Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Notat

Automatiske metoder som hjelp til transkribering av historiske kilder

Notatnr SAMBA/44/10

Forfattere Line Eikvil, Lars Holden, Kåre Bævre

Dato Oktober 2010

© Copyright:
Norsk Regnesentral

Forfatterne

Line Eikvil, sjefsforsker Norsk Regnesentral
Lars Holden, adm. dir. Norsk Regnesentral
Kåre Bævre, forsker, Folkehelseinstituttet

Norsk Regnesentral

Norsk Regnesentral (NR) er en privat, uavhengig stiftelse som utfører oppdragsforskning for bedrifter og det offentlige i det norske og internasjonale markedet. NR ble etablert i 1952 og har kontorer i Informatikkbygningen ved Universitetet i Oslo. NR er et av Europas største miljøer innen anvendt statistikk. Det jobbes med svært mange forskjellige problemstillinger slik som finansiell risiko, jordobservasjon, estimering av torskebestanden og beskrivelse av geologien i petroleumsreservoarer. NR er også ledende i Norge innen utvalgte deler av informasjons- og kommunikasjonsteknologi. Innen IKT-området har NR innsatsområdene e-inkludering, informasjonssikkerhet og multimedia multikanal. NRs visjon er forskningsresultater som brukes og synes.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Tittel	**Automatiske metoder som hjelp til transkribering av historiske kilder**
Forfattere	**Line Eikvil, Lars Holden, Kåre Bævre**
Dato	05.10.2010
År	Oktober 2010
Publikasjonsnummer	SAMBA/44/10

Sammendrag

Gjennom denne rapporten har vi sett på ulike automatiske teknikker som kan være aktuelle som hjelp til transkribering av historiske kilder og hva som er state-of-the-art for teknikker som har vært anvendt på lignende problemstillinger. Vi har sett på oppdeling i skjemastruktur, fargeseparering og gjenkjenning av håndskrevet tekst.

Automatisk oppdeling i skjemastruktur er identifisert som et basiselement i et system som skal tilby automatiske verktøy for hjelp til transkribering. Automatisk gjenkjenning i tradisjonell forstand er mest aktuelt for tall, mens for sammenhengende håndskrift er det foreløpig ikke mulig å oppnå tilstrekkelig høye gjenkjenningsrater. Her foreslår vi derfor i stedet bruk av teknikker som kan være en støtte til den manuelle prosessen. I tillegg foreslår vi at en der det er mulig reduserer problemet ved å bruke informasjon fra andre kilder til å begrense antallet mulige alternativer for et ord eller et tall. I blant kan dette redusere gjenkjenningsproblemet til et verifikasjonsproblem, noe som straks gjør automatiske teknikker mer aktuelle også for tekst.

Gode internettløsninger som en HBR-wiki kan muliggjøre rekruttering av store frivillige manuelle ressurser til å bidra i transkriberingsarbeidet. Hvordan ressursene kan utnyttes vil avhenge av om arkivene er åpne eller lukket.

Emneord	Håndskrevne historiske dokumenter, transkribering, automatiske metoder
Målgruppe	HBR-prosjektet
Tilgjengelighet	Åpen
Prosjektnummer	220468
Satsningsfelt	Bildeanalyse og mønstergjenkjenning
Antall sider	25

© Copyright Norsk Regnesentral

Innhold

1 Innledning

Full automatisk transkribering av håndskrevne historiske dokumenter er foreløpig utenfor rekkevidde. Automatiske teknikker i kombinasjon med manuelle ressurser kan likevel være svært verdifulle hjelpemidler som kan bidra til å redusere kostnadene forbundet med transkribering av historiske kilder. Dette krever imidlertid at en finner gode totalløsninger for kombinasjon av manuelle og automatiske prosesser og finner fram til automatiske teknikker og metoder som egner seg for slik kombinasjon.

Fokus i denne rapporten er på strategier og metoder koblet til selve analysen av dokumentbildet og mindre på relaterte temaer knyttet til for eksempel datarepresentasjon, databaser, sikkerhet, brukergrensesnitt osv. Dette vil likevel være viktige komponenter i et totalsystem. Kapittel 2 gir en generell oversikt over litteraturen på dette området, mens kapittel 3 og 4 diskuterer hvordan disse metodene eventuelt kan utnyttes i forbindelse med et historisk befolkningsregister og hvilke besparelser og effektiviseringer som da kan oppnås. Kapittel 5 gir et konkret eksempel på en automatisk metode utviklet i prosjektet for uttrekking av tabellstruktur i skjemaer fra Fob1950. Kapittel 6 og 7 diskuterer datakilder som er aktuelle i prosessen for å etablere et norsk befolkningsregister, hvordan disse eventuelt kan lenkes og utnytte åpne wiki-løsninger for hjelp i transkriberingen.

2 Automatiske metoder

Dette avsnittet vil i korthet beskrive aktuelle automatiske teknikker og hva som er state-of-the-art for teknikker som har vært anvendt på lignende problemstillinger. Vi ser her på historiske håndskrevne skjemaer generelt uten fokus på spesifikke typer av dokumenter, men antar at skjemaene har en fast struktur som er kjent på forhånd og som kan utnyttes i prosessen.

2.1 Oppdeling i skjemastruktur

For dokumenter med skjemastruktur har det vært anvendt ulike automatiske metoder for å dele dokumentbildet opp i henhold til skjemastrukturen. Det finnes metoder for å finne tabeller i det generelle tilfellet der en ikke har noe kunnskap om hva slags tabeller som kan forekomme, men bedre resultater vil oppnås om strukturen for de aktuelle tabellene er kjent på forhånd. For dokumentene det er snakk om her vil det imidlertid finnes såpass mange bilder av samme skjema at det lønner seg å gjøre den manuelle jobben med å definere de aktuelle tabelltemplatene.

Automatisk bestemmelse av tabellstruktur når en har informasjon om hvilke tabeller som kan forekomme, kan foregå gjennom følgende trinn:

1. Finne tabellstruktur fra bildet. Dette gjøres gjerne ved at en først finner orienteringen på dokumentet og deretter finner tabellseparatorer (f.eks. linjer) ved analyse av bildet.
2. Identifisere tabelltype. Dette er bare nødvendig dersom det innenfor et sett av dokumenter kan forekomme flere typer av tabeller. Tabelltypen må identifiseres fra strukturen funnet i steg 1.

3. Verifikasjon og tilpasning av identifisert tabell i henhold til tabelltemplat.
4. Uttrekking av rader, kolonner og celler basert på identifisert tabell i bildet. Når dette er bestemt, vil en vite hvilket område i bildet en rad/kolonne eller celle svarer til og subbilder kan refereres til eller klippes ut.

Tilnærmingene baserer seg gjerne på en initiell manuell fase der tabellene som kan forekomme defineres og hvor systemet læres opp til å gjenkjenne tabeller hvis det er behov for det. Her er det da behov for spesifikasjon av tabellstruktur og innhold samt en egnet representasjon av denne informasjonen. Automatiske metoder kan være verktøy også i denne fasen, ved at de foreslår en struktur basert på et sett av dokumenter med samme skjema (Nielson og Barrett, 2003).

Nøyaktigheten av en slik algoritme vil avhenge av hvor mye som er kjent og hvor mange typer av tabeller som kan forekomme samtidig. Er det bare en tabell som forekommer, er problemet enklere (med mindre det er veldig dårlig kvalitet eller store distorsjoner i bildene). Kapittel 5 beskriver en slik løsning utviklet for skjemaer fra FoB1950.

2.2 Fargeseparering

For noen skjemaer som for eksempel folketellingen fra 1950, fins det informasjon skrevet i forskjellige farger. For å skille slike typer av informasjon er det behov for fargeseparering. Slike teknikker baserer seg gjerne på å transformere fargeinformasjonen fra en RGB representasjon til et fargerom som er bedre egnet for segmentering. Hvor bra fargene kan skilles avhenger både av fargene som er benyttet og av kvaliteten på dokumentene.

2.3 Gjenkjenning av håndskrevne tall

Metoder for automatisk gjenkjenning av håndskrevne tall har hatt ganske stor suksess for gjenkjenning av håndskrevne postnummer. For enkeltsiffer ligger de rapporterte gjenkjenningsratene her ofte på 98-99% (Liu et al., 2003). Dette forutsetter at tall enkelt kan deles opp i enkeltsiffer. Generelt er dette enklere for håndskrevne tall enn håndskrevet tekst. For mange applikasjoner vil en også kjenne syntaksen og vite hvor mange siffer et tall skal bestå av (som for postnummer, årstall o.l.).

2.4 Gjenkjenning av sammenhengende håndskrift

Ved gjenkjenning av trykket tekst (OCR) er det vanlig å bryte opp ord i enkeltbokstaver og så gjenkjenne bokstav for bokstav. For sammenhengende håndskrift derimot er man nødt til å ha gjenkjent bokstavene for å kunne gjøre denne oppdelingen riktig. Samtidig kan en vanligvis ikke gjenkjenne bokstavene uten først å dele opp ordet. Dette refereres gjerne til som Sayres paradoks. Metodene som er brukt for gjenkjenning av sammenhengende håndskrift forsøker derfor på ulike måter å komme seg rundt dette.

To hovedtyper av tilnærminger har typisk vært benyttet (i) kombinert segmentering og gjenkjenning (vanligvis basert på HMM-modeller hentet fra talegjenkjenning) og (ii) holistisk metoder som ser på hvert ord som en helhet. Den første typen av teknikker bygger modeller basert på karakterer som systemene trenes opp til å gjenkjenne. Denne treningsfasen kan være kompleks og krever gjerne en ekspert på automatisk gjenkjenning. Treningen av de holistiske metodene er enklere og krever ikke på samme

måte eksperter. Til gjengjeld trengs det mange flere modeller (en for hvert ord) som kan ta tid å trene. Et stort antall ord vil også påvirke både nøyaktighet og beregningskostnader ved klassifikasjonen.

For gjenkjenning i håndskrevne historiske dokumenter ser de fleste ut til å mene at holistiske tilnærminger er best egnet (Bilane et al., 2009). Likevel er fullstendig transkripsjon av historiske kilder ved hjelp av automatiske teknikker foreløpig ansett å være utenfor rekkevidde. Dagens gjenkjenningsteknikker har typisk feilrater på over 50% ved gjenkjenning av håndskrevne ord (Rath, Manmatha, Lavrenko, 2004). Fokus har derfor i det siste vært mer på indeksering og gjenfinning i slike dokumenter framfor gjenkjenning. Den vanligste tilnærmingen her, kalt "word spotting", ble først foreslått av Manmatha et al. i 1996 og har siden vært mye brukt. En variant av dette, inspirert av CBIR (content based image retrieval), har også dukket opp. Vi kaller denne varianten for "word retrieval". Disse to tilnærmingene er beskrevet nærmere nedenfor.

2.4.1 Word spotting

I (Rath, Manmatha, 2007) er denne teknikken anvendt på håndskrevne historiske manuskripter. Her identifiseres først enhetene som tilsvarer ord, deretter beregnes det bildebaserte egenskaper. På basis av egenskapene gjøres det en clustring der ordene deles inn i clustre (grupper) basert på et likhetsmål. Tanken er da at et slik cluster vil inneholde forskjellige forekomster av det samme ordet. Deretter kan clusterne manuelt tilordnes riktig ord. På denne måten reduseres det manuelle arbeidet ved at et helt cluster kan gis en merkelapp samtidig, i stedet for at ordene må behandles ett for ett.

2.4.2 Word retrieval

På samme måte som for word spotting trekkes det her ut egenskaper for et helt ord, og ofte er det også samme likhetsmål som benyttes her som ved word spotting. Den overordnede tilnærmingen er likevel noe forskjellig. Med denne tilnærmingen lagrer man for hvert segmentert ord et sett med bildebaserte egenskaper (men i utgangspunktet ingen merkelapp med det transkriberte ordet). Dermed får man en bildebasert indeks der det kan gjøres søk basert på et bilde av et ord. Resultatet av et slikt søk vil være alle indekserte ord (ordbilder) som ligner på søkeordet. Selve søkeordet består da altså av et bilde. Det vanligste er at brukeren/operatøren finner søkeordet ved å plukke det ut fra dokumentbildene, mens noen også har sett på hvordan en kan lage syntetiske søkebilder basert på ASCII-tekst og eksempler fra dokumentene (Leydier et al., 2009).

2.4.3 Betydning av et redusert vokabular

For de fleste av tilnærmingene over vil en reduksjon av vokabularet (eller ordlisten) kunne redusere klassifikasjonsproblemet og gi bedre resultater. Med reduksjon av vokabularet mener vi da en reduksjon i mulige ord for et gitt ordbilde. Her finnes det flere ulike tilnærminger både generelle og mer spesifikk som ofte også kan kombineres. De mer problemspesifikke metodene vil benytte kunnskap om egenskaper ved dokumentene til å definere temaer og så ha ulike ordlister for ulike temaer. For skjemabaserte registre kan sånne temaer for eksempel være fornavn, etternavn, yrke eller lignende, hvor hvert av disse temaene har sin egen ordliste. De mer generelle tilnærmingene bruker karakteristikker ved ordbildet til å redusere ordlisten. Dette kan

baseres på enkle egenskaper som for eksempel ordlengde, men også mer avanserte holistiske metoder har blitt brukt til dette.

3 Mulige tilnærminger til problemet

I det følgende vil vi se på muligheter for hvordan automatiske verktøy kan utnyttes til å forenkle transkriberingsprosessen. Bruk av slike verktøy kan bety at en også tenker litt annerledes på hvordan prosessen organiseres.

3.1 Utnytte skjemastruktur

Automatiske teknikker kan brukes til å bestemme skjemastruktur i et bilde og videre til automatisk å trekke ut informasjonen knyttet til en enhet (tilsvarende en person, en husstand eller lignende avhengig av type skjema og type register). Informasjonen knyttet til en enhet kan da lagres som et sett med bilder tilsvarende cellene i skjemaet som har med denne enheten å gjøre.

På denne måten vil informasjonen ligge inne, selv om den ikke er søkbar. En kan da i første omgang fokusere på å få transkribert det som skal være søkbart (og eventuelt det som trengs til beregning av statistikker). Den andre informasjonen vil fortsatt ligge der som et bilde som vil være manuelt lesbart og som er lenket til enheten, men som ikke vil være søkbar i seg selv. Med et system som flagger enheter i forhold til hvor mye som er transkribert, vil en likevel til en hver tid kunne holde oversikt over hva som er ferdig og hvor mye som gjenstår.

Lagring av bildene på denne måten gir også muligheter for å se på dataene kolonnevis eller cellevis. Transkribering kan da for eksempel gjøres kolonnevis. Da er mer kunnskap tilgjengelig om mulig innhold og ordlister. Dette kan være til hjelp både i den manuelle transkriberingen og gi større muligheter for utnyttelse av semi-automatiske og automatiske teknikker. Frikobling av celler på sin side kan gi muligheter for anonymisering som kanskje kan gi grunnlag for tilgang til andre og større manuelle ressurser i transkriberingsprosessen.

Kvaliteten på dokumentene og hvor veldefinert skjemaet er, vil avgjøre hvor godt skjemastrukturen i bildet kan bestemmes. Dersom det er få skjemakandidater og disse er veldefinerte, vil strukturen i de fleste tilfeller kunne bestemmes greit. Utfordringen kan da være knyttet til at tekst og tall ikke holder seg innenfor et felt, men krysser over i andre felter. Det betyr at bildeutsnittene kanskje må gjøres større en selve feltet i skjemaet. Tekst som overlapper med linjer eller annen tekst vil også være vanskeligere å gjenkjenne både manuelt og automatisk.

3.2 Automatiske gjenkjenningsstrategier

Automatisk gjenkjenning vil trolig ha mest for seg når det gjelder gjenkjenning av tall og eventuelt også gjenkjenning av tekstfelter med kategorisk informasjon med svært begrenset vokabular (eks: nasjonalitet, trosretning, sivil status eller lignende), ettersom det bare er her gjenkjenningsratene kan forventes å være tilstrekkelige høye. Et svært begrenset vokabular kan også oppnås for mer generelle felter dersom det fins annen

informasjon, for eksempel fra andre kilder, som kan bidra til å begrense utfallsrommet. I en automatisk prosess vil det ellers være viktig å kunne forkaste tvilstilfeller slik at feilraten blir så liten som mulig.

3.3 Semi-automatiske gjenkjenningsstrategier

For sammenhengende håndskrevet tekst vil automatiske teknikker brukt som semi-automatiske tilnærminger til hjelp i det manuelle arbeidet være mest relevante. Teknikker basert på *word spotting* kan være aktuelle i denne sammenheng. En slik tilnærming består, som nevnt i avsnitt 2.4.1, i at ordbildene først clustres og at en deretter kan manuelt tilordne merkelapper til hele clustre framfor enkeltord. Denne tilnærmingen kan forenkle og effektivisere den manuelle prosessen gjennom at en jobber med grupper av ord framfor enkeltord.

Samtidig visualisering av alle kandidatene innenfor et cluster kan her bidra til å gjøre det enkelt å luke ut eventuelle instanser som har havnet i feil cluster. Erfaring fra andre anvendelser som for eksempel gjenkjenning av håndskrevne tall i kart, har vist at dette kan være et nyttig hjelpemiddel i en manuell prosess ettersom det menneskelige øyet fort fanger opp ting som skiller seg ut. Slik visualisering kan kanskje også bidra til å gjøre det enklere å avgjøre hva som er riktig transkribering for et ord når en kan se flere forekomster samtidig.

Teknikker basert på *word retrieval* kan være interessante i situasjoner der en bruker har transkribert et enkeltfelt. Disse teknikkene kan da brukes for å finne ord som ligner blant de som ennå ikke er transkribert slik at en kan velge å gi samme merkelapp også til en del av disse. Dette kan for eksempel være interessant for åpne arkiver med mange brukere. Der kan brukeren være interessert i å transkribere enkeltfelter. Med en slik løsning vil en da også kunne få merket ikke bare det aktuelle ordet, men også andre ord i databasen som er tilstrekkelig like til at en med rimelig sikkerhet kan anta at de representerer det samme ordet.

Det kan eventuelt også være interessant å kombinere *word retireval* med *word spotting*, der en først gjør clustering og merking på et begrenset sett og deretter bruke *word retrieval* på et større ikke-indeksert sett for å finne flere ord som tilhører samme cluster.

Teknikkene nevnt her baserer seg alle på likheter mellom bilder av samme ord. Siden håndskrift varierer kan likevel samme ord se forskjellig ut avhengig av hvem som har skrevet det. Hvis alle dokumentene er skrevet av forskjellige personer, og alle disse har veldig forskjellig håndskrift, kan det derfor være lite å hente effektviseringsmessig med disse teknikkene.

3.4 Utnyttelse av kontekst og andre kilder

Kontekst og kunnskap om dokumentet som analyseres samt tilleggsinformasjon fra andre kilder, kan bidra til at automatiske teknikker kan utnyttes i større grad og det kan ofte også effektivisere den manuelle prosessen. Dette fordi slik informasjon kan bidra til å redusere utfallsrommet. I beste fall kan en gjøre problemer om fra et gjenkjenningsproblem til et verifikasjonsproblem (fra et ”hva står her?”-problem til et ”står det Ole Olsen?”-problem).

Kunnskap om dokumentet og feltene som analyseres kan fortelle hvilken ordliste som skal brukes (eks. stedsnavn, personnavn osv) eller det kan finnes innbyrdes avhengigheter mellom feltene (eks. barn født etter foreldre) som også kan utnyttes. Der en har mulighet til å linke til andre kilder som allerede er tekstlig søkbare, kan det også være mer slik informasjon å hente.

3.5 Utnyttelse av internettbrukere

Der en har mulighet til det kan utnyttelse av frivillige gjennom gode internettløsninger utgjøre en betydelig ressurs. Hvordan ressursene kan utnyttes vil avhenge av om arkivene er åpne eller lukket. For åpne arkiver bør slike ressurser greit kunne utnyttes. For lukkede arkiver er dette mer usikkert, men her kan i hvert fall en skjemabasert struktur og organisering av bilder være en nøkkel til anonymisering av informasjonen. Dette kan gi muligheter for å presentere løsrevne subbilder med stedsnavn, fornavn, årstall eller lignende, gjerne som et randomisert utvalg. Informasjonen som presenteres vil da ikke ha noen kobling til person.

4 Mulige effektiviseringer og besparelser

I det følgende vil vi se litt på hvilke besparelser som kanskje kan oppnås med de foreslåtte metodene.

4.1 Automatisk oppdeling i skjemastruktur

Oppdeling i skjemastruktur kan åpne for nye måter å jobbe med de digitaliserte dokumentene på. Selv om all annen transkribering gjøres manuelt, vil denne tilnærmingen gjøre at en kan fokusere på det viktigste først i stedet for å transkribere alt sekvensielt. Dette igjen vil gjøre at informasjon som er viktig blir tigjengelig fortere. Dersom bare om lag halvparten av informasjonen som er registrert om en enhet er viktig for indeksering og analyser, vil denne halvparten bli tilgjengelig dobbelt så fort som hvis alt skulle tas sekvensielt.

Det at informasjonen, også den som ikke er transkribert, er ordnet i en skjemastruktur kan også gi andre nye muligheter til effektivisering og besparelser. En slik organisering gjør det mulig å løsrive enkeltfelter fra resten av informasjonen slik at den kan anonymiseres. Med sikre bakenforliggende systemer kan dette gi muligheter til å utnytte frivillige ressurser i den manuelle transkriberingen.

4.2 Automatisk gjenkjenning

Her bør en fokusere på de oppgavene der gjenkjenning kan gjøres med relativt stor sikkerhet. Dette vil typisk gjelde tall og tekst der kontekst eller informasjon fra andre kilder kan bidrar til at ordlisten er veldig begrenset. Besparelsene som kan oppnås, vil avhenge av de enkelte dokumenttypene, og hvor stor andel av feltene som vil kunne gjenkjennes med tilstrekkelig sikkerhet til at det vil være en gevinst.

I rapporterte eksperimenter for gjenkjenning av tall, som i Liu et al, opererer de med gjenkjenningsrater på opp mot 99% for enkeltsiffer (som for et helt årstall med fire siffer tilsvarer 96%) , men da er samtidig feilraten 1% og ingen forkastet. I en

anvendelse som dette ønsker man feilrater som er svært mye lavere enn dette. Da må en stor andel forkastes og andelen gjenkjente tall vil da vanligvis synke betraktelig. På den annen side kan annen kunnskap, som for eksempel at et tall faktisk er et årstall, bidra til å redusere problemet noe. Med et stort antall tall som skal transkriberes, kan likevel en relativ lav andel gjenkjente gi en stor besparelse så lenge feilraten er tilstrekkelig lav.

For automatisk verifisering av tekstlig informasjon eller gjenkjenning med veldig begrenset ordliste, vil suksessraten trolig være enda høyere enn for tall. For skjemaer med mye slik informasjon, bør derfor en ganske stor del av prosessen kunne automatiseres.

4.3 Støtte til den manuelle prosessen

Clustering og indeksering gjør det mulig å behandle flere ord samtidig og kan gjennom dette effektivisere prosessen. Effektiviseringsmessig er det mer å hente når det fins større grupper av dokumenter med lite variasjon i håndskriften.

Visualisering av flere like ord samtidig kan gjøre det lettere å transkribere riktig. Basert på kontekst, informasjon fra andre kilder og egenskaper fra ordene vil en også kunne få opp forslag til transkriberinger, presentert for eksempel som en kort drop-down liste. Å velge et ord fra en kort liste kan være mer effektivt enn å skrive inn ordet.

5 Oppdeling i skjemastruktur for FoB1950

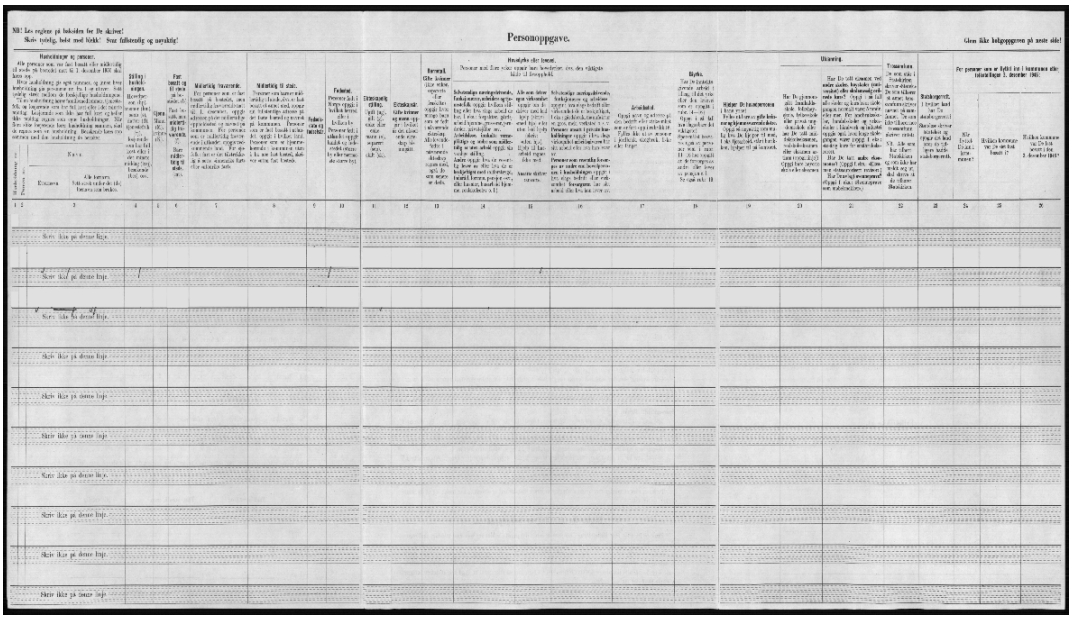
Som vist gjennom beskrivelsene i de foregående avsnittene, vil et basiselement i et system som skal tilby automatiske verktøy for hjelp til transkribering, være verktøy som kan dele opp bildene i henhold til skjemastrukturen. I det følgende har vi sett på hvordan dette kan gjøres for husstandsskjemaene fra folketellingen i 1950.

Vi har valgt å se på disse skjemaene fordi:

- De fins i et stort volum slik at metoder utviklet spesielt for disse skjemaene lett kan forsvares samtidig som enhver støtte i transkriberingsprosessen vil være til stor hjelp.
- Folketellingen fra 1950 er høyt prioritert fordi den vil utvide dagens elektroniske befolkningsregister med 10 år.
- Skjemaene har en veldefinert struktur.

5.1 Data

Dataene programmet er utviklet for er husstandskjemaene fra 1950. Disse skjemaene består av en utside og en innside, der innsiden inneholder skjemaet vi har sett på her. Skjemaet har en størrelse tilsvarende tre A4-ark og er inndelt i 27 kolonner og 20 rader. Radene og kolonnene i skjemaene er separert med svarte linjer. Bildene vi har mottatt er scannet ved Riksarkivet med en oppløsning på 300 dpi som RGB-fargebilder. For skjemaanalysen har vi bare benyttet det blå båndet. Bildet nedenfor gir et inntrykk av layouten på skjemaene.



5.2 Beskrivelse av metoden

Analysen for å bestemme skjemastrukturen skjer gjennom fire hovedtrinn:

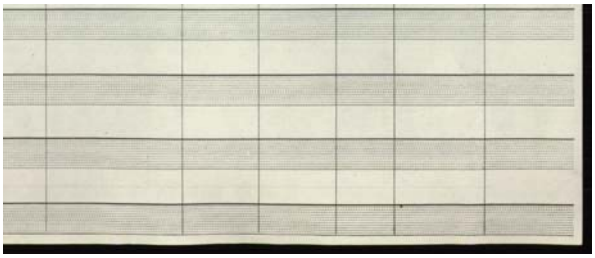
- Finner orientering og posisjon for skjemaet
- Gjør en omtrentlig posisjonering av gridet
- Gjør en mer nøyaktig posisjonering basert på bildet
- Kontroll og justering av gridpunkter

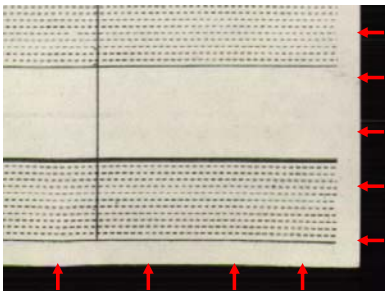
Disse fire trinnene er forklart i mer detalj nedenfor.

5.2.1 Finne orientering og posisjon

Alle skjemaene er scannet samme vei, men posisjoneringen og vinkelen kan variere litt fra bilde til bilde og må derfor først bestemmes. Skjemaene er scannet slik at bakgrunnen (utenfor skjemaet) vil være svart. Vi ønsker å finne både den horisontale og den vertikale kanten skjemaet, og for å oppnå dette har vi valgt å bruke høyre kant og nedre kant av skjemaet. Dette fordi det ser ut til å være minst slitasje på disse kantene. Bildet nedenfor viser hvordan disse kantene kan se ut.

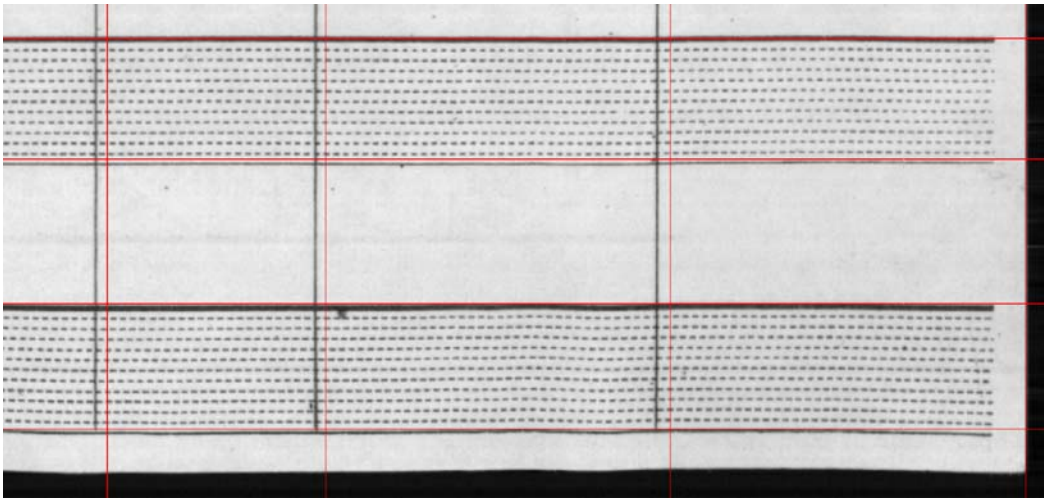
Først finnes den høyre kanten av skjemaet ved å lete innover fra bildets høyre ytterkant etter en gradient (overgang) fra mørkt til lyst (se figur nedenfor). Dette gjøres for hver linje i bildet og resulterer i et sett med punkter langs kanten av skjemaet. Fra disse punktene estimeres den rette linjen som best beskriver denne kanten. Til dette benyttes en metode for robust regresjon (LMS; Rousseeuw og Leroy, 1987). Tilsvarende operasjon gjøres deretter for den nedre kanten av skjemaet. Når disse to linjene er bestemt har vi da posisjonen og orienteringen for skjemaet.





5.2.2 Omtrentlig posisjonering av grid

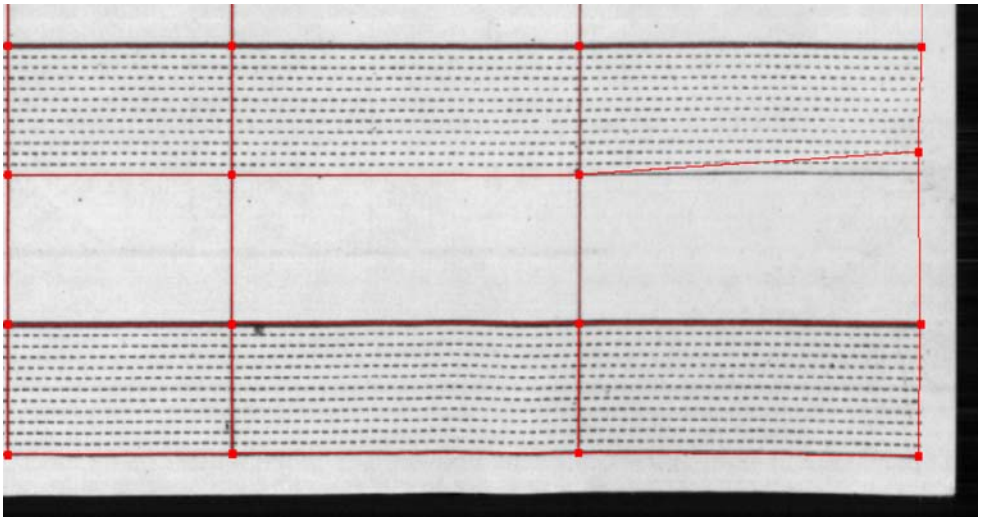
Siden skjemastrukturen er kjent og er den samme for alle dokumentene, samtidig som alle skjemaene er scannet i samme oppløsning, vil kolonnene og radene i skjemaene ha omtrent den samme relative posisjonen for alle bildene. Vi kan derfor på forhånd definere de relative posisjonene til gridlinjene i skjemaene. Når så posisjonen og orienteringen for et nytt skjema er funnet, kan de relative posisjonene til gridlinjene benyttes til å finne omtrentlig posisjon for disse i det aktuelle bildet. Disse linjene vil ikke stemme helt med linjene i bildet, men er et godt utgangspunkt for den videre analysen. Figuren nedenfor viser et utsnitt av bildet der disse forhåndsdefinerte linjene er tegnet inn med rødt.



5.2.3 Posisjonering av grid i henhold til bildet

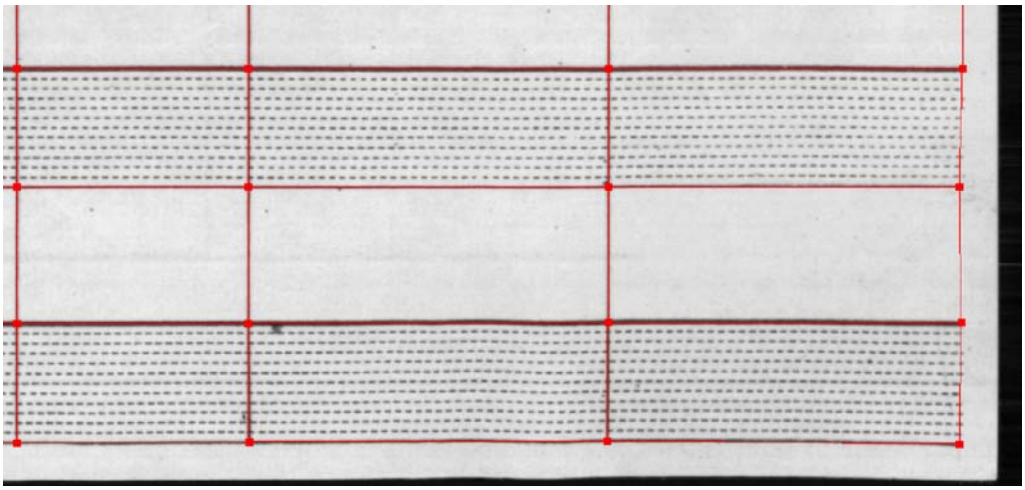
De forhåndsdefinerte gridlinjene vil sjelden stemme helt med gridlinjene i bildet. Dette skyldes at linjene i bildet på grunn av ulike distorsjoner sjelden er perfekte rette linjer. Det er derfor behov for en mer nøyaktig posisjonering som utnytter informasjonen i selve bildet. Vi gjør dette ved å ta utgangspunkt i krysningspunktene estimert fra de forhåndsdefinerte linjene og bruke templatmatching til å søke etter de tilsvarende punktene i bildet i et lite område rundt utgangspunktene. Ulike templater defineres for kryss og endepunkter langs høyre, venstre og nedre kant. Resultatet av templatmatchingen er et sett med krysningspunkter som stemmer bedre med bildeinformasjonen, men templatmatchingen kan iblant la seg lure av andre elementer i bildet. Figuren nedenfor viser samme utsnitt av bildet som tidligere der krysnings- og

endepunkter funnet ved templatmatching samt linjene mellom dem er merket med rødt. En kan her se at templatmatchingen har feilet for ett av endepunktene langs høyre marg.



5.2.4 Kontroll og justering av gridpunkter

Fordi templatmatchingen kan bli lurt av annen informasjon i bildet og velge et galt punkt, gjør vi en siste kontroll og justering av de estimerte punktene. På dette punktet i analysen vil de fleste krysningspunktene være riktig definert. Dette sammen med det faktum at vi vet at punktene skal beskrive et grid der alle cellene i en rad er like høye og alle cellene i en kolonne er like brede, benyttes til å gjøre denne justeringen. For kolonnene og radene finner vi median bredde og høyde, og eventuelle celler med avvik identifiseres. Gridpunktet som antas å forårsake avviket identifiseres gjennom traverseringsrekkefølgen, og nye koordinater bestemmes fra nabopunktene der disse er av god nok kvalitet. Dette gjøres i flere iterasjoner inntil alle punkter med avvik antas å være justert. Figuren nedenfor viser resultatet av denne justeringen.



5.3 Resultater

Metoden beskrevet over er implementert i et program som er testet ut på et antall skjemaer. Prototypprogrammet baserer seg på C og fritt tilgjengelige programvareverktøy (Xite og R) og kjører under Linux. På testplattformen tar analysen cirka 30 sek per bilde.

Totalt har vi mottatt bilder av 425 skjemaer fra Riksarkivet. Vi har kjørt analysen og sett på resultatene for et tilfeldig utvalg av noen titalls skjemaer. Analysen lar seg greit kjøre på alle skjemaene, men vi har bare gjort stikkprøver siden den manuelle inspeksjonen av så mange skjemaer blir for tidkrevende innenfor denne begrensede studien. Resultatene for alle skjemaene vi har inspisert er gode og viser at metoden er robust for de variasjonene som opptrer innenfor dette datasettet.

6 Bruk av automatiske metoder i å etablere et norsk befolkningsregister

6.1 Data

Som utgangspunkt for vurderingene i dette notatet er følgende datakilder mest aktuelle:

- Folketellinger: FoB 1950, 1891, 1930, 1920
- Kirkebøker: ca 1850-1930,1930-1964
- Grunnlagsmateriale fra Folkemengdens Bevegelse 1925-1964

En av de viktigere datakildene er Folketellingen fra 1950 (FoB 1950). Nedenfor følger en litt mer detaljert beskrivelse av denne.

6.1.1 Folketellingen 1950.

Denne folketellingen har 3,3 mill. registrerte personer hvorav 1,4% er født i utlandet. Både SSB og RHD gir en detaljert beskrivelse av denne folketellingen. De viktigste feltene er:

- **Gårds- og bruksnummer.**
- **Navn**. Viktig felt som vil kreve mye manuelt arbeid. I Fob 1900 vil 150 navn for hvert kjønn dekke 70-75% av alle fornavn mens det trengs ca 1700 navn for tilsvarende dekning av etternavn.
- **Fødselsdato og –år**. Dette er bare tall, men skrevet i meget lite rubrikk. Feltet er viktig for lenking
- **Fødested.** Kan være viktig, som oftest er dette stedsnavn i nærheten av nåværende bosted. Hvis variasjonsmulighetene er små, gir det større muligheter for utnyttelse av automatiske metoder.
- **Stilling i familien**. Viktig for å avgjøre familieforhold (betegnes ved hovedperson=hp, husmor=hm, sønn=s, osv.). Med få mulige alternativer kan automatisk lesing være en mulighet.
- **Felt om midlertidig bosted**. De aller fleste felt er tomme. Automatisk identifikasjon av tomme felt kan redusere arbeidet mye
- **Kjønn, husholdning, ekteskapelig stilling.** Meget få alternative svar gjør at vi kan håpe på at automatisk lesing klarer det meste.

- **Trossamfunn, Statsborgerskap.** Nesten alle har samme svar (”Statsk”), så det å gjenkjenne de som kanskje har et annet svar, vil være en stor hjelp.
- **Rød kontrollskrift** fylt ut av SSB i ettertid som i hovedsak består i tallkoder. Den kan være lettere å gjenkjenne da den er i annen farge, tydelig og liten variasjon i skrift. Den vil ikke bidra til å gjenkjenne personer, men kan hjelpe i statistiske analyser. For sosio-økonomisk informasjon er denne vel så viktig som den rent identifiserende kjerneinformasjonen. Her kan det ligge en stor gevinst i bruk av automatiske teknikker.

6.2 Problemstillinger

1. **Hvilke kilder er tilgjengelig scannet?** Arkivverket opplyser at fullskala prosess på FoB1950 ikke er aktuelt før tidligst årsskifte 2011/2012, men prøvekommuner kan være aktuelt før. Da er det naturlig å velge Rendalen, Sula eller Asker der vi har kommet vesentlig lengre. Det antas å ta 3 årsverk å brette ut og tilbake og 1,5 årsverk å scanne skjemaene. FoB 1891 er planlagt scannet fra august 2010 til august 2011 og krever ca et årsverk. Mange kirkebøker er allerede scannet.

2. **Hva er prioriteringsrekkefølgen?** Foreløpig prioriteres FoB 1950 for hele befolkningen høyt for å kunne utvide dagens befolkningsregister med 10 år. Det vil kreve både transkribering, lenking (samme person flere kilder) og kobling (mellom generasjoner/familiemedlemmer). Det er også høyt prioritert å tilrettelegge folketellingen 1891 for slektsforskere fordi vi her kan få mye gratis arbeidskraft til transkribering, lenking og kobling. For mange statistikker er det ikke nødvendig å ha hele befolkningen. Det er viktig å prioritere felt som kan brukes i søking, lenking og kobling, dernest det som brukes i viktige statistikker. Mange andre felt er ikke så viktig. Det vil være enkelt å transkribere enkeltfelt senere eller for et representativt utvalg dersom det scannede bildet er knyttet til hver post. Systemer for behandling av kirkebøkene kan være viktige for å virke koordinerende i transkriberingsdugnad.

3. **Kostnadene med transkribering** av de forskjellige kildene. Det som kan gjøres automatisk eller der det er lett å mobilisere frivillige, vil være vesentlig billigere enn andre felt. Det bør derfor legges til rette for at dette tas tidlig. Det taler for at bilder for FoB 1891 der enkeltfelt er identifisert tas tidlig.

4. **Bruk av andre kilder** kan ofte bidra til å gi mer informasjon slik at antallet mulige tolkninger av et ord eller et tall reduseres. Dette kan lette manuell transkribering og kan også gis tørre muligheter for automatisk gjenkjenning ettersom verifikasjon eller gjenkjenning mot en kort ordliste er enklere enn gjenkjenning med en lang ordliste. Slik informasjon kan dermed bidra til at automatiske metoder kan brukes på en større andel av feltene.

5. **Hvordan foreta lenking** av transkriberte verdier, dvs. gjenfinne samme person i flere kilder? Ved bruk av verifikasjon kan man løse lenkingsproblemet samtidig som man gjør transkriberingen. Det kan for eksempel skje gjennom at en gjenkjenner et navn i folketellingen 1950 ved at navnet ligner på navnet til personen som er registrert som eier av boligen i matrikkelen fra 1950 eller bor i huset i folketellingen i 1964 (kilder som allerede er transkribert). Der man ikke lenker samtidig med transkribering, må man først transkribere de feltene som er viktige for lenkingen og

så forsøke å lenke med samme person i andre kilder. For åpne kilder som legges inn i HBR-wiki vil slektsforskere kunne utføre denne lenkingen. For lukkede kilder er det nødvendig med arbeidskraft med taushetserklæring. Disse må ha innsyn i alle felt og det er sannsynlig at dette må være betalt arbeidskraft. For denne perioden har man imidlertid fødselsdato slik at automatiske metoder kan utnyttes i større grad og det manuelle arbeidet kan fokuseres på feil, manglende data eller andre spesielle problemer.

6. **Bruk av konfidensestimater** ved automatisk klassifisering. Der en benytter automatisk gjenkjenning er det mulig å velge en tilnærming der systemet for hvert felt angir estimert sikkerhet for klassifikasjonen sammen med foreslått tolkning. Dermed kan en begynne med manuell transkribering av feltene med størst usikkerhet og eventuelt sette en grense for hva som må tas manuelt og hva som er tilstrekkelig sikkert. Alternative sannsynlige verdier kan være av interesse også der manuell transkribering skal gjøres for å effektivisere det manuelle arbeidet.

7. **Hvordan kan man mobilisere frivillige** eller organiseres betalt arbeidskraft til transkribering? Dette diskuteres i mer detalj i avsnitt 6.4.

6.3 Fra gjenkjenning til verifikasjon ved hjelp av lenking

I det følgende vil vi beskrive noen muligheter for å bruke andre kilder for å omgjøre et gjenkjenningsproblem til et verifikasjonsproblem, der et verifikasjonsproblem består i avgjøre om et tekstelement stemmer med en gitt tolkning eller ikke. I de etterfølgende punktene fokuseres det mest på folketellingen i 1950, men mange av punktene kan også anvendes mer generelt.

Matrikkelen 1950
Det finne en matrikkel fra 1950 som omfatter ca 85.000 eiendommer (byene og Finnmark er ikke med) og som allerede er transkribert. For eiendom er gårds- og bruksnummer samt navn på eier registrert. For folketellingen fra samme år kan en da, når gårds- og bruksnummer er transkribert (manuelt eller automatisk), bruke dette til å foreslå eier gitt i matrikkelen som mulig navn på familiefar registrert på samme eiendom. Dersom det er samsvar, holder det da med en automatisk (eller manuell) verifisering av navnet. For disse 85.000 eiendommene der matrikkel og folketelling er fra samme år, vil det trolig være samsvar for en stor andel av registreringene.

Befolkningsregisteret fra 1960
Befolkningsregisteret fra 1960 som allerede er digitalt tilgjengelig, vil også kunne utnyttes i transkriberingsprosessen. For mange eiendommer, spesielt på landet, vil det være samme personer som bor i boligen i 1950 som i 1960, slik at gårds- og bruksnummer kan utnyttes som kryssreferanse på samme måte som skissert for matrikkelen fra 1950. Dersom fødselsdatoen er transkribert/gjenkjent, vil denne kunne brukes til å gi et redusert utfallsrom av mulige navn som igjen kan reduseres når en kjenner kjønn. Kombinert med egenskaper ved ordbildet kan dette utnyttes til å gjøre automatisk verifisering/gjenkjenning eller til å gi en kort liste med navneforslag til den manuelle transkriberingen.

Folketellingen 1910
Folketellingen fra 1910 er også tilgjengelig digitalt. Selv om det er et langt tidsspenn fra 1910 til 1960, vil i hvert fall personer over 50 år som er med i befolkningsregisteret fra 1960 også være registrert i folketellingen fra 1910. Dette utgjør ca 40% av personene i folketellingen fra 1910. Her kan det da også være mulig å lenke mellom disse ved å bruke kjønn, fødselsdato, navn og region/bolig. Det kan gjøres med automatiske lenkingsprogrammer. Vi vet det er mange feil i fødselsdatoene i 1910-folketellingene. Det er derfor et poeng å få mange slektsforskere til å rette opp disse feilene. Det vil i noen tilfeller være mulig å lenke mellom folketellingen i1910 og det sentrale folkeregisteret. Det kan for eksempel gjøres ved at slektsforskere lenker mellom 1910 folketellingen og lister over døde. Der det er etablert slike lenker vil vi kunne kvalitetssikre dataene i folketellingen 1910 og det sentrale folkeregisteret. Det vil også gjøre det lettere å gjenkjenne og lenke personen i folketellingene imellom; 1920, 1930 og 1950 samt kommunale transkriberte folketellinger.

6.4 Frivillig transkribering

For å få flest mulig frivillige til å bidra til transkribering er det verdt å se på hvordan en kan fange det store publikums interesse. For åpne arkiver kan folk være veldig interessert i å transkribere informasjon de selv jobber med, mens det kan være en større utfordring å få brukere til å transkribere informasjon som ikke er relatert til det de selv er interessert i. For lukkede arkiver må alle feltene som skal transkriberes av frivillige være løsrevet og anonymisert, så her kan utfordringen være enda større.

Vi kan regne med at det blir mange brukere av HBR-wiki, og gjennom denne vil det være mulig å mobilisere personer til frivillig transkribering. Som eksempel var det i 2009 210 millioner visninger av skannede kirkeboksider i Digitalarkivet. Hvis 1% av disse besøkene hadde resultert i en full registrering av en post, ville det gitt 2 millioner transkriberte poster/personer per år. På denne måten kunne alle åpne kirkebøker vært transkribert på 3-4 år. Alternativt, kunne en tenke seg at brukerne for hvert 50. besøk på disse sidene ble bedt om å registrere et (anonymisert) felt fra FoB1950. Da ville en kolonne (ett tema) fra FoB være ferdig transkribert i løpet av et drøyt år (selv om man krevde dobbeltregistrering for validering).

Åpne kilder
For åpne kilder kan en tenke seg å lage en kildetabell med åpne felt tilsvarende hvert felt i kilden, der transkriberte verdier kan skrives inn. Kildetabellen bidrar til å sikre at det er en og bare en person i databasen som er knyttet til den aktuelle registreringen. Hvert felt knyttes til den korresponderende scannede teksten fra kilden. Hver slektsforsker, som leter etter en person, vil lett kunne skrive inn opplysninger fra kilden. Samtidig vil det være lett for senere brukere å verifisere transkriberingen fordi alle vil kunne se det scannede bildet ved siden av. Det vil være viktigst å få transkribert feltene som brukes til søking først og dernest de som brukes til statistikk. Slektsforskere bør kunne få velge å bare skrive inn ett felt eller ta flere felt. Etter hvert som mange vil ønske å finne sin familie i alle kilder, vil vi antagelig relativt raskt få en god dekning.

Lukkede kilder
For lukkede kilder kan en tenke seg å klippe ut utvalgte enkeltfelter for å få disse transkribert av frivillige som løsrevne enheter. På denne måten bør det kunne være

mulig å få transkribert isolerte fornavn, datoer og lignende uten at det skaper et sikkerhetsproblem. Transkribering av etternavn kan være et større problem, men kanskje vil det være mulig å få forståelse for å vise isolerte etternavn. En kunne også tenke seg å legge inn fiktive navn, slik at en bruker ikke kan være sikker på at navnet faktisk finnes i databasen. Det å se et etternavn vil dermed bare gi informasjon om at det sannsynligvis fantes minst en person i landet med det etternavnet da folketellingen ble gjort.

6.4.1 Mobilisering til transkribering

Det er mange måter å organisere frivillig transkribering. I det etterfølgende vil vi beskrive flere alternativer som alle forutsetter at brukeren har logget inn og registrert seg. Når personen logger ut, kan man gi en tilbakemelding om hvor mange felt personen har transkribert og på kvalitet der det er mulig.

Den enkleste måten er å ha et felt der man kan skrive inn transkribert tekst for det tilsvarende feltet i den scannede kilden. Videre er det mulig å gjøre dette mer avansert ved at programmet foreslår verdier og brukeren har mulighet til å bekrefte disse. Det kan også være mulig å utvide dette ytterligere ved å lagre metadata til hvert felt med hvem som har transkribert når, om man valgte foreslått verdi og om andre har bekreftet denne transkriberingen. Den typiske bidragsyter er en slektsforsker som leter etter sine aner og så skriver inn data for denne personen og eventuelt andre personer fra samme kilde.

Alternativt kan man la programmet velge felt. Bruker kan velge om man ønsker navnefelt, tall, navn som ligner på ”Ole” eller spesielt vanskelige felt. I en slik løsning er det mange flere muligheter. Man kan bruke dobbelt registrering ved at det alltid er to uavhengige transkriberinger eller bare ta stikkprøver for å sjekke kvalitet. En slik fremgangsmåte kan også brukes på lukkede kilder da den som transkriberer bare ser ett felt av gangen. Den typiske bidragsyter er en som synes det er morsomt å transkribere. Det stiller større krav til at programmet er brukervennlig.

Etablere ”konkurranser”
Hvis programmet er brukervennlig og vi lager statistikk over hvem som har bidratt mest for hver dag, uke, måned og år, og for hver kilde/type felt kan dette kanskje motivere flere til å transkribere. Noen vil like å transkribere fødselsdatoer, andre fornavn, steder eller yrke, og en kan her også tenke seg å gjøre det mulig å velge å arbeide for eksempel bare med navn som ligner på ”Ole”. Kvaliteten kan kontrolleres ved å jevnlig sjekke mot felt der transkriberingen er kjent og la flere personer transkribere samme felt.

”Tvungen” transkribering
For en HBR-wiki kan en tenke seg å legge inn muligheter for at brukerne med jevne mellomrom blir spurt om å transkribere et felt (for eksempel hvert 10. minutt). Man kan ha en ”snill” løsning der brukeren kan slå av dette. Kvaliteten av arbeidet med en slik løsning vil sannsynligvis være noe mer variabel enn for mer frivillige løsninger.

Et annet alternativ som har vært nevnt er å se på muligheter for samarbeid med organisasjoner som benytter såkalte captchas[1]. Dette er vanligvis et bilde av en forvrengt tekst som krever manuell tolkning og som brukes som en verifikasjon ved pålogging til en tjeneste. Tekster fra håndskrevne dokumenter kunne trolig fungere som captchas. En vanlig captcha må imidlertid ha en tolkning som er kjent på forhånd, men en kan tenke seg at det benyttes to hver gang, en kjent og en ukjent.

Betaling
Det vil også være mulig å lage en betalingsløsning knyttet til den frivillige transkriberingen, der alle som ønsker det internasjonalt kan få betaling pr transkribert felt av tilfredsstillende kvalitet. Her kan det ligge en mulighet til å få transkribert mange felt, kanskje spesielt tall, til en lav kostnad og med liten administrativ overhead. Transkribering av fødselsdato kunne for eksempel være veldig nyttig. Løsningen kan eventuelt senere utvides til også å gjelde felt der det kan bli aktuelt å kreve taushetserklæring, for eksempel etternavn. Aktive bidragsytere kan kontaktes for å høre om de er interessert. Det er få, om noen, lignende muligheter til å tjene penger på internett. Bidragsytere kan komme fra for eksempel Øst-Europa og India, men også norske pensjonister, skoleelever og uføre kan være aktuelle. En utfordring kan være å få publisert denne muligheten i de riktige miljøene. I løpet av flere år vil imidlertid informasjonen spre seg. Med en betaling på 10 øre pr transkribert felt, vil man da kunne få transkribert de viktigste feltene i de mest aktuelle kildene for 1-2 mill. kr.

Før en betalingsløsning velges er det imidlertid viktig å tenke igjennom hvordan dette vil påvirke de som bidrar i en ubetalt transkriberingsdugnad. Vil de fleste som bidrar i dugnaden da også ønske betaling? Og vil muligheten til betaling virke stimulerende eller vil den virke demotiverende? Usikkerheten knyttet til dette, kan bety at man bør starte opp uten betaling og vurderer betaling etter hvert og eventuelt på et annet nettsted.

Statistikker basert på lukket del
Selv om informasjonen om enkeltpersoner fra de lukkede registrene ikke kan vises, vil aggregerte data og statistikk fra disse registrene kunne vises. Dette vil kanskje kunne bidra til å øke interessen for transkriberingen, selv om det ikke er en direkte kobling mellom en enkelt transkribering og bedre statistikk. Mulige statistikker kan bestå i antall transkriberte eller lenkede i forskjellige deler av landet, sammenhengen mellom levealder til foreldre og barn, utvikling av familieforhold, mobilitet osv.

7 Bruk av HBR-wiki

Alle kilder fram til og med 1910 kan brukes i HBR-wiki. Hvis vi kan utvide dette, vil det lette lenkingen av personene. Men det er meget viktig å ikke strekke seg så langt at man får problemer med personvernet og må søke godkjenning fra Datatilsynet. Her er noen forslag:

[1] CAPTCHA (kort for "Completely Automated Public Turing test to tell Computers and Humans Apart"), er en type turingtest for å finne ut om brukeren er en datamaskin eller menneske. Ofte brukes bildet av en forvrengt tekst til dette.

Liste over døde
Ved å tillate lenking av personer i HBR-wiki til lister over døde, vil dette lette lenkingen til folketellingene. Kanskje kan slektsforskere skrive inn bosted på tidspunkt for de aktuelle folketellingene 1920, 1930 og 1950 for personer som er koblet til listen over døde. Det vil gjøre transkriberingen og lenkingen av disse folketellingene mye lettere.

Lenking til åpne kilder
Noen kommunale folketellinger og kirkebøker etter 1910 er publisert. Det er mulig å tillate lenking av disse personene, men for eksempel bare for personer som også lenkes til listen over døde.

Private wikier
WeRelate planlegger å utvide et eksisterende program for private wikier slik at det skal være lett å lage en kobling mellom HBR-wiki og den private wikien. Dette gir en bruker mulighet til å lage et oppdatert slektstre for hele familien. Hvis det er utviklet pen grafikk for dette, kan det bli meget populært. Det kan være mulig å tillate at private wikier kan sendes inn og bidra til lenkingen i den lukkede delen. Vi må være forberedt på at kvaliteten på slike private wikier er blandet, men det vil være mulig å ta hensyn til dette med bruk av verifikasjon.

8 Oppsummering og konklusjon

Gjennom denne rapporten har vi sett på ulike automatiske teknikker som kan være aktuelle som hjelp til transkribering av historiske kilder og hva som er state-of-the-art for teknikker som har vært anvendt på lignende problemstillinger. Vi har sett på oppdeling i skjemastruktur, fargeseparering og gjenkjenning av håndskrevet tekst.

Automatisk oppdeling i skjemastruktur er identifisert som et basiselement i et system som skal tilby automatiske verktøy for hjelp til transkribering. Uavhengig om det er manuelle eller automatiske trinn som benyttes videre i prosessen, kan slik oppdeling være nyttig fordi den gir en mulighet til å jobbe med transkribering av skjemaer både rad for rad, kolonne for kolonne og celle for celle. Vi har derfor også sett på hvordan dette kan løses og implementert en prototype beregnet for analyse av husstandsskjemaene fra folketellingen i 1950.

Automatisk gjenkjenning i tradisjonell forstand er mest aktuelt for tall, mens for sammenhengende håndskrift er det foreløpig ikke mulig å oppnå tilstrekkelig høye gjenkjenningsrater. Her foreslår vi derfor i stedet bruk av teknikker som kan være en støtte til den manuelle prosessen som for eksempel automatisk gruppering eller gjenfinning av lignende ord. I tillegg foreslår vi at en der det er mulig reduserer problemet gjennom å begrense ordlister. Bruk av informasjon fra andre kilder og lenking med denne informasjonen er en måte å oppnå reduserte ordlister på og i blant kan dette redusere gjenkjenningsproblemet til et verifikasjonsproblem. Noe som straks gjør automatiske teknikker mer aktuelle også for tekst.

I tillegg til løsningene nevnt over, kan gode internettløsninger som en HBR-wiki muliggjøre rekruttering av store frivillige manuelle ressurser til å bidra i

transkriberingsarbeidet. Hvordan ressursene kan utnyttes vil avhenge av om arkivene er åpne eller lukket. For åpne arkiver bør slike ressurser greit kunne utnyttes. For lukkede ressurser kan oppdeling av skjemaer i isolerte og anonymiserte celler være en nøkkel til å utnytte slike ressurser. For mobilisering av slike ressurser er noen ulike alternativer foreslått som tvungen transkribering ved bruk av ressursene, frivillig transkribering med et konkurranseelement eller for å få tilgang til statistikker eller frivillig transkribering mot betaling.

Konklusjonen fra dette arbeidet er at en ved valg av de rette automatiske teknikkene kan få verktøy som kan bidra til å lette det manuelle arbeidet og dermed redusere kostnadene ved transkribering av historiske kilder. Gode verktøy og internettløsninger kan også gi muligheter for å utnytte store frivillige ressurser i transkriberingen, noe som både kan redusere kostnaden og ikke minst tiden det vil ta å transkribere store arkiver.

Referanser

[1] Bilane, P.; Bres, S.; Challita, K,; Emptoz, H. *A segmentation free approach for indexing digitized syriac manuscripts.* 17th European Signal Processing Conference, Glasgow, August 2009.

[2] Leydier, Y.; Ouji, A; LeBourgeois, F.; Emptoz, H. *Towards an omnilingual word retrieval system for ancient manuscripts.* Pattern Recognition, Vol 42, No. 9, September 2009.

[3] Liu, C-L.; Nakashima, K.; Sako H.; Fujisawa H. *Handwritten digit recognition: benchmarking of state-of-the-art techniques.* Pattern Recognition (2003), pp. 2271-2285.

[4] Nielson, H.E; Barrett, W.A. *Consensus-Based Table Form Recognition*. 7th International Conference on Document Analysis and Recognition , ICDAR 2003.

[5] Rath, Tony M.; Manmatha, R. *Word spotting for historical documents*. 2007

[6] Rath, Toni M.; Manmatha, R; Lavrenko, V. *A Search Engine for Historical Manuscript Images.* SIGIR 2004.

[7] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Wiley, 1987

[8] *Folketellingen 1950*, http://www.ssb.no/emner/historisk_statistikk/artikler/art-2009-08-11-01.html Statistisk Sentralbyrå

[9] *Folketellingen 1950*, http://www.rhd.uit.no/census/ft1950.html Registreringssentralen for historiske data