation. *Journal of Document and Text Management*, 1(3): 215-241.

Luhn, H. P. (1958): The automatic creation of literature abstracts. I: IRE National Convention, pages 60-68, 1958. Også i: Mani, Inderjeet og Mark T. Maybury (red): *Advances in automatic text summarization*. MiT Press, 1999.

Mani, Inderjeet (2001): Summarization evaluation: An overview. I: *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization.*

Mani, Iderjeet (2001): *Automatic text summarization.* John Benjamins publishing company, 2001.

Mani, Inderjeet og Mark T. Maybury (red): *Advances in automatic text summarization*. MiT Press, 1999.

Mazdak, Nima (2004): FarsiSum. A Persian text summarizer. Master-oppgave. Tilgjengelig på http://www.dsv.su.se/~hercules/papers/FarsiSum.pdf Sitert 23.06.04.

McKeown, K., D. Jordan og V. Hatzivassiloglou (1998): Generating Patient-Specific Summaries of Online Literature. I *Working Notes of the Workshop on Intelligent Text Summarization*, 34-43.

Norfa - Nordisk forskerutdanningsakademi: Tilgjengelig på http://www.norfa.no/ Sitert 27.07.04

Norges forskningsråd: Tilgjengelig på http://www.forskningsradet.no/ Sitert 27.07.04

NorSum (2003): NorSum-demo på internett. Tilgjengelig på http://swesum.nada.kth.se/index.html

Radev, Dragomir R. og W. Fan (2000): Automatic summarization of search engine hit lists. I: *Proceedings of the Workshop on Recent Advances in NLP and IR*, 99-109.

Radev, Dragomir R., Eduard Hovy og Kathleen McKeown (2002): Introduction to the Special Issue on Summarization. I: *Computational Linguistics*, Volume 28, Number 4, pp.399-408.

Rath, G. J., A. Resnick og T. R. Savage (1961): The formation of Abstracts by the selection of sentences. I: *American Documentation*, 12(2). Også i: Mani, Inderjeet og

Mark T. Maybury (red): *Advances in automatic text summarization*. MiT Press, 1999.

Rosén, Victoria og Koenraad de Smedt (1999): SCARRIE. Scandinavian Proofreading Tools. Tilgjengelig på http://www.ling.uib.no/~desmedt/scarrie/ Sitert 08.06.04

ScandSum (2003): ScandSum – Summarization network in Scandinavia. Tilgjengelig på http://www.dsv.su.se/~hercules/scandsum.html Sitert 29.07.04

Spärck Jones, Karen (1999): Automatic summarizing: Factors and directions. I: Mani, Inderjeet og Mark T. Maybury (red): *Advances in automatic text summarization*. MiT Press, 1999.

# Appendiks A: Programmeringskode

Programmeringskode for generering av referansesammendraget, skrevet i Perl

```
#!/usr/local/bin/perl -w

# bestemmer hvilken versjon av progr som skal brukes; metode I eller II
my $internal_matching = 0; # 1 = true (metode I), 0 = false (metode II)
my $freq_file = $ARGV[0];
my $summ_file = $ARGV[1];

# leser fra innputtfilene og putter de inn i arrays
my @freq_content = read_file($freq_file);
my @summ_content = read_file($summ_file);
my $num_summaries = $#summ_content + 1;


# frekvensdataene deles opp og settes inn i en to-dimensjonal array
my $sum_sentences = 0;
my @freq_data = ();

foreach $line (@freq_content) {
   # henter setningsnummer og frekvens fra hver linje
   my ($sentence, $frequency) = split(/ /, $line);

   # adderer frekvenstallene
   $sum_sentences += $frequency;

   # dropper setninger med 0 i frekvens
   if ($frequency >= 1) {
        # legger til en array på slutten av @freq_data, som dermed blir en to-dim array
        push @freq_data, [$sentence, $frequency];
   }
}


# beregner gjennomsnittslengden (målt i antall setninger) av sammendragene og runder av
til nærmeste heltall
my $average_length = int(($sum_sentences / $num_summaries) + 0.5);

# sjekker om siste setnings frekvens er lik neste setnings frekvens
my $last_index = $average_length-1;
my $last_frequency = $freq_data[$last_index][1];
my $next_frequency = $freq_data[$last_index+1][1];

if ($last_frequency != $next_frequency) {
   my @certain = map { $_->[0] } @freq_data[0..$last_index];
# henter ut setningsnummerne og lagrer disse i en vanlig array
   print_result(@certain);
   exit;
```

```
}

# markerer setningene som "sikre" eller "usikre" basert på frekvens og
gjennomsnittslengde
my @certain = ();
my @uncertain = ();
my %uncertain_scores = ();
my $limit = $last_frequency; # grenseverdi

foreach $element (@freq_data) {
   my ($sentence, $frequency) = @{$element};

   # hvis frekvensen er over grenseverdien markeres setningen som "sikker"
   if ($frequency > $limit) { push @certain, $sentence; }

   # hvis frekvensen er lik grenseverdien markeres setningen som "usikker", med score 0
   elsif ($frequency == $limit) {
         push @uncertain, $sentence;
         $uncertain_scores{$sentence} = 0;
   }

   # hvis frekvensen er under grenseverdien har vi kommet for langt og kan avbryte
løkken
   else { last; }
}

# deler sammendragsdata inn i en hash-tabell med sammendr.nr som nøkkel og setn.nr
som verdi
my %summ_data = ();

foreach $line (@summ_content) {
   # henter sammendragsnr. og setningsnr. fra hver linje
   my ($summary, @sentences) = split(/ /, $line);

   # setter inn et nøkkel/verdi-par i %summ_data hash
   $summ_data{$summary} = [@sentences];
}

# sjekker om metode I (true) eller metode II (false) skal brukes
if ($internal_matching) {
   # metode I - finner ut hvilke usikre setninger som internt matcher hverandre best
   # den doble for-løkken representerer indeksering av @uncertain i både x- og y-retning
   for (my $i = 0; $i <= $#uncertain; $i++) {
         for (my $j = 0; $j <= $#uncertain; $j++) {
            # unngår doble sammenligninger (unødvendig å finne ut om både i er lik j og j
er lik i - det holder den ene veien)
            if ($i <= $j) { next; }

            # sjekker hvor mange sammendrag kombinasjonen av setning i og j oppstår i
            my $combinations = find_combinations($uncertain[$i], $uncertain[$j]);
```

```
            # gi i- og j-setningene +1 i score
            $uncertain_scores{$uncertain[$i]} += $combinations;
            $uncertain_scores{$uncertain[$j]} += $combinations;
        }
    }
} else {
    # metode II - finner ut hvike usikre setninger som matcher de sikre setningene best
    # den doble for-løkken representerer indeksering av @certain i x-retning og @uncertain
i y-retning

for (my $i = 0; $i <= $#uncertain; $i++) {
        for (my $j = 0; $j <= $#certain; $j++) {
            # sjekker hvor mange sammendrag kombinasjonen av setning i og j oppstår i
            my $combinations = find_combinations($uncertain[$i], $certain[$j]);

            # gi i-setningene (fra @uncertain) +1 i score
            $uncertain_scores{$uncertain[$i]} += $combinations;
        }
    }
}

# sorterer de usikre setningene etter scoren de har oppnådd
# my @sorted_sentences = map { "$_ " . $uncertain_scores{$_} } sort sort_uncertain
keys %uncertain_scores; # her blir også score gitt til variabelen
my @sorted_sentences = sort sort_uncertain keys %uncertain_scores;
# her blir bare setningene gitt til variabelen


# Her er alle subrutinene
# her blir @uncertain sortert etter score etter at matching er gjort
sub sort_uncertain {
    my $cmp = $uncertain_scores{$b} <=> $uncertain_scores{$a};
    if ($cmp == 0) { return $a <=> $b; }
    return $cmp;
}

# beregner hvor mange setninger vi mangler i @certain og flytter tilsvarende best rangerte
setninger fra @sorted_sentences til @certain
my $missing = $average_length - ($#certain+1);
push @certain, @sorted_sentences[0..$missing-1];
print_result(@certain);
#print "\n\n";
#print_result(@sorted_sentences);
exit;

# i de to neste sub'ene utføres selve matchingen av setningene mot sammendragene
sub find_combinations {
    my $sentence1 = $_[0];
    my $sentence2 = $_[1];
    my $num_matches = 0;
```

```
    foreach $key (keys %summ_data) {
        my @sentences = @{$summ_data{$key}};
        if (array_contains($sentence1, @sentences) && array_contains($sentence2,
@sentences)) {
            $num_matches++;
        }
    }

    return $num_matches;
}
sub array_contains {
    my $match = $_[0];
    my @array = @_[1..$#_];

    foreach $element (@array) {
        if ($element == $match) { return 1; } # true
    }

    return 0; # false
}

# refereansesammendraget skrives ut
sub print_result {
    foreach $element (@_) {
        print "$element\n";
    }
}

# leser fil inn i array og lukker filen etterpå
sub read_file {
    my $file = $_[0];

    open (FILE, $file) or die ("Can not open file.\n");
    my @content = <FILE>;
    close (FILE);

    chomp @content;
    return @content;
}
```

# Appendiks B: Innputtfiler til programmet

Eksempel på innputtfil til programmet (Appendiks A).

Første tallkolonne i filen er setningsnummer med deres tilhørende frekvens.

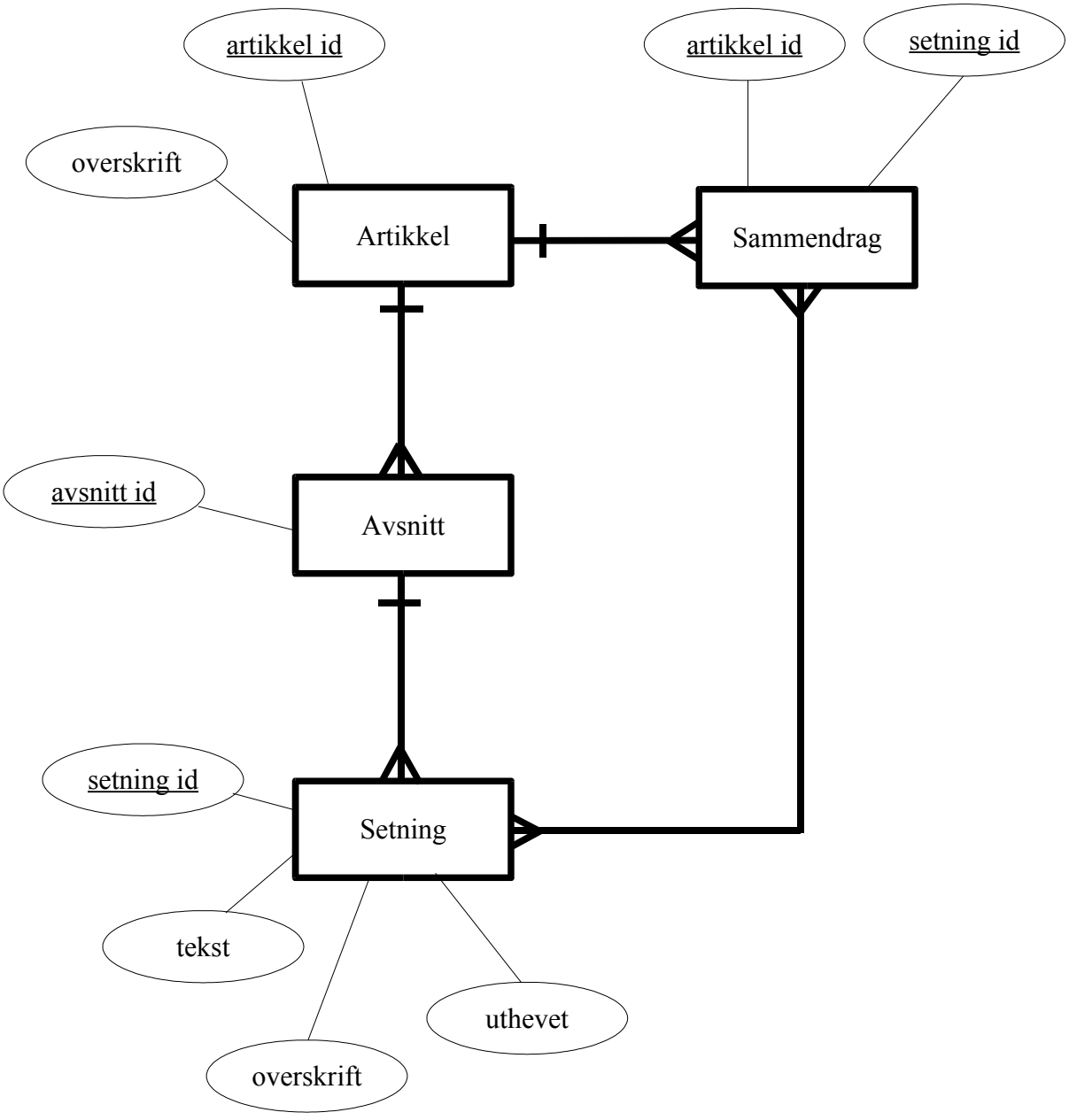
560 9
556 9
574 8
584 7
577 7
557 7
567 6
581 4
564 4
562 4
561 4
555 4
553 4
552 4
580 3
575 3
563 3
593 2
590 2
576 2
573 2
572 2
566 2
565 2
588 1
585 1
582 1
579 1
571 1
569 1
592 0
591 0
589 0
587 0
586 0
583 0
578 0
570 0

I den neste filen viser første tallkolonne til nummeret på sammendraget, mens resten viser til setningsnummer. Det er altså ett sammendrag per rad.

36 553 556 557 560 561 562 563 564 565 566 567 574 575 580 581 584 588 590 593
52 553 556 557 560 563 574 584
100 556 557 560 574 577 584
193 552 553 555 556 557 560 562 564 566 567 572 573 574 577 581 584
128 555 556 560 567 572 573 574 577 580 581 584
139 552 556 557 561 574 577 579 584
165 553 555 556 560 562 564 569 571
211 560 561 563 567 574 575 576 577 580 581
246 552 555 556 557 560 562 564 565 567 575 577 582 584 585 590 593
253 552 556 557 560 561 567 574 576 577

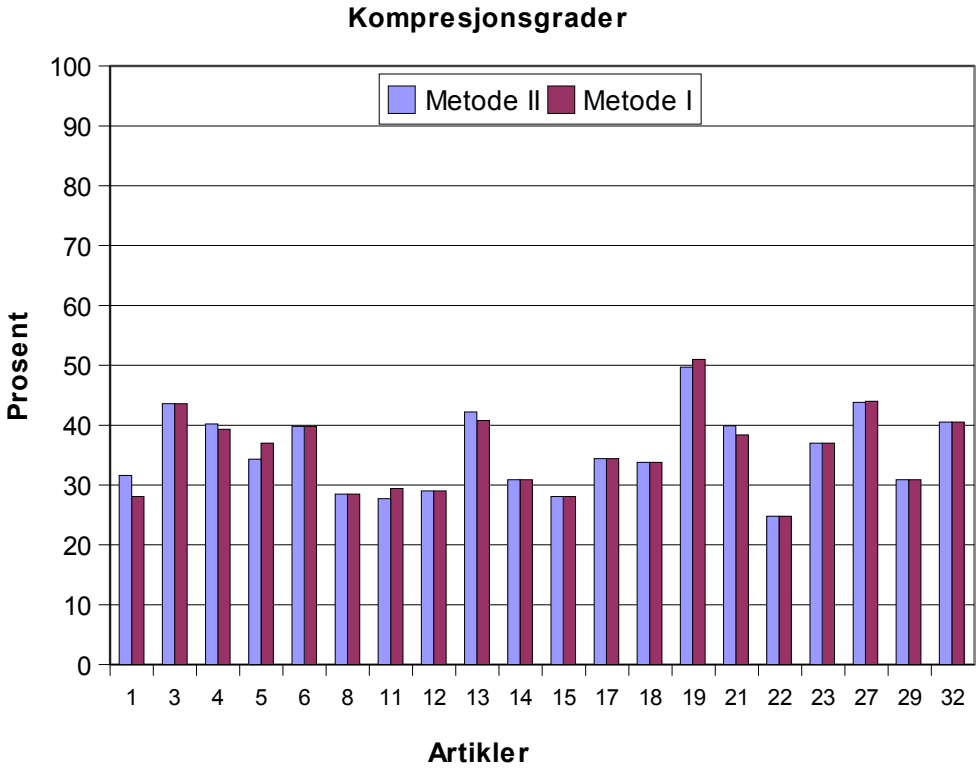
# Appendiks C: ER-diagram over databasen

ER-diagram over databasen



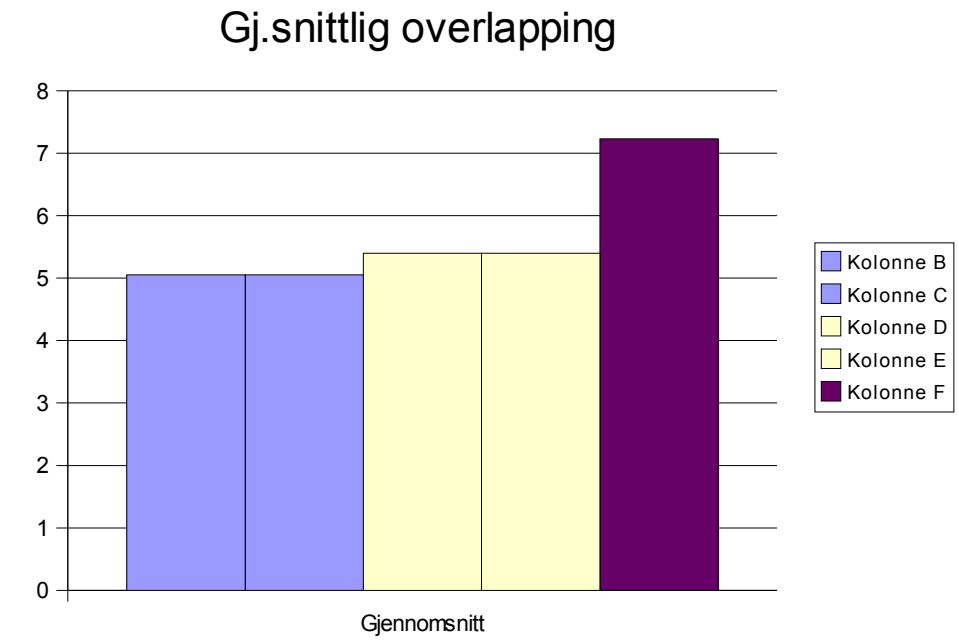
| Artikkel id | Metode II | Metode I |
|---|---|---|
| 1 | 31,6 | 28,1 |
| 3 | 43,6 | 43,6 |
| 4 | 40,2 | 39,3 |
| 5 | 34,3 | 37,0 |
| 6 | 39,8 | 39,8 |
| 8 | 28,5 | 28,5 |
| 11 | 27,7 | 29,4 |
| 12 | 29,0 | 29,0 |
| 13 | 42,2 | 40,8 |
| 14 | 30,9 | 30,9 |
| 15 | 28,1 | 28,1 |
| 17 | 34,4 | 34,4 |
| 18 | 33,8 | 33,8 |
| 19 | 49,7 | 51,0 |
| 21 | 39,9 | 38,4 |
| 22 | 24,8 | 24,8 |
| 23 | 37,0 | 37,0 |
| 27 | 43,8 | 44,0 |
| 29 | 30,9 | 30,9 |
| 32 | 40,5 | 40,5 |
| Gjennomsnitt | 35,54 | 35,47 |
| Standardavvik | 6,70 | 6,79 |

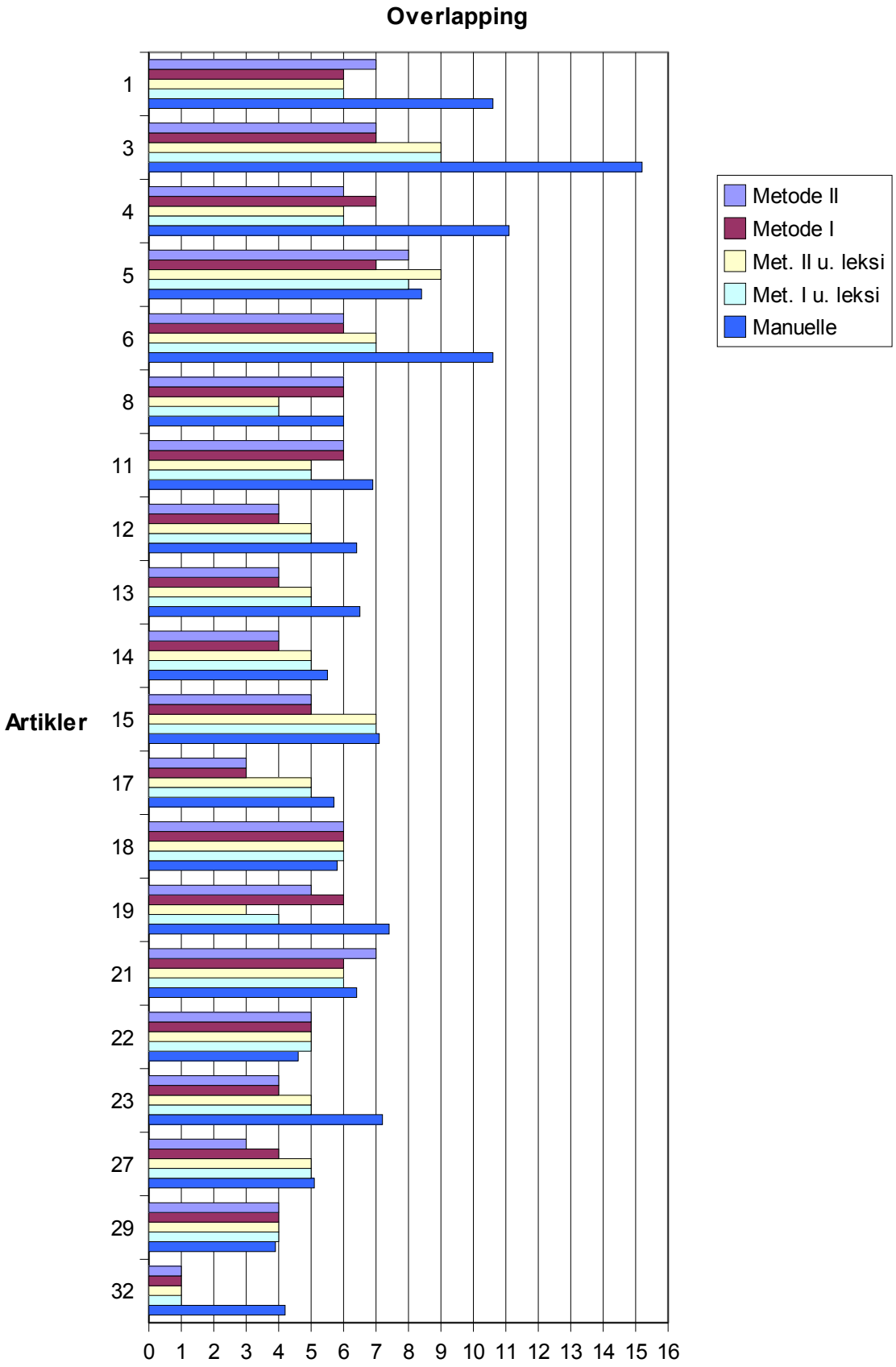
Gjennomsnittlig kompresjonsgrad av RS generert med metode I og II



| Artikkel id | Metode II | Metode I | Met. II u. leksikon | Met. I u. leksikon | Manuelle |
|---|---|---|---|---|---|
| 1 | 7 | 6 | 6 | 6 | 10,6 |
| 3 | 7 | 7 | 9 | 9 | 15,2 |
| 4 | 6 | 7 | 6 | 6 | 11,1 |
| 5 | 8 | 7 | 9 | 8 | 8,4 |
| 6 | 6 | 6 | 7 | 7 | 10,6 |
| 8 | 6 | 6 | 4 | 4 | 6 |
| 11 | 6 | 6 | 5 | 5 | 6,9 |
| 12 | 4 | 4 | 5 | 5 | 6,4 |
| 13 | 4 | 4 | 5 | 5 | 6,5 |
| 14 | 4 | 4 | 5 | 5 | 5,5 |
| 15 | 5 | 5 | 7 | 7 | 7,1 |
| 17 | 3 | 3 | 5 | 5 | 5,7 |
| 18 | 6 | 6 | 6 | 6 | 5,8 |
| 19 | 5 | 6 | 3 | 4 | 7,4 |
| 21 | 7 | 6 | 6 | 6 | 6,4 |
| 22 | 5 | 5 | 5 | 5 | 4,6 |
| 23 | 4 | 4 | 5 | 5 | 7,2 |
| 27 | 3 | 4 | 5 | 5 | 5,1 |
| 29 | 4 | 4 | 4 | 4 | 3,9 |
| 32 | 1 | 1 | 1 | 1 | 4,2 |
| Gjennomsnitt | 5,05 | 5,05 | 5,4 | 5,4 | 7,23 |
| Standardavvik | 1,70 | 1,54 | 1,82 | 1,67 | 2,77 |

Gjennomsnittlig overlapping av setninger mellom RS og automatiske sammendrag
Siste kolonne viser overlapping gj.snittlig mellom RS og manuelle sammendrag

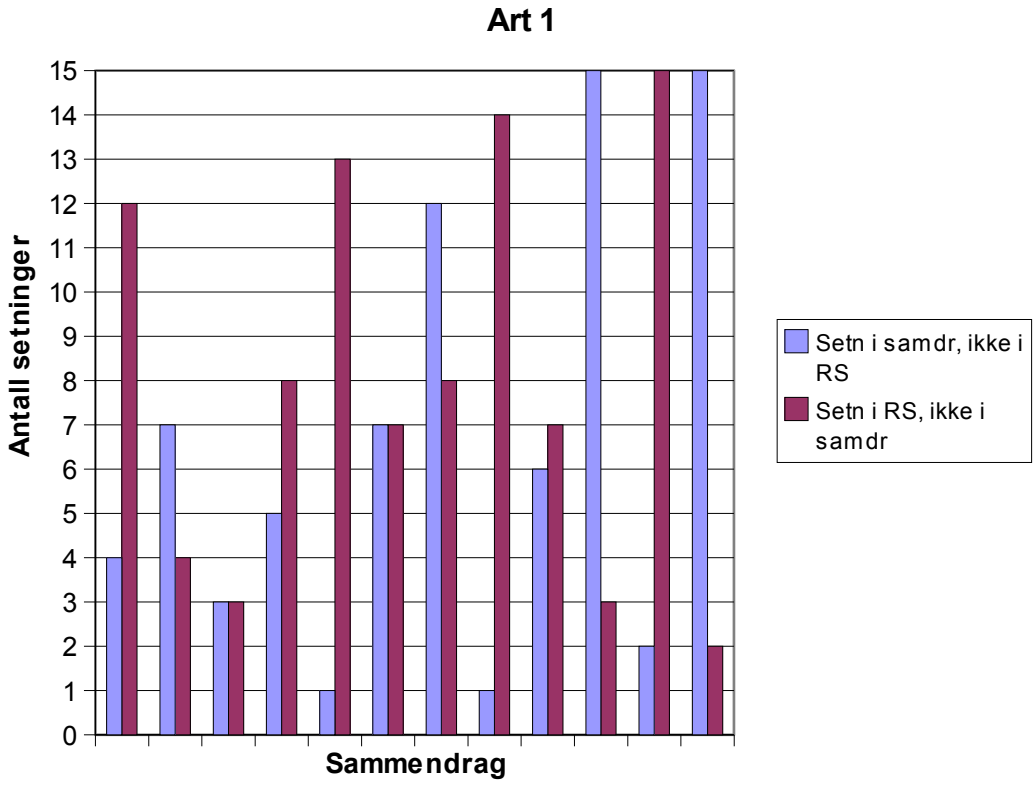



Overlapping
Artikler
1
3
4
5
6
8
11
12
13
14
15
17
18
19
21
22
23
27
29
32
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Metode II
Metode I
Met. II u. leksi
Met. I u. leksi
Manuelle

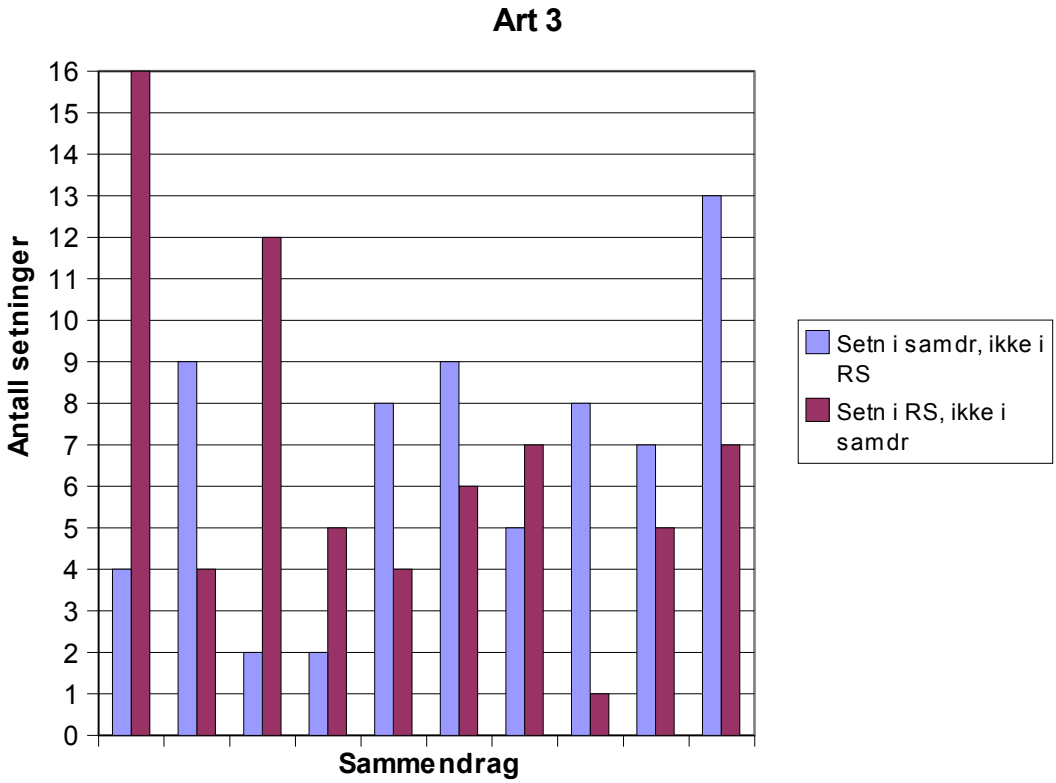
| MS | RS | AS m leksi | RS m leksi | AS u leksi | RS u leksi |
|---|---|---|---|---|---|
| 6,5 | 8 | 17 | 12 | 16 | 13 |
| 6,7 | 6,7 | 21 | 13 | 20 | 15 |
| 5,5 | 6,5 | 16 | 11 | 17 | 11 |
| 5,1 | 5,5 | 12 | 7 | 14 | 6 |
| 5 | 5,4 | 16 | 10 | 16 | 9 |
| 3,5 | 4 | 8 | 4 | 12 | 6 |
| 2,9 | 5,9 | 8 | 8 | 8 | 9 |
| 4,2 | 4,4 | 10 | 7 | 10 | 6 |
| 4,8 | 6,1 | 14 | 9 | 13 | 8 |
| 3,9 | 4,5 | 9 | 6 | 9 | 5 |
| 4,2 | 3,9 | 11 | 6 | 9 | 4 |
| 4,1 | 4 | 11 | 7 | 9 | 5 |
| 3,1 | 3,2 | 7 | 3 | 6 | 3 |
| 3,7 | 5,1 | 11 | 7 | 13 | 9 |
| 3,5 | 5,7 | 10 | 6 | 8 | 6 |
| 3,6 | 3,4 | 6 | 3 | 6 | 3 |
| 2,9 | 2,8 | 10 | 6 | 7 | 5 |
| 2,9 | 3,4 | 10 | 5 | 6 | 4 |
| 3,9 | 4,1 | 8 | 4 | 9 | 4 |
| 2,7 | 2,8 | 7 | 6 | 6 | 6 |
| Gjennomsnitt | | | | | |
| 4,14 | 4,77 | 11,10 | 7 | 10,7 | 6,85 |

MS = Setninger som finnes i manuelle sammendrag, men ikke i RS
RS = Setninger som finnes i RS, men ikke i manuelle sammendrag
AS = Setninger som finnes i automatisk sammendrag, men ikke i RS
RS = Setninger som finnes i RS, men ikke i automatiske sammendrag

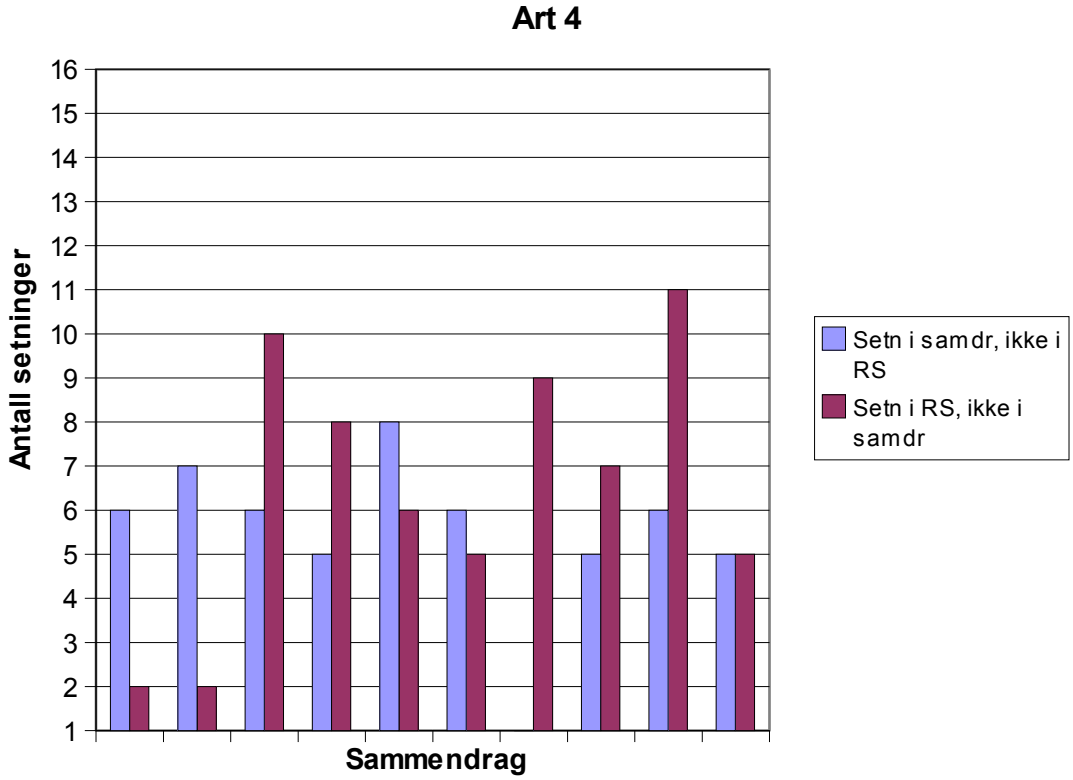
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 199 | 4 | 12 |
| 13 | 7 | 4 |
| 16 | 3 | 3 |
| 176 | 5 | 8 |
| 119 | 1 | 13 |
| 131 | 7 | 7 |
| 137 | 12 | 8 |
| 208 | 1 | 14 |
| 270 | 6 | 7 |
| 271 | 15 | 3 |
| 292 | 2 | 15 |
| 304 | 15 | 2 |
| | | |
| Gjennomsnitt | 6,5 | 8 |
| Standardavvik | 5,02 | 4,57 |
| | | |
| Autosam med leksikon | 17 | 12 |
| Autosam uten leksikon | 16 | 13 |
| Gjennomsnitt | 16,5 | 12,5 |



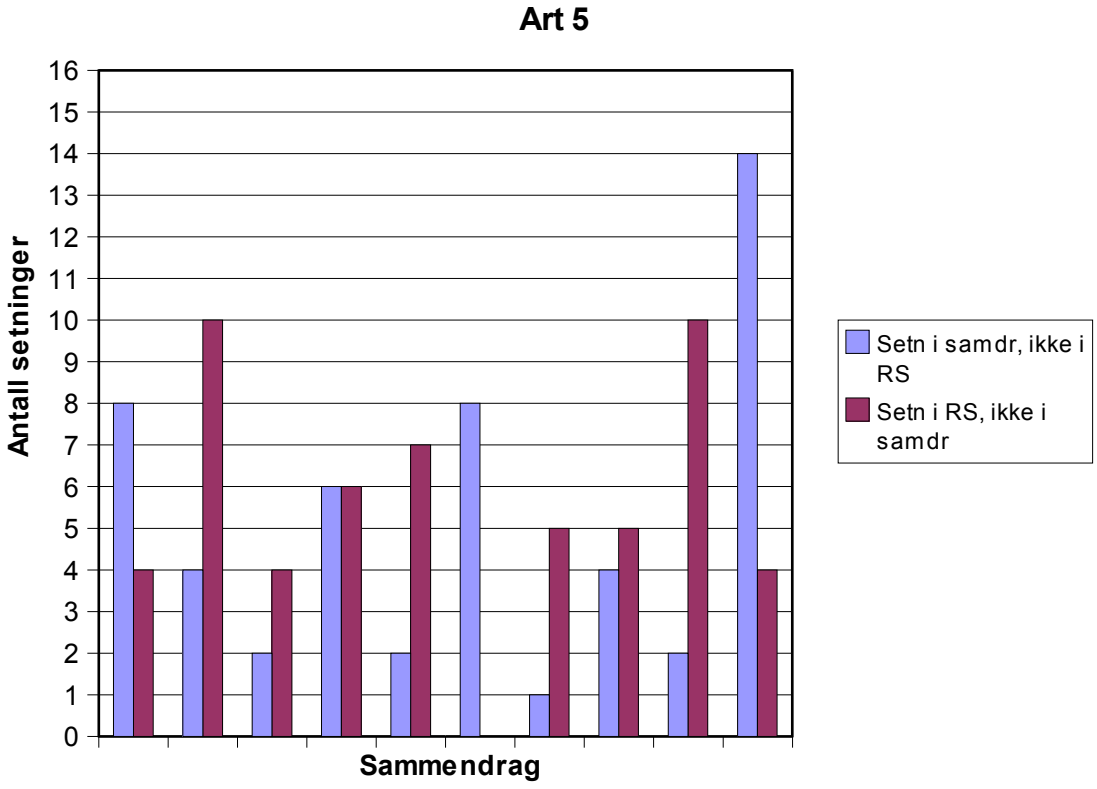
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 205 | 4 | 16 |
| 31 | 9 | 4 |
| 54 | 2 | 12 |
| 61 | 2 | 5 |
| 132 | 8 | 4 |
| 143 | 9 | 6 |
| 215 | 5 | 7 |
| 239 | 8 | 1 |
| 272 | 7 | 5 |
| 316 | 13 | 7 |
| | | |
| Gjennomsnitt | 6,7 | 6,7 |
| Standardavvik | 3,47 | 4,32 |
| | | |
| Autosam med leksikon | 21 | 13 |
| Autosam uten leksikon | 20 | 15 |
| Gjennomsnitt | 20,5 | 14 |



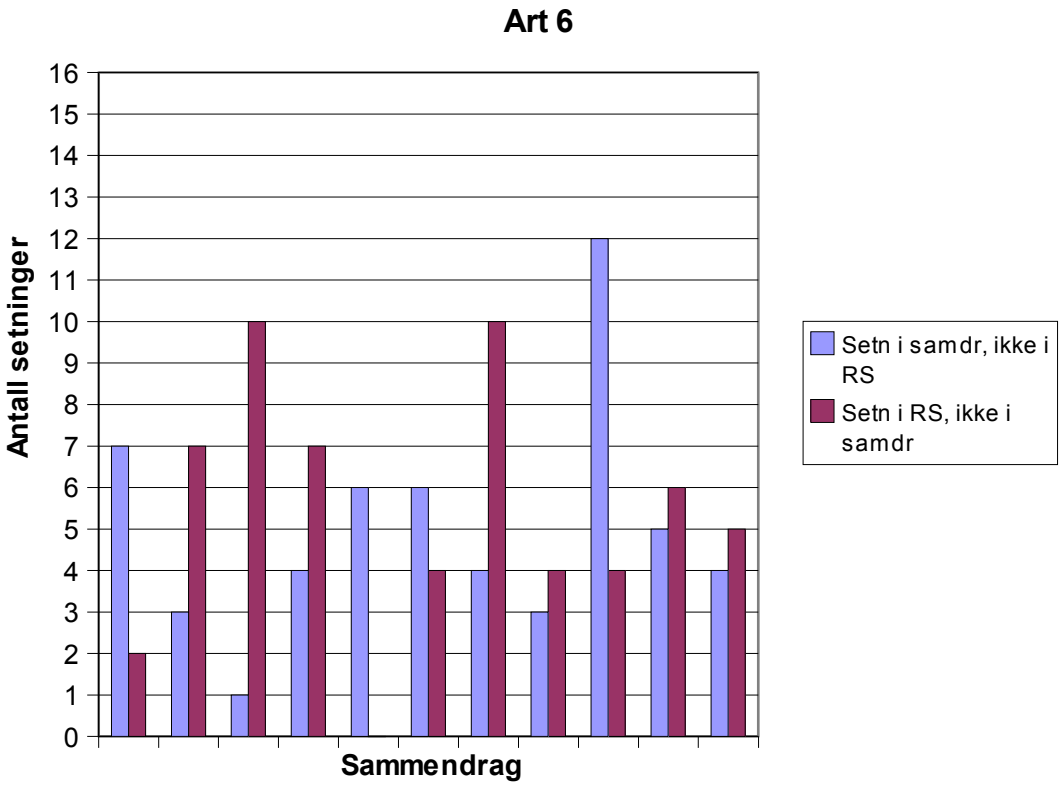
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 11 | 6 | 2 |
| 55 | 7 | 2 |
| 110 | 6 | 10 |
| 177 | 5 | 8 |
| 133 | 8 | 6 |
| 144 | 6 | 5 |
| 210 | 1 | 9 |
| 234 | 5 | 7 |
| 235 | 6 | 11 |
| 289 | 5 | 5 |
| | | |
| Gjennomsnitt | 5,5 | 6,5 |
| Standardavvik | 1,84 | 3,1 |
| | | |
| Autosam med leksikon | 16 | 11 |
| Autosam uten leksikon | 17 | 11 |
| Gjennomsnitt | 16,5 | 11 |



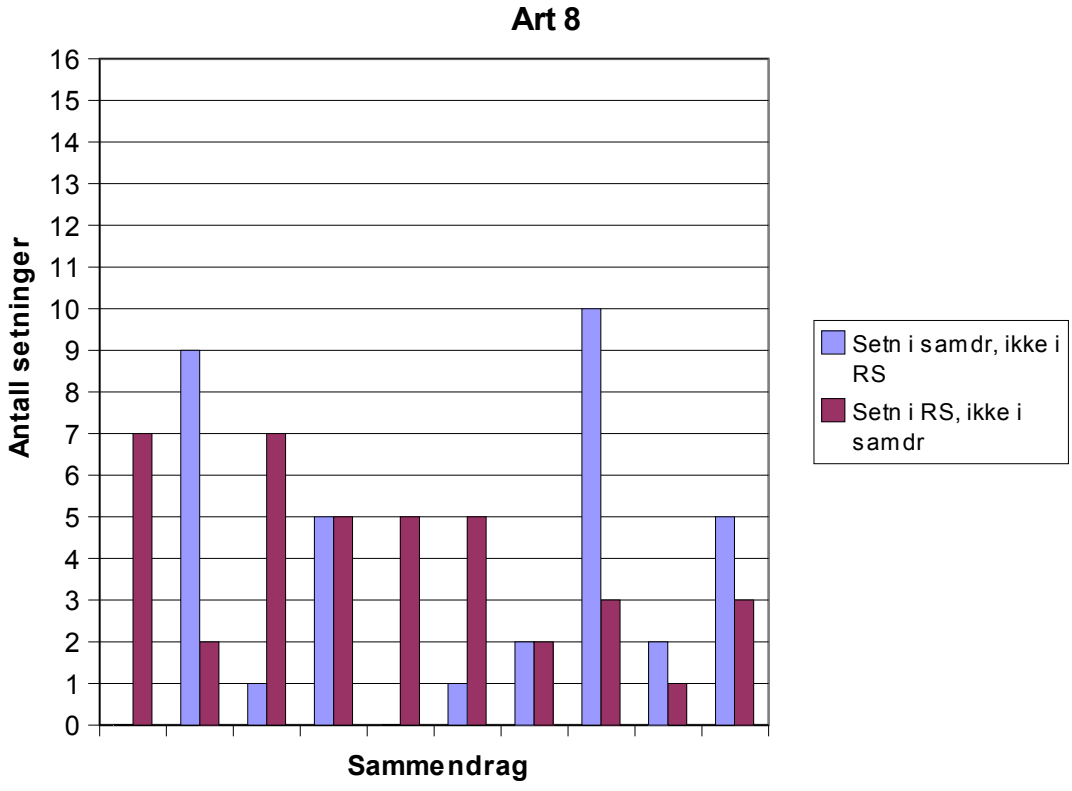
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 29 | 8 | 4 |
| 53 | 4 | 10 |
| 71 | 2 | 4 |
| 134 | 6 | 6 |
| 145 | 2 | 7 |
| 146 | 8 | 0 |
| 207 | 1 | 5 |
| 252 | 4 | 5 |
| 267 | 2 | 10 |
| 317 | 14 | 4 |
| | | |
| Gjennomsnitt | 5,1 | 5,5 |
| Standardavvik | 4,01 | 2,99 |
| | | |
| Autosam med leksikon | 12 | 7 |
| Autosam uten leksikon | 14 | 6 |
| Gjennomsnitt | 13 | 6,5 |



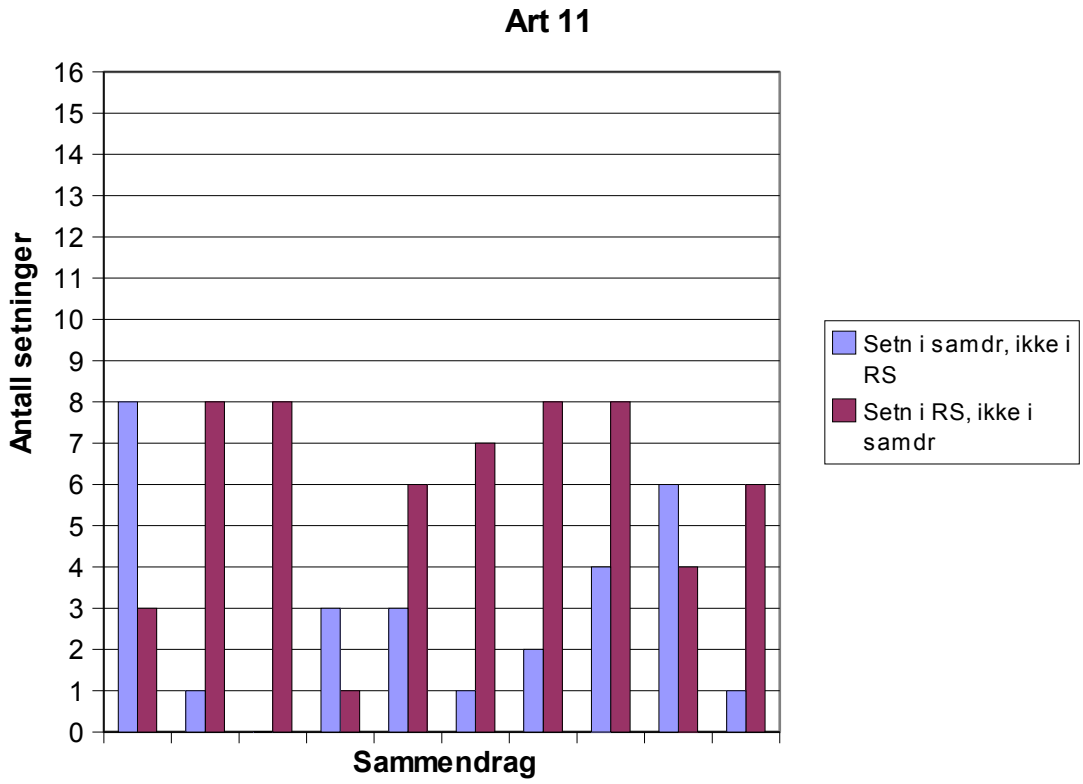
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 34 | 7 | 2 |
| 70 | 3 | 7 |
| 116 | 1 | 10 |
| 129 | 4 | 7 |
| 136 | 6 | 0 |
| 147 | 6 | 4 |
| 206 | 4 | 10 |
| 214 | 3 | 4 |
| 242 | 12 | 4 |
| 256 | 5 | 6 |
| 308 | 4 | 5 |
| | | |
| Gjennomsnitt | 5 | 5,4 |
| Standardavvik | 2,86 | 3,07 |
| | | |
| Autosam med leksikon | 16 | 10 |
| Autosam uten leksikon | 16 | 9 |
| Gjennomsnitt | 16 | 9,5 |



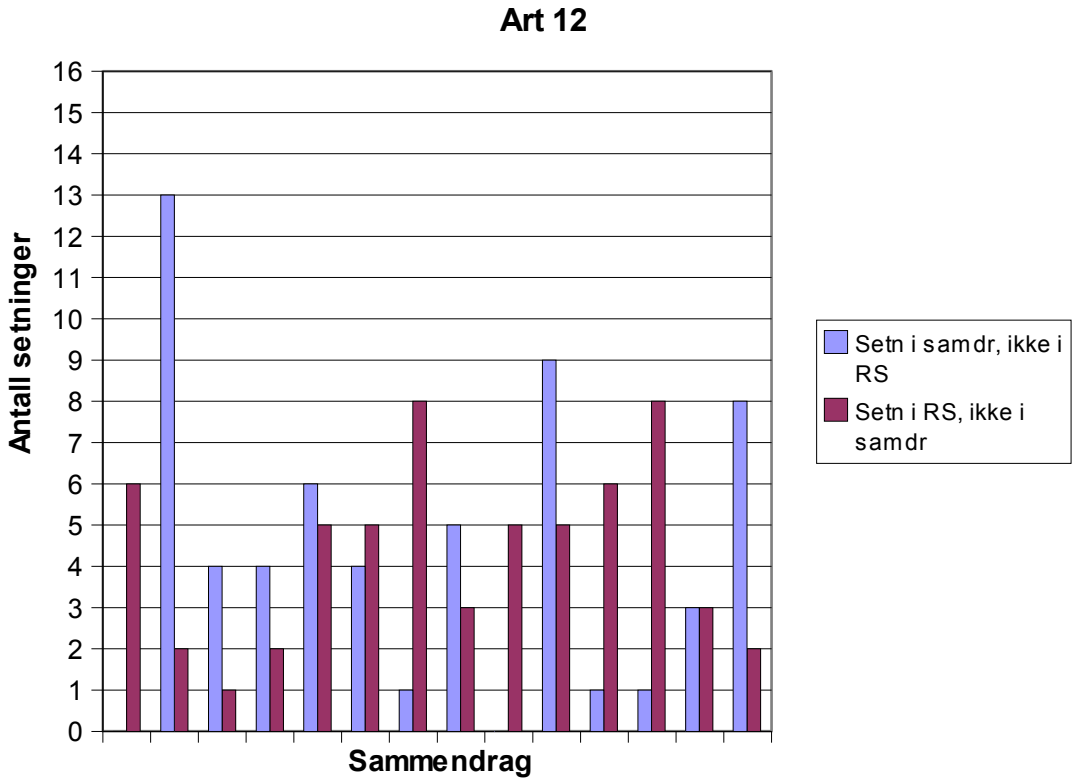
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 6 | 0 | 7 |
| 59 | 9 | 2 |
| 78 | 1 | 7 |
| 97 | 5 | 5 |
| 123 | 0 | 5 |
| 138 | 1 | 5 |
| 209 | 2 | 2 |
| 241 | 10 | 3 |
| 254 | 2 | 1 |
| 311 | 5 | 3 |
| | | |
| Gjennomsnitt | 3,5 | 4 |
| Standardavvik | 3,63 | 2,11 |
| | | |
| Autosam med leksikon | 8 | 4 |
| Autosam uten leksikon | 12 | 6 |
| Gjennomsnitt | 10 | 5 |



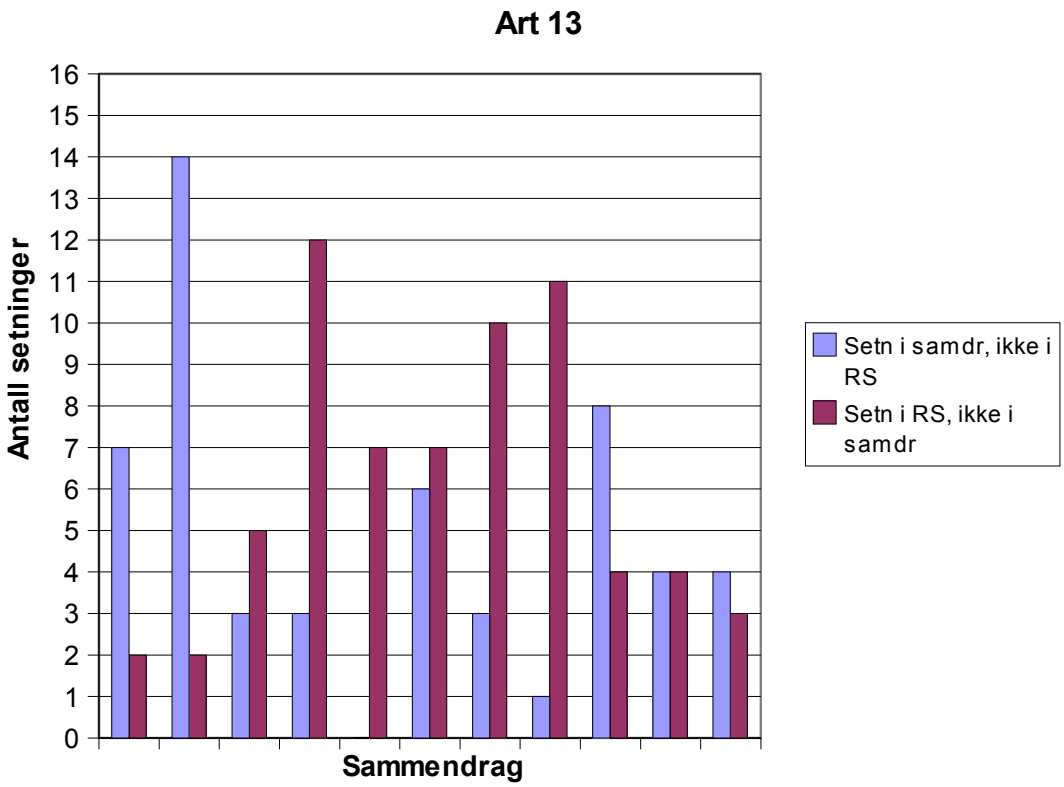
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 36 | 8 | 3 |
| 52 | 1 | 8 |
| 100 | 0 | 8 |
| 193 | 3 | 1 |
| 128 | 3 | 6 |
| 139 | 1 | 7 |
| 165 | 2 | 8 |
| 211 | 4 | 8 |
| 246 | 6 | 4 |
| 253 | 1 | 6 |
| | | |
| Gjennomsnitt | 2,9 | 5,9 |
| Standardavvik | 2,51 | 2,47 |
| | | |
| Autosam med leksikon | 8 | 8 |
| Autosam uten leksikon | 8 | 9 |
| Gjennomsnitt | 8 | 8,5 |



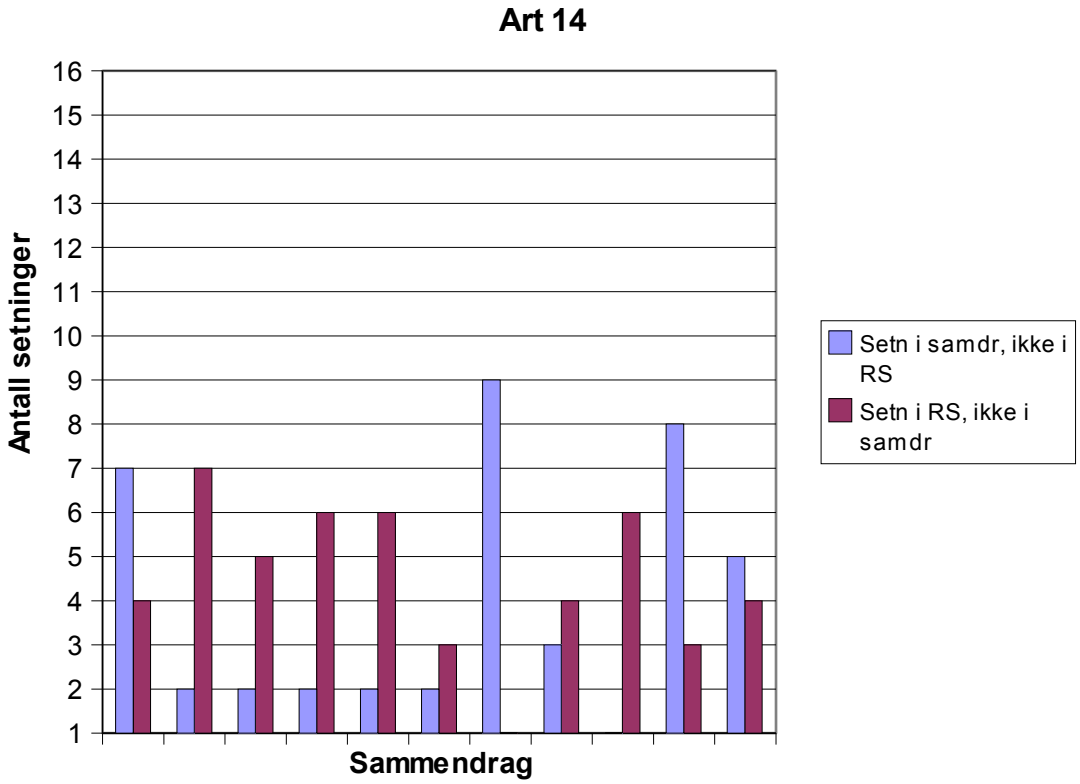
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 2 | 0 | 6 |
| 196 | 13 | 2 |
| 15 | 4 | 1 |
| 22 | 4 | 2 |
| 32 | 6 | 5 |
| 75 | 4 | 5 |
| 77 | 1 | 8 |
| 88 | 5 | 3 |
| 91 | 0 | 5 |
| 92 | 9 | 5 |
| 101 | 1 | 6 |
| 122 | 1 | 8 |
| 140 | 3 | 3 |
| 187 | 8 | 2 |
| | | |
| Gjennomsnitt | 4,2 | 4,4 |
| Standardavvik | 3,79 | 2,24 |
| | | |
| Autosam med leksikon | 10 | 7 |
| Autosam uten leksikon | 10 | 6 |
| Gjennomsnitt | 10 | 6,5 |



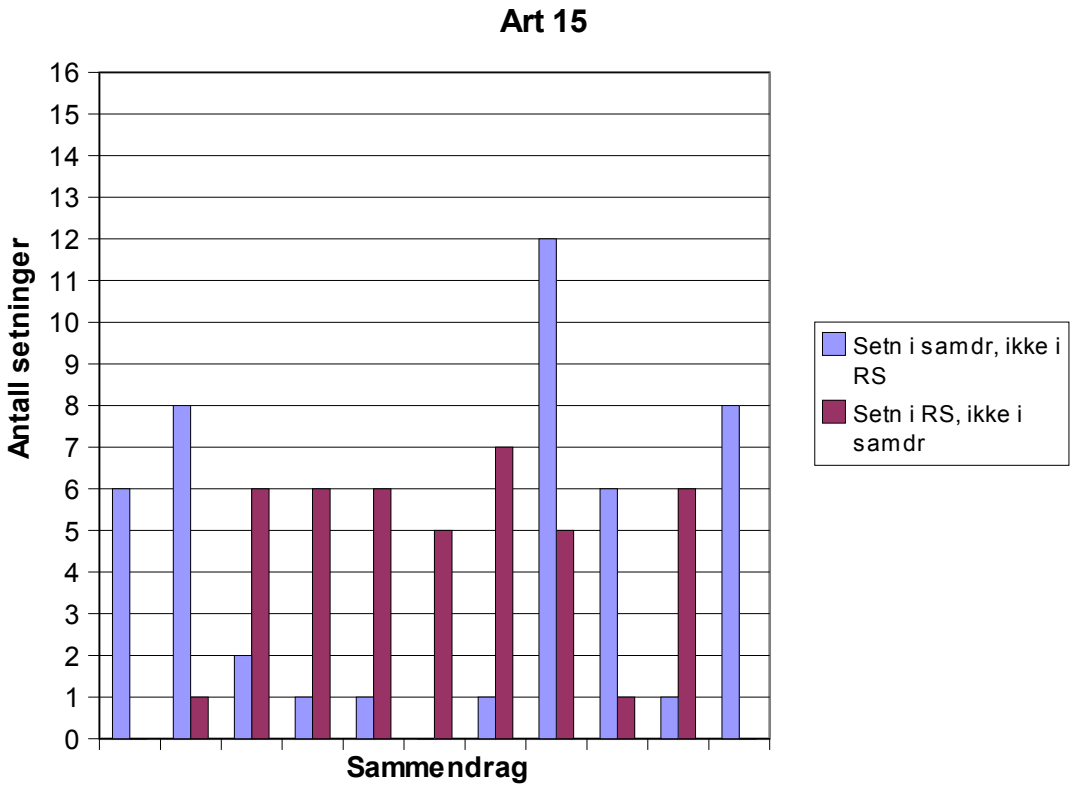
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 194 | 7 | 2 |
| 41 | 14 | 2 |
| 69 | 3 | 5 |
| 114 | 3 | 12 |
| 124 | 0 | 7 |
| 141 | 6 | 7 |
| 154 | 3 | 10 |
| 168 | 1 | 11 |
| 245 | 8 | 4 |
| 258 | 4 | 4 |
| 294 | 4 | 3 |
| | | |
| Gjennomsnitt | 4,8 | 6,1 |
| Standardavvik | 3,87 | 3,59 |
| | | |
| Autosam med leksikon | 14 | 9 |
| Autosam uten leksikon | 13 | 8 |
| Gjennomsnitt | 13,5 | 8,5 |



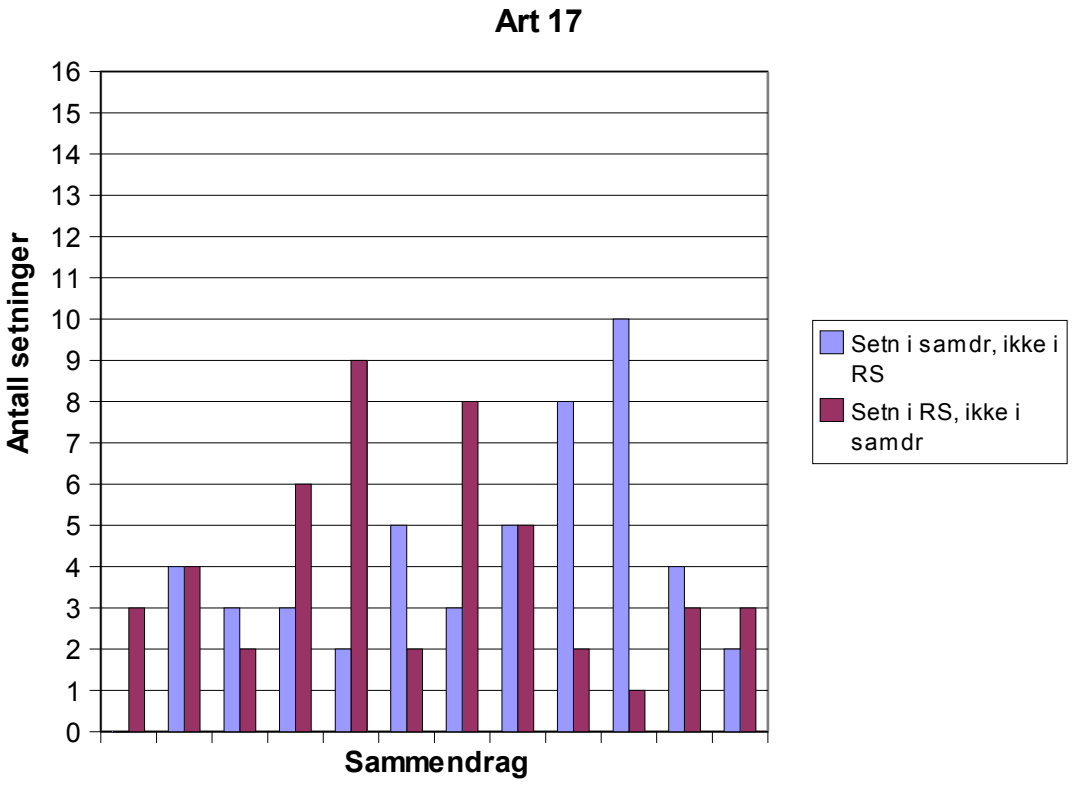
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 27 | 7 | 4 |
| 68 | 2 | 7 |
| 121 | 2 | 5 |
| 142 | 2 | 6 |
| 155 | 2 | 6 |
| 204 | 2 | 3 |
| 220 | 9 | 1 |
| 251 | 3 | 4 |
| 259 | 1 | 6 |
| 280 | 8 | 3 |
| 309 | 5 | 4 |
| | | |
| Gjennomsnitt | 3,9 | 4,5 |
| Standardavvik | 2,84 | 1,75 |
| | | |
| Autosam med leksikon | 9 | 6 |
| Autosam uten leksikon | 9 | 5 |
| Gjennomsnitt | 9 | 5,5 |



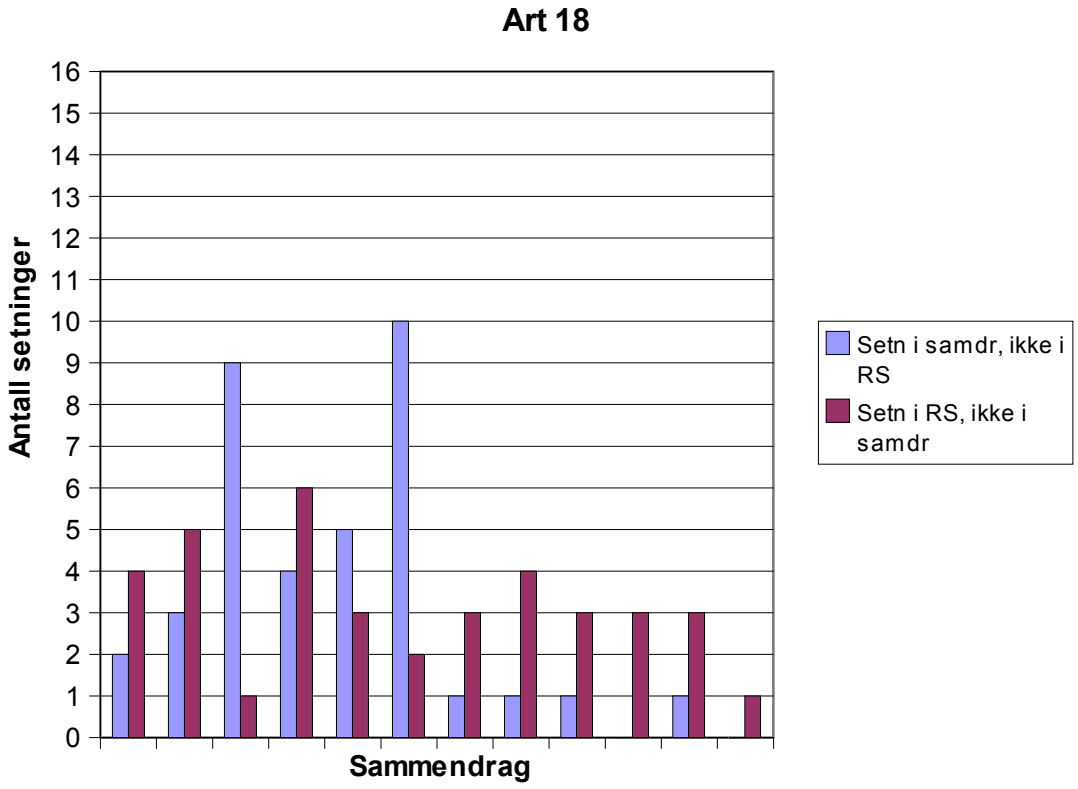
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 195 | 6 | 0 |
| 181 | 8 | 1 |
| 48 | 2 | 6 |
| 90 | 1 | 6 |
| 113 | 1 | 6 |
| 156 | 0 | 5 |
| 225 | 1 | 7 |
| 244 | 12 | 5 |
| 281 | 6 | 1 |
| 302 | 1 | 6 |
| 314 | 8 | 0 |
| | | |
| Gjennomsnitt | 4,2 | 3,9 |
| Standardavvik | 4,00 | 2,77 |
| | | |
| Autosam med leksikon | 11 | 6 |
| Autosam uten leksikon | 9 | 4 |
| Gjennomsnitt | 10 | 5 |



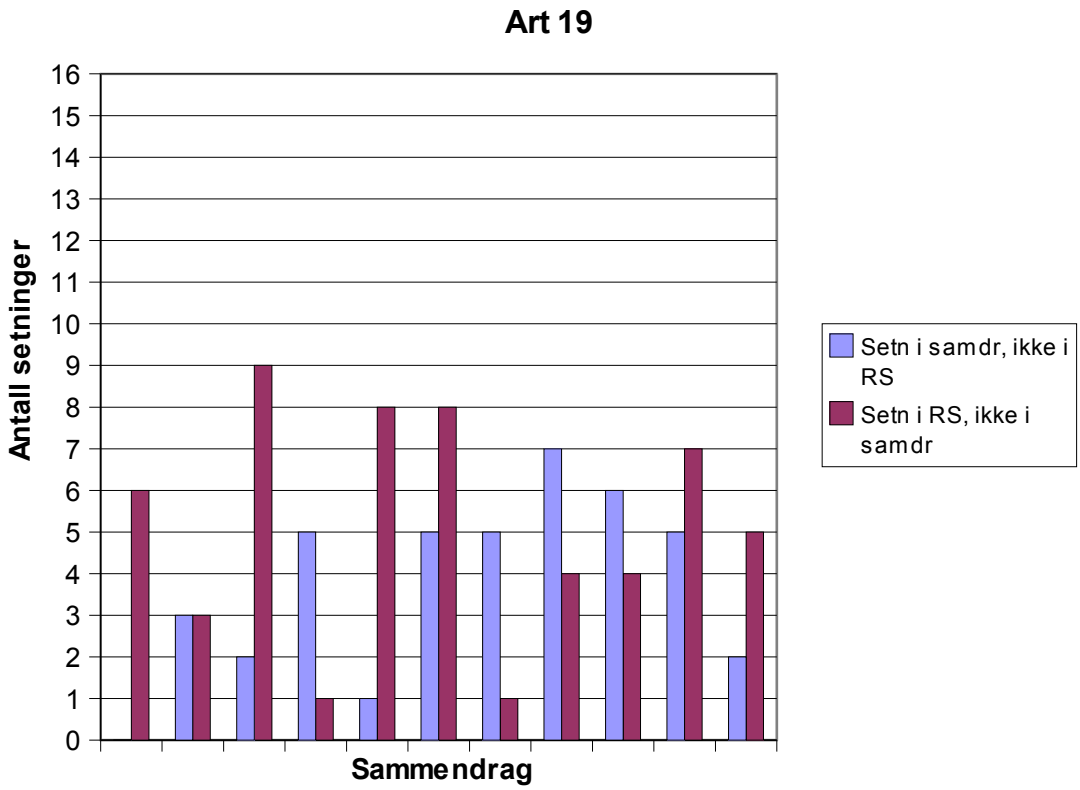
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 17 | 0 | 3 |
| 38 | 4 | 4 |
| 73 | 3 | 2 |
| 105 | 3 | 6 |
| 115 | 2 | 9 |
| 127 | 5 | 2 |
| 164 | 3 | 8 |
| 203 | 5 | 5 |
| 233 | 8 | 2 |
| 255 | 10 | 1 |
| 307 | 4 | 3 |
| 321 | 2 | 3 |
| | | |
| Gjennomsnitt | 4,1 | 4,0 |
| Standardavvik | 2,71 | 2,52 |
| | | |
| Autosam med leksikon | 11 | 7 |
| Autosam uten leksikon | 9 | 5 |
| Gjennomsnitt | 10 | 6 |



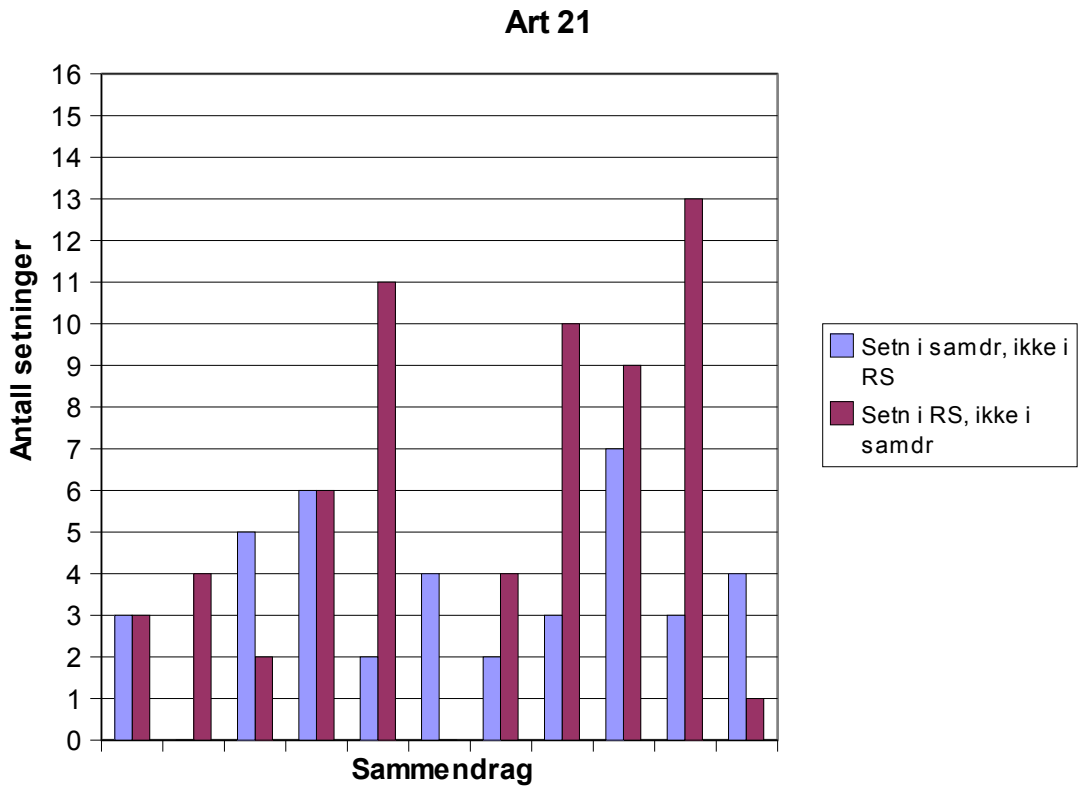
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 46 | 2 | 4 |
| 66 | 3 | 5 |
| 185 | 9 | 1 |
| 108 | 4 | 6 |
| 192 | 5 | 3 |
| 178 | 10 | 2 |
| 157 | 1 | 3 |
| 224 | 1 | 4 |
| 249 | 1 | 3 |
| 269 | 0 | 3 |
| 306 | 1 | 3 |
| 313 | 0 | 1 |
| | | |
| Gjennomsnitt | 3,1 | 3,2 |
| Standardavvik | 3,37 | 1,47 |
| | | |
| Autosam med leksikon | 7 | 3 |
| Autosam uten leksikon | 6 | 3 |
| Gjennomsnitt | 6,5 | 3 |



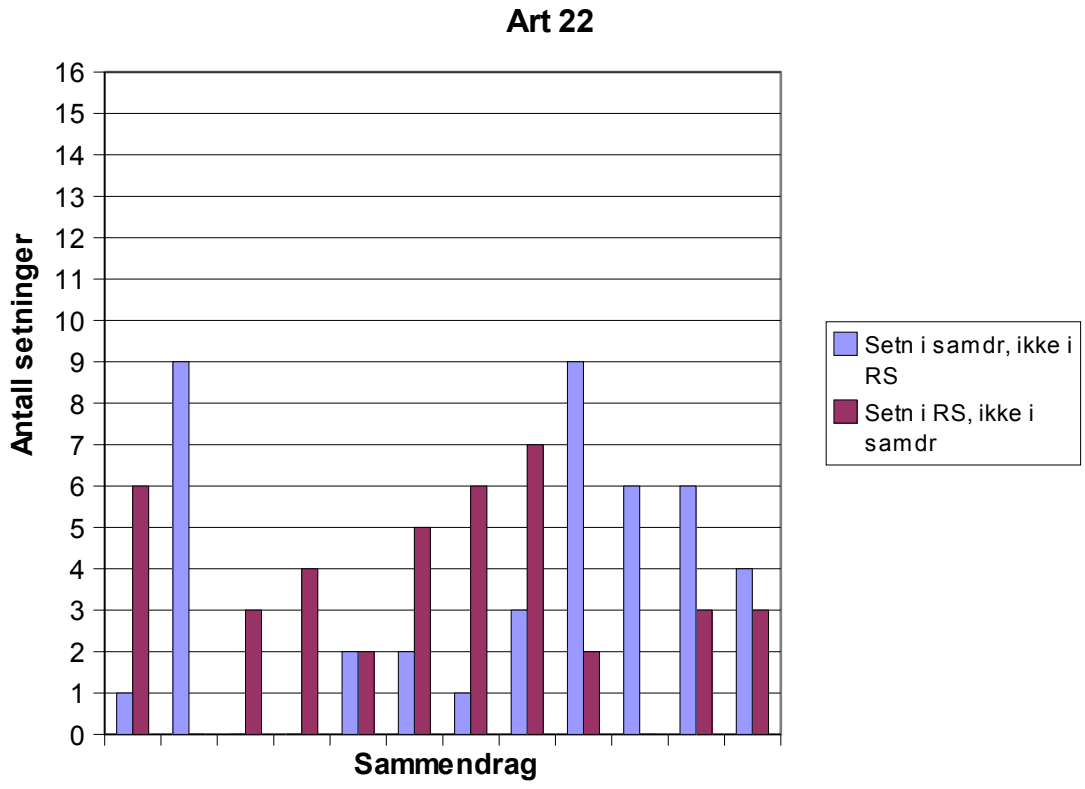
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 45 | 0 | 6 |
| 57 | 3 | 3 |
| 80 | 2 | 9 |
| 182 | 5 | 1 |
| 117 | 1 | 8 |
| 202 | 5 | 8 |
| 222 | 5 | 1 |
| 237 | 7 | 4 |
| 257 | 6 | 4 |
| 261 | 5 | 7 |
| 301 | 2 | 5 |
| | | |
| Gjennomsnitt | 3,7 | 5,1 |
| Standardavvik | 2,24 | 2,77 |
| | | |
| Autosam med leksikon | 11 | 7 |
| Autosam uten leksikon | 13 | 9 |
| Gjennomsnitt | 12 | 8 |



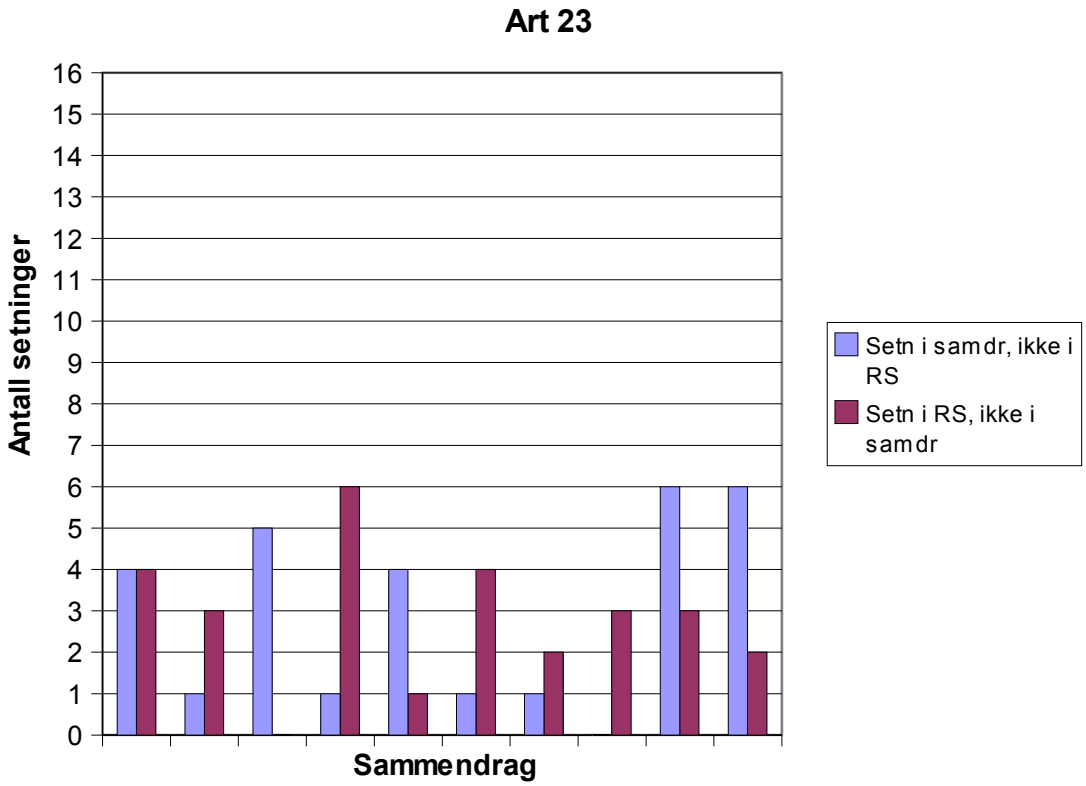
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 10 | 3 | 3 |
| 18 | 0 | 4 |
| 26 | 5 | 2 |
| 74 | 6 | 6 |
| 76 | 2 | 11 |
| 86 | 4 | 0 |
| 99 | 2 | 4 |
| 191 | 3 | 10 |
| 197 | 7 | 9 |
| 166 | 3 | 13 |
| 300 | 4 | 1 |
| | | |
| Gjennomsnitt | 3,5 | 5,7 |
| Standardavvik | 1,97 | 4,38 |
| | | |
| Autosam med leksikon | 10 | 6 |
| Autosam uten leksikon | 8 | 6 |
| Gjennomsnitt | 9 | 6 |



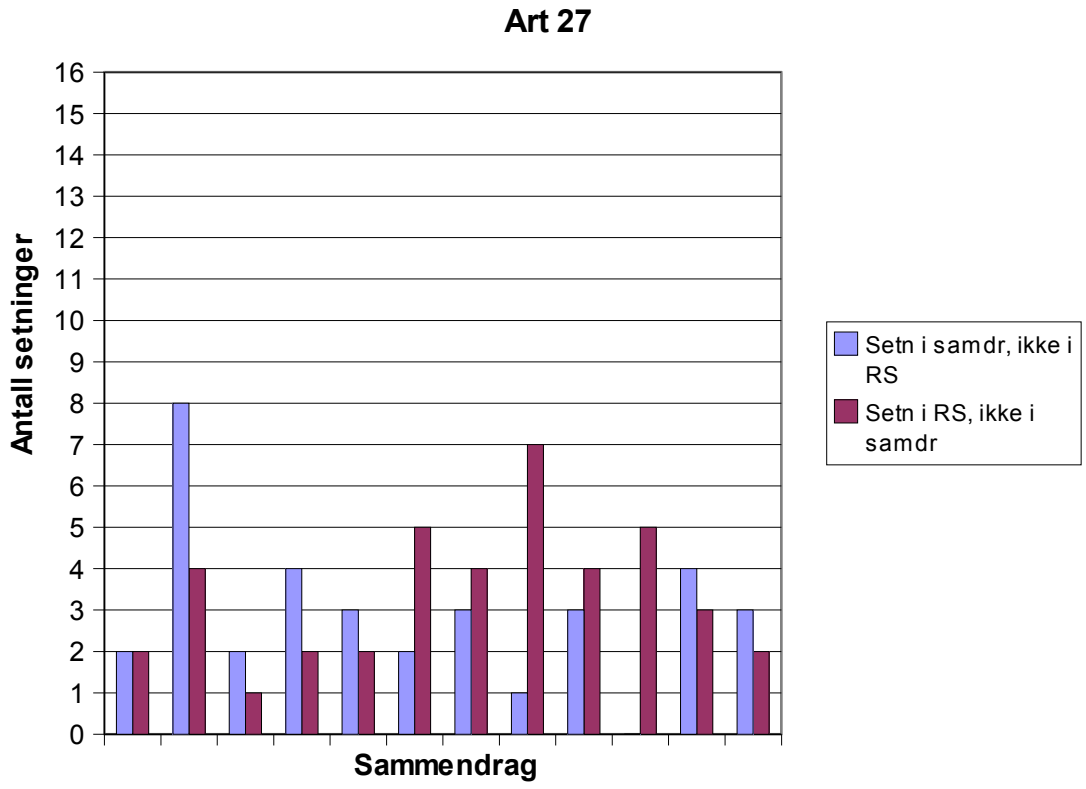
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 20 | 1 | 6 |
| 184 | 9 | 0 |
| 51 | 0 | 3 |
| 81 | 0 | 4 |
| 98 | 2 | 2 |
| 158 | 2 | 5 |
| 161 | 1 | 6 |
| 167 | 3 | 7 |
| 236 | 9 | 2 |
| 287 | 6 | 0 |
| 299 | 6 | 3 |
| 315 | 4 | 3 |
| | | |
| Gjennomsnitt | 3,6 | 3,4 |
| Standardavvik | 3,23 | 2,27 |
| | | |
| Autosam med leksikon | 6 | 3 |
| Autosam uten leksikon | 6 | 3 |
| Gjennomsnitt | 6 | 3 |



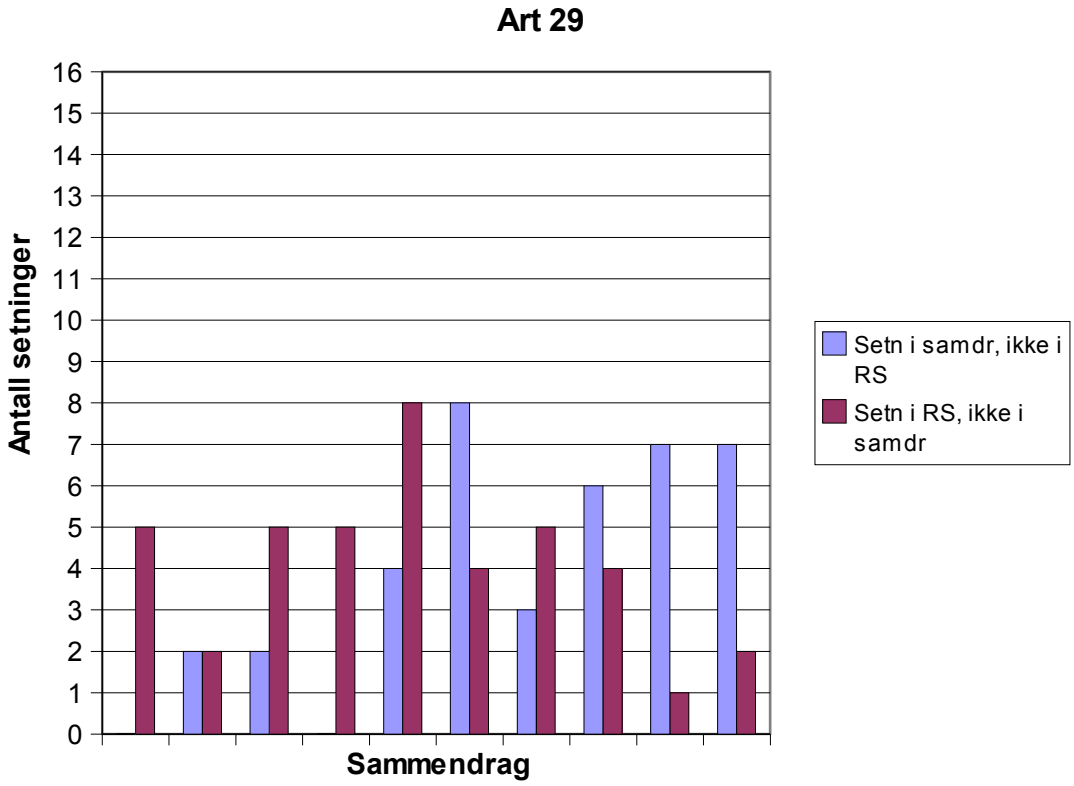
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 19 | 4 | 4 |
| 50 | 1 | 3 |
| 183 | 5 | 0 |
| 111 | 1 | 6 |
| 126 | 4 | 1 |
| 162 | 1 | 4 |
| 163 | 1 | 2 |
| 173 | 0 | 3 |
| 286 | 6 | 3 |
| 298 | 6 | 2 |
| | | |
| Gjennomsnitt | 2,9 | 2,8 |
| Standardavvik | 2,33 | 1,69 |
| | | |
| Autosam med leksikon | 10 | 6 |
| Autosam uten leksikon | 7 | 5 |
| Gjennomsnitt | 8,5 | 5,5 |



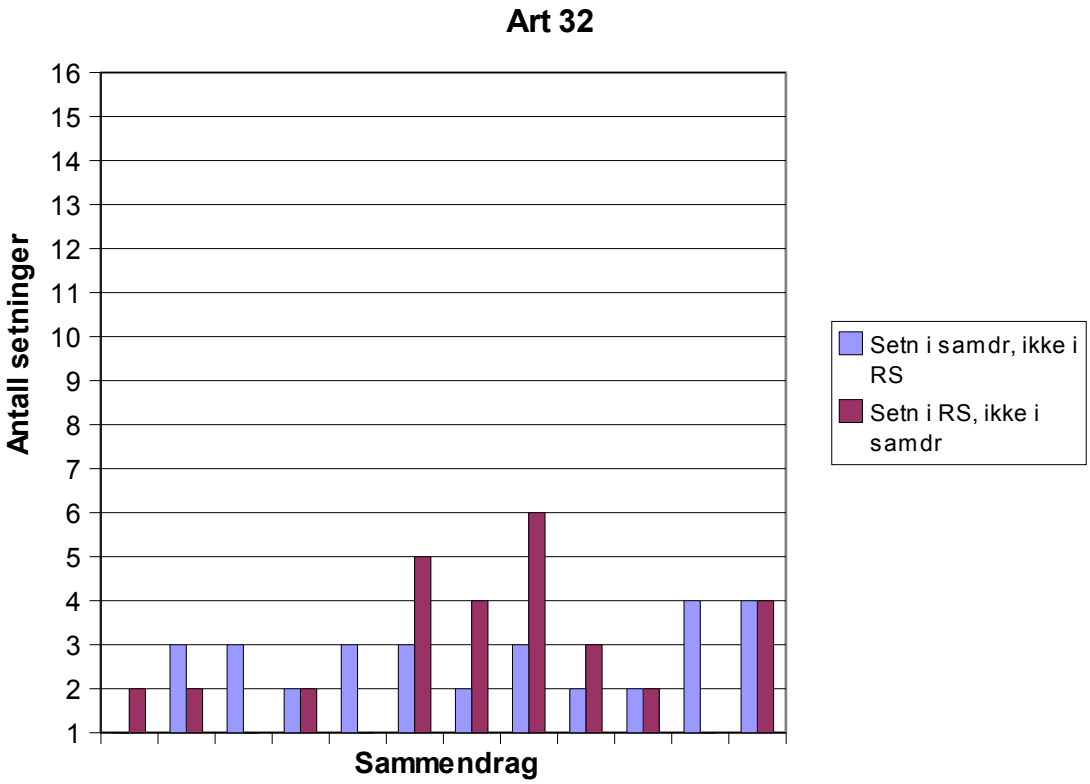
| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 5 | 2 | 2 |
| 190 | 8 | 4 |
| 179 | 2 | 1 |
| 56 | 4 | 2 |
| 72 | 3 | 2 |
| 82 | 2 | 5 |
| 95 | 3 | 4 |
| 109 | 1 | 7 |
| 174 | 3 | 4 |
| 229 | 0 | 5 |
| 297 | 4 | 3 |
| 310 | 3 | 2 |
| | | |
| Gjennomsnitt | 2,9 | 3,4 |
| Standardavvik | 1,98 | 1,73 |
| | | |
| Autosam med leksikon | 10 | 5 |
| Autosam uten leksikon | 6 | 4 |
| Gjennomsnitt | 8 | 4,5 |



| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 42 | 0 | 5 |
| 47 | 2 | 2 |
| 60 | 2 | 5 |
| 112 | 0 | 5 |
| 213 | 4 | 8 |
| 223 | 8 | 4 |
| 227 | 3 | 5 |
| 250 | 6 | 4 |
| 318 | 7 | 1 |
| 296 | 7 | 2 |
| | | |
| Gjennomsnitt | 3,9 | 4,1 |
| Standardavvik | 2,96 | 2,02 |
| | | |
| Autosam med leksikon | 8 | 4 |
| Autosam uten leksikon | 9 | 4 |
| Gjennomsnitt | 8,5 | 4 |

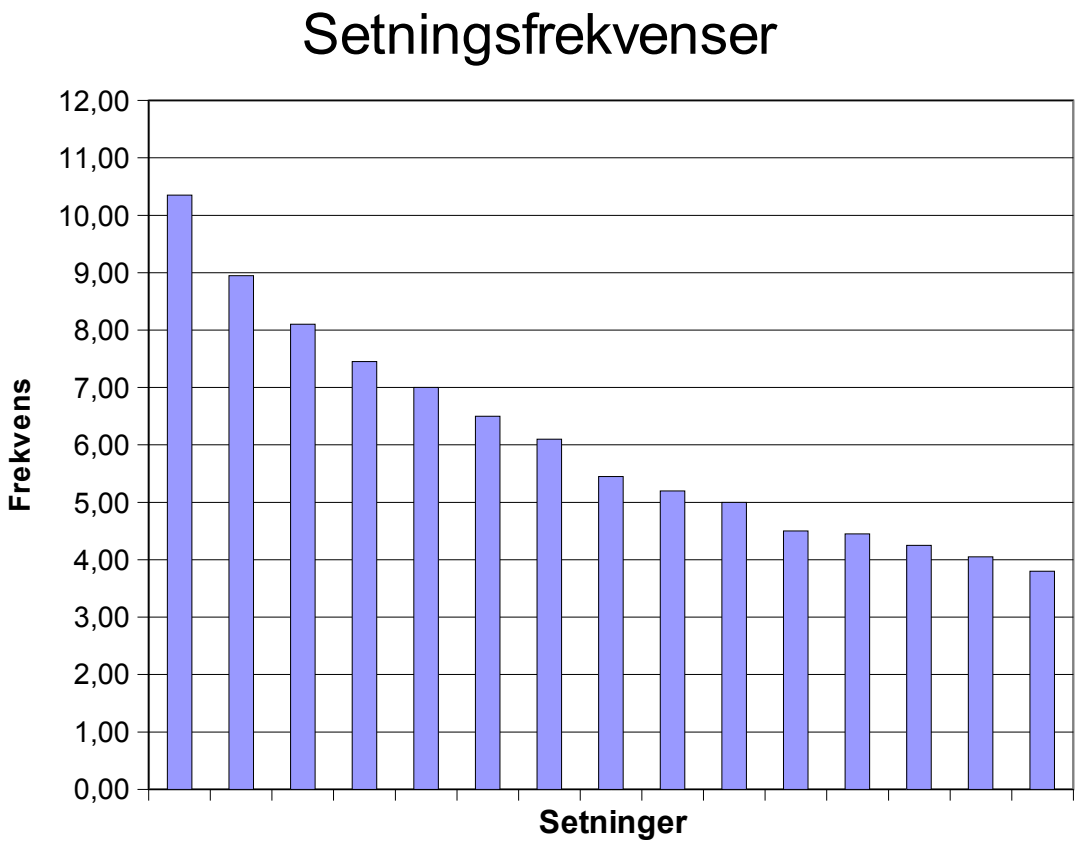


| Sammendragsnr | Setn i samdr, ikke i RS | Setn i RS, ikke i samdr |
|---|---|---|
| 1 | 1 | 2 |
| 62 | 3 | 2 |
| 79 | 3 | 1 |
| 89 | 2 | 2 |
| 94 | 3 | 1 |
| 96 | 3 | 5 |
| 103 | 2 | 4 |
| 104 | 3 | 6 |
| 135 | 2 | 3 |
| 228 | 2 | 2 |
| 291 | 4 | 1 |
| 295 | 4 | 4 |
| | | |
| Gjennomsnitt | 2,7 | 2,8 |
| Standardavvik | 0,89 | 1,66 |
| | | |
| Autosam med leksikon | 7 | 6 |
| Autosam uten leksikon | 6 | 6 |
| Gjennomsnitt | 6,5 | 6 |



Frekvensliste for de 15 mest frekvente setningene i hver artikkel

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Art 1 | 10 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 |
| Art 3 | 10 | 9 | 8 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 6 |
| Art 4 | 10 | 10 | 9 | 8 | 8 | 7 | 7 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 |
| Art 5 | 9 | 9 | 8 | 7 | 7 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 |
| Art 6 | 11 | 10 | 10 | 9 | 9 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| Art 8 | 9 | 9 | 8 | 6 | 6 | 5 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 |
| Art 11 | 9 | 9 | 8 | 7 | 7 | 7 | 6 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 |
| Art 12 | 12 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 7 | 6 | 5 | 5 | 4 | 4 | 4 |
| Art 13 | 11 | 8 | 8 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 |
| Art 14 | 11 | 8 | 7 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 3 | 3 |
| Art 15 | 11 | 10 | 9 | 9 | 8 | 7 | 6 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 3 |
| Art 17 | 11 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 5 | 4 | 4 | 4 | 3 | 3 |
| Art 18 | 12 | 12 | 9 | 9 | 8 | 6 | 5 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 3 |
| Art 19 | 10 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 |
| Art 21 | 9 | 8 | 8 | 7 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| Art 22 | 12 | 8 | 7 | 6 | 6 | 6 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 3 | 3 |
| Art 23 | 10 | 9 | 9 | 9 | 7 | 7 | 6 | 5 | 5 | 5 | 4 | 4 | 3 | 3 | 2 |
| Art 27 | 11 | 9 | 8 | 8 | 7 | 7 | 7 | 5 | 5 | 5 | 4 | 4 | 4 | 3 | 3 |
| Art 29 | 8 | 6 | 5 | 5 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Art 32 | 11 | 9 | 7 | 7 | 6 | 6 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 2 |
| Gj.snitt | 10,35 | 8,95 | 8,11 | 7,45 | 7,00 | 6,50 | 6,10 | 5,45 | 5,20 | 5,00 | 4,50 | 4,45 | 4,25 | 4,05 | 3,80 |



Her skal du lage et sammendrag av teksten du har valgt. Sammendraget vil bli presentert fortløpende nederst på siden. Klikk på setningene i teksten for å ta de med i sammendraget. Hvis du vil fjerne en setning fra sammendraget, klikk på den i sammendraget. Ta med så mange setninger du mener er nødvendig for å bevare innholdet i teksten, men ikke flere enn 10.

**Artikkelnavn: Funnet død på Karmøy**

http://caeneus.org/anja/swesum.php?artikkel_id=32

---

# Funnet død på Karmøy.

Det er trolig den 43 år gamle Per Ove Vea som er funnet død i et sumpområde på Karmøy. En mann og en kvinne er siktet for legemsbeskadigelse mot Vea.

Tre teknikere fra Kripos kom til Karmøy i går for å bistå politiet i etterforskningen. **Politiet forteller at den døde har ytre skader.** I løpet av torsdagen vil han bli sendt til obduksjon slik at man får fastslått dødsårsaken. Per Ove Vea ble meldt savnet siden han sist ble sett søndag morgen. Mannskaper fra politiet og Røde Kors har lett etter ham i Åkrahamn-området vest på Karmøy. Det var også letemannskaper fra Røde Kors som fant den døde i en sumpområde ved Tjøsvollvatnet. **Området er svært utilgjengelig.** Det er mye siv der den døde ble funnet.

Politiet hadde i går kveld ikke fått identifisert den døde, men kjente ikke til at andre personer var savnet i området. Liket ble hentet ut i 20-tiden i går kveld, og blir sendt til obduksjon i dag.

En mann og en kvinne ble onsdag avhørt av politiet på Karmøy. De to skal ha vært i slåsskamp med Per Ove Vea søndag morgen. De to, som begge er i 40-årene, ble pågrepet tirsdag. På grunn av slåsskampen søndag morgen er de siktet for legemsbeskadigelse mot den savnede 43-åringen. De to skal være bekjente av den savnede. Det var disse to som mandag meldte 43-åringen savnet.

Området der liket ble funnet, ligger bare noen få hundre meter fra bolighuset der slåsskampen skal ha skjedd søndag morgen.

Sammendraget ditt (5 setninger):

Det er trolig den 43 år gamle Per Ove Vea som er funnet død i et sumpområde på Karmøy.

I løpet av torsdagen vil han bli sendt til obduksjon slik at man får fastslått dødsårsaken. Per Ove Vea ble meldt savnet siden han sist ble sett søndag morgen.

En mann og en kvinne ble onsdag avhørt av politiet på Karmøy. De to skal ha vært i slåsskamp med Per Ove Vea søndag morgen.

Lagre

**SweSum - Automatisk Textsammanfattare av Martin Hassel och Hercules Dalianis**
**Lokalisering, gränssnitt och svensk pronomenresolution av Martin Hassel**

In English, please!

Var så god och skriv eller klistra in egen text att sammanfatta:

```
- Jeg tenkte bare på Markus

- Jeg så bare noe stort, grått som kom mot oss. Og så sa det bang. Et
langt brak som aldri tok slutt. Jeg tenkte bare på Marcus.

Slik opplevde Marianne Sæther Rekk sammenstøtet mellom bussen og
```

Du kan även ladda upp en text/HTML-fil från din egen dator:

Eventuella nyckelord som är viktiga i texten.

Välj typ av text: Tidningstext

Välj te: No

Sammanfattning av orginaltext: 30 procent

Skriv ut nyckelord och statistik Antal nyckelord: 10

Använd pronomenresolution (endast för svenska)

Ställ in viktning av diskursparametrar:

| Första raden | Fetstil | Numeriska värden | Nyckelord | Anv. nyckelord |
|---|---|---|---|---|
| 1000 | 10 | 1.133 | 0.360 | 500 |

Sammanfatta

Läs mer om textsammanfattning

Kommentarer till Hercules?
Kommentarer till Martin?



Not.
Textsammanfattning för svenska, danska, norska och engelska anses vara state-of-the-art och sammanfattning för franska, spanska och tyska befinner på prototypstadiet.
Generisk sammanfattning tar ej hänsyn till vilket språk texten är skriven på.
Pronomenresolution är endast (delvis) implementerat för Svenska och befinner sig ännu på prototypstadiet.

Not II.
Vi vill tacka Dorte Haltrup (CST), Paul Meurer (UiB), Pascal Vaillant (ENST) och Horacio Rodríguez (UPC) för deras hjälp med danska, norska, franska respektive spanska lexikon och kommentarer.

This page was last modified: 14/9 2004

- Jeg tenkte bare på Markus

- Jeg så bare noe stort, grått som kom mot oss. Og så sa det bang. Jeg tenkte bare på Marcus.

Slik opplevde Marianne Sæther Rekk sammenstøtet mellom bussen og lastebilen i Fyllingsdalen. 14 personer er skadet, og fem er sendt på sykehus, etter at en buss og containerbil kolliderte ved 10.00-tiden.

Lengre framme i bussen hørtes skrik og jammer.

- Jeg tar bussen på jobb hver dag, og har ofte tenkt på at dette krysset er trafikkfarlig. Bussen rundet svingen, og der kom den svære lastebilen imot. På høyre side i bussen hadde hun god utsikt til lastebilen som kom mot, og innså at sammenstøtet ikke var til å unngå.

- Bussen kom fra Oasen og lastebilen kom fra sentrum. De har truffet hverandre front mot front i krysset.

Deler av bussen er helt smadret etter den voldsomme smellen. Fronten er trykket inn, frontruten og andre ruter knust og inngangsdøren regelrett borte.

Også fronten på lastebilen er smadret.
**Lexikon:** Norska
**Ord före** 654
**Ord efter** 186
**Sammanfattningsgrad:** 28%
**Typ av text:** tidningstext
**Nyckelord:** *busse front lastebil* tenke *gammel pressemelding holdt behandling ulykke skrik*

<-- Back