Sissel Marie Vatne

Kilder til bildebeskrivelse

En undersøkelse av bildetekstens potensial som grunnlag for automatisk indeksering av bilder i digitale bøker

Masteroppgave

Avdeling for journalistikk, bibliotek og informasjonsfag

Sammendrag

Denne oppgaven undersøker mulighetene for automatisk tekstbasert indeksering av bilder i digitale bøker. Oppgaven har et særlig fokus på bildetekstens potensial som indekseringskilde, og tre hovedspørsmål ble stilt: hva skal til for å identifisere bildetekster i de digitale bøkene, hva skal til for å identifisere mulige indekstermer av typen personnavn, stedsnavn og årstall i bildetekstene og hvor godt indekserer disse termene bildene i forhold til de tre emnetypene termene representerer? Fordi det ble funnet lite forskning om bildeindeksering i digitale bøker, ble det nødvendig å utvikle metoder for identifisering og indeksering spesielt for denne oppgaven. Fremgangsmåten som ble utviklet for å identifisere bildetekster, ga lovende resultater, med en gjenfinning på 88% for 10 bøker. Det er allikevel utfordringer knyttet til problemstillingen som det gjenstår å løse i videre forskning. Fremgangsmåten for identifisering av personnavn ga veldig gode resultater med en presisjon på 0,9 og en fullstendighet på 0,88. Metoden for å identifisere stedsnavn derimot må utvikles videre. Når det gjelder hvor godt indekstermene indekserer bildene, indikerer resultatene at de genererte indekstermene i mange tilfeller er gode indekstermer, men de indikerer også at bildeteksten ikke er en like god indekseringskilde i alle tilfeller. Det vil derfor være interessant videre å vurdere andre kilder for bildebeskrivelse i digitale bøker, som et supplement til bildeteksten.

Masteroppgave ved Høgskolen i Oslo, Avdeling for journalistikk, bibliotek- og informasjonsfag

Oslo 2009

Forord

Da jeg begynte på masterstudiet ved Høgskolen i Oslo, var masteroppgaven det jeg gruet meg mest til. Jeg så for meg en slitsom prosess, hvor hver side og hvert fremskritt ville være et resultat av strevsomme, mentale kamper. Nå har ikke prosessen vært fullstendig uten frustrerende øyeblikk, men i det store og det hele har det å jobbe med denne oppgaven vært spennende, lærerikt og ikke minst: veldig gøy. Det er nok mye på grunn av temaet for oppgaven; bildeindeksering er en svært interessant retning innenfor kunnskapsorganisering og gjenfinning, og det skal bli spennende å følge med på hvordan denne retningen utvikler seg videre.

Det er mange jeg må takke for at denne oppgaven kom i mål. Takk til Nasjonalbiblioteket for at jeg fikk lov til å bruke dokumenter fra Nordområdesamlingen i mine undersøkelser, og en særlig takk fortjener Svein Arne Brygfjeld for å bistå meg med mine mange forespørsler om datatyper og dokumentasjon. Takk fortjener også Michael Preminger og Kjetil Iseli for uvurderlig it-støtte og hjelp med alle de tekniske utfordringene jeg møtte på veien. Torhild Bjerkreim hos Gyldendal Akademisk svarte villig på mine spørsmål om forlagets bildebruk, noe jeg er svært takknemlig for. Jeg må også takke mine medstudenter for gode faglige og ufaglige diskusjoner som skapte et godt miljø på masterrommet. Min korrekturleser, Karen Emilie Vatne, fortjener også takk for å peke ut mange av de utallige kommafeilene som har sneket seg inn i språket mitt siden min siste norsktime på ungdomsskolen. Og sist, men ikke minst, tusen takk til min veileder Ragnar Nordlie, som til stadighet hjalp meg i riktig retning når jeg ikke visste hvor jeg var, eller hvor jeg skulle.

Sissel Marie Vatne
Oslo, 23. juni 2009

Innholdsfortegnelse

Figurliste

Tabelliste

1 Innledning

1.1 Bakgrunn

De senere årene har det vært en betydelig innsats fra flere aktører innenfor det å gjøre fulltekstbøker digitalt tilgjengelig. Blant kommersielle aktører har Google alliert seg med en rekke biblioteker for å kunne gjøre bøker tilgjengelig i Google Books, og Microsoft prøvde seg på det samme med Microsoft Live Book Search. I forskningsmiljøet er særlig Carnegie-Mellon universitetet i Pittsburgh involvert i digitalisering av bøker med sitt Universal Digital Library (UDL), et digitaliseringsprosjekt som i utgangspunktet hadde et mål om å nå 1 million digitaliserte bøker, noe de oppnådde i april 2007.

Nasjonalbibliotekene har gjort en innsats i digitaliseringen av sine respektive lands litteratur, og de europeiske nasjonalbibliotekene har gått sammen i The European Library for å øke tilgjengeligheten av det digitale materialet. Nasjonalbiblioteket har digitalisert rundt 100. 000 bøker for bevaring, og har som mål å digitalisere hele den fysiske boksamlinga bestående av 450.000 titler (Nasjonalbiblioteket 2009). For å kunne gjøre det digitaliserte materialet tilgjengelig for allmennheten, lanserte i 2007 Nasjonalbiblioteket, i samarbeid med representanter for norske rettighetshavere, et pilotprosjekt kalt Nordområdeprosjektet. I Nordområdeprosjektet ble noen hundre nyere bøker knyttet til Nordområdene lagt ut i fulltekst (En hale uten ende 2008). Pilotprosjektet ble ansett som så vellykket at et større prosjekt kalt Bokhylla.no ble lansert i 2009. I dette prosjektet er 12.000 rettighetsbelagte bøker i fulltekst, i tillegg til bøker som er falt i det fri, gjort tilgjengelig på internett gjennom nettsiden bokhylla.no.

Digitale fulltekstdokumenter er spennende for forskning innenfor informasjonsgjenfinning fordi de åpner for å utnytte dokumentets innhold i automatiske gjenfinningssystem. I tidligere system har man vært nødt til å forholde seg til metadata som tittel, forfatter og emneord. Ved å ha tilgang til digitale fulltekstdokumenter, har man nå muligheten til å trekke ut enkeltelementer som man kan indeksere for seg. Mange bøker har illustrasjoner som man kan tenke seg at det vil være nyttig og ønskelig å finne tilbake til gjennom søk. Tidligere ville dette vært utenkelig, fordi den eneste måten bildene var tilgjengelige på, var ved å bla fysisk gjennom bøkene. Gjennom digitaliseringen vil bilder i bøkene kunne indekseres og dermed gjøres søkbare.

Formålet med denne oppgaven er å undersøke hvilket potensial som ligger i å indeksere bilder i digitale bøker gjennom automatisk tekstlig indeksering. Det fokuseres på noen få muligheter som presenteres og diskuteres i forhold til fordeler, ulemper og utfordringer knyttet til tilnærmingen.

1.2 Begrepsavklaring

I denne oppgaven skilles det mellom tekstbasert bildeindeksering (text-based indexing) og innholdsbasert bildeindeksering (content-based indexing). Tekstbasert indeksering kalles også for deskriptorbasert (descriptor based) eller konseptbasert (concept-based) indeksering. Denne typen indeksering baserer seg på å uttrykke bildets innhold gjennom språk, enten kontrollert vokabular som ved bruk av emneordslister og tesauri, eller fri tekst. Innholdsbasert indeksering kan defineres som enhver teknologi som i prinsippet bidrar til å organisere digitale bildearkiv etter bildenes visuelle innhold, det vil si elementer som farge, form og tekstur (Datta, Joshi, Li og Wang 2008).

I forbindelse med tekstbasert og innholdsbasert indeksering snakkes det ofte om lavnivå-egenskaper (low level features) og høynivå-egenskaper (high level features) ved et bilde. Lavnivå-egenskaper, også kalt primitive egenskaper, er egenskaper som er knyttet til bildets visuelle attributter, altså egenskaper som bildets farge, form og tekstur. Høynivå-egenskaper er egenskaper som er knyttet til bildets semantiske innhold og vil for eksempel være gjenstander, aktiviteter og personer som er avbildet. Det inkluderer også abstrakte begrep som kan knyttes til bildets innhold, slik som stemninger, følelser og symboler. I gjenfinningssystem vil lavnivå-egenskaper ofte kun behandles som datakoder, mens høynivå-egenskaper vil registreres som tekst.

Siden oppgavens materiale er tekst i bøker, er det nødvendig å definere noen typografiske begrep som blir brukt utover i teksten. Bildetekst er tekst som ledsager et bilde i boken, og som er tydelig markert som bildeledsagende tekst i bokens layout gjennom tekstens formatering og/eller plassering. Brødtekst er betegnelsen på den løpende teksten i en bok, i motsetning til for eksempel overskrifter og bildetekster. (”brødtekst” 2009).

1.3 Problemstilling

Temaet for denne oppgaven er automatisk tekstlig indeksering av bilder i digitale bøker. Dette er et stort og omfattende tema som til dels er uutforsket, i og med at det er få undersøkelser som har

dette temaet som hovedfokus. I den grad noen tidligere har forsøkt å bruke automatiske metoder for tekstlig indeksering, har det som regel vært i kombinasjon med innholdsbaserte metoder. Noe av grunnen til at det ikke har vært fokus på automatisk tekstbasert indeksering av bilder, er at det materialet som har vært tilgjengelig for indeksering, i liten grad har vært egnet til denne typen indeksering. Bilder i bildebaser har sjelden tilknyttet tekst som ikke en indekserer manuelt har tilordnet bildene, og forskningsfokuset for automatisk bildeindeksering har derfor vært å utvikle innholdsbaserte metoder. De multimodale dokumentene som har vært best egnet til bildeindeksering, har vært nettdokumenter som inkluderer både bilder og tekst i html-filer. Dette har blant annet kommersielle aktører som Google og Altavista utnyttet til sine bildesøk. Nå som multimodale dokumenter også finnes i form av digitale bøker, vil det være interessant å undersøke om teksten i disse dokumentene kan brukes til å indeksere bilder i bøkene, slik teksten i html-filer brukes for å indeksere bilder på nettsider.

Fordi mulighetene på området er store og uutforskede, ville det vært en massiv, for ikke å si umulig, oppgave å gjøre en uttømmende undersøkelse av disse mulighetene. Det har derfor vært nødvendig å innsnevre fokuset for oppgaven. For å bestemme hva som mest sannsynlig ville være et fruktbart fokus ble to spørsmål vurdert; Hva slags informasjon er det ønskelig å bruke til å beskrive bilder, og hvor vil man mest sannsynlig kunne finne denne informasjonen i bøkene?

For å kunne få til korrekt indeksering av bildene, er man avhengig av at teksten har en relasjon til bildet. Teksten i en bok vil ha forskjellig grad av relasjoner til de ulike bildene i boken, og det ble derfor regnet som viktig i denne oppgaven å identifisere et tekstlig analysegrunnlag som man ville tro har en sterk relasjon til bildet. Intuitivt vil man forvente at et bildes bildetekst vil være den biten av tekst i boken som har den sterkeste tilknytningen til det bildet. Teorier om forholdet mellom tekst og bilder støtter opp om dette, og det ble derfor besluttet at bildeteksten til hvert bilde skulle brukes som indekseringsgrunnlag for det aktuelle bildet. Dessverre er det ikke kodet i de digitale filene hvilken tekst som er bildetekst og hvilken tekst som er vanlig brødtekst. Det ble derfor nødvendig å utvikle en fremgangsmåte for å identifisere hva som er bildetekst, og dette ble ett av hovedfokusene for oppgaven.

Det finnes mange teorier for hvordan man kan beskrive bilder i forbindelse med indeksering. Fordi denne oppgaven baserer seg på automatisk tekstlig indeksering, var det naturlig å fokusere på beskrivelser knyttet til høynivå-egenskaper, fordi det er denne typen egenskaper som er

forbundet med tekstlige beskrivelser. Forskning på brukeratferd ved bildesøk har vist at brukere som regel søker på generelle eller navngitte objekter og hendelser (se kapittel 3.4), og det ble derfor besluttet å lete etter denne typen beskrivelser. Videre ble det besluttet å snevre inn beskrivelsestypene ytterligere ved å fokusere på beskrivelser som har en lett gjenkjennelig syntaks, noe som ofte er tilfelle for navngitte objekter og hendelser. Beskrivelsestypene det ble bestemt å bruke er egennavn, av typen personnavn og stedsnavn, og år.

Basert på dette ble fokuset i denne oppgaven å undersøke bildetekstens potensial som grunnlag for automatisk indeksering av bilder i digitale bøker. Oppgaven ble dermed delt i tre delundersøkelser:

- Hva skal til for å identifisere bildetekster i de digitale bøkene?
- Hva skal til for å identifisere mulige indekstermer av typen personnavn, stedsnavn og årstall i bildetekstene?
- Hvor godt indekserer disse termene bildene i forhold til de tre emnetypene termene representerer?

1.4 Oppgavens disposisjon

I kapittel 2 presenteres og drøftes forskjellige teorier om bildeindeksering og forholdet mellom bilde og tekst i bøker. Det diskuteres også kort hvorvidt automatisk tekstbasert indeksering av bilder er mulig eller ikke sett fra et teoretisk standpunkt.

I kapittel 3 gjennomgås tidligere forskning relatert til bildeindeksering.

I kapittel 4 diskuteres og begrunnes de valgene som er tatt i forhold til oppgavens fokus i lys av presentert teori og forskning.

Kapittel 5 presenterer datamaterialet og gjennomgår hvordan materialet er behandlet for å forberede det til bruk i undersøkelsene.

Kapittel 6 tar for seg denne oppgavens forsøk med å identifisere bildetekster i bøkene. Fremgangsmåten som er brukt blir gjennomgått og resultatene av undersøkelsene blir presentert.

Kapittel 7 tar for seg identifisering av egennavn og årstall i bildetekster. Her blir den grunnleggende fremgangsmåten forklart samt de videre teknikkene som er brukt i denne

oppgaven. Resultatene av denne analysen blir presentert med en diskusjon av hva som må utvikles videre.

Kapittel 8 diskuterer muligheter for videre forskning innen området, generelt og basert på det som er gjort i denne oppgaven.

2 Teoretisk grunnlag for automatisk tekstbasert bildeindeksering

2.1 Indekseringsteori

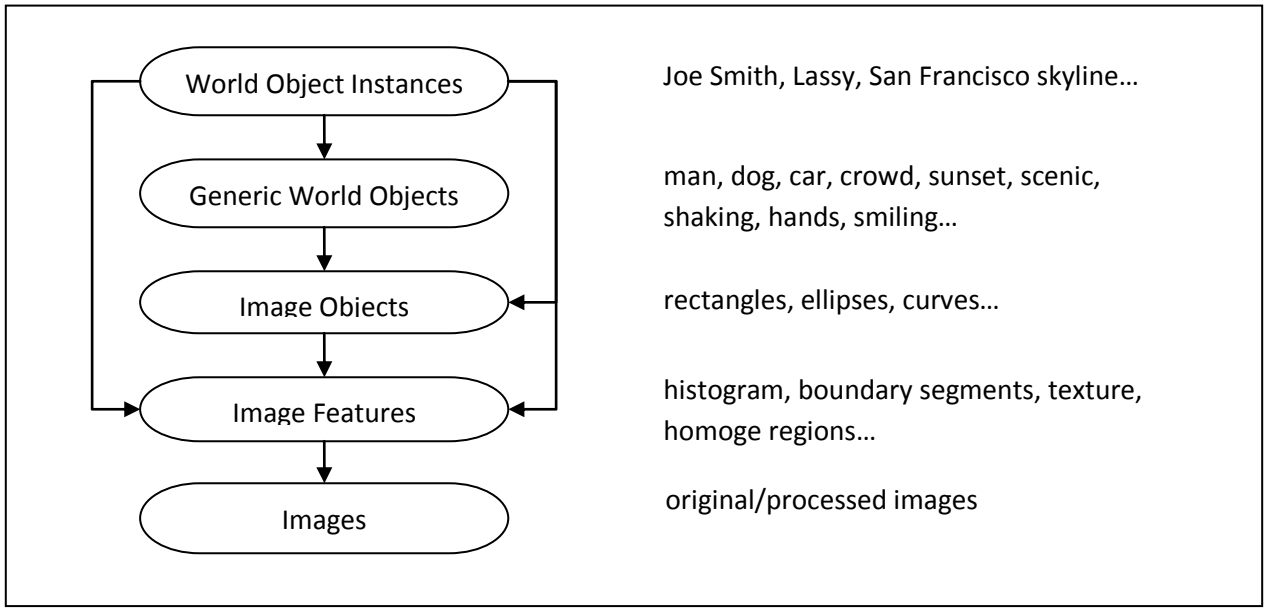
Et sentralt spørsmål i indekseringslitteraturen handler om hvordan bilder skal beskrives. Det finnes en rekke attributter som kan brukes for å beskrive bilder, og det er stor spredning i hvordan forskjellige personer oppfatter samme bilde. Attributtene som kan brukes til å beskrive et bilde, kan også være av vidt forskjellig karakter, og det har derfor vært et behov for å identifisere ulike klasser eller nivåer av bildebeskrivelse. I dette kapittelet vil tre av de vanligste tilnærmingene beskrives: Shatfords tolkning av Panofsky, Mehrotras abstraksjonsnivå og Eakins bildeinnholdsnivå.

Panofskys teori om betydning i kunstverk innebærer å skille mellom pre-ikonografisk beskrivelse, ikonografisk analyse og ikonologisk tolkning. Den pre-ikonografiske beskrivelsen handler om det som er avbildet i bildet, mens den ikonografiske analysen trekker inn historier og kjente temaer som eksisterer utenfor bildet. Panofskys ikonologiske tolkning går et skritt lenger og trekker inn symbolikken i bildet (Rasmussen 1997). Shatfords (1986) tolkning av Panofsky er mye brukt i indekseringsteorien. Hun skiller mellom hva et bilde er *av* (of) og hva det er *om* (about); med andre ord: hva som er avbildet og hva bildet handler om, eller bildets objektive og subjektive mening. Basert på dette skiller hun mellom tre ulike nivå av emneord: de generiske avbildningene, de spesifikke avbildningene og bildets symbolikk eller tema. Disse nivåene har hver fire fasetter: hvem, hva, hvor og når (Figur 1). I Shatfords tolkning handler både det pre-ikonografiske og det ikonografiske nivået om beskrivelser av det som er avbildet, noe som hos Panofsky kun gjaldt den pre-ikonografiske beskrivelsen. Shatfords pre-ikonografiske og ikonografiske nivåer tilsvarer bildets *ofness*, den objektive beskrivelsen av det som er avbildet, mens det ikonologiske nivået trekker inn bildets *aboutness* og dets subjektive mening.

Fasett	*Pre-ikonografi*	*Ikonografi*	*Ikonologi*
HVEM?	Type person, ting	Navngitt individ, ting	Mytisk skikkelse
HVA?	Handling, situasjon	Navngitt hendelse/arrangement	Følelse, abstraksjon
HVOR?	Type plass, geografi	Navngitt sted	Symbolisert sted
NÅR?	Årstid, syklus, tidspunkt	Dato, periode	Følelse, symbolisert tid

Figur 1. Shatfords indekseringsnivå etter Enser (1995)

Mehrotra er inne på noe av det samme i sine abstraksjonsnivå, men hans tilnærmingsmåte er noe forenklet, selv om han trekker inn enda flere aspekter (Lancaster 2003). Han starter med bildets attributter, som er egenskaper av typen farge og tekstur. Det neste nivået er bildeobjekter, som omfatter figurer som kurver og rektangler. Disse to nivåene tar med andre ord for seg bildets lavnivå-egenskaper. Deretter kommer generiske objekter (Generic World Objects) som kan tilsvare Shatfords pre-ikonografi, og til slutt har Mehrotra nivået tilfeller (World Object Instances) som tilsvarer Shatfords ikonografi. Denne måten å dele bildets bestanddeler på tar hensyn til de aspektene den innholdsbaserte indekseringen baserer seg på, i motsetning til Panofsky/Shatford som utelukkende fokuserer på det ikoniske i bildene, høynivå-egenskapene, som oftere er forbundet med tekstbasert indeksering.



Figur 2. Mehrotras abstraksjonsnivå etter Lancaster (2003)

En siste tilnærming som er mye referert er Eakins tre nivåer (Enser 2008):

1. Visuelle primitive egenskaper
2. Logiske (deriverte) egenskaper
3. Induktiv tolkning

Eakins første nivå tilsvarer Mehrotras to nivåer bildeobjekter og bildeattributter, og inkluderer for eksempel farge, tekstur og form. Eakins andre nivå er egenskaper som finnes i bildet og som kan trekkes ut direkte, uten noen form for tolkning. Dette tilsvarer de to øverste nivåene til Mehrotra og Shatfords pre-ikonografi og ikonografi, og inkluderer objekter, aktiviteter og hendelser. Eakins siste nivå består av abstrakte egenskaper som krever en tolkning av bildets innhold utover det som objektivt sett er avbildet. Dette nivået tilsvarer Shatfords ikonologi.

De forskjellige teoriene har fokus på forskjellige egenskaper ved et bilde. Selv om det også er den enkleste tilnærmingen, er Eakins tilnærming er den eneste som spenner fra lavnivå-egenskaper som farge og form og helt til abstrakte egenskaper ved et bilde. Mehrotra inkluderer også lavnivå-egenskaper, men tar bare med noen typer høynivå-egenskaper; abstrakte egenskaper er ikke inkludert i hans nivåer. Shatford ignorer lavnivå-egenskapene og fokuserer utelukkende på bildets semantiske egenskaper, eller høynivå-egenskapene. Fordi denne oppgaven fokuserer på tekstlig indeksering vil Shatfords nivå brukes når det skal refereres til indekseringsnivå. Dette er fordi Shatford behandler de ulike høynivå-egenskapene grundigere enn de andre teoriene som er nevnt, og det anses som viktigere å kunne referere presist til de semantiske egenskapene ved et bilde, enn å kunne referere til nivå som har med lavnivå-egenskaper å gjøre, når indekseringen som foretas er tekstbasert.

2.2 Bilder og tekst i kontekst

Når bildetekstene i bøkene brukes som indekseringsgrunnlag i denne oppgaven, så hviler dette på den grunnleggende antakelsen om at bildene har en form for relasjon til teksten. Er dette en akseptabel antakelse? Blant undersøkelsene som nevnes i kapittelet om tidligere forskning, er det flere som fremhever at teksten i bildets kontekst kommer langt ned på listen av indekseringskilder fordi andre kilder var bedre egnet til å uttrykke bildeinnholdet (Frankel, Swain og Athitsos 1996; Mukherjea og Cho, 1999). Det kan jo tyde på at forholdet mellom tekst og bilde ikke er så sterkt som man kanskje skulle tro. Og hvis antakelsen om at bilder og tekst har en relasjon, er feil, og bildet er plassert der det er, uavhengig av hva teksten er, kan man ikke forvente at teksten skal kunne brukes for å si noe om bildet. Intuitivt virker dette lite sannsynlig.

Det vil være nærliggende å tro at om et bilde er inkludert i en bok, så er det med et formål, og formålet vil på en eller annen måte være knyttet til teksten den illustrerer. Men det betyr ikke at alle illustrasjoner i bøker har samme formål. Selv i ordet "illustrere" ligger det en tvetydighet: å illustrere kan være å opplyse og forklare med eksempler, eller det kan være å pryde og utsmykke(”Illustrere” 2004). En illustrasjon med formål å opplyse vil ha en annen relasjon til teksten enn det en illustrasjon med formål å dekorere vil ha.

Når man ønsker å indeksere bilder basert på nærliggende tekst, er det en del spørsmål som dukker opp. Vil et bilde som er en illustrasjon i en bok, ha en direkte relasjon til teksten den illustrerer, eller snudd på hodet: vil teksten kunne si noe om bildet? Og selv om det er en klar relasjon mellom teksten og bildet, vil det være forskjell i meningsinnholdet som uttrykkes i bildet og meningsinnholdet som uttrykkes med tekst? Hva er forskjellen, og hva vil det bety om indekseringen ikke fanger opp dette? Kan man forvente at automatisk tekstlig indeksering i det hele tatt vil være formålstjenlig?

2.2.1 Forholdet mellom bilde og tekst

I sine artikler om bilders budskap kommer den franske litteraturteoretikeren Roland Barthes (1977a; 1977b) inn på hvordan tekst og bilder forholder seg til hverandre i kommunikasjonen. Hans oppfatning er at bilder og tekst alltid vil være knyttet til hverandre. Et bilde vil aldri kunne stå alene, men vil alltid stå i kontekst til teksten. Bildet og teksten vil aldri kunne uttrykke det samme, de kan ikke være identiske, men bildets og tekstens innhold og budskap vil være beslektede og komplementære(Barthes 1977a). Han understreker at bilder er flertydige av natur, de vil ha budskaper som teksten ikke har. Teksten vil representere en kontekst som binder bildet til et bestemt budskap. Denne forankringen av bildets betydning vil føre til en svekking av bildets andre betydninger(Barthes 1977b). I dette tilfellet har bildet et meningsinnhold som er mer mangfoldig enn tekstens, men teksten legger restriksjoner på bildets meningsinnhold. Den konteksten teksten tilfører bildet, kan være en understreking av bildets hovedinnhold, dets denotasjon, men den kan også tilføre bildet konnotasjoner som ikke vil vært åpenbare om bildet sto for seg selv. Hvor teksten befinner seg, kan ha noe å si for om teksten understreker denotasjonene i bildet eller tilfører det konnotasjoner: en bildetekst vil gjerne speile et bildes denotasjon, mens teksten for øvrig og overskrifter kan gi større konnotasjoner avhengig av hva som er fokus i teksten og fremhevet i overskriftene(Barthes 1977a). I de tilfellene hvor

konteksten gir bildet konnotasjoner som ikke er iboende i bildet, vil tekstens og bildets primære meningsinnhold være betydelig forskjellig, selv om bildet og teksten er beslektet.

Ifølge Barthes finnes det altså et uløselig forhold mellom bilder og tekst, men det er ikke nødvendigvis en positiv ting. Konteksten kan begrense bildet slik at meningsinnholdet i beskrivelsen ikke representerer det primære meningsinnholdet i bildet. Kan man separere et bilde fra konteksten når teksten bildet beskrives ut fra, er det som skaper konteksten i utgangspunktet? Tekst og bilde er ikke identisk, og teksten vil aldri kunne fange opp alle dimensjoner av bildet. Men det kan heller ikke vanlig manuell indeksering. All tekstbasert indeksering vil sette bildet inn i en kontekst fordi all tekst vil ha en innskrenkende effekt på bildets meningsinnhold, om beskrivelsen er trukket ut automatisk eller tilordnet manuelt. Erfaring viser at forskjellige indekserere beskriver samme dokument ulikt, og dette gjelder særlig når dokumentet som beskrives, er et bilde. Fordi ”et bilde sier mer enn 1000 ord”, ser ikke indekserne ut til å bli enige om hvilke ord som skal beskrive bildet. Hva hver enkelt får ut av bildet, er avhengig av øyet som ser, og derfor vil et bilde aldri kunne få en uttømmende indeksering gjennom tekst. Det at automatisk tekstlig indeksering bare vil kunne beskrive bildet i en gitt kontekst, er dermed ikke et særfenomen for den automatiske indekseringen, selv om fenomenet vil være mer fremtredende ved automatisk indeksering enn ved manuell indeksering.

Det er ikke mulig å separere et bilde fra konteksten ved automatisk indeksering. Konteksten teksten skaper vil ved automatisk indeksering være den eneste konteksten bildet kan indekseres i. Bildebeskrivelsen vil dermed kun kunne representere den ene konteksten teksten representerer. Dette vil føre til en snevrere indeksering enn det man kunne oppnå ved manuell indeksering. Men så lenge bildets og tekstens meningsinnhold overlapper, må en snevrere indeksering kunne ha verdi med tanke på ressurssparingen automatisk indeksering oppnår. Det vil allikevel måtte betegnes som problematisk med en snever beskrivelse når beskrivelsen legger vekt på konnotasjoner ved bildet som ikke representerer det primære meningsinnholdet.

2.2.2 Bildets funksjon

I forordet til *The history of the illustrated book* skriver John Harthan (1981):”Illustration serves a variety of purposes. It can be instructional, didactic, documentary, literary, recreational, decorative and probably much else”. Det finnes forskjellige årsaker for å illustrere bøker med bilder. De forskjellige årsakene til å illustrere gjør også at bildene fyller forskjellige funksjoner i

forhold til teksten. Kan funksjonen bildet har i tilknytning til teksten kunne påvirke hvordan bildet vil bli beskrevet ved automatisk indeksering? Man kan for eksempel tenke seg at et bilde med en dokumentarisk funksjon vil få en mer konkret beskrivelse enn et bilde som har en ren dekorativ funksjon. Men vil det bety at det vil være en bedre beskrivelse? Et annet spørsmål er om det i det hele tatt vil være mulig å få til en konsistent indeksering over hele dokumentsamlingen når bilder kan ha mange forskjellige funksjoner/relasjoner til teksten.

2.2.3 Taksonomier

A Functions expressing little relation to the text	B Functions expressing close relation to the text	C Functions that go beyond the text
A1 Decorate	*B1 Reiterate*	*C1 Interpret*
A1.1 Change pace	B1.1 Concretize	C1.1 Emphasize
A1.2 Match style	B1.1.1 Sample	C1.2 Document
A2 Elicit emotion	B1.1.1.1 Author/Source	*C2 Develop*
A2.1 Alienate	B1.2 Humanize	C2.1 Compare
A2.2 Express poetically	B1.3 Common referent	C2.2 Contrast
A3 Control	B1.4 Describe	*C3 Transform*
A3.1 Engage	B1.5 Graph	C3.1 Alternate progress
A3.2 Motivate	B1.6 Exemplify	C3.2 Model
	B1.7 Translate	C3.2.1 Model cognitive process
	B2 Organize	C3.2.2 Model physical process
	B2.1 Isolate	C3.3 Inspire
	B2.2 Contain	
	B2.3 Locate	
	B2.4 Induce perspective	
	B3 Relate	
	B3.1 Compare	
	B3.2 Contrast	
	B3.3 Parallel	
	B4 Condense	
	B4.1 Concentrate	
	B4.2 Compact	
	B5 Explain	
	B5.1 Define	
	B5.2 Complement	

Figur 3. Marshs taksonomi over bildefunksjoner

Det har vært gjort flere forsøk på å kartlegge bildefunksjoner. Innenfor fagområdet bibliotek og informasjonsvitenskap har blant annet Berinstein (1997) skrevet om informasjonsnytten ved bilder, og hun har også presentert en taksonomi over bildefunksjoner. I taksonomien skiller hun ikke klart mellom bildetyper og bilderoller, hvor bildetyper sier noe om hva slags bilde det er snakk om, mens bilderoller er den funksjonen bildet oppfyller i teksten. For eksempel har hun kategorier som identifiserer bildetyper som *instruerende*, *forklarende* og *lokaliserende* og kategorier som identifiserer roller av typen *dekorerende*, *erklærende*, *emosjonell*, *fortellende* og *erstattende*. Marsh og White (2003) går grundigere til verks i sin taksonomi, som er basert på en litteraturgjennomgang av forskning som direkte eller indirekte behandler spørsmålet. Taksonomien består av 17 funksjoner som er delt i tre hovedkategorier etter hvor sterk relasjonen mellom bildefunksjonen og teksten er: bildefunksjoner som uttrykker liten grad av relasjon til teksten, sterk grad av relasjon til teksten og funksjoner som overstiger teksten(Marsh og White, 2003; 653). Oversikten over Marshs kategorier er gjengitt i Figur 3. Bilder som har en funksjon med liten tilknytning til teksten, kan være vanskelig å indeksere, men samtidig er ikke kategoriene gjensidig utelukkende. Et bilde som oppfyller en emosjonell funksjon, kan samtidig ha en konkretiserende funksjon, og det vil dermed være en relasjon mellom bilde og tekst som kan utnyttes i indekseringen. Men hvis det finnes tilfeller hvor bildets eneste funksjon er for eksempel dekorativ, vil dette kunne påvirke indekseringen negativt.

2.2.4 Tekst- og bilderelevans

Levin og Mayer (1993) har sett på illustrasjonens evne til å fremme læring, med fokus på hvorvidt bilder kan oppfattes som relevante i forhold til teksten. Man kan tenke seg en parallell mellom det at en leser oppfatter bildet som relevant for teksten og god tekstlig indeksering av bildene. Hvis en leser ikke oppfatter et bilde som relevant i forhold til teksten, så vil neppe indekseringen basert på teksten kunne oppfattes som en god beskrivelse av bildet. Det betyr ikke at et bilde som blir oppfattet som relevant av en leser, automatisk vil få en god indeksering basert på den samme teksten: bildet kan bli oppfattet som relevant i konteksten, mens beskrivelsen vil kunne bli mindre relevant når hele konteksten ikke er tilgjengelig. Levin og Mayers prinsipp om tekstrelasjon kan bli oppfylt av forskjellige typer bilder. De legger ikke uventet vekt på kategorier som tilsvarer Marshs kategorier B1 *gjenta* og B2 *organisere,* altså innenfor kategorien for funksjoner som har en sterk relasjon til teksten, men også funksjoner som havner i kategorien for funksjoner som overstiger teksten (C3) vil kunne oppfylle tekstrelasjonsprinsippet. Levin og

Mayer understreker også at illustrasjoner med et rent dekorativt formål ikke vil kunne oppfattes som en relevant illustrasjon.

2.2.5 Illustreringspraksis

Illustrasjonsteori er en interessant inngang til hvordan man kan forvente at bilder forholder seg til tekst, men samtidig kan det være interessant å vite hvordan forlagene tenker når de velger å illustrere sine bøker. Torhild Bjerkreim (personlig kontakt), forlagsdirektør i Gyldendal Akademisk, en avdeling av forlaget som ofte produserer bøker med illustrasjoner, legger vekt på at illustreringspraksisen varierer etter hvilken målgruppe, nivå, fagområde og salgsforventninger forlaget har. Hovedfunksjonene Bjerkreim beskriver som forlagene legger vekt på, er funksjoner som faller inn under Marsh sine kategorier B1 *gjengi* og C1 *tolkning* (for eksempel illustrasjoner i naturfagsbøker og instruksjonsbøker), B5 *forklare* (modeller for å synliggjøre teori) og A1 *dekorere* og A2 *emosjonell funksjon* (levendegjøring og inspirasjon).

2.3 Tekst som kilde til bildebeskrivelse

Det har vært en liten grad av diskusjon i fagmiljøene om hvorvidt tekst fra et teoretisk standpunkt kan brukes til automatisk indeksering av bilder. Størstedelen av bildeindekseringen har vært knyttet til bildebaser hvor bildene som registreres, er bilder uten noen form for tekst knyttet til seg før indekseringen. All tilknyttet tekst i basen vil være manuelt generert, og dermed har det aldri vært behov for å diskutere hvorvidt det er mulig teoretisk å bruke tekst i automatisk indeksering av bilder. Det har kanskje vært mer aktuelt for forskere som bruker tekstbaserte metoder i kombinasjon med innholdsbaserte metoder, særlig i forhold til indeksering av bilder i html-filer, men fordi fokuset blant disse er mer rettet mot strukturert tekst enn ustrukturert tekst, har det heller ikke vært nødvendig for disse å se om det finnes teoretisk grunnlag for å utnytte nærliggende tekst i bildeindeksering. Gjennomgangen over har vist at det finnes problematiske element som det kan være greit å være klar over og oppmerksom på, ved automatisk tekstbasert indeksering. Tekst vil til enhver tid representere en kontekst som vil begrense indeksering av flertydige bilder. Samtidig er det ingen garanti at bildet har en sterk nok relasjon til teksten, eller at teksten har en sterk nok relasjon til bildet, til at teksten kan brukes for å beskrive bildet. Men i det store og det hele er det en bred enighet om at selv om den ikke er konsistent og lik i alle tilfeller, så er det en relasjon mellom tekst og bilde. Det burde derfor kunne være mulig å utnytte teksten til automatisk indeksering. Indekseringen vil ikke kunne være like bra som manuell

tekstlig indeksering, men manuell tekstlig indeksering er ikke en reell mulighet fordi det er så ressurskrevende. Og i forhold til innholdsbasert indeksering kan tekstbaserte indeksering gi søkeinnganger som brukerne ønsker og som de innholdsbaserte systemene ikke kan tilby. Kombinasjoner av innholdsbaserte og tekstbaserte metoder kan kanskje styrke tilbudet enda mer ved å utnytte styrkene ved de to forskjellige tilnærmingsmåtene.

3 Forskning innen bildeindeksering

Forskning på indeksering av bilder kan i grove trekk deles inn i tre leire, automatisk innholdsbasert indeksering og gjenfinning, manuell tekstbasert indeksering og gjenfinning, og brukerfokuserte studier som undersøker hvilke behov brukere har i bildegjenfinningssituasjoner. Det finnes også forskning som fokuserer på teoretiske aspekter ved bildegjenfinning, uavhengig av hvilke metoder som brukes til indekseringen. Denne typen forskning er opptatt av hvordan bildeindeksering skiller seg fra indeksering av tekstlig materiale. Shatford (Shatford 1986; Layne 1994) er opptatt av at man må forstå hva slags innhold/hvilke attributter bilder har, og hva slags innhold det er viktig å beskrive for å kunne lage gode gjenfinningssystem for denne typen materiale. Hun bruker emneanalyse for å analysere og klassifiserer ulike emnetyper et bilde kan representere, for å etablere et teoretisk grunnlag for hvordan bilder kan indekseres. Hun skriver også om forskjellen mellom bildets *ofness* og *aboutness*, noe også Ornager (gjengitt etter Tsai 2006) tar opp. Bildets to innholdsaspekt, det objektive og det ekspressive (ofness og aboutness) må gjenspeiles i bildebeskrivelsen, mener Ornager.

Av den forskningen som fokuserer på utvikling av indekseringsmetoder og praktisk indeksering, fokuserer størstedelen av forskningen på indeksering og gjenfinning av bilder i bildebaser (Bjarnestam 1998; Constantopoulos og Doerr 1995; Srihari 1995), eller indeksering og gjenfinning av bilder i html-filer på nett (Frankel, Swain og Athitsos 1996; Chen 1999; Mukherjea og Cho 1999). I tillegg finnes det eksempler på arbeid med bilder i andre dokumenttyper, slik som Maderlechner, Panyr og Suda (2006) som har jobbet med å annotere bilder i pdf-dokumenter.

3.1 Tekstbasert indeksering og gjenfinning

Tekstbasert indeksering kalles også for deskriptorbasert(descriptor-based) eller konseptbasert (concept-based) indeksering. Denne typen indeksering baserer seg på å uttrykke bildets innhold

gjennom språk, enten kontrollert vokabular som ved bruk av emneordslister og tesauri, eller fri tekst. I noen oversiktsartikler om forskning på bildeindeksering settes det likhetstegn mellom tekstbasert indeksering og manuell indeksering, slik Chu (2001) gjør, sannsynligvis fordi tekstbasert indeksering tradisjonelt sett kun har blitt utført manuelt. Ofte er det bilder i bildebaser uten tilknyttet tekst som blir indeksert og i slike tilfeller må det en manuell indeksering til for å knytte tekst/termer til hvert bilde. Det at det behøves manuell indeksering for å bruke tekstbasert indeksering og gjenfinning, har vært et av ankepunktene mot tekstbasert indeksering, fordi manuell indeksering er svært ressurskrevende. Ved store samlinger vil det være tilnærmet umulig å indeksere bildene manuelt.

Forskning på tekstbasert indeksering har vært mest utbredt i bibliotekfaglige miljø (Enser 2008). Mye av forskningen fokuserer særlig på språkene som brukes ved denne typen indeksering. Jörgensen (2003) har studert blant annet tesaurusene; Library of Congress Thesaurus for Graphic Materials (LCTGM), Art & Architecture Thesaurus (AAT) og ICONCLASS for å finne ut hvordan termene er fordelt på ulike fagemner for å se i hvilken grad disse verktøyene kan brukes for generelle bildesamlinger. AAT til bruk ved bildeindeksering blir også grundig gjennomgått av Peterson (Tsai 2006). Systemer som bruker tekstbasert indeksering blir beskrevet blant annet av Bjarnestam (1998), som har sett på et system for gjenfinning av tekstlig indekserte bilder i bildebaser for kommersielt bruk (stock photography). Systemet legger vekt på bruk av hierarkisk ordnet kontrollert vokabular og mulighet for gjenfinning av både abstrakt og konkret innhold på flere språk. Når det gjelder indeksering av bilder på nett, bruker store kommersielle aktører som Google og Altavista tilsynelatende hovedsakelig tekstbasert indeksering ved å analysere den strukturerte og ustrukturerte teksten på nettsidene i indekseringen, men de eksakte algoritmene er ikke tilgjengelig for offentligheten, og det kan også være at det er innholdsbaserte metoder involvert uten at aktørene oppgir dette. (Google 2009; AltaVista 2007). En grundigere gjennomgang av studier innenfor tekstbasert bildeindeksering kan man finne hos Goodrum (2000) og Winget (2002).

3.2 Innholdsbasert indeksering og gjenfinning

Et alternativ til tekstbasert indeksering er innholdsbasert indeksering. Datta et al. (2008) definerer innholdsbasert bildegjenfinning som enhver teknologi som i prinsippet bidrar til å organisere digitale bildearkiv etter bildenes visuelle innhold, det vil si elementer som farge, form og tekstur.

Eakins (gjengitt etter Tsai 2006) kaller det visuelle innholdet de primitive egenskaper ved et bilde, mens andre bruker betegnelsen lavnivå-egenskaper (low level features). Hvilke elementer som er utnyttet i innholdsbasert bildeindeksering, beskrives i en taksonomi utviklet av Gudivada og Raghavan (1997). Eksempler på systemer som bruker ren innholdsbasert indeksering, finnes hos Zachary og Iyengar (2001) og Lieberman, Rosenzweig og Sing (2001). Fordi de primitive egenskapene ved et bilde er velegnete for automatisk indeksering, er innholdsbasert indeksering en metode som kan håndtere store bildearkiv som det ikke er mulig å indeksere manuelt. Innholdsbaserte bildegjenfinningsbaserte systemer i dag har forskjellige innganger til søking, de vanligste er innledende nøkkelord, skisser, eksempelbilder, ikoner eller en kombinasjon av disse (Lew 2000). En oversikt over ulike innholdsbaserte bildesøkesystemer finnes hos Datta et al. (2008), og Enser (2008) har en oversikt over forskning innenfor dette feltet.

3.3 Kombinasjoner av tekstbasert og innholdsbaserte metoder

Det finnes også eksempler på kombinasjoner av tekstbaserte og innholdsbaserte metoder. Bilder i bildebaser har ofte hatt lite tekst knyttet til seg som ikke er tilordnet manuelt, men det finnes eksempler på at man kan kombinere tekstbaserte og innholdsbaserte metoder i bildebaser når det finnes tekst som allerede er assosiert med bildene (Srihari 1995; Paek et al. 1999; Barnard og Forsyth 2001 og Barnard 2003). Selv om det finnes eksempler på bildebaser med tilknyttet tekst, har dette vært så sjeldent at det har vært naturlig å ha et skille mellom tekstlig manuell indeksering og automatisk innholdsbasert indeksering. Men med utbredelsen av multimodale nettsider, strukturerte websider med både bilder og tekst, har forskningsområdet fått tilgang på datamateriale som kombinerer tekst og bilde på en måte som ikke har vært til stede i bildebasene. Dette har gjort at mange har ønsket å kombinere innholdsbaserte og tekstbaserte metoder til indekseringen, hvor teksten på nettsidene analyseres for å finne tekst som kan assosieres til bildene. Eksempler på dette finner man i Frankel, Swain og Athitsos (1996), Chen (1999) og Mukherjea og Cho (1999). Felles for de fleste av disse systemene er at den tekstbaserte indekseringen er et supplement til den innholdsbaserte indekseringen, og brukes ofte for å generere en tekstlig inngang til innholdsbasert gjenfinning (Dunlop og Rijsbergen 1993), eller for å bidra til å avklare tvetydighet i innholdsbasert gjenfinning (Chen et al. 2001; Mukherjea og Cho 1999). Tekstbaserte metoder kan også brukes for å knytte tekstlig informasjon sterkere til den visuelle informasjonen. Et eksempel er Srihari (1995) som kombinerer tekstbaserte metoder med

innholdsbaserte metoder for å kjenne igjen ansikter og mennesker for å kunne knytte navn i teksten til menneskene på bildene.

Veldig mange av systemene for gjenfinning av bilder på nett gjør som Frankel, Swain og Athitsos (1996), og utnytter det at html-filer er strukturerte når de skal analysere teksten på nettsidene. Tekst hentet fra bildets filnavn, bildetaggens alt-tekst, sidens tittel, og lenker er blant gjengangerne i systemene. Flere av disse systemene anser også den generelle, ustrukturerte teksten på nettsidene som relevant, men den blir sjelden regnet som mest viktig for å indeksere bildene, fordi det rett og slett er mye vanskeligere å analysere hvilken spesifikk del av teksten på siden som er mest relevant for hvert bilde(Mukherjea og Cho 1999). Løpende tekst blir derfor ofte nedprioritert i forhold til mer strukturert tekst knyttet til html-taggingen. Her skiller Rowe seg ut ved å konsentrere seg om den løpende teksten (Rowe og Guglielmo 1992; Rowe 1993; Rowe og Frew 1998). Rowe er opptatt av å kunne analysere tekst uten å ty til avansert språkanalyse, og han diskuterer blant annet hvordan lingvistiske nøkkelord i bildeteksten kan bidra til å oppklare syntaktiske og semantiske tvetydigheter (1993). Senere bruker han slike nøkkelord for å finne tekstavsnitt som kan være aktuelle som bildetekster (Rowe og Frew 1998). Systemet som beskrives, er utviklet for å finne ord relatert til bilder på nettsider og finner nøkkelord ved å analysere tekst nær bildereferanser. Nøkkelordene er ikke det eneste som vurderes, også Rowe og Frew bruker andre, tidligere nevnte, aspekter, som betydning(overskrift/tittel etc.), alt-tekst, lenker og bildefilens navn, i tillegg til layoutmessige aspekter som skriftstil og plassering.

De aller fleste systemene som indekserer bilder på nett, bruker html-filer som datagrunnlag, men det finnes unntak. Maderlechner, Panyr og Suda (2006) har undersøkt hvordan man kan identifisere bildetekster i pdf-dokumenter på nettet, og fokuserer da særlig på bruk av strukturelle element, som tekstens plassering i dokumentet i forhold til bildet.

Kherfi, Ziou og Bernardi(2004) diskuterer grundigere problemer og utfordringer ved design og implementering av søkemotorer for bilder på nett, inkludert datainnsamling og behandling, indeksering, spørsmålsspesifisering, gjenfinning, likhet, dekningsområde, prestasjonsevne og evaluering. De presenterer også en grundig gjennomgang av hvordan disse utfordringene er behandlet av eksisterende søkemotorer.

3.4 Brukerstudier

En stor forskjell mellom innholdsbasert og tekstbasert indeksering er hvordan metodene tilrettelegger for gjenfinning av bilder. Der innholdsbaserte metoder legger vekt på primitive egenskaper, vil ofte tekstbaserte metoder indeksere det semantiske innholdet ved bildet, og dette vil påvirke hvordan en bruker kan søke etter bilder i systemet. Enkelte forskere har ønsket å undersøke hvordan brukere faktisk søker etter bilder, hva de bruker bildene til og hva de ønsker å kunne søke etter når de leter etter bilder. Armitage og Enser (gjengitt etter Tsai 2006) sin undersøkelse og analyse av brukerbehov ved søk i bildebaser, fant at få er ute etter gjenfinning via lavnivå-egenskaper. Markkula og Sormunen(1998)gjorde en mer spesialisert studie av journalisters behov i forhold til gjenfinning av bilder ved å studere deres søkeadferd når de leter etter bilder i deres vanlige hverdag. Deres funn støtter Armitage og Enser ved at journalister som brukergruppe hadde stort behov for å finne bilder av navngitte personer og steder og generiske ting på den ene siden, og abstrakte temaer som følelser og atmosfære på den andre siden. Choi og Rasmussen (gjengitt etter Tsai 2006) har også undersøkt brukeratferd i bildebaser og de fant også at de fleste brukerne søkte på det konseptuelle nivået (tilsvarende Shatfords pre-ikonografiske og ikonografiske nivå). Dette kan tyde på at selv om innholdsbasert indeksering har mye for seg når det gjelder å finne liknende bilder, ansiktsgjenkjennelse og innenfor lignende bruksområder, er det viktig å gi brukerne en tekstlig inngang ved vanlige bildegjenfinning.

3.5 Dette prosjektet sett i lys av tidligere forskning

I den grad noen tidligere har forsøkt å bruke automatiske metoder for tekstlig indeksering, har det som regel vært i kombinasjon med innholdsbaserte metoder, og det finnes få forsøk hvor det er den tekstbaserte indekseringen som er den viktigste. De forsøkene som har brukt tekstbasert indeksering av bilder, har i all hovedsak brukt datamaterialer som består av html-filer fordi det er den typen multimodale data som har vært tilgjengelig. Med den økte satsningen på digitalisering av bøker har det kommet en ny type data som har både tekst og bilder. Å indeksere bilder i digitale bøker gjennom teknikker utviklet for html-dokumenter vil være problematisk av flere grunner. Fordi dokumentene er ulikt oppbygd vil enkelte av teknikkene ikke være overførbare, og samtidig vil teknikker utviklet for html-dokumenter ikke ta hensyn til spesielle aspekter ved de digitale bøkene som kan utnyttes for å gi en bedre indeksering. Digitale bøker skiller seg fra

html-dokumenter på flere måter, men særlig viktig er den markante forskjellen i markeringsstrukturen i de ulike dokumenttypene.

I html-dokumenter grupperes innhold sammen etter hva slags type tekst eller innhold som befinner seg mellom taggene. Det finnes tagger for overskrifter på forskjellige nivå, avsnitt, bilder, tabeller osv. I Figur 4 er det et utdrag av html-koden til en bok fra Gutenberg-prosjektet. Her er et bilde tagget med bildetaggen <img>, avsnittet er kodet med avsnitt-taggen <p> og overskrifter er kodet med taggene <h2> og <h3>, som sier noe om det hierarkiske forholdet mellom overskriftene. I tillegg er det i dette tilfellet brukt et class-attributt definert som *caption* for å markere bildeteksten. Taggene i html, slik som <p>, <h2> osv. er felles for alle html-filer. Class-attributt er det opptil den enkelte forfatter å definere, og det er derfor ikke gitt at alle bilder vil ha en slik markering av hva som er bildetekst. Mange av studiene anser derfor <img>-taggens alt-attributt for å være bildets bildetekst (Frankel, Swain og Athitsos 1996; Chen 1999; Mukherjea og Cho 1999). I eksempelet i Figur 4 er innholdet i dette attributtet ”HENRY IN MORNING DRESS, WITH GREAT HAT”. Markeringen i html følger den logiske strukturen i et dokument. Med logisk struktur menes struktur som følger innhold og ikke form, med andre ord sier det noe om hva slags innhold det er snakk om, og ikke hvordan innholdet ser ut. Faktisk vil hvordan html-filen se ut, være avhengig av innstillinger hos hver enkelt leser, som hvilken størrelse skriften skal ha og hvilken nettleser siden vises i. Innenfor avsnittene som er markert med <p> vil linjeskiftene tilpasse seg disse innstillingene.

```
<p class="figcenter"><a name="illus006"></a>
<img src="./images/illus006.jpg"
alt="HENRY IN MORNING DRESS, WITH GREAT HAT."
title="HENRY IN MORNING DRESS, WITH GREAT HAT." /></p>
<p class="figcenter caption">HENRY IN MORNING DRESS, WITH GREAT HAT.</p>
<p>If Henry, who in the last six years had only once left Sagres, to knight
Don Pedro's eldest son at Coimbra<span class='pagenum'><a name="Page_259"
id="Page_259">[Pg 259]</a></span> in 1445, had now been able, in
presence as well as writing, to stand by his brother in this crisis, the
Regent might have been saved
[…]
</p>
[…]

<h2><a name="CHAPTER_XVII" id="CHAPTER_XVII"></a>CHAPTER XVII.</h2>

<h3>CADAMOSTO.</h3>
```

Figur 4. Eksempel på html-koding. Fra Beazley (2006)

I motsetning til html-dokumentene, som er designet for å være nettdokumenter, er de digitale bøkene en gjengivelse av de fysiske eksemplarene av bøkene. Derfor gjengir markeringen i de digitale filene hvordan de ulike elementene av teksten og det øvrige innholdet er plassert i forhold til hverandre. Kodingen gjenspeiler bokens fysiske struktur og taggene er delt inn etter hvordan de ulike elementene er inndelt på siden, for eksempel i blokker <TextBlock >, linjer <TextLine> og ord <String />. Bildene er markert med taggen <ComposedBlock>. Hver av disse taggene har attributter som identifiserer hvor på siden det aktuelle elementet befinner seg og hvilken utstrekning det har.

Mens teknikker for indeksering av bilder i html-filer kan bruke den logiske strukturen i dokumentene til å identifisere hva som er bildetekst og hva som er overskrifter, må en tilsvarende indeksering av bilder i digitale bøker fokusere på den fysiske strukturen. På grunn av dette kan ikke de teknikkene som finnes for indeksering av bilder i html-dokumenter direkte overføres til et materiale bestående av digitale bøker. Det er derfor nødvendig å utvikle nye teknikker som tar hensyn til egenskapene og strukturen til de digitale bøkene.

```
<TextLine ID="P81_TL00004" HPOS="57" VPOS="2046" WIDTH="732" HEIGHT="34">
<String ID="P81_ST00037" HPOS="57" VPOS="2046" WIDTH="66" HEIGHT="34" CONTENT="jakt." WC="0.97"
CC="20000"/>
<SP ID="P81_SP00033" HPOS="123" VPOS="2080" WIDTH="14"/>
<String ID="P81_ST00038" HPOS="137" VPOS="2048" WIDTH="76" HEIGHT="26" CONTENT="Foto:" WC="0.99"
CC="00001"/>
<SP ID="P81_SP00034" HPOS="213" VPOS="2080" WIDTH="14"/>
<String ID="P81_ST00039" HPOS="227" VPOS="2046" WIDTH="114" HEIGHT="34" CONTENT="Fridtjof" WC="1.00"
CC="00000000"/>
<SP ID="P81_SP00035" HPOS="341" VPOS="2080" WIDTH="11"/>
<String ID="P81_ST00040" HPOS="352" VPOS="2049" WIDTH="120" HEIGHT="31" CONTENT="Nansen," WC="0.99"
CC="0000001"/>
<SP ID="P81_SP00036" HPOS="472" VPOS="2080" WIDTH="13"/>
<String ID="P81_ST00041" HPOS="485" VPOS="2048" WIDTH="304" HEIGHT="32" CONTENT="Nasjonalbiblioteket."
WC="0.99" CC="01000010000000000000"/>
</TextLine>
</TextBlock>
<ComposedBlock ID="P81_CB00001" HPOS="67" VPOS="236" WIDTH="1330" HEIGHT="710" STYLEREFS="TXT_2
PAR_LEFT" TYPE="Illustration"><GraphicalElement ID="P81_CB00001_SUB" HPOS="67" VPOS="236" WIDTH="1330"
HEIGHT="710"/>
```

Figur 5. Eksempel på koding av digitale bøker i alto xml. Fra Solli (2002)

I denne oppgaven brukes informasjon om den fysiske strukturen i bøkene til å identifisere hva som er et bildes bildetekst. I tillegg brukes semantisk analyse for å identifisere egennavn og årstall i bildeteksten. Selv om semantisk analyse i denne oppgaven bare brukes på tekst som allerede er identifisert som bildetekst, kan det være nyttig å utforske bruken av semantisk analyse videre i forbindelse med bildeindeksering i digitale bøker. Fordi man kan utnytte den logiske strukturen til dokumentene når man indekserer bilder i html-dokumenter, er semantisk analyse av teksten lite utbredt i forskningen. Rowe (Rowe og Guglielmo, 1992; Rowe, 1993; Rowe og Frew, 1998) er en av de få som argumenterer for bruken av semantisk analyse. Når det gjelder å bruke digitale bøker som datamateriale, har man ikke i samme grad muligheten til å bruke logisk struktur, og semantisk analyse kan derfor være en retning som det kan være nyttig å utforske i forhold til å finne den teksten som vil være mest relevant som grunnlag for bildeindekseringen.

4 Grunnpremisser for undersøkelsene

Formålet med denne oppgaven er å undersøke hvilket potensial som ligger i å indeksere bilder i digitale bøker gjennom automatisk tekstlig indeksering. Oppgaven må sees på som et utforskende bidrag, som så vidt skraper i overflaten på de mulighetene som eksisterer innenfor dette temaet. I denne oppgaven er det fokusert på noen få muligheter som presenteres og diskuteres i forhold til fordeler, ulemper og utfordringer knyttet til tilnærmingen, basert på de indikasjonene de utførte undersøkelsene har gitt. De valgene som er foretatt i forhold til oppgavens fokus diskuteres i avsnittene under.

4.1 Valg av indekseringsnivå

Formålet når man indekserer bilder må være å gjøre bildene tilgjengelige og gjenfinnbare for brukere. De ord og begrep som brukes til indekseringen, må derfor være ord og begrep som det kan tenkes at brukerne vil finne på å søke på, for å finne akkurat det bildet. I teorikapittelet (kapittel 2) ble det lagt frem teorier for hvordan bilder beskrives og på hvilke nivå beskrivelsene kan defineres, fra innholdselementer som farge og form til abstrakte konseptuelle begrep som følelser og stemninger, og i gjennomgangen av tidligere forskning ble det presentert undersøkelser som viste at brukere som regel søker på de høyere nivåene, som navngitte personer og steder, generiske objekter eller abstrakte begrep. Som nevnt tidligere, er Shatford opptatt av hva slags innhold det er viktig å beskrive når man indekserer bilder (kapittel 2.1). I forhold til

automatisk tekstbasert indeksering er det to sider av dette som det er viktig å ha et bevisst forhold til: hva slags innhold er det ønskelig å beskrive, og hva slags beskrivelse er det mulig å finne i teksten? I forhold til manuell indeksering mener Shatford at alle nivåene hun beskriver (pre-ikonografisk, ikonografisk og ikonologisk) bør være med der det passer, for å få en best mulig beskrivelse av bildene med tanke på gjenfinning. Formålet med dette er å favne flest mulige nyttige innganger, og forhindre at mulige relevante emner blir oversett (Jörgensen 2003). Selv om Shatford var fokusert på manuell indeksering, gir nivåene hennes noen holdepunkter også i forhold til automatisk indeksering i forhold til hva slags informasjon det er ønskelig å kunne trekke ut fra teksten.

Når man snakker om tekstlig indeksering vil lavnivå-egenskaper gjerne oppfattes som mindre interessante. Egenskaper som form og retning er egenskaper som kan brukes i gjenfinningsalgoritmer, men som brukere sjelden søker på tekstlig. Det er selvsagt mulig å argumentere for at det kan være nyttig å representere visse lavnivå-egenskaper også i tekstlig form, for eksempel farger, men her vurderes disse som mindre viktige og velges derfor bort. Når det gjelder Shatfords øverste nivå ikonologi, kan det diskuteres hvorvidt det er ønskelig å inkludere denne typen beskrivelse. Spørsmålet er om det er viktigere å indeksere bildet ut ifra alle mulige beskrivelser som kan oppfattes som nyttig eller er det viktigere å indeksere bildet med beskrivelser som alle kan enes om. Shatfords ikonologi er det nivået som er mest utsatt for subjektive vurderinger i forhold til relevans versus støy. Igjen er dette en debatt som hovedsakelig er interessant knyttet til manuell indeksering, men problemstillingen kan også angå automatiske indekseringssystem. For at et automatisk indekseringssystem skal kunne inkludere ikonologiske beskrivelser, er det avhengig av at de finnes i den eksisterende teksten. Hvorvidt dette vil være tilfellet, avhenger av forholdet mellom bildene og teksten i hver bok. Hvis man ser på taksonomien til Marsh, så finnes det tilfeller der teksten kan representere en ikonologisk beskrivelse. Et eksempel vil være Marshs funksjon B1.1 *Concretize* som Marsh definerer som ”mak[ing] explicit a textual reference to a thing or abstract concept” (2002 s. 161). Også kategoriene hvor bildets innhold går utover tekstens innhold, vil kunne gjøre at teksten gir bildet en ikonologisk beskrivelse. Marsh (2002 s.194) skriver:”In general the modeling codes function as metaphors. (…) The purpose of metaphor is to make new experiences and ideas more understandable by expressing them in terms of what is already known”. Bilder og tekst kan ha et gjensidig forhold av konkretisering og fortolkning. Tekst kan tolke bilder og bilder kan tolke

tekst. I disse tilfellene kan man tenke seg at teksten vil kunne ha en ikonologisk beskrivende funksjon i forhold til bildet. Det er derfor et tema som bør diskuteres også i forbindelse med automatisk tekstlig indeksering. Hvis et automatisk system finner ikonologiske beskrivelser, er det da interessant å inkludere dette i indekseringen, og er det mulig å identifisere beskrivelsen som ikonologisk automatisk, slik at man kan behandle beskrivelsen slik det er ønskelig? Denne diskusjonen faller litt utenfor denne oppgavens fokus, men det er spørsmål det kan være viktig å være klar over hvis noen skulle ta opp tråden fra denne oppgaven og utvikle automatisk tekstbasert indeksering ytterligere.

Til syvende og sist tyder det på at det er Shatfords nivå pre-ikonografi og ikonografi som er de klareste kandidatene til hvilke elementer det er ønskelig å beskrive bilder med gjennom automatisk tekstbasert indeksering. Dette understøttes også av brukerstudiene som viser at det er disse nivåene brukere som regel søker etter, og som det derfor er viktig å inkludere i bildebeskrivelsene (Markkula og Sormunen 1998).

Når man har slått fast at det er ønskelig å finne beskrivelser på Shatfords pre-ikonografiske og ikonografiske nivå, gjenstår spørsmålet om slike beskrivelser kan finnes i teksten, og i så fall hvordan man skal gå frem for å finne disse beskrivelsene.

4.2 Valg av indekseringskilde

Teksten i en digital bok kan deles inn i flere kategorier etter hvilken funksjon de har i teksten. Tre gode kandidater til å være indekseringsgrunnlag er bildetekst, brødtekst og overskrifter. Andre kandidater vil for eksempel være indekser og innholdsfortegnelser. Barthes understrekte at hvor nært eller fjernt en tekst var fra bildet, ville påvirke hvordan bildet ble påvirket av teksten (se kapittel 2.2.1). Det ville være mer sannsynlig at tekst nær bildet, slik som en bildetekst, understrekte denotasjonene i bildet fremfor å tilføre bildet konnotasjoner.

Hvor tett knyttet brødteksten er til et bilde vil også avhenge av bildets funksjon. Dette gjelder i forhold til relevans, og om man kan forvente å finne en beskrivelse av bildet i teksten, men man kan også forestille seg at bildets funksjon påvirker plasseringen av bildet i forhold til teksten. Enkelte av funksjonstypene kan kreve en tettere fysisk kobling mellom tekst og bilde enn andre. Hvis teksten for eksempel er en analyse av et maleri, vil det være formålstjenlig at maleriet enten er plassert fysisk nær den analyserende teksten, eller at det i teksten forekommer en referanse til

bildet det snakkes om. Andre bildefunksjoner krever kanskje ikke en like klar fysisk kobling mellom bilde og tekst, og det kan gjøre det vanskeligere å identifisere hvilke avsnitt som er sterkest knyttet til et bilde. I eksisterende forskning er ikke brødteksten brukt til indeksering i veldig utstrakt grad. Rowe (1992; 1993; 1998) er som sagt en av de få som har konsentrert seg om denne typen tekst.

Mens forholdet mellom brødtekst og et bilde vil være avhengig av hvilken funksjon bildet har i forhold til teksten, så vil bildetekst per definisjon ha en direkte tilknytning til bildet. På engelsk heter bildetekst *caption*, og Merriam-Webster definerer *caption* som "the explanatory comment or designation accompanying a pictorial illustration" ("caption" 2009). Vi forbinder en bildetekst med en tekst som beskriver, forklarer eller tilføyer informasjon til et bilde. Frankel, Swain, Athitsos (1996), Rowe og Frew (1998), Srihari og Zhang (1999), Paek et al. (1999), Mukherjea og Cho(1999), Barnard et al. (2003) og Maderlechner, Panyr og Suda (2006) bruker alle på en eller annen måte bildetekster i indeksering eller gjenfinning av bilder.

4.3 Evaluering

Undersøkelsene i denne oppgaven er delt i to: først prøves en metode for å identifisere bildetekstene i bøkene, og deretter blir det undersøkt om det er mulig å identifisere egennavn i bildetekstene, og hvorvidt disse kan brukes til å indeksere bildene. Dette blir sett på som to ulike undersøkelser og de vurderes derfor hver for seg i forhold til hvor holdbare metodene er. I tillegg blir det forsøkt å si noe om hvorvidt de identifiserte indekstermene kan brukes til å beskrive bildene.

For den første undersøkelsen ble det vurdert hvor holdbar metoden for å finne bildetekster i bøkene er. I denne undersøkelsen bestod datamaterialet av 10 bøker. For hver av bøkene ble avsnitt i teksten vurdert som mulige bildetekstkandidater, og basert på vektingen, ble de fire øverst rangerte avsnittene hentet og koblet til den bildeblokken de var vurdert mot. Vurderingen foregikk ved at det ble telt opp hvor mange av de manuelt identifiserte bildetekstene i bøkene som ble funnet igjen i én av de fire avsnittene som var hentet automatisk. I tillegg ble det telt opp hvor mange av bildetekstene som ble gjenfunnet i de forskjellige rangeringsposisjonene.

Den andre undersøkelsen gikk ut på å undersøke, gitt at man var i stand til å finne bildeteksten, om det var mulig å identifisere egennavn i bildeteksten, og hvorvidt disse egennavnene ville være

gode indekstermer for bildene. Derfor ble datamaterialet som ble brukt i vurderingen, plukket blant de bøkene som oppnådde en god gjenfinningsprosent i den første undersøkelsen. I denne delen ble det bestemt å bruke målene fullstendighet og presisjon for å vurdere holdbarheten til metoden for å identifisere egennavn, og egennavnenes egnethet som indekstermer for bildene. Funksjonen for å måle presisjon for et gjenfinningssystem er som regel antall relevante dokumenter som ble gjenfunnet av systemet, delt på antall gjenfunne dokumenter. Overført til denne undersøkelsen ble presisjonen til navnegjenkjenningsfunksjonen vurdert som antall korrekt identifiserte egennavn delt på det antallet termer som ble foreslått som egennavn. Hvis metoden foreslo at det fantes 4 personnavn i teksten, mens egentlig bare 3 av disse var personnavn, ville presisjonen bli ¾ =0,75. Tilsvarende er funksjonen for å måle fullstendighet for et gjenfinningssystem som regel antall gjenfunne relevante dokumenter delt på antall relevante dokumenter i samlingen. Overført til denne undersøkelsen ble fullstendighet målt ut i fra antall korrekt identifiserte egennavn delt på antall egennavn som fantes i teksten. Hvis det for eksempel fantes 5 egennavn i teksten, og metoden kun identifiserte 3 av disse, ville fullstendigheten bli 3/5=0,6.

I tillegg ble det målt presisjon og fullstendighet for å vurdere hvorvidt indekseringstermene kunne brukes til å indeksere bildet. Presisjonen ble målt ved å dele antall korrekte indekstermer delt på antall indekstermer som var foreslått. Et portrett som forestilte Fridtjof Nansen, men som hadde fått tildelt indekstermene ”Fridtjof Nansen” og ”Eva Nansen”, ville få en presisjon på ½=0,5. Korrekt indeksterm ble vurdert som en term som var direkte relatert til det som var avbildet på bildet. Et personnavn måtte derfor tilhøre en av personene på bildet, et stedsnavn måtte være navnet på stedet bildet var tatt, og et årstall måtte tilsvare det året bildet var blitt tatt. Fullstendigheten ble målt ved å dele antall korrekte indekstermer på antall mulige indekstermer av samme type. Hvis et bilde viste tre personer, men kun to av dem var blitt identifisert med indekstermer, ville fullstendigheten blir 2/3=0,67.

Det finnes metodiske problem ved å bruke presisjon og fullstendighet i forhold til indeksering på denne måten. I forhold til å identifisere egennavn i teksten, vil disse målene fungere bra, fordi det er enkelt å definere hva som er personnavn i teksten og som derfor bør gjenfinnes. Men i forhold til å avgjøre hvordan et bilde bør beskrives, er det vanskeligere å definere en klar fasit. I hvilke tilfeller bør man forvente at personer skal indekseres ved navn? Ved portrett og klare

gruppebilder er dette forholdsvis enkelt, fordi ved slike bilder bør alle personer indekseres med navn. Men hva med bilder hvor det er tydelig at det er personer på bildet, men det er på en slik avstand at det ikke er mulig å identifisere hvem det er? Skal navnene på disse personene regnes med i tallene som er basis for fullstendighetsmål? Selv om det ikke er lett å definere noen fasit for bildeindeksering, er det allikevel noe enklere å lage en fasit for egennavn som indekseringstermer enn det ville vært for andre, mer generelle termer, hvor det ofte er uenighet blant indekserere om hvilke termer som regnes som relevante og ikke. I denne undersøkelsen er fasiten basert på hvilke navngitte elementer av de ulike fasettypene det ville være mulig å identifisere ved en manuell indeksering. Nærmere kriterier for fasitbestemmelsene er beskrevet i kapittel 7.3.1.

I undersøkelsene er det brukt veldig små utvalg. De vurderingene som er foretatt av metodene, er derfor ikke representative i forhold til hvordan metodene ville ha gjort det for større utvalg og for andre typer bøker. De vurderingsmetodene som er beskrevet er ressurskrevende metoder fordi det for alle disse metodene er nødvendig å manuelt utarbeide en fasit som det automatiske resultatet kan måles mot. Det var derfor innen rammene av denne oppgaven nødvendig å bruke små utvalg, for å kunne ha en mulighet til å vurdere indikasjoner ved resultatene, selv om disse indikasjonene ikke ville være representative for et større utvalg. I den andre undersøkelsen er i tillegg datamaterialet ikke tilfeldig utvalgt, men plukket i forhold til hvilke kvaliteter boken hadde. Dette kan ha påvirket resultatene noe, noe som diskuteres grundigere i kapittel 7.4.

5 Datamaterialet

Til denne undersøkelsen har Nasjonalbiblioteket gitt tilgang til et utvalg digitale bøker fra deres Nordområdesamling. Bøkene i denne samlingen er tilgjengeliggjort gjennom en avtale mellom Den norske forleggerforening, Den norske Forfatterforening, Norsk faglitterær forfatter- og oversetterforening, Norsk kritikerlag, forvaltningsorganisasjonen LINO og Nasjonalbiblioteket. Avtalen innebærer at rettighetsbelagt materiale innenfor emnet nordområdene blir gjort tilgjengelig i digital fulltekstversjon. Denne avtalen gjør at det i denne undersøkelsens datamateriale finnes bøker som er utgitt så sent som i 2006.

Et utvalg på 101 bøker ble funnet ved å velge bort alle bøker i samlingen som ikke var registrert som illustrerte, og som ikke var på norsk. For å kunne undersøke automatisk bildeindeksering må

materialet naturlig nok ha bilder som kan indekseres, og fordi indekseringen benytter seg av tilknyttet tekst, ville det være en fordel om bøkene var på samme språk. Selv om alle bøkene er på norsk er det fremdeles noen ulikheter i bøkenes språkform: 86 av bøkene er skrevet på bokmål, 4 er skrevet på nynorsk og 2 er skrevet på både bokmål og nynorsk. 9 av bøkene mangler informasjon om hvilken språkform som er brukt.

Bøkene er utgitt mellom 1964 og 2006, 50 % av utvalget er gitt ut etter 1996. De dekker ulike fagområder inkludert historie, kulturhistorie, reiselitteratur, biografier, friluftsliv og biologi, og er skrevet for ulike brukergrupper, fra populærvitenskap til offisielle rapporter. Bøkene har et omfang fra 7 til 600 sider, med et gjennomsnitt på 180 sider. Rundt halvparten av bøkene har under 75 bilder, gjennomsnittet er 85 bilder. Rundt 20 prosent av bøkene har under 50 bilder, mens rundt 13 prosent har mer enn 150 bilder. Totalt er det 6455 bilder i utvalget. Selv om samlingen inneholder 101 bøker er undersøkelsene utført på små samlinger på rundt 10 bøker. Resultatene er derfor ikke generaliserbare, men gir en pekepinn på hvilke potensialer som ligger i å bruke tekst som grunnlag for automatisk indeksering av bilder i digitale bøker.

De digitale bøkene er kodet med alto xml, mens metadatafilene er kodet i mets/mods. Alto er et markeringsspråk for ocr-scannede dokumenter, som også brukes av Library of Congress. Det er designet for å være en utvidelse av mets. Mets er Library of Congress sitt markeringsspråk for metadata i digitale dokument. Generelt kan man si at mets blir brukt for å beskrive metadata og strukturell informasjon, mens alto blir brukt for å beskrive fysisk struktur og innhold.

5.1 Forberedende behandling

For å kunne bruke teksten i de digitale bøkene som grunnlag, er det essensielt at den scannede teksten er korrekt scannet og tilsvarer virkelig ord. I de digitale bøkene fra Nasjonalbiblioteket var det generelt flere feilscanninger som gjorde det vanskelig å arbeide med dokumentene. Det var derfor nødvendig å bruke tid på å renske opp i kodingen slik at den scannede teksten stemte mer overens med den faktiske teksten i kildedokumentet.

I mellomtiden gjorde Nansen det beste ut av det ufrivillige opp-holdet i Godthåb ved å studere eskimoene og deres kultur, jakt- og fangstmetoder og overlevelsesteknikk. Det skulle han få god nytte av seinere. Han fikk dessuten stor respekt for eksimo-kultu-ren, og mente det var en tragedie at den hvite mann presset sin vestlige, materialistiske kultur på dette selvforsynte fangtfolket.	I mellomtiden gjorde Ilanßen 6et deßte ut av 6et utrivillige opp nol6et i (io6tnad ve 6 a Btu6ere eßkimoene og 6ereß kultur, jakt og tangßtmeto6er og overleveißeßteknikk. vet Bkulle nan ta go6 nytte av Beinere. tan tikk 6eßßuten Btor reßpekt tor eßkimo-kultu ren, og mente 6et var en tragedie at 6en nvite mann preßßet zin veßtlige, materialißtißke kultur pa dette Belvtorßvnte	I mellomtiden gjorde Ilansen det beste ut av det ufrivillige opp holdet i ve 6 å studere eskimoene og deres kultur, jakt og fangstmetoder og overlevelsesteknikk. vet skuffe han ta god nytte av Beinere. tan tikk dessuten stor respekt for eskimo-kultu ren, og mente det var en tragedie at den hvite mann presset gin høstlige, materialistiske kultur pa dette selvforsynte fangstfolket.

Figur 6. Opprinnelig tekst, ocr-scannet tekst og renset ocr-tekst

Rensingen foregikk ved at hvert ord som ble hentet fra ocr-dokumentet ble sammenlignet med ordlister av norske ord og navn. Hvis ordet ble identifisert på en av listene, ble det akseptert, hvis ikke ble det sendt gjennom et program for å teste for mulige feil. Programmet bestod av rundt 80 av de vanligste feilscanningene som var blitt identifisert ved en manuell gjennomgang av et utvalg eksempeldokumenter. Listen over mulige feilscanninger ble for enkelthetsskyld organisert alfabetisk etter riktig bokstav. Hvis for eksempel feilen var at bokstaven ”d” ble identifisert som et 6-tall, ville denne feilen havne under ”d*”* i listen. Noen av feilene var av en slik karakter at hvis tegnene dukket opp i teksten, ville det være svært sannsynlig at det var en feil, og den kunne rettes uten å måtte teste om det var en feil eller ikke. En vanlig feil, som ikke vises i Figur 6, var at bokstaven ”s” ble erstattet med en stor B. Hvis det i teksten ble oppdaget et ord som hadde en stor B midt i ordet, for eksempel ”NanBen”, så ville B-en bli endret til en s med en gang. I andre tilfeller var det vanskeligere å vite om tegnene var korrekt eller feil. For eksempel ville ofte bokstavene ”n” og ”h” byttes om i scanningen. I disse tilfellene kunne man ikke anta at det var en feilscannet ”h” hver gang man traff på en ”n” i teksten. Når slike tegn ble funnet ble tegnet byttet ut med det mulige erstatningstegnet og testet for om det nye ordet fantes på ordlistene. Hvis så var tilfellet, ville det nye ordet erstatte det opprinnelige ordet, hvis ikke ville det opprinnelige ordet beholdes. Men fordi hvert ord kan inneholde flere feilscanninger, som i ordet ”6eββuten” i Figur 6, var det ikke mulig å forkaste det nye ordet med en gang. 6-tallet kunne for eksempel korrekt endres til en ”d”, men det nye ordet ville ikke finnes på noen liste fordi s-ene fremdeles

var feilkodet som β. Hvert nye ord som ikke ble anerkjent som et korrekt ord med en gang, måtte derfor sendes gjennom rensingsprogrammet i sin nye form, slik at eventuelle andre feilscanninger kunne rettes opp. Denne teknikken gjør at programmet kunne kjørt i uendelig tid uten å finne et korrekt ord. Kombinasjonsmulighetene for ord med mange tegn som er identifisert som mulige feilkodinger, er enorme. Det ble derfor besluttet å sette en grense på opptil fire feilkodinger per ord. Hvis det ikke var funnet et ord som stod på en av ordlistene etter fire erstatninger, ville erstatningen bli forkastet og det opprinnelige ordet ble beholdt og sjekket mot neste erstatningsmulighet på listen.

Denne måten å rense teksten på åpner for feiltolkninger. I Figur 6 er for eksempelet ordet ”veβtlige” i rensingen feiltolket til å være ordet ”høstlige”. Dette er på grunn av den alfabetiske ordningen av de mulige feilene og fordi renseprogrammet blir avsluttet i det øyeblikket et nytt ord finnes på en av ordlistene. En vanlig feilscanning var at bokstaven ”h” ble scannet som ”v”. Derfor testes ordet ”veβtlige” mot ”heβtlige”, og deretter blir ”heβtlige” testet mot andre feil til resultatet blir ”høstlige”, som finnes i ordlisten. Det finnes flere måter dette kunne vært unngått på. En måte er å enten teste hvert ord mot én og én endring, og kun sjekke for flere feil hvis den første gjennomgangen ikke resulterte i et korrekt ord. Hvis det er mulig å identifisere et ord med bare én erstatning, ville dette sannsynligvis være nærmere det opprinnelige ordet enn et ord som ble funnet gjennom å erstatte flere av tegnene. En annen måte en slik feil kunne vært unngått på ville vært å lagre mulige identifiserte kandidatord i en liste, og deretter velge det kandidatordet som ligner mest på det ocr-scannede ordet, slik at det ordet som trengte færrest endringer for å bli anerkjent som ord, blir valgt som erstatningsord.

Å rense teksten var ikke en prioritet i oppgaven, men fordi utgangspunktet var så dårlig, var det helt nødvendig å gjøre noe. Selve renskningen kunne derfor vært gjort bedre hvis dette hadde vært en prioritet, for eksempel ved å bruke et av de forslagene som er presentert i det forrige avsnittet. Som man kan se i Figur 6 er den endelige teksten fremdeles ikke så korrekt som den burde være, og nye feil har kommet til, men det endelige resultatet representerer en betydelig forbedring i forhold til utgangspunktet, og opprenskningen bidro til en betydelig forbedring av undersøkelsenes resultat.

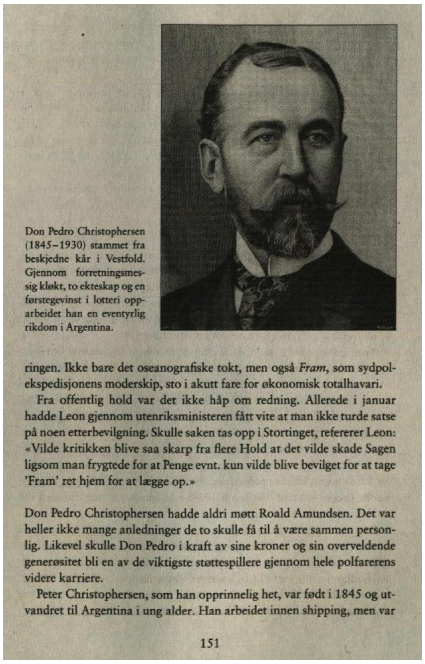
6 Identifisering av bildetekst

Av de undersøkelsene som ble nevnt blant relatert forskning, og som har brukt bildetekster i indekseringen, har de fleste brukt html-dokumenter, og de har som regel fokusert på nærhet i koding og tekst i bildets alt-attributt, noe som ikke vil være overførbart til et materiale bestående av digitale bøker som er kodet annerledes enn html-dokumenter. Maderlechner, Panyr og Suda (2006) lokaliserte bildetekster som var posisjonert enten over eller under bilder i pdf-dokumenter på nettet, og deres metode for å identifisere mulige bildetekster var å analysere sidens layout, bildene og tekstblokkenes størrelse og geometrisk plassering på siden. Ideen vil være overførbar fra pdf-dokumenter til digitale bøker, men fremgangsmåten vil ikke være overførbar siden et pdf-dokument er et bilde av en side, mens sidene i de digitale bøkene er kodet med xml. I litteratursøkene til denne oppgaven ble det ikke funnet undersøkelser som har fokusert på å finne bildetekster i digitale bøker. Det var derfor heller ikke funnet noen teknikker som var direkte relaterbare til denne undersøkelsen. Fremgangsmåten som blir presentert under er derfor utarbeidet til denne oppgaven og presenteres som et forslag til hvordan man kan identifisere bildetekster i digitale bøker.

6.1 Grunnantakelser angående bildetekst

I utviklingen av metoden for å finne bildetekstene i bøkene er det gått ut ifra noen grunnantakelser om egenskaper ved bildetekster.

1) *Bildetekster vil være posisjonert nær bildet teksten skal beskrive*
   For at man skal oppfatte hvilken bildetekst som hører til hvilket bilde, vil bildetekster som regel bli plassert rett under, rett over, eller rett ved siden av bildet de hører til. Det finnes unntak, for eksempel hender det at bildetekstene plasseres i margen mens teksten fyller hovedsiden, og en bildetekst kan dermed havne i venstre marg mens bildet er på neste side, og nærhetskriteriet vil dermed ikke være oppfylt. Men antakelsen i denne undersøkelsen er at majoriteten av bildetekster vil være plassert i umiddelbar nærhet av bildet.



Don Pedro Christophersen (1845–1930) stammet fra beskjedne kår i Vestfold. Gjennom forretningsmessig kløkt, to ekteskap og en førstegevinst i lotteri opparbeidet han en eventyrlig rikdom i Argentina.

ringen. Ikke bare det oseanografiske tokt, men også *Fram*, som sydpolekspedisjonens moderskip, sto i akutt fare for økonomisk totalhavari.

Fra offentlig hold var det ikke håp om redning. Allerede i januar hadde Leon gjennom utenriksministeren fått vite at man ikke turde satse på noen etterbevilgning. Skulle saken tas opp i Stortinget, refererer Leon: «Vilde kritikken blive saa skarp fra flere Hold at det vilde skade Sagen ligsom man frygtede for at Penge evnt. kun vilde blive bevilget for at tage 'Fram' ret hjem for at lægge op.»

Don Pedro Christophersen hadde aldri møtt Roald Amundsen. Det var heller ikke mange anledninger de to skulle få til å være sammen personlig. Likevel skulle Don Pedro i kraft av sine kroner og sin overveldende generøsitet bli en av de viktigste støttespillere gjennom hele polfarerens videre karriere.

Peter Christophersen, som han opprinnelig het, var født i 1845 og utvandret til Argentina i ung alder. Han arbeidet innen shipping, men var

151

Figur 7. Eksempelside: bildetekst. Fra Bomann-Larsen (1995)

2) *Bildetekster vil ofte være omgitt av mer luft enn annen tekst*

   Det skal være lett for en leser å kunne raskt identifisere hva som er en bildetekst. Det er derfor nødvendig å gi bildeteksten et annet utseende enn teksten rundt. En måte å gjøre dette på er å separere bildeteksten og omgi den med mer luft enn det annen tekst har. På Figur 7 er det lett å se at bildeteksten i dette tilfellet har mer luft både over og under teksten enn det de vanlige avsnittene har.

3) *Bildetekst vil ofte ha en annen formatering enn teksten rundt*

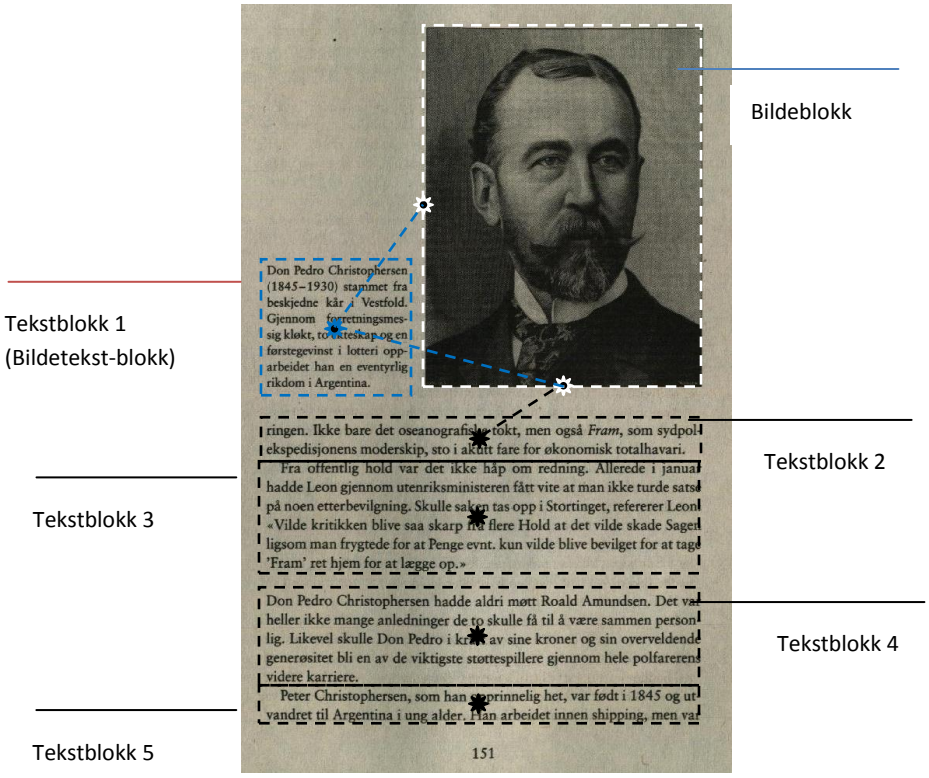
   En annen måte å hjelpe leseren med å identifisere bildetekst er å gi den en annen formatering enn den øvrige teksten, noe som signaliserer at teksten har en annen funksjon. Det er vanlig at bildeteksten har en mindre fontstørrelse (som på Figur 7), og av og til har den en annen skriftstil som for eksempel kursiv eller fet skrift.

4) *Bildetekst vil ofte være en kortere tekst enn avsnittene i resten av teksten*

   Bildeteksten skal ofte være en kort beskrivelse eller kommentar til bildet og er sjelden like lang som vanlige avsnitt. I eksempelet på bilde 1 har hvert av de to hele avsnittene på siden dobbelt så mange ord som bildeteksten. Også her finnes det unntak, særlig ved

gruppebilder med mange mennesker hvor alle menneskene skal identifiseres i bildeteksten.

6.2 Fremgangsmåte



Figur 8. Eksempelside: avstand tekstblokker – bilde. Fra Bomann-Larsen (1995)

6.2.1 Avstand mellom tekst og bilde

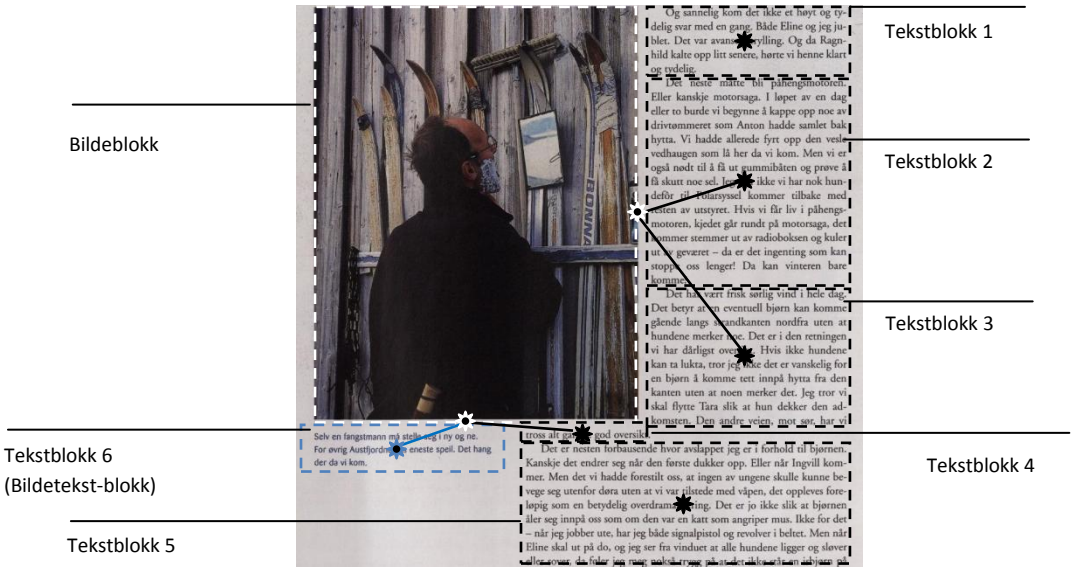
I digitaliseringsprosessen blir innholdet på hver side kodet som blokker med koordinater som beskriver hvor på siden de befinner seg. Hvert bilde blir en blokk, og teksten lagres i flere blokker som vist på Figur 8. For å finne avstanden mellom bildeblokken og hver av tekstblokkene, ble det valgt å måle avstanden mellom midtpunktet av bildeblokkens kanter og midtpunktet av hver av tekstblokkene ved å bruke Pythagoras' læresetning:

$$avs_{ji_k} = \sqrt{\left(hTB_j - hBB_{i_k}\right)^2 + \left(vTB_j - vBB_{i_k}\right)^2}$$

hvor *hTBj* er den horisontale koordinaten for tekstblokkens midtpunkt, og *vTBj* er den vertikale koordinaten for tekstblokkens midtpunkt. Tilsvarende er $hBBi_k$ den horisontale koordinaten for midtpunktet til bildeblokkens sidekant *k*, og $vBBi_k$ er den vertikale koordinaten for det samme

midtpunktet. Avstanden ble målt for hver av bildeblokkens fire sider, og det målet som var minst for hver tekstblokk, ble brukt som avstandsmål.

Ved å måle avstand fra midtpunkt til sidekant kan man differensiere mellom tekster hvor mye av teksten er plassert nært bildet, og tekster hvor kun starten eller slutten på en tekst er plassert nær bildet.



Figur 9. Eksempelside: avstand tekstblokker – bilde. Fra Aasheim (2003)

I eksempelet på Figur 9 har tekstblokkene 1, 2, 3 og 4 alle lik avstand til bildets høyre kant fra sin nærmeste blokkside som den egentlige bildeteksten, blokk 7, har til bildets nedre kantlinje fra sin nærmeste kantlinje. Men fordi bildeteksten er en kort tekst hvor så mye av teksten som mulig er plassert nær bildet, vil midtpunktet i tekstblokken være nærmere bildeblokkens kant enn det midtpunktet til tekstblokk 2 vil være. Ved å bruke avstanden som direkte utgangspunkt for vektingen, så vil tekstblokker med lav vekt ha større sannsynlighet for å være bildetekst enn tekstblokker med høy vekt.

Et alternativ som først ble utprøvd var å regne avstanden mellom midtpunktene i bilde- og tekstblokkene. Dette ga også gode resultater, men ikke så gode som å måle fra bildets sider. En mulig grunn til dette kan være at bildets fasong kan påvirke hvor bildets midtpunkt er. Illustrasjonen på siden i Figur 9 er i portrettformat, noe som gjør at tekst på siden av bildet i utgangspunktet vil være nærmere midtpunktet enn tekst under eller over, fordi avstanden mellom

midtpunktet og bildets øvre og nedre kant er lengre enn avstanden mellom midtpunktet og bildets høyre og venstre kant. Dette veies til en viss grad opp ved at tekstblokkene på høyre side er mye bredere enn bildeteksten er høy, fordi vanlig tekst ofte vil forsøke å fylle ut sidene, som på Figur 9. Dette vil påvirke tekstblokkens fasong, mens det at bildeteksten ofte vil følge bildets kant så mye som mulig, og dermed vil påvirkes av bildetekstblokkens fasong, vil gjøre at bildetekstens midtpunkt kommer nærmere bildet. Et annet alternativ som ikke ble utprøvd, ville vært å måle fra bildeblokkens sider til hver side av alle tekstblokkene. En ulempe ved en slik tilnærming ville vært at om et tekstavsnitt stod tett inntil bildet over bildet, og bildeteksten stod tilsvarende tett til bildet under bildet, ville disse tekstblokkene bli vurdert som like nær bildet, selv om tekstavsnittet startet langt fra bildet og bildeteksten stoppet nær bildet. Ved å bruke tekstblokkenes midtpunkt blir tekstavsnittets størrelse også en faktor i å avgjøre hvor nær bildet teksten er. Et mulig problem med denne tilnærmingen er det fenomenet som illustreres i Figur 9. Tekstblokk 4 tilhører avsnittet i tekstblokk 3, men fordi layouten er som den er, oppleves den som separat fra tekstblokk 3 og i skanningen har den blitt tolket som en egen tekstblokk. Dette gjør at halve avsnitt, som tekstblokk 4 i Figur 9 og 2 i Figur 8, i noen tilfeller, vil være nærmere bildeblokken enn det tekstblokk 6 er. Dette bidrar til at det ikke er mulig å bruke kun avstand til å identifisere bildetekstene.

6.2.2 Avstand mellom tekstblokker

Den andre antakelsen om bildetekst var at bildetekst er mer omgitt av luft enn andre tekstblokker, for at det skal være lett å identifisere teksten som en bildetekst. Det vil si at bildetekstens nærmeste nabotekstblokk bør være lengre unna enn det den nærmeste tekstblokken vil være for et vanlig avsnitt. I dette tilfellet er det interessant å måle avstand mellom kantlinjer og ikke midtpunkt, siden størrelsen på tekstblokken ikke har noe betydning for hvor nær den neste tekstblokken kan regnes å være. Spørsmålet er om tekstblokken tilhører en kontinuerlig tekst, eller om den er plassert med en viss avstand til annen tekst. Det enkleste målet vil da være å måle avstanden fra hver av tekstblokkens kanter til hver av kantene til tekstblokkene over og under. Den avstanden som er minst vil representere hvor nær tekstblokken er andre tekstblokker. Hvis det ikke finnes tekstblokker over eller under er det et tegn på at tekstblokken kan være en bildetekst, og tekstblokken vektes deretter.

6.2.3 Formatering

Den tredje antakelsen var at bildetekst har annen formatering enn annen tekst, og da særlig når det gjelder fontstørrelse. Ved å bruke fontverdien som en vekt, får man en vekt som ikke er lik for hele dokumentsamlingen fordi ulike bøker bruker ulike fontstørrelser, men dette vil ikke påvirke vektingen. Fordi det viktigste er å vekte avsnittene relativt til avsnittene i nærheten, vil det å bruke fontverdien som vekt oppnå det ønskede resultatet; nemlig å vekte avsnitt med mindre fontstørrelse mer fordelaktig enn tekst i nærheten som har større fontstørrelse.

6.2.4 Avsnittslengde

Den siste grunnantakelsen var at bildeteksten vil være et kortere avsnitt enn avsnittene i den kontinuerlige teksten. Det vil allikevel bli feil å gi mer fordelaktig vekting jo kortere et avsnitt er, for selv om bildetekster er kortere enn annen tekst, vil ikke en bildetekst på 45 ord være en bildetekst i mindre grad enn en bildetekst på 10 ord. Hvis det finnes to relativt korte avsnitter i nærheten av bildet, er det ikke gitt at det er den korteste som er bildeteksten. Men det vil være mer sannsynlig at et avsnitt på 20 ord er en bildetekst, enn at et avsnitt på 150 ord er en bildetekst, så det finnes en sammenheng mellom lengden på avsnittet og sannsynligheten for at teksten er en bildetekst. Med et estimat på hva som er lengden til den ”ideelle” bildetekst, vil man kunne fastslå hvor nær den ideelle bildeteksten et avsnitt er i lengde. Dermed kan man vekte slik at jo nærmere en tekst er det ideelle målet, jo mer fordelaktig vekt får avsnittet. Samtidig kan man tenke seg at det finnes et vindu rundt det ideelle målet, hvor avsnitt som har en tekstlengde som hører til i vinduet ikke bør differensieres for mye. For å ta et eksempel: hvis den ideelle avsnittslengden for en bildetekst regnes for å være 30 ord, så er det ikke nødvendigvis ønskelig å si at et avsnitt på 35 ord er mer sannsynlig en bildetekst enn et avsnitt på 20 ord. Hvis man sier at alle avsnitt som ligger nær den ideelle lengden har like stor sannsynlig for å være bildetekst, vil det være naturlig å ikke vekte disse veldig ulikt. Det er like sannsynlig at alle avsnittene i vinduet er en bildetekst basert på lengden, fordi de har en lengde som er typisk for bildetekster, og det å differensiere proporsjonalt med differensen fra det ideelle målet vil være å misrepresentere sannsynligheten for at tekstene er bildetekster. Men samtidig vil det utenfor vinduet til en grad være slik at jo fjernere en tekstblokks lengde er fra det ideelle målet, jo mindre sannsynlig er det at teksten er en bildetekst. Hvis det ikke er ønskelig å differensiere mye mellom avsnitt på for eksempel 20 og 35 ord, kan det være desto mer ønskelig å differensiere disse avsnittene fra

avsnitt på 75 ord. Det er også et poeng å merke seg at om en tekst har 200 eller 300 ord, så er det tilnærmet like lite sannsynlig at teksten er en bildetekst.

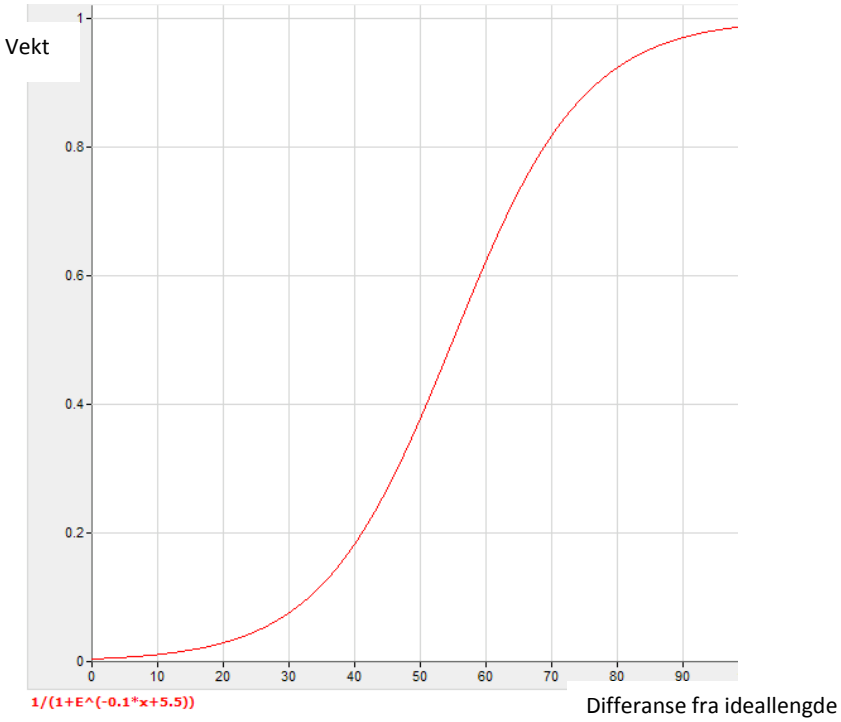
Det finnes flere tilnærmingsmåter for å oppnå en slik differensiering. Den enkleste måten vil være å vekte proporsjonalt med avstanden fra det ideelle målet, men som forklart i det forrige avsnittet vil dette kunne differensiere mellom avsnitt som i utgangspunktet er like gode bildetekstkandidater. En annen måte er å opprette kategorier hvor avsnittene får vekt etter hvilken kategori de tilhører. Avsnitt med en lengde på mellom 15 og 30 ord får en vekt, avsnitt med en lengde på mellom 30 og 50 får en annen osv. Ulempen med dette er at det blir store hopp mellom avsnitt som har bare ett ords forskjell i lengde hvis avsnittslengdene befinner seg i utkanten av kategoriverdiene. Til denne oppgaven ble det utviklet en funksjon for å ta mest mulig hensyn til disse utfordringene. Det ble gjort ved å vekte med en variant av en logistisk funksjon hvor d_j står for differensen mellom tekstblokklengden og den ideelle bildetekstlengden, og α og β er konstanter som påvirker akselerasjon og vindustørrelse:

$$v_j = \frac{1}{1 + e^{(-\alpha d_j + \beta)}}$$

Ved å bruke α=0.1 og β=5.5 får man vektingskurve som vises i Figur 10.

Ved å bruke denne vektingskurven vil det i liten grad bli differensiert mellom avsnitt som har opptil 10 ord differanse fra den ideelle lengden, avsnitt med opptil 20 ords differanse vil bli vektet noe mindre fordelaktig, mens avsnitt som har over 20 ord mer eller mindre enn den ideelle bildetekstlengden vil få en stadig høyere og dermed mindre fordelaktig vekting. Når differensen nærmer seg 100 ord, vil det i liten grad bli differensiert mellom avsnittene lenger.

Ved å gå igjennom 10 bøker manuelt og telle lengden på bildetekstene i disse bøkene ble estimatet for en gjennomsnittslengde for bildetekster 26. Fordi det ofte kan oppstå i feil i skanningen slik at enkle ord deles i to, ble det operert med et litt større lengdeestimat på 28.



Figur 10. Vektingskurve for avsnittslengde

Vekten til avsnitt *j* for bilde *i* ble dermed beregnet med følgende formel:

$$v_{ji} = \frac{min(avs_{ji_k})}{1000} + \frac{0{,}5}{min(avs_{jt})} + \frac{s_j}{100} + \frac{1}{1 + e^{(-\alpha d_j + \beta)}}$$

Hvis det ikke var mulig å regne avstand mellom avsnitt *j* og andre tekstblokker *t*, fordi det ikke fantes andre tekstblokker i nærheten, ble verdien 0 brukt som delvekt. Tilsvarende ble verdien 11 brukt hvis det ikke fantes informasjon om avsnittets tekststørrelse *s*. Hvis tekstblokk *j* ikke befant seg på samme side som bildet, ble $\frac{min(avs_{ji_k})}{1000}$ erstattet med antall sider mellom tekstblokksiden og bildeblokksiden.

6.3 Resultater

Denne metoden for å finne bildetekster ble testet ut på 10 tilfeldig valgte bøker fra datamaterialet. I disse bøkene var det totalt 903 bilder som hadde bildetekst og som ble brukt som grunnlag for vurderingen. Fordelingen av bildene i bøkene varierte fra 18 bilder i boken med færrest bilder til 202 og 204 bilder i bøkene med flest bilder. De resterende 7 bøkene hadde mellom 30 og 90 bilder hver. De tallene som refererer til antall bilder her, refererer til antall bildeblokker i ocr-

filene. Som det kommenteres lenger ned i teksten, kan det bety at enkelte av "bildene" det refereres til, egentlig består av flere bilder som er tolket som ett bilde i scanningsprosessen. I slike tilfeller er bildeteksten vurdert som gjenfunnet hvis teksten som er funnet, er bildeteksten til ett av bildene i bildeblokken. Dette kan være en svakhet ved statistikken som det bør tas høyde for. Teksten i de 10 bøkene ble vektet i forhold til bildene, og avsnittene ble rangert fra lavest til høyest vekting, hvor lav vekting betydde at teksten hadde stor sannsynlighet for å være bildets bildetekst. For hvert bilde ble de 4 øverst rangerte avsnittene hentet ut til vurdering. For hvert bilde ble det registrert om avsnittet med bildeteksten var blant de vurderte avsnittene, og eventuelt på hvilken plass i rangeringen bildeteksten hadde havnet. Bilder som ikke hadde bildetekst i boken ble ekskludert fra resultatene. Fordelingen av gjenfunne og ikke gjenfunne bildetekster vises i tabell 1 og 2. Selv om materialet er for lite til å trekke noen fullstendige slutninger, kan statistikken som blir presentert under kanskje gi en indikasjon på om metoden fungerer eller ikke, eller kanskje mer interessant: under hvilke omstendigheter metoden synes å gjøre det bra, og under hvilke omstendigheter den synes å gjøre det mindre bra.

	Antall funnet	Ikke funnet	Antall bilder
Tall	799	104	903
Prosent	88 %	12 %	100 %

Tabell 1. Gjenfunne bildetekster

	Rang 1	Rang 2	Rang 3	Rang 4	Antall bilder
Tall	666	87	34	12	799
Prosent	83 %	11 %	4 %	2 %	100 %

Tabell 2. Gjenfunne bildetekster etter rangering

Av de 903 bildetekstene knyttet til bildene ble 799 gjenfunnet blant ett av de fire hentede avsnittene. Totalt sett ble bildeteksten dermed gjenfunnet i 88 % av tilfellene. Blant de gjenfunne bildetekstene er 83 % av dem funnet i det avsnittet som ble rangert høyest. 11 % ble funnet i det avsnittet som ble rangert nest høyest. Det vil si at bildeteksten ble funnet blant de to høyest rangerte avsnittene i 94 % av tilfellene. Dette må sies å være gode resultater, men det er fremdeles mulighet for forbedring, ikke minst fordi det er stor forskjell mellom bøkene for hvor gode resultatene er.

Tabell 3 og 4 viser spredningen i resultatene i forhold til hvordan de 799 gjenfunne bildetekstene fordelte seg mellom bøkene, og på hvilken rangering de ble gjenfunnet. I halvparten av bøkene (2, 3, 5, 6 og 7) ble alle eller nesten alle bildetekstene gjenfunnet i ett av de fire vurderte avsnittene. Tre av bøkene (1, 8 og 9) har en gjenfinningsprosent på rundt 80 %. I disse bøkene blir altså 4 av 5 bildetekster gjenfunnet. Det dårligste resultatet er for bok 10 hvor kun 70 % av bildetekstene ble funnet.

	Tall			**Prosent**	
Bok	Totalt	Funnet	Ikke funnet	Funnet	Ikke funnet
1	82	68	14	83 %	17 %
2	18	18	0	100 %	0 %
3	78	78	0	100 %	0 %
4	55	42	13	76 %	24 %
5	202	202	0	100 %	0 %
6	91	90	1	99 %	1 %
7	37	37	0	100 %	0 %
8	46	38	8	83 %	17 %
9	204	163	41	80 %	20 %
10	90	63	27	70 %	30 %

Tabell 3. Gjenfunne bildetekster fordelt på bøker

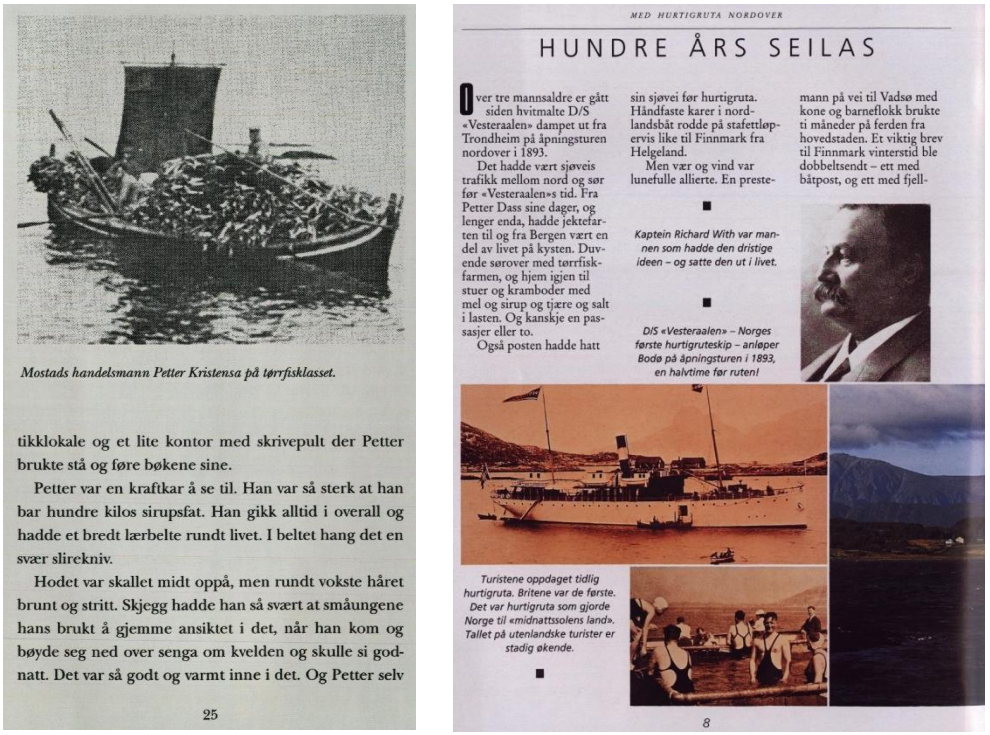
Av de 799 bildetekstene som ble gjenfunnet, ble 666 gjenfunnet i det høyest rangerte avsnittet, dette tilsvarer 83 %. I åtte av de ti bøkene ble over 80 % av bildetekstene høyest rangert, i tre av disse bøkene ble bildeteksten funnet i det høyest rangerte avsnittet i over 96 % av tilfellene. I de to bøkene med dårligst resultat ble kun 60 og 70 % av bildetekstene rangert øverst. Dette viser at hvor god denne teknikken er til å finne bildetekster, avhenger av hvilken bok som indekseres.

	Tall					Prosent			
Bok	Totalt	Rang 1	Rang 2	Rang 3	Rang 4	Rang 1	Rang 2	Rang 3	Rang 4
1	68	56	6	4	2	82 %	9 %	6 %	3 %
2	18	18	0	0	0	100 %	0 %	0 %	0 %
3	78	75	3	0	0	96 %	4 %	0 %	0 %
4	42	35	6	1	0	83 %	14 %	2 %	0 %
5	202	176	20	6	0	87 %	10 %	3 %	0 %
6	90	87	2	0	1	97 %	2 %	0 %	1 %
7	37	33	3	0	1	89 %	8 %	0 %	3 %
8	38	31	2	4	1	82 %	5 %	11 %	3 %
9	163	117	30	10	6	72 %	18 %	6 %	4 %
10	63	38	15	9	1	60 %	24 %	14 %	2 %

Tabell 4. Gjenfunne bildetekster etter rangering fordelt på bøker

Det finnes indikasjoner på hvilke faktorer som kan påvirke dette resultatet. Den mest åpenbare faktoren er sidens layout. Bøker med enkel layout gjør det mye enklere å finne bildeteksten enn bøker med komplisert layout. På Figur 11 er det for eksempel enklere å finne bildeteksten til bildet på siden til venstre (hentet fra bok 2 i tabellene), mens det er vanskeligere å identifisere hva som er bildeteksten for bildene på siden til høyre (hentet fra bok 4 i tabellene).

Det er to hovedgrunner til at komplisert layout gjør det vanskeligere å identifisere bildetekster. Den ene årsaken er at på en side med flere bilder, og dermed ofte flere bildetekster, vil det å identifisere bildeteksten ikke bare være å skille blokker med bildeteksttrekk fra vanlige tekstblokker, men også å kunne identifisere hvilke av bildetekstblokkene som tilhører hvilket bilde. Portrettet av kaptein Richard With til venstre på Figur 11 har for eksempel to mulige bildetekstblokker med lik formatering og noenlunde lik avstand til bildet.



Mostads handelsmann Petter Kristensa på tørrfisklasset.

tikklokale og et lite kontor med skrivepult der Petter brukte stå og føre bøkene sine.

Petter var en kraftkar å se til. Han var så sterk at han bar hundre kilos sirupsfat. Han gikk alltid i overall og hadde et bredt lærbelte rundt livet. I beltet hang det en svær slirekniv.

Hodet var skallet midt oppå, men rundt vokste håret brunt og stritt. Skjegg hadde han så svært at småungene hans brukt å gjemme ansiktet i det, når han kom og bøyde seg ned over senga om kvelden og skulle si godnatt. Det var så godt og varmt inne i det. Og Petter selv

25

MED HURTIGRUTA NORDOVER

HUNDRE ÅRS SEILAS

Over tre mannsaldre er gått siden hvitmalte D/S «Vesteraalen» dampet ut fra Trondheim på åpningsturen nordover i 1893.

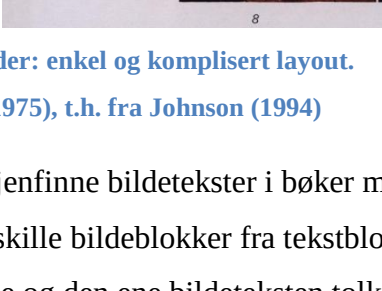
Det hadde vært sjøveis trafikk mellom nord og sør før «Vesteraalen»s tid. Fra Petter Dass sine dager, og lenger enda, hadde jektefarten til og fra Bergen vært en del av livet på kysten. Duvende sørover med tørrfiskfarmen, og hjem igjen til stuer og kramboder med mel og sirup og tjære og salt i lasten. Og kanskje en passasjer eller to.

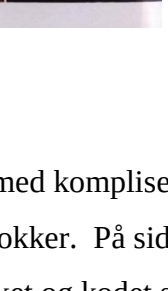
Også posten hadde hatt sin sjøvei før hurtigruta. Håndfaste karer i nordlandsbåt rodde på stafettløpervis like til Finnmark fra Helgeland.

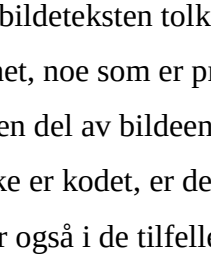
Men vær og vind var lunefulle allierte. En prestemann på vei til Vadsø med kone og barneflokk brukte ti måneder på ferden fra hovedstaden. Et viktig brev til Finnmark vinterstid ble dobbeltsendt – ett med båtpost, og ett med fjell-

Kaptein Richard With var mannen som hadde den dristige ideen – og satte den ut i livet.

D/S «Vesteraalen» – Norges første hurtigruteskip – anløper Bodø på åpningsturen i 1893, en halvtime før ruten!







Turistene oppdaget tidlig hurtigruta. Britene var de første. Det var hurtigruta som gjorde Norge til «midnattssolens land». Tallet på utenlandske turister er stadig økende.

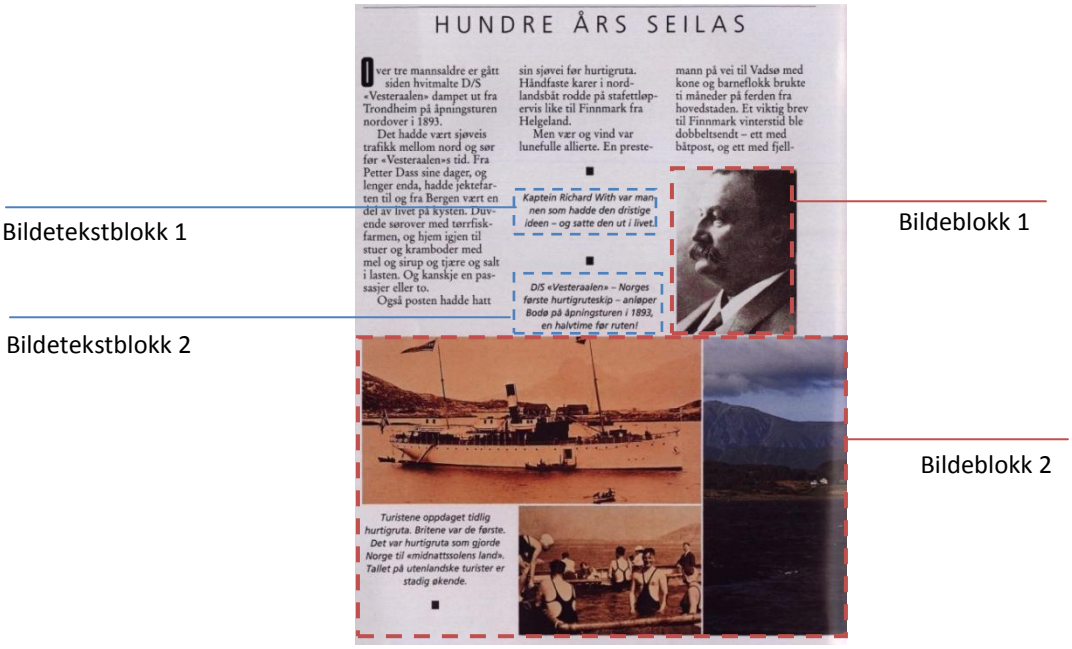
8

Figur 11. Eksempelsider: enkel og komplisert layout.

T.v. fra Johnson (1975), t.h. fra Johnson (1994)

Den andre grunnen til at det er vanskelig å gjenfinne bildetekster i bøker med komplisert layout er at OCR-scanningen har problemer med å skille bildeblokker fra tekstblokker. På siden som er avbildet på Figur 12 er de tre nederste bildene og den ene bildeteksten tolket og kodet som én bildeblokk. Dette gjør at alle bildene oppfattes som én enhet, noe som er problematisk i seg selv, men det andre problemet er at bildeteksten oppfattes som en del av bildeenheten, og teksten blir dermed ikke kodet i sidens xml-dokument. Når teksten ikke er kodet, er den heller ikke tilgjengelig for gjenfinning. Denne problematikken gjelder også i de tilfellene bildeteksten er lagt oppå bildet, noe som gjør at teksten sjelden blir oppdaget og kodet.

I Figur 12 kan man også til høyre i bildeblokk 2 se et eksempel på et bilde som strekker seg over to sider. Når bilder strekker seg over flere sider, er det vanskelig å vite ut i fra kodingen om det er det samme bildet som strekker seg over flere sider, og som derfor skal ha samme bildetekst, eller om det er to forskjellige bilder på hver sin side. I dette tilfellet er bildet også oppfattet som en del av en annen bildeblokk, noe som kompliserer det hele enda mer.



Figur 12. Eksempelside: bilde- og bildetekstblokker. Fra Johnson(1994)

6.4 Oppsummering

Disse forsøkene med å identifisere hva som er bildetekst i digitale bøker, har gitt lovende resultater. Det er fremdeles forbedringspotensial i metoden, men hvis den utvikles videre bør den kunne bli en god metode for dette formålet. Blant områdene som bør utforskes mer er mulige løsninger til å indeksere bilder som går over flere sider, og bedre løsninger for de tilfellene hvor bildeteksten står på en annen side enn det bildet gjør. I tillegg bør det testes videre hvilke faktorer som bør ha størst betydning for den endelige vekten. I dette forsøket er det heller ikke tatt hensyn til tekststil i formateringen. Bildetekster vil ofte være i kursiv, og da vil det kunne være gunstig å vekte tekst som er registrert med denne formateringen.

7 Indeksering basert på bildetekst

7.1 Navnegjenkjenning

Beskrivelser på Shatfords ikonografiske nivå vil ofte være egennavn. Det å identifisere egennavn i tekster er nyttig i mange forskjellige sammenhenger, og det er derfor et område det har blitt forsket mye på. Det blir ofte kalt Named entity recognition eller navnegjenkjenning, og har vært delemne for flere internasjonale konferanser (Røyneberg 2004). Det finnes derfor flere systemer som er skapt for å gjenkjenne ulike egennavn. De fleste systemene er laget for det engelske språket, men det finnes også system som er utviklet for nordiske språk. Nomen Nescio Project er et skandinavisk samarbeidsprosjekt hvor det er utarbeidet seks system for navnegjenkjenning, hvorav tre av systemene er laget for tekster på norsk (Røyneberg 2005). Grunnprinsippet for å finne egennavn i tekster er et enkelt prinsipp, og likt for så å si alle systemer: Ord med stor bokstav er egennavn, med mindre ordet er det første ordet i en setning De fleste av systemene som er laget for navnegjenkjenning er allikevel systemer som bruker avansert språkanalyse, og det er flere grunner til at det å identifisere og klassifisere egennavn ikke er så enkelt som grunnprinsippet skulle tilsi.

7.1.1 Språkanalyse

Med språkanalyse menes her språkanalyse i forbindelse med datalingvistikk, som går ut på å analysere naturlig språk i forhold til semantikk og/eller syntaks for å kunne utnytte det semantiske innholdet i en tekst i automatisk tekstprosessering. Syntaktisk språkanalyse bruker grammatiske regler til å analysere ord i en tekst, mens semantisk språkanalyse bruker ordenes betydning i analysen. Språkanalyse kan variere i kompleksitet fra enkle teknikker som å identifisere store forbokstaver og bruke lister for å identifisere personnavn, stedsnavn osv., til mer avanserte teknikker som krever større grad av lingvistisk forståelse. I denne oppgaven er det valgt å kun bruke enkle språkanalyseteknikker, men det kunne vært et alternativ å bruke mer avanserte former for språkanalyse for å identifisere egennavn i teksten.

Et eksempel på en språkanalyseteknikk som er mer avansert enn det som brukes i denne oppgaven, er det å bruke lingvistiske regler for å identifisere ordklasser. Ved å kunne identifisere hvilke ord som er substantiver, pronomener og verb, vil man kunne behandle disse ordene og ordene rundt etter hvilken ordklasse de tilhører. Dette vil for eksempel være nyttig for å

identifisere pre-ikonografiske beskrivelser, fordi slike beskrivelser ofte vil være substantiver eller verb, mens determinativer, artikler og konjunksjoner sjelden vil ha verdi alene i en slik situasjon. Basert på dette vil det kunne være mulig å skille mellom forskjellige betydninger av homonymer, særlig der de ulike betydningene av homonymet tilhører forskjellige ordklasser. For eksempel vil det, ved å bruke syntaktiske regler, være mulig å skille mellom verbet og substantivet ”måke”, og mellom adjektivet og substantivet ”lyst”. Ved å ha den grammatiske syntaksen tilgjengelig, vil det være mulig å analysere setninger og identifisere relasjoner mellom ordene i setningen, som at ”Nansen” og ”Krefting” er subjektet i setningen ”Nansen og Krefting flår isbjørn”. En teknikk innen språkanalyse som blir brukt innenfor navnegjenkjenning, er en form for semantisk analyse hvor visse substantiver og verb blir brukt som indikatorer knyttet til semantiske roller. For eksempel vil verb som ”jobbe”, ”kjøpe” og ”eie” være forbundet med personer, mens ”reise” og ”bo” vil være eksempler på verb som er knyttet til sted. Koblet med syntaktisk analyse vil det være mulig å bruke denne typen informasjon til å klassifisere ulike typer egennavn (Røyneberg 2004).

7.1.2 Utfordringer ved navnegjenkjenning

Grunnprinsippet er som sagt at alle ord som begynner med stor bokstav er et egennavn, med mindre ordet står først i en setning. Et første problem med denne forutsetningen er at egennavn også kan være det første ordet i en setning. Det er derfor nødvendig å vurdere om et ord som står først i en setning, er et egennavn eller et vanlig ord. For å komplisere saken ytterligere kan egennavn også være identiske med vanlige ord. Dette gjelder både fornavn (Bjørn, Stein osv.) og etternavn (Fjell, Berg osv). En måte å skille mellom ordet og egennavnet på, er å se etter de tilfellene der flere ord med stor bokstav står etter hverandre. Hvis ordet som har stor forbokstav følges av et ord med stor forbokstav, er et sannsynligvis et navn. Det finnes allikevel eksempler på at navn ikke alltid følger strukturen ”Fornavn Etternavn” eller ”Fornavn Mellomnavn Etternavn”. Navn kan ha preposisjoner som skrives med liten bokstav (Ludwig van Beethoven), de kan refereres med titler (doktor Hansen) og de kan refereres med kun ett av navnene (”Nansen er kjent for…”, ”Eva syntes…”).

Til tross for at det finnes mange utfordringer knyttet til å identifisere egennavn, er nok hovedgrunnen til at mange av systemene krever avansert språkanalyse at de skal kunne håndtere å klassifisere egennavnene, og ikke bare identifisere dem. I mange situasjoner er det nødvendig å

vite hvilken type egennavn det er snakk om i hvert tilfelle, og det å klassifisere egennavnene er en mye mer komplisert prosess enn kun å identifisere dem.

7.1.3 Grunnleggende prinsipp

Rowe og Frew (1997) argumenterer med at i forbindelse med bildeindeksering kan selv enkel språkanalyse være nyttig og gi gode resultater, uten å måtte bruke veldig avanserte språkanalysesystem. Det vil i denne oppgaven derfor brukes enkle språkanalyseteknikker for å utforske muligheten for å identifisere bildebeskrivelser på ikonografisk nivå i de gjenfunne bildetekstene.

Til tross for at det bare brukes enkle språkanalyseteknikker, er det allikevel gjort forsøk på å delvis differensiere mellom ulike egennavnstyper. Grunnen til det er at formålet med undersøkelsen er å indeksere bilder. En person kan refereres til ved navn på forskjellige måter; ved fornavn, etternavn eller ved fullt navn. Når personnavnet skal brukes til å indeksere et bilde er det ønskelig å bruke det fulle navnet, og det bør derfor være mulig å gjenkjenne personnavn slik at disse kan behandles annerledes enn for eksempel stedsnavn. Til dette formålet er enkel språkanalyse brukende, men det er ingen tvil om at det i fremtiden vil være interessant å prøve ut et avansert språkanalysesystem i indekseringen.

Som tidligere nevnt finnes det flere aspekter ved språkets natur som skaper utfordringer når det kommer til å identifisere egennavn og skille dem fra andre ord. Det mest typiske eksempelet er å identifisere om første ordet i en setning er et navn eller om det er et ord. Det finnes flere måter å løse dette på, men en av de vanligste metodene, som også krever minst språkanalyse, er å bruke lister (Røyneberg 2005). Ved å sjekke om termen det er snakk om står på en liste over ord, personnavn eller stedsnavn, vil det være enklere å skille mellom egennavn og ord. En fordel med å bruke lister er at det i seg selv ikke krever veldig avansert språkanalyse. De fleste systemene kombinerer riktignok lister med andre språkanalyseteknikker (Røyneberg 2005), men prinsippet bak lister gjør at det kan brukes til enkel språkanalyse også i seg selv. Termer man er usikker på om er et ord eller et egennavn sjekkes mot listene, og i mange tilfeller vil dette kunne være nok for å stadfeste om det er snakk om et egennavn eller et alminnelig ord. Et minus ved å bruke lister er at de kan være krevende å lage og å vedlikeholde fordi egennavn endrer seg med tiden; nye egennavn kommer til og stavemåte kan endre seg. I tillegg kan ikke bruken av lister alene fange

opp ambiguøse tilfeller hvor en term teoretisk sett kan være både et ord og et navn, og hvor det kun er sammenhengen som viser hva som er tilfellet i en gitt situasjon.

7.1.4 Ordlister

Det ble laget tre lister til denne oppgaven: en ordliste med norske ord, en personnavnliste med norske og internasjonale navn, og en stedsnavnsliste med norske og internasjonale stedsnavn.

Ordlisten er lastet ned fra Norsk ordbank[1] og er i liten grad modifisert for oppgaven. Det som er gjort er å trekke ut ordene i ren form, slik at all grammatisk informasjon utover selve ordet er strippet bort. Hvert ord forekommer både i ubøyd form og i alle sine bøyde former. I tillegg er enkle bokstaver som er definert som ord, fjernet fra listen. Dette er gjort for å hindre at initialer blir avfeiet som navn fordi de er registrert som ord. Totalt er det 634 344 ord i ordlisten.

Navnelisten ble laget ved å kombinere data fra Statistisk Sentralbyrås (SSB) navnestatistikk og US Census Bureau[2]. Statistikken som ble hentet fra SSB, var alle for- og etternavn i Norge som er brukt av 200 eller flere. Dette ble slått sammen i samme liste. I tillegg ble det ansett som nødvendig å hente inn navneinformasjon om navn fra andre land enn Norge, fordi mange av tekstene omhandler personer som ikke er nordmenn. Det ble besluttet å hente data fra USA, fordi USA som et multikulturelt land, vil ha navneforekomster som stammer fra mange forskjellige språkgrupper. Navnedataene fra USA er hentet fra folketellingsbyrået sine lister over kvinnenavn, mannsnavn og etternavn og ble sist oppdatert i 1995. Den totale listen med norske og amerikanske navn er på 94 950 navn.

Stedsnavnslisten ble laget av en kombinasjon av stedsnavn fra yr.no sin stedsoversikt og data fra geonames.org. Yr.no sin tilgjengelige stedsoversikt består kun av rundt 6000 navn som yr.no anser som viktige stedsnavn i Norge. For å få en mer utfyllende stedsliste ble det hentet data fra geonames.org. Det ble hentet stedsnavn fra hele verden, og for å få en liste som det var mulig å håndtere i programmene, måtte listen modifiseres kraftig. Alle stedsnavn som ikke ble ansett for å være et geografisk sted ble slettet, for eksempel ble alle bygninger og konstruksjoner, som hotellnavn, brannstasjonsnavn osv., slettet fra listen. I tillegg ble alle stedsnavn som inneholdt tegn det ikke er mulig å skrive med et vanlig norsk tastatur slettet fra listen, med den

[1] http://www.edd.uio.no/prosjekt/ordbanken/ [hentet 2009-04-02]
[2] http://www.census.gov/genealogy/names/names_files.html [hentet 2009-04-02]

begrunnelsen at det var mindre sannsynlig at disse navnene ville forekomme i en norsk bok, enn navn med vanlige latinske bokstaver. Den endelige listen over stedsnavn endte på 553 134 navn.

7.2 Fremgangsmåte

7.2.1 Personnavn

Det første steget for å identifisere personnavn knyttet til bilder, var å finne alle hele personnavn i boken. Med hele personnavn menes personnavn som består av mer enn ett ledd. Dette ble gjort for å få en oversikt over hvilke hele navn som forekom, og med hvilken frekvens, slik at denne oversikten kunne brukes senere for å avgjøre hvilke hele navn enkeltnavnreferanser mest sannsynlig viste til. For å finne personnavnene ble grunnprinsippet for å finne egennavn fulgt. Alle ord som hadde stor forbokstav ble vurdert som egennavnskandidat. Hvis ordet var det første ordet i en setning, ble det vurdert opp mot de ulike listene for å avgjøre om hvorvidt det var et ord eller et egennavn. Hver gang et egennavn ble identifisert, ble det også søkt etter sekvenser av egennavn ved å se om ordene umiddelbart etter også var egennavn.

Når alle hele personnavn var identifisert var neste steg å finne personnavn knyttet til bildene i boken. For hvert bilde ble den teksten som var vurdert som bildetekst for det bildet, analysert på samme måte for å se om teksten inneholdt personnavn. Hele personnavn ble funnet på samme måte som for å finne personnavn i hele boken. I tillegg ble det gjort forsøk på å identifisere enkle navn, altså navn som bare bestod av ett ledd.

7.2.1.1 Enkle navn

Når det i teksten ble funnet termer som ble identifisert som personnavn, men som ikke var en del av en sekvens, ble det gjort forsøk på å identifisere mulige hele navn som kunne knyttes til det enkle navnet.

7.2.1.2 Navnesekvenser

Basert på oversikten over navn som ble laget for hele teksten, ble hele navn som inkluderte det enkle navnet, vurdert som kandidater. Disse ble vektet og rangert og den høyest vektede kandidaten ble valgt som indekseringsterm. Det ble ikke funnet eksempler på mulige vektingsfunksjoner til et slikt formål i eksisterende forskning, og det ble derfor utformet en funksjon som tok hensyn til de aspektene som ble vurdert som viktig for vektingsformålet.

Vektingen ble gjort med følgende formel for kandidatnavn *i* til enkeltnavn *j*:

$$v_i = \log(f_i) + \frac{1}{a_{ij}}$$

Vekten består av frekvensen (f) av kandidatnavnet i hele boken pluss avstanden (a) mellom enkeltnavnet og den nærmeste forekomsten av kandidatnavnet i teksten, målt i antall sider. Det ble brukt en cutoff-verdi på 0,3 for å minimalisere forekomstene av feil. Fordi datamaterialet inneholder mange ord som er feilscannede, kan det forekomme at enkelte ord får store bokstaver, selv om de ikke har det i den opprinnelige teksten. Ved slike tilfeller er det vanskelig å skille dem fra egennavn fordi de da ofte ikke tilsvarer ord som finnes i ordlisten. Ved å bruke en cutoff fjernes de kandidatene som bare forekommer én gang i hele teksten, noe som luker ut mange av feilene forårsaket av feilscanning. Hvis det i boken ikke finnes hele navn som inkluderer det enkle navnet, eller hvis kandidatene ikke kommer over cutoff-verdien, brukes det enkle navnet som indekseringsterm, med mindre det også finnes på stedsnavnslisten. Argumentet for dette er at hvis en term finnes både på personnavnslisten og stedsnavnslisten, vil det mest sannsynlig være et sted hvis det bare forekommer som et enkeltnavn i teksten, og ikke i en navnesekvens.

Dette er bare et forslag til hvordan en slik vekting kan foregå, og forskjellige vektings- og normaliseringsfunksjoner bør prøves ut for å finne den mest egnede måten for å knytte enkeltnavnsforekomster sammen med de hele navneformene de mest sannsynlig skal representere.

7.2.1.3 Familiereferanser

I noen av tilfellene der navn forekommer som et enkeltnavn og ikke et fullt navn, er det fordi det er en del av en setning hvor flere medlemmer av familien ramses opp. Eksempel på dette er ”Leon, Gustav og Roald Amundsen” og ”Eva og Fridtjof Nansen”. I disse tilfellene ville det vært ønskelig å få indekseringstermene ”Leon Amundsen”, ”Gustav Amundsen” og ”Roald Amundsen”, og ”Eva Nansen” og ”Fridtjof Nansen”. Ved å bare se etter sekvenser ville man få indekseringstermene ”Leon”, ”Gustav” og ”Roald Amundsen” i det første tilfellet og ”Eva” og ”Fridtjof Nansen” i det andre tilfellet. Oppramsinger kan kjennes igjen på strukturen ved at de bruker ordet ”og” og eventuelt kommategn ved flere navn. Dette er selvfølgelig kjennetegn på alle oppramsinger og gjelder også i tilfeller som ”Nansen og Amundsen”, hvor det ikke er

ønskelig å behandle personene som familie. Det er allikevel en viss forskjell i de to typene oppramsing. Når det gjelder personer med samme etternavn vil ofte strukturen være slik at de første navnene oppføres med fornavn, mens det siste oppføres med fullt navn. I oppramsing av andre personer er det vanligere å bruke samme navneform for hver person; enten ”Nansen og Amundsen” eller ”Fridtjof Nansen og Roald Amundsen”. Hvis man kommer over en sekvens med strukturen ”Navn, Navn og Fullt Navn” eller ”Navn og Fullt Navn” vil det derfor være nærliggende å tro at det er snakk om et familieforhold, og at den første personen vil ha samme etternavn som den siste personen i oppramsingen. I de tilfellene et enkeltnavn blir etterfulgt av et komma eller termen ”og” ble det derfor vurdert om det var snakk om en familiereferanse, og i så fall ble enkeltnavnet koblet til et fullt navn som indekseringsterm. Hvis termene etter ordet ”og” viste seg å tilhøre en navnesekvens ble siste bit av sekvensen koblet sammen med enkeltnavnet slik at det ble et fullt navn.

7.2.2 Stedsnavn

I tillegg til å se etter personnavn, har det blitt sett etter forekomster av stedsnavn i bildeteksten. På samme måte som ved personnavn blir grunnprinsippet for å finne egennavn brukt, men istedenfor å bruke en liste med personnavn brukes en liste over stedsnavn. En term som har stor forbokstav har blitt vurdert som stedsnavn hvis det finnes i stedsnavnslisten, eller hvis termen hverken er et ord eller et navn og oppfyller et av disse kriteriene:

- *Termen har en bestemt endelse.*
  Hvis termen ikke finnes i ordlisten eller i navnelisten og ender på kombinasjoner som ”-en”, ”-et” blir det regnet som et stedsnavn. Dette kriteriet fanger opp termer som gjelder bestemte steder av typen ”Oslofjorden”, ”Galdhøpiggen” osv, i de tilfellene disse ikke er inkludert i stedsnavnslisten.
- *Termen starter med en retningsviser*
  Tanken bak dette kriteriet er at termer som har stor forbokstav og som begynner med en retningsviser som ”nord-”, ”sør-”, ”øst-”, ”vest-”, ”aust-” og ”syd-” mest sannsynlig referer til et sted.
- *Termen ender med stedsendelser*
  Forekomster av typen ”havn”, ”by”, ”land”, ”stad”, ”nes”, ”sand” kombinert med stor forbokstav har i denne oppgaven blitt tolket som en indikator for et stedsnavn.

7.2.3 År

I tillegg til å analysere bildetekster for å finne visse typer egennavn, ble det lett etter tidsbenevnelser som årstall og perioder. Dette ble gjort ved å se etter strukturer som ligner på årstall. Enten ved å se etter årstall som er regnet for å være 4 tall etter hverandre, eller tall etterfulgt av tidsbenevnende tekst: ”-tallet”, ”-århundret”.

7.3 Resultater

Forutsetningen for at tekstanalysen som er beskrevet skal fungere, er at analysen blir gjort på bildets bildetekst. I gjennomgangen av resultatene har derfor eksempler fra boken *Jegeren Fridtjof Nansen* av Svein Solli (2002) blitt brukt. Dette var en av bøkene som fikk best resultat på gjenfinning av bildetekster. Alle bildetekstene ble funnet og 96 prosent av bildeteksten (75 av 78 mulige) ble funnet øverst i rangeringen. Siden tekstanalysen er foretatt på den øverst rangerte bildetekstkandidaten, vil indekseringsresultatene fra denne boken i all hovedsak være basert på en analyse av det som faktisk er bildeteksten i boken.

Eksempelpostene som blir presentert her, er poster som er generert basert på indekseringen som er beskrevet i forrige kapittel. Postene er lagret som Dublin Core XML-filer og inkluderer Dublin Core (DC)-elementene coverage, subject, description, identifier og creator. Indekseringstermer som er enten stedsnavn eller tidsangivelse, er angitt som dc:coverage, og blir i eksemplene kalt “Tid/Sted”. Personnavn er lagt til dc:subject og blir kalt ”Avbildede personer”. I dc:description er bildeteksten, som er kilden til tekstanalysen, lagt. Denne kalles ”Beskrivelse” i eksemplene. I tillegg er det en kategori kalt opphav, der informasjonen som er lagt i dc:creator, blir vist. Det har ikke vært et fokus å trekke ut opphavsinformasjon i denne oppgaven, men fordi rettighetsinformasjonen i enkelte bøker har blitt lagt i bildeteksten, ble det tatt et valg om å skille denne informasjonen fra den innholdsbeskrivende delen av bildeteksten der det var mulig.

Tid/Sted	Lyngdal
	Strand
	Buskerud
	1912,
Avbildede personer	Axel Heiberg
	Oscar Heiberg
	Hjalmar Krag
	Georg Sibbern
	Kong Haakon
	Fridtjof Nansen
	Bratt
	Ragnhild Heiberg
	Mimi Krag
	Alexander Nansen
Beskrivelse	Kongejakt 1912, på Axel Heibergs gård Strand i Lyngdal, Buskerud. Fra venstre: Oscar Heiberg, Hjalmar Krag, Georg Sibbern, kong Haakon, Fridtjof Nansen, kaptein Bratt, Ragnhild Heiberg, Axel Heiberg, Mimi Krag. Også Alexander Nansen var med på denne jakta, og har sannsynligvis knipset bildet. Foto: Fridtjof Nansen, Nasjonalbiblioteket.
Opphav	Fridtjof Nansen,
	Nasjonalbiblioteket.

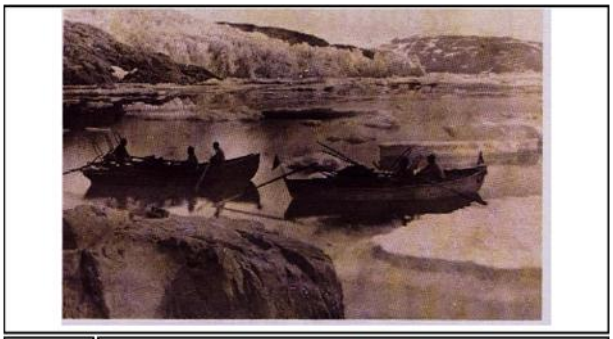
Figur 13. Eksempelpost

7.3.1 Statistikkgrunnlag

Boken har 311 bildeblokker som er identifisert i ocr-scanningen. Av disse ble én av blokkene utelukket fra statistikken fordi det er et bibliotekstempel som har blitt vurdert som et bilde, og derfor vurdert som irrelevant for det som undersøkes.

I statistikkberegningen ble det gjort visse valg som kan påvirke resultatene. I indekseringen ble informasjon knyttet til opphav forsøkt skilt ut fra resten av teksten. Der dette har vært vellykket, har ikke opphavsteksten blitt vurdert i forhold til om teksten inneholder et navn eller ikke. Disse navnene har derfor blitt utelukket fra statistikken knyttet til hvor vellykket teknikken har vært for

å identifisere navn.. I de tilfellene disse navnene ikke har blitt skilt ut som opphavsinformasjon, har de derimot blitt analysert på lik linje med annen tekst og dermed også inkludert som en del av statistikken. Hvis for eksempel en fotograf er inkludert i kategorien avbildede personer, vil dette navnet bli inkludert som korrekt identifisert navn i analysen av navnegjenkjennelse i teksten, mens det vil bli vurdert som feilindeksert navn i analysen av hvor god indekseringen er. Hvis fotografen ikke inkluderes som avbildet person, men plassert i opphavkategorien, vil ikke dette navnet bli inkludert i noen av statistikkene. Navn som blir identifisert som navn i teksten, og som er feilstavet, blir tolket som korrekt identifisert navn i teksten hvis det skal være et personnavn, og som korrekt indeksert hvis personen det gjelder er avbildet på bildet. Dette kan være en svakhet i statistikken fordi noen av feilstavingene kan bidra til at det ikke er mulig å få treff på indekseringstermene ved søk i et gjenfinningssystem. Samtidig vil det i de fleste tilfellene være mulig å identifisere det korrekte navnet ved å se på beskrivelsen, men dette vil ikke hjelpe hvis det korrekte navnet ikke er søkbart. For eksempel vil ”Hans Ivioirke Nansen” være regnet som en korrekt identifisering av et personnavn i teksten, og som korrekt indeksert navn til tross for at navnet er ”Hans Moltke Nansen”. Derimot er navn som er tydelig feil identifisert, regnet som feil i statistikken. Når for eksempel ”Kåre Nansen” blir identifisert som ”Kåre Berg”, blir dette regnet som feil både i forhold til navnegjenkjenning i tekst, og i forhold til indekseringen.



Beskrivelse	På grunn av vanskelige isforhold drev ekspedisjonen langt sørover langs kysten før den kunne komme i land, og verdifull tid gikk tapt til å ro nordover igjen. Foto: Fridtjof Nansen, Nasjonalbiblioteket.
Opphav	Fridtjof Nansen, Nasjonalbiblioteket.

Tid/Sted	Godthåb
Avbildede personer	Otto Sverdrup
	Fridtjof Nansen
Beskrivelse	Sverdrup og Nansen underveis til Godthåb med båten de laget av vidjekvister og seilduk. Foto: Fridtjof Nansen, Nasjonalbiblioteket/Gyldendals arkiv.
Opphav	Fridtjof Nansen, Nasjonalbiblioteket/Gyldendals arkiv.

Figur 14. Eksempelpost: Personer på avstand

I de tilfellene fotografiene viste personer på en slik avstand, eller av en slik karakter, at det ikke var mulig å identifisere personene, ble ikke disse personene regnet med i statistikken over indekserte personer, med mindre teksten nevner personene på bildet. For eksempel ble ikke personene på bildet til venstre i Figur 14 regnet som personer som burde vært indeksert med navn, mens personene på bildet til høyre i samme figur ble vurdert som personer som burde vært indeksert, fordi navnene står i bildeteksten. Bilder hvor personene er i hovedfokus, enten som portrett eller i gruppebilder, ble alltid vurdert som personer som burde vært indeksert.

Personnavn som forekom flere ganger i en bildetekst ble bare regnet én gang.

7.3.2 Personnavn

I forhold til gjenkjenning av personnavn i teksten, har metoden gitt et godt resultat. I bildetekstene til de 310 bildeblokkene som ble undersøkt fantes det 181 navnereferanser. Dette inkluderer alle typer personnavn, inkludert hele navn, fornavn, etternavn og familiereferanser. Teknikken som har blitt presentert identifiserte 178 personnavn, og av disse var 160 korrekte identifiseringer. Det gir en presisjon på 0,9 og en fullstendighet på 0,88. Når det gjelder hvorvidt de identifiserte personnavnene var egnet som indekseringstermer, så er ikke resultatene like gode. Med egnete indekseringstermer menes indekseringstermer som kan direkte relateres til bildet. Når det gjelder personnavn menes det at den foreslåtte indekseringstermen må være navnet på en av personene på bildet. Mulige forklaringer til presisjons- og fullstendighetsverdiene blir presentert i de kommende underkapitlene. I tillegg blir det presentert noen forslag til mulige endringer og videreutviklinger som kan bidra til et bedre resultat.

	I teksten	Foreslåtte	Korrekt identifiserte	Presisjon	Fullstendighet
Alle navn	181	178	160	0,90	0,88

Tabell 5. Navnegjenkjenning: alle personnavn

	På bildet	Foreslåtte persontermer	Korrekte persontermer	Presisjon	Fullstendighet
Personer	168	178	131	0,74	0,78

Tabell 6. Indeksering: personnavn

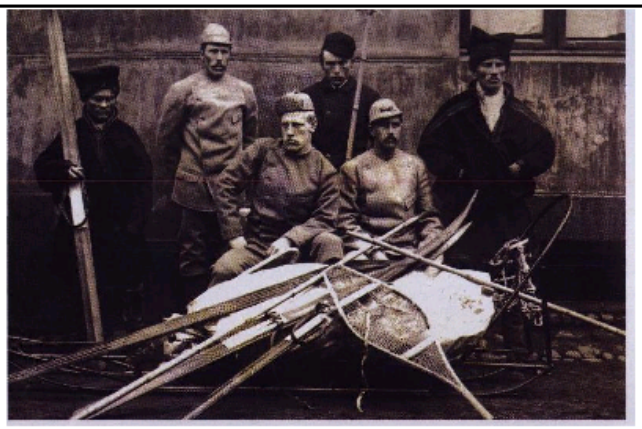
7.3.2.1 Gjenkjenning av hele navn

Fremgangsmåten for å gjenkjenne personnavn fungerer veldig bra når det gjelder å identifisere fulle navn i teksten. Som man kan se av Tabell 7, er både presisjonen og fullstendigheten i gjenkjenning av hele navn veldig høy. Av 110 mulige hele navn i teksten, ble 105 korrekt identifisert, noe som gir en fullstendighet på 0,95. Presisjonen er noe dårligere med en verdi på 0,9, men dette er allikevel også et meget godt resultat.

	I teksten	Foreslåtte	Korrekt identifiserte	Presisjon	Fullstendighet
Hele navn	110	117	105	0,90	0,95

Tabell 7. Navnegjenkjenning hele navn

Et eksempel på gjenkjenning av hele navn kan man se i Figur 15. I denne posten har alle personnavnene i denne bildeteksten blitt identifisert som personnavn, selv der hvert navn har mange ledd. Som nevnt i innledningen til kapittel 7.3., blir navn knyttet til opphavsinformasjon skilt bort fra navn som har med bildets innhold å gjøre. I dette, og de fleste andre tilfeller, er det signalordet ”Foto:” som brukes for å identifisere hvilke navn som ikke skal regnes som indekstermer. Andre signalord som ble brukt til dette, er ordene ”Fotograf:”, ”Kilde:” og ”Tegning:”, men det kan også tenkes at det finnes andre slike signalord som bør inkluderes i denne listen. Et punkt det er verdt å merke seg er at det ved gruppebilder, som vist i Figur 15, er nødvendig å ha en beskrivelse/bildetekst knyttet til bilde for å identifisere hvor personene befinner seg på bildet. Hvis bare indekstermene var blitt brukt til bildebeskrivelsen, ville det vært vanskeligere å identifisere de ulike personene på bildet. Beskrivelsen kommer også godt med i de tilfellene bildeteksten inneholder informasjon som ikke relaterer seg direkte til bildet, fordi den kan oppklare hvorfor tilsynelatende gale indekstermer er tilordnet bildet. Slike indekstermer vil ofte bli regnet som støy i søkeresultat, men i det minste vil beskrivelsen kunne forklare hvorfor indekstermen er tilordnet bildet. Feil indekstermer bør selvfølgelig unngås så langt som mulig, men i de tilfellene det allikevel forekommer, vil det være bedre å vise hvorfor indekstermen er satt til bildet enn å ikke gi brukeren et slikt svar.



Avbildede personer	Ole Nielsen Ravna
	Oluf Christian Dietrichson
	Fridtjof Nansen
	Kristian Kristiansen Trana
	Otto Neumann Knoph Sverdrup
	Samuel Johansen Barto
Beskrivelse	Ekspedisjonen før avreisen med en del av utrustningen. Fra venstre: Ole Nielsen Ravna, Oluf Christian Dietrichson, Fridtjof Nansen, Kristian Kristiansen Trana, Otto Neumann Knoph Sverdrup, Samuel Johansen Barto. Foto: Siems c Co, Nasjonalbiblioteket/Gyldendals arkiv.
Opphav	Siems c Co,
	Nasjonalbiblioteket/Gyldendals arkiv.

Figur 15. Eksempelpost: Gjenkjenning av personnavn

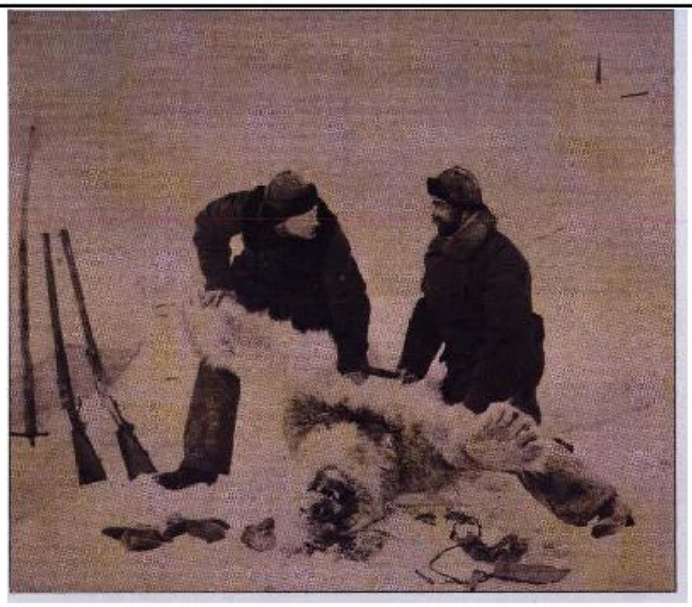
7.3.2.2 Gjenkjenning av enkle navn

	I teksten	Foreslåtte	Korrekt identifiserte			Presisjon	Fullstendighet
			Enkle	Hele	Totalt		
Enkle navn	71	61	9	45	54	0,89	0,76

Tabell 8. Navnegjenkjenning enkle navn

I Tabell 8 presenteres tallene for gjenfinning av enkle personnavn. I forhold til å gjenkjenne enkle navn og helst koble dem til riktig hele navn, oppnår metoden en dårligere presisjon og fullstendighet enn ved gjenkjenning av hele navn. Dette er som forventet, fordi ved gjenkjenning av hele navn er det nok å kjenne igjen sekvensene til navnet i teksten. For enkle navn må det også undersøkes om det finnes mulige hele navn som det enkle navnet kan kobles til. Derfor må en presisjon på 0,89 regnes som et meget godt resultat. Fullstendigheten er noe dårligere med en verdi på 0,76. Av de 54 enkle navnene som ble korrekt identifisert gjennom metoden, ble 45

koblet til et korrekt helt navn, mens 9 ble beholdt som et enkelt navn. Her må det allikevel presiseres at disse tallene ikke er representative. For det første er datagrunnlaget for lite til å kunne si noe generelt om metoden, og for det andre er det en svakhet at dokumentet som er brukt, er en biografi med fokus på Fridtjof Nansen. Det betyr at hovedpersonen ofte blir referert i teksten, og ofte referert med kun etternavn. Fordi algoritmen for å bestemme enkeltnavn baserer seg på frekvens og nærhet, så vil "Nansen" alltid kobles med "Fridtjof Nansen", og dermed gi en korrekt identifisering. Av de 45 enkle navnene som er korrekt koblet til et helt navn, står koblingen "Nansen" – "Fridtjof Nansen" for 16. Det er vanskelig å si noe om hvor godt metoden ville fungert for en bok som ikke er så konsentrert om én person. Dette bør derfor undersøkes videre, og metoden bør testes på andre typer bøker.



Avbildede personer	Fridtjof Nansen
	Axel Krefting
Beskrivelse	Nansen og Krefting flår isbjørn, og diskuterer sikkert skuddvirkningen. De fant at solide fugler var bedre enn hulspisskuler av ekspress-typen. Foto: Fridtjof Nansen, Nasjonalbiblioteket.
Opphav	Fridtjof Nansen,
	Nasjonalbiblioteket.

Figur 16. Eksempelpost: Gjenkjenning av enkle personnavn (etternavn)

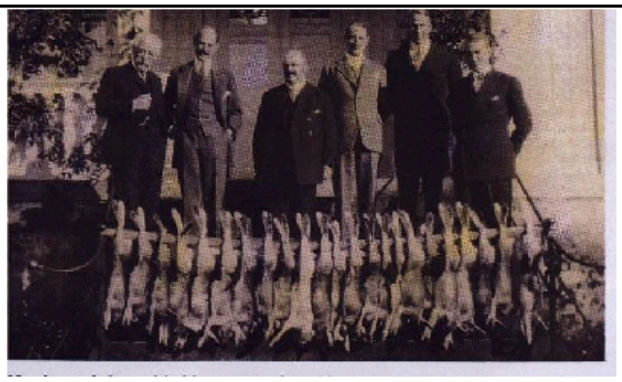
Postene i Figur 16 og Figur 17 gir gode eksempler på hvilke enkeltnavnsreferanser som finnes i bøkene. I Figur 16 refereres de to mennene med etternavn, mens i Figur 17 refereres ekteparet Nansen med deres fornavn. I begge tilfellene søkes det i de fulle navnene som er gjenfunnet i boken som helhet. Som nevnt i forrige avsnitt har boken metoden er testet på, et fokus på Fridtjof Nansen, og hans navn forekommer derfor svært ofte. Når termen ”Nansen” eller ”Fridtjof” forekommer alene vil disse mest sannsynlig referere til Fridtjof Nansen. Fordi vektingen av mulige kandidater bruker frekvens og nærhet som grunnlag, er det derfor ikke så overraskende at det fulle navnet ”Fridtjof Nansen” blir valgt både når ”Fridtjof” og når ”Nansen” står alene. Men i tillegg ser vi at enkeltnavnene ”Eva” og ”Krefting”, som forekommer sjeldnere i boken, blir korrekt koblet med ”Eva Nansen” og ”Axel Krefting”. Dette viser at presisjons- og fullstendighetsverdiene ikke utelukkende er gode på grunn av ”Fridtjof Nansen”.

Tid/Sted	Nordpolen
Avbildede personer	Fridtjof Nansen
	Eva Nansen
Beskrivelse	Fridtjof og Eva som nygifte. Han gjorde det helt klart fra starten at han skulle en tur oppom Nordpolen før han slo seg til ro. Det skulle bli verre enn som så ... Foto: Gihbson, Nasjonalbiblioteket/Gyldendals arkiv.
Opphav	Gihbson,
	Nasjonalbiblioteket/Gyldendals arkiv.

Figur 17. Eksempelpost: Gjenkjenning av enkle personnavn (fornavn)

7.3.2.3 Familiereferanser

En annen variant av enkeltnavnsreferansene er referanser knyttet til familierelasjoner. Et eksempel på dette finnes i Figur 18. Her er to av mennene på bildet brødre: Fritz Anker-Rasch og Ole Anker-Rasch. I teksten står det ”Fritz og Ole Anker-Rasch”, og i utgangspunktet er det dermed kun Ole som blir registrert med fullt navn. Men fordi ”Fritz” etterfølges av ordet ”og” og en navnesekvens, blir etternavnet til Ole Anker-Rasch også identifisert som etternavnet til Fritz.



Tid/Sted	Halden
Avbildede personer	Fridtjof Nansen
	Kong Haakon
	Peter M Anker
	Ragnvald Graff
	Fritz Anker-Rasch
	Ole Anker-Rasch
	Svein Norheim
Beskrivelse	Her har jaktlaget kledd seg om til middag, og poserer med dagen fangst på hagetrappa på Rød - totalt 23 harer. Fra venstre: Fridtjof Nansen, kong Haakon, Peter M. Anker, Ragnvald Graff, Fritz og Ole Anker-Rasch. Foto: Stiftelsen De Ankerske Samlinger, Rød Herregård. Repro: Svein Norheim, Halden historiske samlinger.
Opphav	Stiftelsen De Ankerske Samlinger,
	Rød Herregård.

Figur 18. Eksempelpost: Gjenkjenning av personnavn (familiereferanser)

Et problem som denne oppgaven ikke fanger opp, er de tilfellene der et enkeltnavn, som oftests etternavn, brukes for å referere til flere personer. Bildet i Figur 19 forestiller Eva og Fridtjof Nansen, og i teksten omtales de som ”Ekteparet Nansen”. Her er det bare ordet ”Nansen” som oppfattes som et navn, og det antas dermed at navnet refererer til én person. Fordi ”Fridtjof Nansen” har en høyere frekvens enn ”Eva Nansen”, er det Fridtjof som blir indeksert, mens Eva

blir utelatt. Dette er en av grunnene til at fullstendighetsverdien for indeksering av personnavn blir så lav som den er. Når flere personer på et bilde blir referert til med kun et felles navn, blir kun en av personene indeksert, og dette går ut over fullstendighetsverdien. Det er ikke bare for ektepar at denne effekten vil vise seg, det samme vil skje for fraser av typen ”dr. Bruun med familie”, ”familien Mostad”, ”Brødrene Amundsen” osv.

Tid/Sted	1890.
Avbildede personer	Fridtjof Nansen
Beskrivelse	Ekteparet Nansen i skiantrekk anno 1890. Bildet er ikke fra Norefjell, men er arrangert med kunstig snø i fotografen atelier. Foto: L. Szacinski, Nasjonalbiblioteket/Gyldendals arkiv.
Opphav	L. Szacinski, Nasjonalbiblioteket/Gyldendals arkiv.

Figur 19. Eksempelpost: Gjenkjenning av enkle personnavn (ektepar)

Dette er ikke forsøkt løst i denne oppgaven, men man kan tenke seg mulige fremgangsmåter som kan gjøre det mulig å identifisere slike grupper av personer. På samme måte som signalordene ”Foto” og ”Fotograf” er brukt til å identifisere navn som er knyttet til opphavsinformasjon, vil det være mulig å identifisere signalord som identifiser slektsangivelser. Foruten de tidligere nevnte ordene vil ord som ”søsknene” og ”søstrene” være eksempler på slike signalord. Disse signalordene vil kunne brukes for å identifisere de tilfellene hvor enkeltnavn refererer til grupper

av personer, og ikke bare én person. Deretter må de riktige personene knyttes til den rette gruppen. For eksempel vil det være ønskelig å knytte ”Eva Nansen” og ”Fridtjof Nansen” til referansen ”Ekteparet Nansen”, mens det vil være ønskelig å knytte ”Alexander Nansen” og ”Fridtjof Nansen” til referansen ”Brødrene Nansen”.

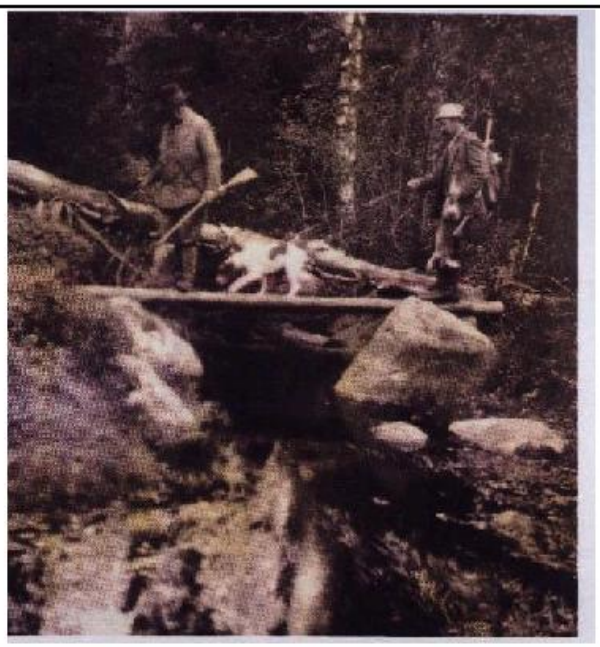
Det vil være flere mulige tilnærmingsmåter for å løse dette. Én måte vil være å kategorisere signalordene etter betydning. Ved å identifisere betydningen av signalordet, vil det være lettere å vite hva man skal se etter. For eksempel vil et ektepar bestå av to personer, mens brødre og søsken vil være alt fra to personer og oppover. Hvis man i tillegg deler opp personnavnslisten slik at fornavnene blir fordelt i to lister avhengig av kjønn, og knytter kjønn til signalordene, vil det være mulig å bruke kjønnsbestemmelse av signalordene til å indikere hvilke navn som vil være aktuelle. Signalordet ”brødrene” vil indikere at gruppen skal bestå av kun menn, mens ”søstrene” vil bestå av kun kvinner, og ”ekteparet” vil betegne én person av hvert kjønn (i hvertfall i bøker utgitt før den nye ekteskapsloven ble vedtatt). Hvis det i indekseringen av alle navn blir knyttet et kjønn til navnet, avhengig av i hvilken liste fornavnet ble identifisert, vil det være enkelt å begrense søkene etter kandidatnavn basert på kjønn.

Ved familiereferanser hvor hver person blir referert ved navn, som ved eksempelet i Figur 18, er det som beskrevet tidligere, mulig å identifisere et helt navn basert på kun et fornavn. Når man først har identifisert at det er snakk om en gruppe eller to personer som har en familierelasjon på denne måten, vil det også være mulig å registrere at disse personene er en del av en gruppe. Kombinert med kjønns- og mengdebestemmelser, vil man kunne tenke seg at man ved senere familiereferanser i teksten, av typen ”brødrene Anker-Rasch”, vil kunne hente frem informasjon om at ”Fritz Anker-Rasch” og ”Ole Anker-Rasch” begge er menn, og referert til som en gruppe ved en tidligere anledning. Basert på dette vil man kunne anta at det er disse det er snakk om også i forbindelse med denne referansen.

7.3.2.4 Feilkilder

Identifisering av personnavn i bildetekster oppnår gode verdier for både presisjon og fullstendighet, men det finnes allikevel unntak. Den vanligste feilkilden er at enkle navn blir koblet til feil fulle navn. Figur 20 viser et eksempel på dette. Bildeteksten er ”På harejakt. Peter Anker og Johan Jeger passerer en klopp”. I dette tilfellet blir ordet ”Jeger” ikke oppfattet som et navn og dermed blir ”Johan” oppfattet som et enkeltnavn. Fordi ”Jeger” ikke oppfattes som et

navn, vil ikke ”Johan Jeger” finnes i oversikten over navn i boken som helhet og andre kandidater blir foreslått. Hjalmar Johansen blir vektet høyest med en verdi over cutoff-verdien og blir valgt som indekseringsterm.



Avbildede personer	Peter Anker
	Hjalmar Johansen
Beskrivelse	På harejakt. Peter Anker og Johan Jeger passerer en klopp. Foto: Fridtjof Nansen, Nasjonalbiblioteket.
Opphav	Fridtjof Nansen,
	Nasjonalbiblioteket.

Figur 20. Eksempelpost: Gjenkjenning av personnavn (feilkilder)

Det er to problem med dette. Det første er at ”Johan Jeger” ikke blir tolket som et navn. Dette er et resultat av å bruke lister til å avgjøre om noe er et navn eller ikke. Det å bruke lister gjør at navn som ikke står på listen, ikke blir oppfattet som navn, noe som igjen fører til feilindekseringer. Dette er delvis omgått ved å si at ord som ikke finnes i navnelisten, skal kunne regnes som navn hvis det ikke finnes i de andre listene heller. Problemet for ”Johan Jeger” er at ”jeger” er et vanlig norsk ord, og dermed vil det ikke kunne regnes som navn når det ikke står på personnavnlisten. I en vanlig tekst vil dette kunne vært omgått ved å si at det kun er det første ordet i en setning som må sjekkes, for å vurdere om det er et ord når det har en stor bokstav. Som regel vil et ord som står midt i en setning, og som har en stor forbokstav, være egennavn. Men

når det gjelder det materialet som denne oppgaven jobber med, ocr-scannet tekst fra bøker, så er det ikke like enkelt. Som man kan se av Figur 6 kan feil i ocr-scanning føre til at vanlige ord midt i en setning får stor forbokstav. Dette gjør at det er vanskelig å vite nøyaktig om stor forbokstav betyr egennavn eller ikke, hvis ikke ordene står i en av listene. Den eneste måten å identifisere ukjente navn på, vil derfor være å bruke språkanalyse som kan forstå ordenes relasjoner i setninger, og som, for eksempel ved å identifisere semantiske roller, kan identifisere at et ord som ikke kan identifiseres som en person basert på ordlistene, faktisk er en person.

Det andre problemet med denne feilindekseringen er at når ”Johan” først tolkes som et enkeltnavn, så returneres navnet ”Johansen”. Man kan argumentere for at det ikke bør søkes trunkert når man skal identifisere et fullt navn fra et enkelt navn. Hvis enkeltnavnet er ”Johan” vil det være intuitivt å forvente at kandidatene er navn hvor ”Johan” utgjør et fullstendig ledd i navnet. Men samtidig er det slik at bøker har en layout som gjør at ord av og til må deles mellom linjer. Ocr-scanning vil ikke oppfatte at det delte ordet er ett ord, men vil behandle det som to. Dermed kan Hjalmar Johansen refereres til som ”Johan” ”sen”, og det vil i slike sammenhenger være ønskelig å kjenne igjen dette navnet som ”Johansen” og ”Hjalmar Johansen”. Det kan tenkes at det vil hjelpe å kunne identifisere slike delte ord, og behandle navnesøkene etter om det er snakk om delte eller udelte ord. Uansett er dette noe som bør undersøkes videre i senere undersøkelser.

7.3.2.5 Oppsummering om gjenfinning av personnavn

	Presisjon	Fullstendighet
Alle navn	0,9	0,88
Hele navn	0,95	0,9
Enkle navn	0,89	0,76

Tabell 9. Navnegjenkjenning: personnavn etter type

Metoden gir gode resultater for identifisering av vanlige hele navn i bildeteksten. Men det finnes allikevel begrensninger ved metoden, og da særlig i forbindelse med ukjente navn som ikke finnes på navnelisten. Når det gjelder å koble enkle navn med deres korresponderende fulle navn gir metoden tilsynelatende gode resultater for presisjon, og noe mindre gode resultater for fullstendighet. Men det er som sagt svakheter ved statistikkgrunnlaget som gjør at metoden bør

testes ut på andre typer bøker. Det bør også jobbes videre med å skille mellom enkle navn som er en del av et delt ord, og enkle navn som er ”hele”, slik som i tilfellet ”Johan” og ”Johan sen”.

Tabell 6 viste også at personnavnene som ble funnet i teksten, ikke alltid fungerer som indekseringstermer. Med en presisjon på 0,74 og en fullstendighet på 0,78, kan det tyde på at bildetekster ofte inneholder beskrivelser som ikke direkte relaterer seg til bildene de hører til. Samtidig er ikke 0,74 og 0,78 veldig dårlige resultat, men tallene tyder på at det kunne vært nyttig å vurdere andre kilder til indekseringen, og sammenlignet bildetekstene som kilde med annen tekst som kilde, for å se om det er mulig å oppnå en bedre indeksering ved å bruke tekst fra andre steder enn bildeteksten. Når det gjelder fullstendighetsverdien i dette tilfellet, så er det verdt å merke seg at dette kan ha en sammenheng med at det her ikke er utviklet en metode for å håndtere slektsreferanser hvor et enkelt navn refererer til flere individer, slik som ”brødrene Amundsen” og ”familien Nansen”. Det kunne derfor vært nyttig å utvikle en metode for å håndtere slike referanser og se om det har noen effekt på fullstendighetsnivået når det gjelder indekseringen.

7.3.3 Stedsnavn

	I teksten	Foreslåtte	Korrekt identifiserte	Presisjon	Fullstendighet
Stedsnavn	54	50	32	0,64	0,59

Tabell 10. Navnegjenkjenning: stedsnavn

	På bildet	Foreslåtte	Korrekt indekserte	Presisjon	Fullstendighet
Stedsnavn	39	48	24	0,5	0,62

Tabell 11. Indeksering: stedsnavn

Metoden for å identifisere stedsnavn fungerer ikke slik den er brukt i denne oppgaven. Når det gjelder å identifisere stedsnavn i teksten, oppnår metoden en presisjon på 0,64 og en fullstendighet på 0,59. I forhold til indeksering er tallet enda lavere med en presisjon på 0,5 og en fullstendighet på 0,62. Bare halvparten av de foreslåtte indekstermene er med andre ord termer som faktisk beskriver stedet på bildet.

Tid/Sted	Lysaker
	Godthåb
	Asker
Avbildede personer	Bjørnstjerne Bjørnson
Beskrivelse	Etter hvert flyttet de nygifte inn i en tømmervilla ved Lysaker, som Bjørn stjerne Bjørnson gav navnet Godthåb. Foto: Nasjonalbiblioteket. Huset ble revet for mange år siden, men tilbygget ble reddet, og står nå på en gård i Asker. Foto: Svein Solli.
Opphav	Svein Solli.

Figur 21. Eksempelpost: Gjenkjenning av stedsnavn

7.3.3.1 Feilkilder

Av de 18 termene som feilaktig er identifisert som sted, er 7 egentlig personnavn, 7 er andre typer egennavn, slik som navn på båter, produkter og ekspedisjonsnavn(se Figur 22). 4 av feilindekseringene er vanlige ord som er tolket som stedsnavn. I de tilfellene indekseringstermene egentlig var personnavn, så er disse termene også identifisert som personnavn. Dette er fordi fremgangsmåten som er brukt behandler personnavnene og stedsnavnene separat. Hvis egennavnene som var identifisert som personnavn, hadde blitt fjernet fra teksten før den ble identifisert som stedsnavn, ville dette problemet vært løst, og presisjonsverdien ville blitt høyere. Ulempen med dette ville vært at stedsnavn som ble identifisert som personnavn, ikke ville kunne blitt vurdert som stedsnavn, og dermed ville dette påvirket fullstendighetsverdien. Et eksempel

vil være stedet ”Franz Josefs land”. ”Franz Josef” blir tolket som et personnavn, og hvis personnavnene ikke skulle kunne bli vurdert som stedsnavn, ville ikke termen ”Franz Josefs land” bli identifisert som et sted. Det kunne vært mulig å vurdere personnavn og stedsnavn sammen, og ikke separat som det er gjort her. Da ville hver term blitt vurdert i forhold til om det var mest sannsynlig at det var et sted eller et personnavn. En regel kunne da vært at hvis kun deler av en navnesekvens tilsvarte et stedsnavn, mens hele sekvensen tilsvarte et personnavn, ville termen bli vurdert til å være personnavn. Hvis derimot personnavnet tilsvarte kun deler av det mulige stedsnavnet, ville stedsnavnet velges. For eksempel ville en navnesekvens som ”Thorvald Gjerdrum” kunne vurderes som personnavn: ”Thorvald Gjerdrum” eller stedsnavn: ”Gjerdrum”. I dette tilfellet tilsvarer hele sekvensen et personnavn, mens kun deler av sekvensen tilsvarer et sted; dermed ville ”Thorvald Gjerdrum” blir valgt som indekseringsterm som personnavn. Mens med en sekvens som ”Franz Josefs land” tilsvarer hele sekvensen et sted, mens kun deler av sekvensen, ”Franz Josef”, tilsvarer en person. I dette tilfellet ville dermed stedsnavnet blitt valgt.

Figur 22 demonstrerer et problem som er mer utbredt ved identifisering av stedsnavn, enn ved identifisering av personnavn; det å skille mellom stedsnavn og andre egennavn som navn på organisasjoner og objekter. I denne posten er det skuten Jason som er identifisert som et sted. Stedsnavn er vanskeligere å identifisere fordi ved personnavn kan man gå ut fra at det er flere ledd, og hvert ledd sjekkes mot personnavnslisten. Stedsnavn har veldig ofte bare ett ledd, dermed er det opptil listene å klassifisere riktig. Det å klassifisere riktig med kun lister er en vanskelig oppgave fordi det ikke er uvanlig at et ord kan representere flere typer egennavn. Organisasjoner kan ha stedsnavn som en del av sitt organisasjonsnavn, og stedsnavn kan ha navn som også brukes om organisasjoner og navngitte objekter. Denne oppgaven har for så vidt ikke hatt som fokus å klassifisere egennavn, men når en del av formålet var å kunne indeksere bilder også med stedsnavn, er det verdt å påpeke at en korrekt indeksering av bildene i forhold til stedsnavn vil være vanskelig uten å bruke mer avanserte teknikker enn det som er brukt her. For eksempel vil bruken av semantiske roller kunne være med på å skille mellom egennavn som tilhører sted, og egennavn som tilhører andre kategorier.

Tid/Sted	Jason
Beskrivelse	Skuta Jason skulle primært drive selfangst, og ekspedisjonen måtte finne seg i å komme i annen rekke. Foto: Fridtjof Nansen, Nasjonalbiblioteket.
Opphav	Fridtjof Nansen, Nasjonalbiblioteket.

Figur 22. Eksempelpost: Gjenkjenning av stedsnavn (sted vs. andre egennavn)

Når det gjelder den lave fullstendighetsverdien for gjenkjenning av stedsnavn, er det flere aspekter som spiller inn på dette. Lav fullstendighet skyldes at stedsnavn i teksten ikke identifiseres som stedsnavn, slik som i Figur 19 og Figur 25. En av hovedgrunnene til dette er at flere av stedsnavnene som ikke blir gjenkjent, ikke står på listen over stedsnavn. Dette gjelder for eksempel viktige stedsnavn som ”Svalbard” og ”Spitsbergen”. Geonames.org ser ikke ut til å ha inkludert Svalbard som en del av Norge i sin liste over norske stedsnavn, og disse stedsnavnene er heller ikke på yr.no sin liste over viktigste steder i Norge. For å kunne fange opp slike steder er det dermed nødvendig å revidere listen som brukes for å identifisere stedsnavn. Men selv om disse termene ikke står på listen over stedsnavn, står de på listen over norske ord. Dette er veldig uheldig siden det er listene som brukes for å vurdere hva som er stedsnavn og hva som er ord. Hvis et stedsnavn er inkludert i ordlisten, og ikke i stedsnavnslisten, er det dobbel grunn til at det ikke identifiseres korrekt som et stedsnavn. Det er derfor også nødvendig å revidere ordlisten og fjerne eventuelle egennavn som er inkludert i den. Dette demonstrerer et av minusene ved å basere seg på lister i arbeidet med å gjenkjenne egennavn. Å få mest mulig fullstendige lister vil være enklere jo mer avgrenset en base er; det ville for eksempel vært en ide å utvikle lister som er mer fullstendige for de emnene og områdene dokumentsamlingen omhandler, og la mer perifere

områder være mindre representert i listene. Men det å utvikle slike lister vil kreve mye mer ressurser enn det å bruke ferdige lister slik det er gjort i dette tilfellet.

Tid/Sted	Godthåb
Avbildede personer	Otto Sverdrup
	Fridtjof Nansen
Beskrivelse	Sverdrup og Nansen underveis til Godthåb med båten de laget av vidjekvister og seilduk. Foto: Fridtjof Nansen, Nasjonalbiblioteket/Gyldendals arkiv.
Opphav	Fridtjof Nansen,
	Nasjonalbiblioteket/Gyldendals arkiv.

Figur 23. Eksempelpost: Gjenkjenning av stedsnavn (feilkilder)

En stor utfordring knyttet til stedsnavn som indekseringstermer, er at stedsnavn, i større grad enn personnavn, ser ut til å refereres til i bildetekster når de ikke har direkte tilknytning til bildet. Dette gjelder for eksempel posten i Figur 17. Her omtales Nansens ferd til Nordpolen, selv om bildet er tatt i Norge. Det samme gjelder i Figur 23, hvor bildet viser Sverdrup og Nansen *underveis* til Godthåb, og dermed indekseres bildet med termen ”Godthåb”. Det kan tenkes at denne koblingen kan oppfattes som relevant selv om stedet ikke direkte er avbildet, fordi det er i forbindelse med Nansens ferd til Godthåb, men det er også like tenkelig at brukere vil oppfatte dette som støy. En variant av dette problemet er når stedsnavnet negeres i teksten, slik det blir i Figur 19. I dette tilfellet ble ikke Norefjell identifisert som et sted, men hvis det hadde blitt identifisert som et sted, ville indekseringen blitt feil, fordi det i dette tilfellet står ”Bildet er ikke fra Norefjell, men er arrangert med kunstig snø”.

7.3.4 År

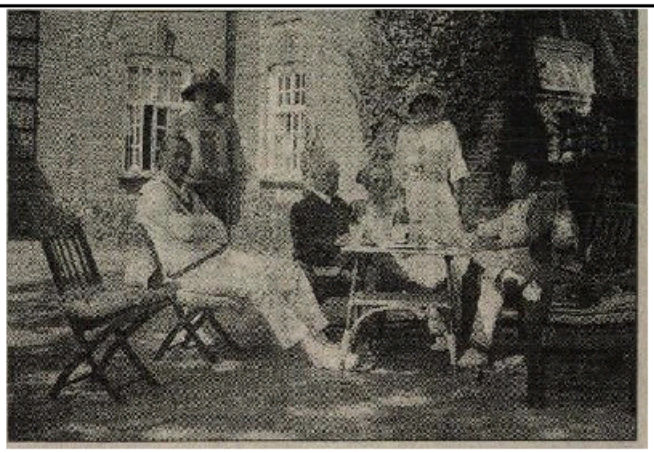
Tid/Sted	1870.
Avbildede personer	Fridtjof Nansen
Beskrivelse	Fridtjof Nansen omkring 1870. Det var på denne tiden han lærte Nord marka å kjenne. Foto: L. Szacinski, Nasjonalbiblioteket/Gyldendals arkiv.
Opphav	L. Szacinski, Nasjonalbiblioteket/Gyldendals arkiv.

Figur 24. Eksempelpost: Gjenkjenning av år

Identifisering av årstall er det som fungerer best av all tekstanalysen. Når det gjelder identifisering av årstall i teksten, oppnår man ved denne metoden en presisjon og en fullstendighet på 1 (Tabell 1). Det vil si at der det finnes årstall i bildeteksten, blir det årstallet gjenkjent og identifisert korrekt, slik man også kan se av Figur 13, Figur 19 og Figur 24. I dette materialet fantes det ingen eksempler på andre tidsangivelser enn rene årstall, men metoden er utviklet for også å oppdage perioder og mer omtrentelige angivelser som 1900-tallet og 1930-årene. Posten som man kan se i Figur 25 er hentet fra en annen bok enn den statistikken forholder seg til. Denne posten viser at tidsangivelser som ”1920-årene” også identifiseres som en tidsangivelse, men dette er som sagt ikke med i statistikken som er vist nedenfor.

	I teksten	Foreslåtte	Korrekt identifiserte	Presisjon	Fullstendighet
År	26	26	26	1	1

Tabell 12. Identifisering av år



Tid/Sted	England
	1920-årene.
Avbildede personer	Champagne
	Yngvar Nielsen
	Eileen
	Charles Peto Bennett
	Astrid
	Trygve
	Kiss Bennett
	Leigh Court
Beskrivelse	Champagne i haven. Fra venstre: ekteparet Niels og Eileen gubbe England ble fa milienavnet skrevet Gude), Charles Peto Bennett med ekteparet Astrid og Trygve gubbe. Bak til høyre: Kiss Bennett. Bildet er tatt på Leigh Court en sommerdag i 1920-årene.

Figur 25. Eksempelpost: Gjenkjenning av tidsperiode.

Fra Bomann-Larsen (1995)

	På bildet	Foreslåtte	Korrekt indekserte	Presisjon	Fullstendighet
År	20	26	20	0,77	1

Tabell 13. Indeksering: år

Når det gjelder å bruke disse årstallene som indekseringstermer for bildene, ser vi de samme tendensene her som ved personnavn og stedsnavn. En presisjon på 0,77 indikerer at også når det gjelder årstall, forekommer det at teksten som er i bildeteksten, ikke direkte relaterer seg til det som faktisk er på bildet. Et eksempel på dette kan vi se i Figur 26, hvor det i teksten er to årsreferanser, 1899 og 1908. Disse årstallene refererer til det årsspennet da Fridtjof Nansen

brukte stedet på bildet, og det er ingen indikasjon på at bildet faktisk er fra noen av de to årene som er referert.

Tid/Sted	Rollag 1899, 1908
Avbildede personer	Fridtjof Nansen
Beskrivelse	Sommerstedet Sørkje i Rollag ble kjøpt av Fridtjof Nansen i 1899, og flittig brukt av familien. Hver eneste høst fram til 1908 var han på jakt i fjellet i disse traktene. Hytta bakerst ble bygget som ei arbeidsstue for Nansen. Foto: A.B. Wilse, Nasjonalbiblioteket.
Opphav	A.B. Wilse, Nasjonalbiblioteket.

Figur 26. Eksempelpost: Gjenkjenning av år (feilkilder)

Det at fullstendigheten er 1, betyr ikke at alle bildene har blitt datert. Det det betyr, er at der det har blitt ansett som mulig å datere et bilde, basert på opplysningene i teksten for øvrig, så har også metoden klart å identifisere årstallet. Dette er et problematisk mål, fordi det baserer seg på at det er teksten som har bestemt kriteriet for om et bilde blir regnet for å være daterbare eller ikke, og det er teksten som brukes som grunnlag for dateringen. I motsetning til personer, hvor det er lettere å avgjøre om et bilde bør indekseres med personnavn eller ikke, så er det vanskeligere å avgjøre om et bilde bør indekseres med årsangivelse eller ikke. Ideelt sett burde alle bilder indekseres med årstall, men det vil i praksis være umulig, fordi mange av bildene ikke er datert i utgangspunktet. Fordi vurderingen av hvilke bilder som burder vært indeksert med år, er gjort som den er, så sier ikke fullstendighetsmålet som er oppgitt her, noe annet enn

fullstendighetsmålet i Tabell 12, nemlig fullstendigheten av årsangivelser i teksten som blir fanget opp av metoden.

Med et datagrunnlag på kun 26 årstall i teksten, og kun 20 bilder som det er mulig å datere, er grunnlaget altfor lite til å kunne si noe definitivt om metoden. Det må allikvel være lov å si at metoden ser lovende ut i forhold til både presisjon og fullstendighet når det gjelder å identifisere spesifiserte tidsangivelser i teksten, men at det at presisjonen for indekseringen er så lav støtter opp om presisjonstallene for personnavn, og for så vidt også stedsnavn, som tyder på at bildeteksten ikke i alle tilfeller vil være et ideelt grunnlag for indeksering.

7.4 Oppsummering

Resultatene for gjenkjenning av personnavn og årsangivelser er meget lovende, med presisjon og fullstendighet på 0,9 og 1, men samtidig er det mye som må gjøres for å kunne identifisere stedsnavn korrekt i teksten. Noen av problemene med metoden er drøftet, og noen forslag har blitt presentert for hva som kan gjøres for å få et bedre resultat innenfor stedsnavngjenkjenningen. Samtidig viser resultatene at det at egennavnene kan gjenkjennes i teksten ikke er ensbetydende med at disse egennavnene vil utgjøre gode indekstermer. I denne undersøkelsen har det vært et fokus på å jobbe med bildetekster som grunnlag for indeksering. Dette var delvis basert på at det har vært et stort fokus i litteraturen og forskningen på å bruke bildetekster til tekstlig indeksering av bilder. Men det finnes kritiske røster som mener at det bør settes spørsmålstegn ved bildetekstens egnethet i indekseringen. Srihari og Zhang er blant de som påpeker de mulige farene ved å bruke bildeteksten som utgangspunkt (1999, s. 498):

> *"Searching captions for keywords and names will not necessarily yield the correct information, as objects mentioned in the caption are not always in the picture. This results in a large number of false positives that need to be eliminated or reduced"*

Dette poenget blir til dels støttet opp om av noen av resultatene fra denne undersøkelsen. Personnavn som indekseringstermer fikk en presisjon på 0,74 og en fullstendighet på 0,78. Stedsnavn oppnådde en presisjon på 0,5 og en fullstendighet på 0,62, mens årstall oppnådde en presisjon på 0,77 og en fullstendighet på 1. Fullstendighetstallet for årstall kan man se bort fra fordi det ble vanskelig å avgjøre hvilke bilder som burde vært indeksert med årstall. De dårlige resultatene for stedsnavn henger sterkt sammen med at metoden for å identifisere stedsnavn i

teksten var så dårlig, at flere av de foreslåtte stedsnavnene ikke ville blitt brukt som indekseringstermer hvis metoden for å kjenne igjen stedsnavn hadde vært bedre. Men selv når det gjelder personnavn og årstall, som oppnådde gode resultater i forhold til å bli gjenkjent i teksten, så er presisjonen nede i 0,74 og 0,77. Mye av grunnen til dette er, som Srihari og Zhang foreslo, at personer, steder og år som blir nevnt i bildeteksten, ikke forekommer på bildet. Men samtidig betyr disse tallene at riktig person blir indeksert på et bilde i 3 av 4 tilfeller.

I presentasjonen av resultatene ble boken *Jegeren Fridtjof Nansen* brukt som eksempel fordi det var en av bøkene med best resultat for identifisering av bildetekster. Denne boken har i tillegg en del egenskaper som gjør at den passer bra til denne undersøkelsens fokus. Bildene som er brukt i boken har til dels en dokumentarisk bildefunksjon og dette reflekteres i bildeteksten. Bildetekstene har i stor grad personnavn, stedsnavn og årstall inkludert i beskrivelsene av bildene. Dette gjorde at boken var velegnet til å teste ut om metoden for å gjenkjenne egennavn fungerte, og om egennavn som forekom i bildeteksten var relevant for det bildet bildeteksten stod til. Det ble besluttet å bruke denne boken til testingen fordi det ble ansett som viktigere å kunne ha et materiale hvor metoden virkelig kunne bli testet ut, enn å bruke en tilfeldig bok for å se hvordan metoden fungerte i et slikt tilfelle. Å bruke kun én bok som kilde vil uansett ikke kunne gi representative data, og da ble det ikke ansett som så nødvendig å etterstrebe representativitet i valg av bok. Dette valget har gjort at det må tas noen forbehold i forhold til resultatene som er presentert. Andre typer bøker vil ikke nødvendigvis ha bildetekster som er like gode kilder, enten fordi de ikke har bildetekster i det hele tatt, eller fordi bildeteksten ikke inneholder egennavn. Det kan også være at bildene har en funksjon i forhold til teksten som gjør at bildeteksten ikke vil beskrive et bilde på samme måte som bildetekstene i den brukte boken gjør. Bildetekster som inneholder mange egennavn, tilhører ofte bilder som har en dokumentarisk funksjon i forhold til teksten. Denne typen bildetekst ligner muligens i større grad på den typen beskrivelse en manuell indekserer ville tilordnet et bilde enn det en tilfeldig bildetekst vil gjøre, og det gjør det lettere å trekke ut informasjon automatisk. Bildetekstene er dermed en god kilde for tekstanalyse fordi bildeteksten i stor grad beskriver bildene med tanke på å beskrive hva bildet viser.

På samme måte som Marsh presenterte ulike funksjoner for bilder i sin taksonomi (kapittel 2.2.3), vil det være nærliggende å tro at bildetekstene vil ha forskjellige funksjoner og karakteristika avhengig av hvilken funksjon bildet skal oppfylle i teksten. Bildetekster som

”«Sjef» stod det skrevet over hele dyret”[3] til et bilde av en isbjørn vil gi en annen type indeksering enn bildetekster til et dokumentarisk bilde. I dette tilfellet referer bildeteksten til en egenskap ved dyret som er åpenbar når man ser bildet, men som ikke overføres til teksten når man ser bort fra bildet. Selv om bildeteksten beskriver en egenskap ved bildet, er det lite sannsynlig at det var en isbjørn brukeren tenkte på når hun søkte på ordet ”sjef”. Men samtidig vil sjef være en ikonologisk beskrivelse av bildet, og visse situasjoner kan det være ønskelig å ha mer abstrakte beskrivelser. Faren med ikonologiske beskrivelser basert på automatisk indeksering er at fordi bildeteksten er skrevet med tanke på at bilde og teksten skal fremstå i en sammenheng, så kan det oppstå en ”lost in translation”-effekt når teksten skal representere bildet på egenhånd. I andre tilfeller hvor bildeteksten har en mer beskrivende funksjon, slik som bildeteksten ”Ekteparet Nansen i skiantrekk anno 1890” i Figur 19, vil nytten av å bruke bildeteksten som indekseringsgrunnlag være mer åpenbar. Denne bildetekst inneholder flere indekstermer man kunne tenke seg ville blitt brukt i en manuell indeksering. Særlig er fokuset på personnavn, stedsnavn og årstall i bildetekstene i denne boken velegnet for den typen automatisk indeksering som er beskrevet. Nå som metoden er testet i forhold til om den kan identifisere egennavn der man vet at det er mange egennavn i bildetekstene, vil det være naturlig å se om metoden kan overføres til andre typer bøker. I tillegg vil det være nødvendig å utvikle metoden videre til å inkludere andre typer beskrivelser. I bøker som bruker bilder på en annen måte enn som dokumentarisk illustrasjon, kan det være nyttigere å indeksere med andre fokus enn det som er lagt til grunn i denne oppgaven, både i forhold til hva som indekseres, og hva som brukes som analysegrunnlag. Slike andre fokus blir diskutert nærmere i kapittelet om videre forskning.

8 Videre forskning

Forsøkene som er gjort her skraper så vidt i overflaten av mulighetene når det gjelder automatisk tekstlig indeksering av bilder i digitale bøker. Fokuset i undersøkelsen har blitt bestemt ut fra hva tidligere forskning på relaterte områder har vist gir resultater og hva brukerbehovet er. Samtidig er det mange aspekter ved dette området som det vil være nyttig og ønskelig å undersøke, som denne oppgaven ikke har vært innom. For det første er ikke bildetekstens potensial som

[3] Fra *Svalbard - fangstfamilien på 79 grader nord* av Stein P. Aasheim

informasjonskilde fullstendig utforsket. For det andre finnes det i teksten for øvrig informasjon som kan utnyttes til bildeindekseringen. De mulighetene som ligger her bør utforskes i senere forskning. I denne oppgaven er det sett på automatisk tekstlig indeksering av bilder basert på omkringliggende tekst. I videre forskning kunne det vært interessant å se utover teksten og benytte andre kilder til indekseringen, dette gjelder både andre kilder til tekstlig informasjon og kilder som gir annen type indeksering. Sist, men ikke minst vil det være ønskelig å generere metadata utover emneindeksering. I dette kapittelet diskuteres det hvilke muligheter som bør utforskes videre innenfor de nevnte tilnærmingsmåtene.

8.1 Videre analyse av bildetekst

Fokus for denne undersøkelsen har vært å finne og bruke bildetekster til indekseringen av bildene. Selv om dette har vært fokus for oppgaven finnes det allikevel mange uutforskete muligheter i å bruke bildetekster som informasjonskilde i forbindelse med bildeindeksering. Metoden som er brukt for å identifisere bildetekster i bøkene, har gitt lovende resultater, noe som gjør at det vil være interessant å undersøke hvordan teksten i bildeteksten kan utnyttes når man først har identifisert den. Denne oppgaven har kun lett etter ikonografisk informasjon, mer spesifikt personnavn, stedsnavn og årstall. Det er ikke gjort forsøk på å identifisere andre typer ikonografisk informasjon som organisasjonsnavn eller navngitte objekter, hendelser og begivenheter. Når det er sagt, så klarer metoden som er presentert, å identifisere mange av egennavnene, men den klarer ikke alltid å skille mellom for eksempel stedsnavn og andre typer av egennavn, som for eksempel skipsnavn og ekspedisjonsnavn. Dette er også ikonografiske beskrivelser, men ikke av de typene denne undersøkelsen hadde som formål å identifisere. Mange av egennavnene som representerer en ikonografisk beskrivelse, er derfor identifisert ved den metoden som er presentert, men det som mangler er en bedre metode for å kunne skille mellom de ulike fasettene til Shatford (Figur 1). Fordi egennavn av typen produktnavn, organisasjonsnavn og stedsnavn har mange fellestrekk i forhold til struktur, er det sannsynlig at man må bruke semantisk analyse for å kunne skille mellom de forskjellige, slik for eksempel Røyneberg gjør i sin oppgave (2004).

I tillegg til ikonografiske beskrivelser bør det undersøkes om det er mulig å finne pre-ikonografiske beskrivelser i bildetekstene. Pre-ikonografiske beskrivelser er, som tidligere nevnt, generelle begrep der ikonografiske beskrivelser er navngitte personer og hendelser. Mens skuta

”Jason” i Figur 22 tilsvarer en ikonografisk beskrivelse for fasetten ”hvem”(personer og ting), vil ”båten” i Figur 23 være et eksempel på en pre-ikonografisk beskrivelse for samme fasett. Ifølge brukerstudier er det ikonografiske og pre-ikonografiske termer flest brukere søker etter når de leter etter bilder (Tsai, 2006), og når denne oppgaven har fokusert på ikonografisk informasjon, vil det være formålstjenlig å senere se på om det er mulig å også finne pre-ikonografisk informasjon i bildetekstene. Dette vil være en større utfordring enn å lete etter ikonografisk informasjon fordi man med pre-ikonografiske beskrivelser ikke har fordelen av tydelig syntaks for å skille ut denne typen informasjon fra annen tekst, slik man har med ikonografiske beskrivelser av typen egennavn.

8.2 Tekst utover bildetekst

Selv om man kan trekke ut enkelte gode indekseringstermer fra bildetekstene, vil det ikke alltid være slik at bildeteksten er det beste analysegrunnlaget. Som tidligere nevnt kan man ikke gå ut i fra at bildeteksten har en funksjon som gjør den egnet til indeksering. I denne undersøkelsen oppnådde indekstermene en presisjon og fullstendighet rundt 0,75. Det kunne derfor vært interessant å se om det var mulig å oppnå bedre resultater ved å bruke annen tekst som indekseringskilde. I tillegg til at bildeteksten kan være upålitelig, må man også ta hensyn til at ikke alle bilder har bildetekst, uten at det gjør det mindre ønskverdig å beskrive bildene for gjenfinning. Det gjelder både bilder som ikke har en bildetekst i boken, men også bildetekster som ikke blir oppfattet i skanneprosessen. Fordi det ikke er gitt at man kan finne en bildetekst for å indeksere et bilde, og fordi bildeteksten ikke alltid vil være det beste grunnlaget, må man kunne indeksere bildene tekstlig på en annen måte i tillegg til å utnytte de mulighetene som finnes ved å finne bildeteksten. I digitale bøker kan man tenke seg at det finnes tekst som ikke kan defineres som bildetekst, men som kan være verdifull som kilde i bildeindekseringen. Dette gjelder avsnitt som kan representere alternative bildetekster ved at de refererer til et bilde, og tekst som ikke har eksplisitte referanser til et bilde, men som allikevel kan inneholde informasjon som kan brukes til å beskrive bildets innhold.

8.2.1 Alternative bildetekster

I bøker finnes det flere eksempler på indirekte bildetekster hvor bildet blir referert til i teksten uten at det er formatert som en bildetekst. Det kan være forskjellige grunner til dette. Av og til gjør bokens layout at bildet og teksten ikke kan plasseres på samme side og det er derfor

nødvendig å referere til den i tekst på andre sider. Andre ganger er bildet så tett knyttet til teksten at det er naturlig å referere til bildet i den løpende teksten. Slike indirekte bildetekster kan være både et tillegg til, og en erstatning for, vanlige bildetekster. En retning det kunne være interessant å undersøke er å finne disse indirekte bildetekstene, og analysere dem på samme måte som de direkte bildetekstene. Det finnes visse elementer det går an å bruke til en slik undersøkelse; hvis bildet og den indirekte bildeteksten er plassert langt fra hverandre i boken kan det finnes en tydelig sidereferanse (”se bilde s. 74”, ”Bildet på side 38 viser…”) som det vil være mulig å utnytte.

Selv når det ikke finnes så eksplisitte referanser kan det finnes elementer som indikerer en bildereferanse. Rowe (1997 s 101-102) bruker termene *depictability*, *reference phrases* og *signal words* om termer som indikerer at de tilhører en bildetekst. Slike referansefraser kan være ”bildet viser”, ”foto”, ”figur”, ”illustrasjon” osv. På Figur 27 kan man i nederste avsnitt se eksempel på dette: ”På bildet er det Alfred Svendsen og Georg Bjørnnes som møtes (eller tar farvel) ved Villa Møen 15. April 1928. Svendsen lå i Rekvika, lengst nord i Wijdefjorden”. Dette eksempelet viser et tilfelle hvor teksten muligens kan regnes som bildetekst, men fordi den er så lang og så lik et vanlig avsnitt, vil den ikke fanges opp av den metoden som er brukt i denne oppgaven. Å inkludere søk etter referansefraser kan fange opp ekstra avsnitt som relaterer seg direkte til bildene. Det å finne slike termer i en setning eller et avsnitt, kan indikere at setningen eller avsnittet referer til et bilde eller en illustrasjon i boken, og teksten rundt disse frasene vil derfor kunne inneholde tekst som kan brukes til å indeksere innholdet i det aktuelle bildet, slik som på Figur 27.

Villa Møen

Når vi opplevde Wijdefjorden som øde, hvordan må det ikke da ha vært i tidligere tider – med hjelpemidler og komfort som lå fjernt fra det vi kunne nyte. Desto mer imponerende er det at Wijdefjorden en gang var et av de mest populære fangstområdene på Svalbard. Fra 1920 til 1930 overvintret det til sammen 19 lag med i alt 47 fangstfolk i Wijdefjorden. Hvert år lå det en håndfull karer fordelt på forskjellige fangsthytter mellom Verlegenhuken og Austfjordneset.

Mest folksomt var det vinteren 1928. Da lå det syv mann fordelt på fire hovedstasjoner i Wijdefjorden: Austfjordneset, Krosspynten, Bangenhuk og Rekvika. (De to siste ligger sør for Verlegenhuken ved utløpet av fjorden.) I tillegg lå to mann på Roosneset i Woodfjorden.

På bildet er det Alfred Svendsen og Georg Bjørnnes som møtes (eller tar farvel) ved Villa Møen 15. april 1928. Svendsen lå i Rekvika, lengst nord i Wijdefjorden.

Figur 27. Eksempelside: signalord. Fra Aasheim **(2003)**

En utfordring knyttet til dette vil være å identifisere hvilken og hvor mye tekst det vil være interessant å hente når man finner slike fraser. På Figur 27 kommer referansefrasen først: ”På bildet er det Alfred Svendsen …”, men frasen kan også komme etter den relevante teksten, som i dette eksempelet fra *Svalbard : vårt nordligste Norge* (1984):

> ”Den ensomste av alle de stasjoner tyskerne opprettet på Svalbard under krigen var allikevel «Haudegen», den som under ledelse av dr. W. Dege, velkjent på Svalbard fra før krigen, ble etablert i Rijpfjorden på Nordaustlandet. Stasjonen var uten dramatikk i effektiv drift frem til høsten 1945 da den ble hentet av et norsk marinefartøy (se bilde s. 218)”.

Hvorvidt frasen står før eller etter den relevante teksten kan kanskje løses ved å se på om frasen er i slutten av en setning eller i begynnelsen av en setning. Men nøyaktig hvor mye tekst som

hører til teksten er vanskeligere å avgjøre. I det siste eksempelet vil både den setningen referansefrasen forekommer i, og setningen som kommer før denne, kunne være relevant for bildeindekseringen. I eksempelet fra Figur 27 vil kun selve setningen referansefrasen tilhører være direkte relevant for bildet. I tillegg vil det være en potensiell utfordring i å skille mellom de tilfellene hvor frasene faktisk viser til et bilde, og de tilfellene hvor bruken av termen er tilfeldig og må regnes som støy.

Det å bruke slike referansefraser kan brukes som et supplement til den metoden som er beskrevet i denne oppgaven på flere måter. En måte kan være å lete etter avsnitt som inneholder referansefraser parallelt med å lete etter bildetekster, og indeksere basert på begge avsnittene hvis det blir funnet avsnitt med referansefraser. Et alternativ vil være å inkludere referansefraser som et element i vektingen slik at avsnitt som inneholder referansefraser, kan konkurrere med de "ekte" bildetekstene om å bli vurdert som bildetekst. Dette kan gjøre at bildetekster som inneholder informasjon som ikke er verdifull i en indekseringssituasjon, kan bli erstattet av avsnitt med referansefraser som kanskje inneholder bedre indekseringstermer. Vektingsalgoritmen som er brukt her måtte i et slikt tilfelle bli bearbeidet slik at avsnitt fra flere sider kan analyseres for referansefraser, siden slike avsnitt ikke trenger å forekomme på samme siden som bildet. I tillegg burde algoritmen utformes slik at avsnitt med referansefraser får en vekt som er i stand til å konkurrere med de kombinerte vektene som avsnitt på samme side som bildet får. Det å bruke avsnitt med referansefraser som eneste indekseringskilde, vil nok ikke være formålsnyttig, fordi det er langt fra alle bilder som blir omtalt på en slik måte, og det må være et mål å indeksere alle bildene i hver bok. Derfor ville denne tilnærmingsmåten nok fungere best som et supplement til andre beskrivelseskilder.

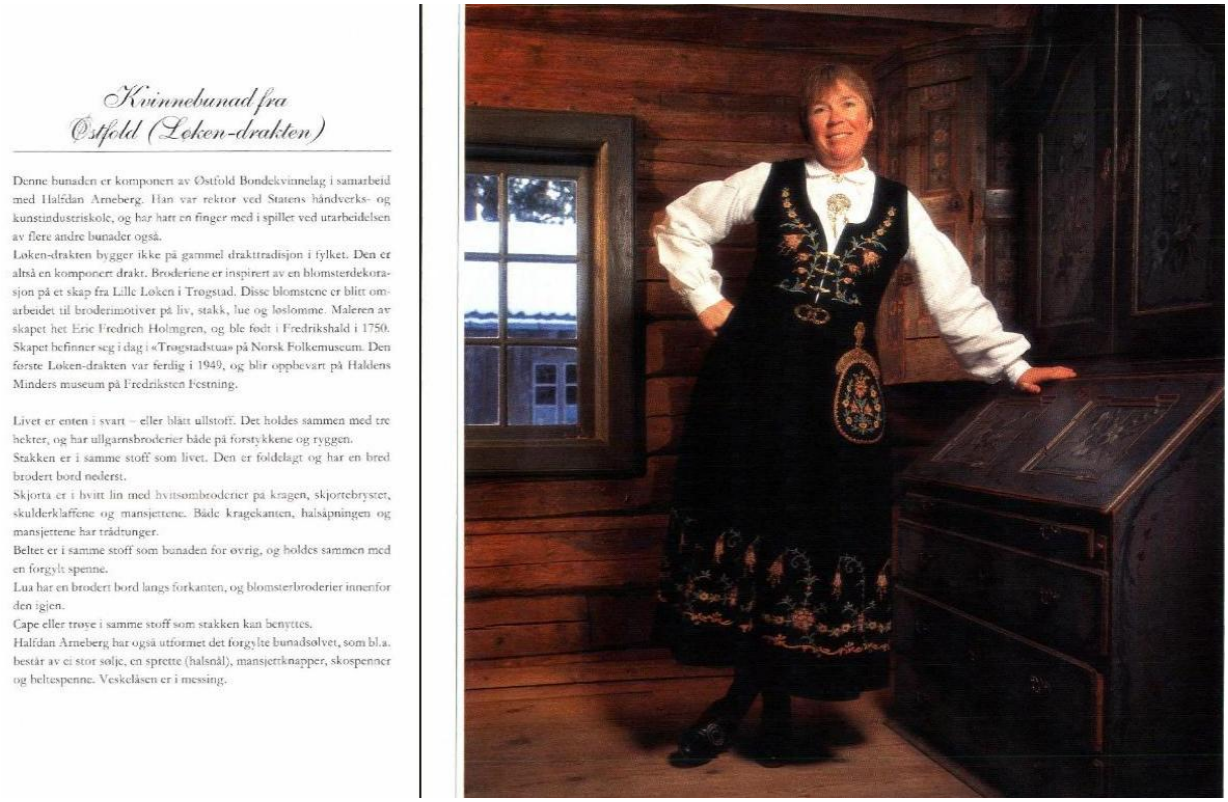
8.2.2 Analyse av annen tekst

En annen tilnærming vil være å undersøke om tekst som ikke har noen klar referanse eller tilhørighet til et bilde, kan brukes til indekseringen av bildene i boken. "Et bilde sier mer enn 1000 ord", sies det, og det er det vanskelig å oppsummere i en bildetekst. Som regel vil bildeteksten fokusere på visse elementer ved bildet, og dermed ignorere andre elementer. Automatisk indeksering vil aldri kunne fange opp alle aspekter ved et bilde, men det kan være at det vil være mulig å fange opp flere aspekter ved å supplere bildetekstindeksering med andre tekstkilder. Enkelte bøker bruker heller ikke bildetekster, direkte eller indirekte, og i disse

tilfellene vil en metode som ikke utelukkende baserer seg på at det eksisterer bildetekst, være nødvendig for å fange opp den best mulige beskrivelsen av bildet.

For å supplere bildetekstindeksering, vil det for eksempel kunne være interessant å bruke bildeteksten som et utgangspunkt for å identifisere mulige syntetiske synonym. Etter å ha identifisert en bildetekstkandidat, kan den funne bildeteksten analyseres for å finne ut hvilke ord som er spesielle for bildeteksten, i forhold til resten av teksten. Denne informasjonen kan så brukes videre, for eksempel ved å bruke latent semantisk indeksering eller andre metoder for å finne syntetiske synonym. Man kan tenke seg at det ved å bruke slike metoder kan være mulig å finne termer som ofte forekommer sammen med termene i den gjenfunne bildeteksten. Det er risiko for å få feilaktige indekseringstermer ved å bruke en slik metode, fordi det ikke er gitt at bildets innhold kan relateres til de syntetiske synonymene, men samtidig er det en mulighet det kan være interessant å utforske for å få en mer fullstendig bildeindeksering enn det som kan oppnås ved å bruke bildetekst alene.

I de tilfellene det ikke eksisterer bildetekst i boken, eller om man ønsker å supplere bildetekstindekseringen, kan man tenke seg at det kan være relevant å også undersøke tekst som befinner seg nær bildet, uten nødvendigvis å være bildetekst eller inneholde bildereferanser. Ved å etablere et vindu på en gitt størrelse rundt bildet, vil man enten kunne analysere teksten for indekstermer direkte, eller behandle den grundigere, for eksempel ved å foreta en statistisk analyse. Den statistiske analysen kan brukes til å identifisere hvilke termer som forekommer oftere rundt bildet enn i bokens tekst for øvrig, og som man dermed kanskje kan anta vil relatere seg til bildet. Bildet på Figur 28 har en bildetekst som lyder ”Fra Trøgstadstua på Norsk Folkemuseum” (bildeteksten var for langt ned på siden til å få med på utsnittet). Selv om en slik beskrivelse kanskje kan være interessant for de som er ute etter bilder tatt på Norsk Folkemuseum, vil kanskje vel så mange synes at en beskrivelse basert på teksten til venstre vil være vel så informativ. Denne teksten beskriver både hva slags bunad det er snakk om, hvor den kommer fra og hva som er særtrekkene ved denne bunaden i forhold til andre bunader.



Kvinnebunad fra Østfold (Løken-drakten)

Denne bunaden er komponert av Østfold Bondekvinnelag i samarbeid med Halfdan Arneberg. Han var rektor ved Statens håndverks- og kunstindustriskole, og har hatt en finger med i spillet ved utarbeidelsen av flere andre bunader også.
Løken-drakten bygger ikke på gammel drakttradisjon i fylket. Den er altså en komponert drakt. Broderiene er inspirert av en blomsterdekorasjon på et skap fra Lille Løken i Trøgstad. Disse blomstene er blitt omarbeidet til broderimotiver på liv, stakk, lue og løslomme. Maleren av skapet het Eric Fredrich Holmgren, og ble født i Fredrikshald i 1750. Skapet befinner seg i dag i «Trøgstadstua» på Norsk Folkemuseum. Den første Løken-drakten var ferdig i 1949, og blir oppbevart på Haldens Minders museum på Fredriksten Festning.

Livet er enten i svart – eller blått ullstoff. Det holdes sammen med tre hekter, og har ullgarnsbroderier både på forstykkene og ryggen.
Stakken er i samme stoff som livet. Den er foldelagt og har en bred brodert bord nederst.
Skjorta er i hvitt lin med hvitsombroderier på kragen, skjortebrysten, skulderklaffene og mansjettene. Både kragekanten, halsåpningen og mansjettene har trådtunger.
Beltet er i samme stoff som bunaden for øvrig, og holdes sammen med en forgylt spenne.
Lua har en brodert bord langs forkanten, og blomsterbroderier innenfor den igjen.
Cape eller troye i samme stoff som stakken kan benyttes.
Halfdan Arneberg har også utformet det forgylte bunadsølvet, som bl.a. består av ei stor sølje, en sprette (halsnål), mansjettknapper, skospenner og beltespenne. Veskelåsen er i messing.

Figur 28. Eksempelside: omkringliggende tekst. Fra Fossnes (1993)

8.3 Andre metadata og eksterne kilder

I denne oppgaven er det kun indeksering eller emnekatalogisering som har vært fokuset. Det betyr ikke at det ikke er behov for å katalogisere bildene også deskriptivt. For at bildene skal kunne brukes og være tilgjengelige, er det nødvendig å generere og registrere metadata som fotograf og rettighetshavere. En viktig retning innen videre forskning må derfor være å undersøke hvordan bildene i de digitale bøkene kan utstyres også med deskriptiv metadata. Noen eksempler på slike metadata er funnet også i denne undersøkelsen, som et biprodukt av å lete etter personnavn. For å skille ut opphavspersoner fra avbildede personer, ble det brukt referansefraser for å identifisere opphavspersoner i bildeteksten (kapittel 7.3.2.1). Dette kan det bygges videre på, ved å finne flere eksempler på slike referansefraser, og ved å lete andre steder enn i bildeteksten. Spørsmål som må stilles innenfor dette området er hvor mye av slik informasjon det er mulig å generere gjennom de digitale bøkene alene, og hva som må hentes fra eksterne kilder. Og hvordan skal man kunne knytte informasjon fra eksterne kilder til de ulike bildene automatisk hvis man ikke kan hente tilstrekkelig identifiserende informasjon fra dokumentene? En type eksterne kilder man kan forestille seg kan være interessante å bruke til et slikt formål er

bildebaser hvor man kan forvente at noen av bildene i bøkene er registrert. I forhold til Nordområdesamlingen finnes det bildebaser som Fridtjof Nansens og Roald Amundsens bildearkiv, hvor mange av bildene som er brukt i bøkene om disse polfarerne også er registrert. For eksempel ligger bildet i Figur 23 i Fridtjof Nansens bildearkiv. Posten i bildearkivet vises i Figur 29. Å utnytte slike databaser kan gjøres både tekstlig, basert på en innledende tekstlig indeksering, eller ved å kombinere tekstbaserte og innholdsbaserte metoder for å identifisere hvilke bilder som er registrert i den eksterne basen.

3b073

Emne:	EKSPEDISJONER OG REISER, Grønland
Personer i bildet:	Sverdrup, Otto Neumann Knoph (1854-1930) Nansen, Fridtjof (1861-1930)
Dato:	Oktober-1888
Sted:	Grønland, Lysefjord (Ameralik)
Motiv:	Otto Neumann Knoph Sverdrup og Fridtjof Nansen i dukbåten, på vei ut av Ameralikfjorden. (Under sees tegninger av åregaflene i båten.) Ett av bildene fra ekspedisjonen over Grønland i perioden juli 1888 til mai 1889. Fridtjof Nansen sammen med 5 norske deltagere krysset Grønland i løpet av en 42-dagers skiferd fra øst- mot vestkysten.
Publisert:	Nansen, Fridtjof: Paa ski over Grønland, Kra 1890, s. 557. (del).
Andre opplysninger:	Billedsamlingen har også to dubletter.
Samling:	BSN; Registrator:WW; Giver:WHH; Tilstand:2

Figur 29. Bildepost fra Fridtjof Nansens bildearkiv

9 Oppsummering

Denne oppgaven har undersøkt hvorvidt tekst kan brukes til automatisk indeksering av bilder i digitale bøker, med særlig fokus på bildetekstens potensial i en slik sammenheng. Utgangspunktet for undersøkelsene var tre spørsmål: Hva skal til for å identifisere bildetekster i de digitale bøkene, hva skal til for å identifisere mulige indekstermer av typen personnavn, stedsnavn og årstall i bildetekstene og hvor godt indekserer disse termene bildene i forhold til de tre emnetypene termene representerer? For å kunne bruke bildetekster som indekseringskilde, var det nødvendig å identifisere dem i de digitale bøkene. Det ble derfor foreslått en fremgangsmåte for å oppnå dette, som klarte å finne bildetekstene til 88 % av bildene i 10 bøker. Dette er et lovende resultat, men det er fremdeles utfordringer knyttet til bildetekstidentifisering som må løses. Fremgangsmåten som ble presentert hadde særlige problem med å håndtere bilder som går over flere sider, og tilfeller hvor bildeteksten står på en annen side enn det bildet gjør. For en av bøkene hvor bildetekstene var blitt identifisert, ble bildetekstene analysert for å identifisere egennavn av typen personnavn og stedsnavn, samt årstall. Fremgangsmåten som ble brukt til dette, ga blandede resultater. Identifisering av personnavn ga gode resultater, med presisjon på 0,9 og fullstendighet på 0,88. Dette er særlig gode resultater med tanke på at navn som forekom i enkel form i teksten, ble forsøkt koblet med dets tilsvarende hele navneform. Metoden for å identifisere stedsnavn må derimot utvikles videre, resultatene bar preg av at bruken av lister i navnegjenkjenning ikke fungerer like bra for sted som for personer. Gitt at identifiseringen av personnavn i teksten ga så gode resultater, er det påfallende at resultatene for hvor godt navnene indekserer bildet, ikke oppnår like gode resultater, med en presisjonsverdi på 0,74 og en fullstendighet på 0,78. Dette indikerer at bildetekstene ikke alltid vil være en god indekseringskilde for bildene de skal beskrive. Til tross for at resultatene av indekseringen er varierende, indikerer allikevel resultatene av denne undersøkelsen, at bildeteksten har et potensial som indekseringskilde for bilder i digitale bøker. Men det kan også være interessant å utforske andre tekstlige kilder, i og utenfor bøkene, for å sammenligne med, og/eller supplere bildeteksten i en automatisk tekstbasert bildeindeksering.

Litteratur

Aasheim, S.P. (2003) *Svalbard : fangstfamilien på 79°N*. Oslo : Cappelen

Altavista (2007) *Hjelp for Altavista : søk : bilde* URL: http://no.altavista.com/help/search/help_img [lesedato: 2009-02-18]

Alto (2004) Tyskland : CCS Content Conversion Specialists GmbH URL: http://tinyurl.com/mw56ru [lesedato: 2009-06-13]

Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. d., Blei, D. M., og Jordan, M. I. (2003). Matching words and pictures. *J. Mach. Learn. Res*., 3, 1107-1135.

Barnard, K. og Forsyth, D. (2001). Learning the semantics of words and pictures. ICCV 2001.

Barthes, R. (1977a). Rhetoric of the Image. I: S. Heath (red.), *Image, Music, Text* 32-51. New York: Hill.

Barthes, R. (1977b). The Photographic Message. I: S. Heath (Ed.), *Image, Music, Text* 15-31. New York: Hill.

Beazley, C.R. (2006, [1894]) *Prince Henry the Navigator, the Hero of Portugal and of Modern Discovery, 1394-1460 A.D.*, URL: http://www.gutenberg.org/files/18757/18757-h/18757-h.htm [lesedato: 2009-06-12]

Berinstein, P. (1997). Moving Multimedia: The Information Value in Images. I: *Searcher*, 5(8), 40-46.

Bjarnestam, A. (1998). Text-based hierarchical image classification and retrieval of stock photography. The Challenge of Image Retrieval Research Workshop.

Bomann-Larsen, T. (1995). *Roald Amundsen – en biografi.* Oslo : Cappelen.

brødtekst (2009). I: *Wikipedia : den frie encyklopedi*. URL:http://no.wikipedia.org/wiki/Br%C3%B8dtekst [lesedato: 2009-06-12]

caption (2009). I: *Merriam-Webster Online Dictionary*. URL: http://www.merriam-webster.com/dictionary/caption [lesedato: 2009-04-06]

Chen, F., Gargi, U., Niles, L., og Schuetze, H. (1999). Multi-modal browsing of images in web documents. IS&T/SPIE Conference on Document Recognition and Retrieval VI.

Chen, Z., Wenyin, L., Zhang, F. og Li, M. (2001). Web mining for web image retrieval. I: *J. Am. Soc. Inf. Sci. Technol.*, 52(10), 831-839.

Chu, H. (2001). Research in image indexing and retrieval as reflected in the literature. *J. Am. Soc. Inf. Sci. Technol.*, 52(12), 1011-1018.

Constantopoulos, P. og Doerr, M. (1995). An Approach to Indexing Annotated Images. I: *Selected Papers from the Third International Conference on Hypermedia and Interactivity in Museums (ICHIM'95 / MCN '95)*. URL: http://tinyurl.com/lnrnxs [lesedato: 2009-02-18]

Datta, R., Joshi, D., Li, J. og Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. I: *ACM Comput. Surv.*, 40(2), 1-60.

Dunlop, M. D. og Rijsbergen, C. J. v. (1993). Hypermedia and free text retrieval. I: *Inf. Process. Manage.*, 29(3), 287-298.

En hale uten ende (2008) Oslo : Nasjonalbiblioteket. URL: http://tinyurl.com/ks6jop [lesedato: 2009-06-12]

Enser, P. (1995). Pictorial information retrieval. I: *Journal of Documentation.*. 51(2)

Enser, P. (2008). "The evolution of visual information retrieval". I: *J. Inf. Sci.*, 34(4), 531-546.

Fossnes, H. (1993). *Norges bunader og samiske folkedrakter*. Oslo : Cappelen.

Frankel, C., Swain, M. J. og Athitsos, V. (1996). *WebSeer: An Image Search Engine for the World Wide Web*. Chicago : University of Chicago.

Goodrum, A. A. (2000). "Image Information Retrieval: An Overview of Current Research" I: *Informing Science*, 3(2).

Google (2009) *OSS for Google bildesøk*. URL:http://www.google.no/help/faq_images.html [lesedato: 2009-02-18]

Gudivada, V. N. og Raghavan, V. V. (1997). Modeling and retrieving images by content. I: *Inf. Process. Manage.*, 33(4), 427-452.

Harthan, J. (1981). Introduction. I: J. Harthan (Ed.), *The history of the illustrated book : the western tradition* 7-8. London: Thames and Hudson.

illustrere (2004). I: *Ordnett*. URL:http://ordnett.no [lesedato: 2009-03-01]

Johnson, T.E. (1975) *Alt for Norge : historien om et utvær.* Oslo : Cappelen.

Johnson, T.E. (1994). *Med Hurtigruta nordover*. Oslo : Boksenteret.

Jörgensen, C. (2003). *Image retrieval: Theory and research*. Lanham, MD: Scarecrow Press.

Kherfi, M. L., Ziou, D., & Bernardi, A. (2004). Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. I: *ACM Comput. Surv.*, 36(1), 35-67.

Lancaster, F.W. (2003). *Indexing and abstracting in theory and practice*. London : Facet.

Layne, S. S. (1994). Some issues in the indexing of images. I: *J. Am. Soc. Inf. Sci.*, 45(8), 583-588.

Levin, J. R. og Mayer, R.E (1993). Understanding illustrations in text. I: B. K. Britton, Woodward, A., Binkley, M. (red.), *Learning from Textbooks: Theory and Practice* 95-113. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lew, M. S. (2000). Next-Generation Web Searches for Visual Content. *Computer*, 33(11), 46-53.

Lieberman, H., Rozenweig, E. og Singh, P. (2001). Aria: An Agent for Annotating and Retrieving Images. I: *Computer*, 34(7), 57-62.

Maderlechner, G., Panyr, J. og Suda, P. (2006). Finding Captions in PDF-Documents for Semantic Annotations of Images. I: *Structural, Syntactic, and Statistical Pattern Recognition* 422-430.

Markkula, M. og Sormunen, E. (1998). Searching for photos--Journalists' practices in pictorial IR. The Challenge of Image Retrieval, (Electronic Workshops in Computing, eWIC). URL:http://http//www.ewic.org.uk/ewic/workshop/view.cfm/CIR-98 [lesedato: 2009-02-18].

Marsh, E.E. (2002). *Rhetorical relationships between images and text in web pages*. University of Maryland. Doktorgradsavhandling.

Marsh, E. E., & White, M. D. (2003). A taxonomy of relationships between images and text. I: *Journal of Documentation*, 59(6), 647-672.

Mukherjea, S. og Cho, J. (1999). Automatically Determining Semantics for World Wide Web Multimedia Information Retrieval. I: *Journal of Visual Languages & Computing*, 10(6), 585-606.

Nasjonalbiblioteket (2009). *Bokhylla.no* http://www.nb.no/bokhylla [2009-06-12]

Paek, S., Sable, C. L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B. H., Chang, S. F., et al. (1999). Integration of visual and text based approaches for the content labeling and classification of Photographs. ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval.

Rasmussen, E. M. (1997). Indexing Images. I: *Annual Review of Information Science and Technology (ARIST)*, Volume 32, 169-196.

Rowe, N. C. (1994). Inferring depictions in natural-language captions for efficient access to picture data. I: *Inf. Process. Manage*., 30(3), 379-388.

Rowe, N. C. og Frew, B. (1998). Automatic caption localization for photographs on World Wide Web pages.I: *Inf. Process. Manage*., 34(1), 95-107.

Rowe, N. C. og Guglielmo, E. J. (1993). Exploiting captions in retrieval of multimedia data. I: *nf. Process. Manage.*, 29(4), 453-461.

Røyneberg, E. (2004). *Tekstanalyse for geografisk informasjonsgjenfinning*. Trondheim : NTNU.

Røyneberg, E. (2005). *AIDaS – Automatisk identifisering av stedsnavn i nyhetstekster*. Trondheim : NTNU. - Masteroppgave

Shatford, S. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. I: *Cataloging & Classification Quarterly*, 6(Spring 1986), 39-62.

Solli, S. (2002) *Jegeren Fridtjof Nansen*. Oslo : Gyldendal.

Srihari, R. K. (1995). Automatic Indexing and Content-Based Retrieval of Captioned Images. I:*Computer*, 28(9), 49-56.

Svalbard, vårt nordligste Norge (1984). Oslo: Det beste.

Tsai, C.F. (2007). A review of image retrieval methods for digital cultural heritage resources. I:*Online Information Review*, 31(2), 185-198.

Winget, M. (2002). *Intellectual Access to Images : An Overview* URL:http://www.unc.edu/~winget/research/mini_lit_review1.html [lesedato: 2009-03-18]

Zachary, J., & Iyengar, S. S. (2001). Informaton theoretic similarity measures for content based image retrieval. I:*J. Am. Soc. Inf. Sci. Technol*., 52(10), 856-857.

Personlig kontakt

Bjerkreim, T. Forlagsdirektør, Gyldendal Akademisk Forlag