Gunnar Thorvaldsen
Registreringssentral for historiske data
Universitetet i Tromsø

NORMALISERING AV PERSONNAVN.

Historiske navnedata som folketellinger og kirkebøker blir brukt til mange forskningsformål. Ofte består mye av kildestrevet i å finne igjen personer fra den ene kilda til den andre for å få oversikt over individers livsløp. Skal dette arbeidet bli overkommelig, er det nesten en forutsetning at kildene fins i maskinleselig versjon. Gjenfinning kan tenkes foretatt automatisk ved hjelp av EDB-programmer, men hittil har den vanlige metoden vært å sortere kildene etter ulike nøkler, og finne fram i datalistene manuelt.

Det viktigste kriterium for identifikasjon er gjerne for- og etternavn. Den som prøver metoden, finner snart ut at på 1800-tallet var skrivemåten av navn langt mindre konsekvent enn nå. Vi kan si at skrivemåten snarere fulgte skriveren enn den som bar navnet. Presten eller folketelleren skrev Niels eller Nils etter eget hode.

Dette medfører at arbeidet med identifikasjon tar svært mye tid. Skal vi se etter Anne eller Ane? Heter hun Ni(e)lsdatter til etternavn kan vi i alle fall tenke oss 4 kombinasjoner. En form for normalisering eller standardisering vil dermed kunne spare oss for mange oppslag i listene. Og skal automatisk lenking ha noe for seg, må EDB-programmene bli fortalt hvilke navneformer som sannsynligvis er synonyme. Siden kildene brukes bl.a. i navneforskning er det ikke akseptabelt at navneformene endres før de legges inn i maskinen.

Her har vi emnet for denne artikkelen: Hvordan kan vi rasjonalisere identifikasjon ved å redusere antall navneformer uten å ta bort noe vesentlig av den informasjon som ligger i at navn skal være ulike? Siden siktemålet er praktisk, skal vi ta forholdsvis lett på fonetisk teori omkring navneformer. Vi legger heller til grunn en mer intuitiv forståelse av hvilke skrivemåter av navn som er synonyme, og konsentrerer oppmerksomheten om det tekniske. Det betyr på ingen måte at kommentarer til denne artikkelen bør følge samme oppskrift!

Vi kan i utgangspunktet tenke oss å utføre normaliseringen på to ulike hovedmåter. Enten ved å formulere formelle kriterier for hvilke tegn og tegnsekvenser som skal forandres (alle forekomster av -ie- forandres til -i-?). Eller vi lar EDB-programmet slå opp i ei ordliste hvor det

Normalisering av personnavn
Gunnar Thorvaldsen
Proceedings of NODALIDA 1981, pages 32-37

for hvert navn finner hvilken form som er standard. Jeg har selv prøvd begge disse opplegg på et par folketellinger. En vurdering av resultatene følger her.

Utgangspunktet var folketellinga for Alta i 1875. Her fins navn og andre opplysninger om 2419 individer. Et EDB-program ble skrevet som går igjennom tellinga, plukker ut alle fornavn, sorterer og oversetter hvert av dem ifølge formelle kriterier etter førstnevnte modell.

Hvilke kriterier var innbakt i programmet?

1. To like tegn reduseres til ett. (Anne blir Ane)
2. h fjernes, men ikke i begynnelsen av ord. (Alethe blir Alete)
3. c blir til k, foran e og i til s. (Carl blir Karl, Cecilie blir Sesilie)
4. aa blir til å (unntak fra regel 1).
5. ou blir til au. (Poul blir Paul)
6. w blir til v.
7. ph blir til f. (Philip blir Filip)
8. I endelsene -na, -ia, -ta, -ka, -la blir a erstattet med e. Mia blir Mie.
9. I endelsene -rd, -id, -ld, -nd blir d fjernet. Sigrid blir Sigri.
10. æ blir til e (Bæret blir Beret)
11. Endelsen -af og -au blir til -av. (Olaf blir Olav)

Kriteriene er resultat av mange navnelister og mye prøving og feiling. En nyttig metode var å sortere navna baklengs for å få oversikt over endelsene. Flere algoritmer måtte fjernes, f.eks. hadde det liten hensikt å erstatte ie med i i Niels når Daniel samtidig ble Danil.

Resultatet ble ei liste med ustandardiserte og standardiserte navneformer hvor programmet etter mitt syn har gjort mer nytte enn ugagn. Dvs. at for hvert kriterium er antall vellykkede standardiseringer større enn antall mislykkede. For hvert navn har programmet også satt på antall forekomster av navnet i kilda og et merke for type standardisering.

Den oppmerksomme leser vil forstå at lista inneholder en rekke navn som har blitt forandret uten grunn. F.eks. vil det bare forvirre når Johan finnes som Joan. Dessuten fins det mange standardiseringer som det er tungvint å skrive algoritmer for fordi de er vanskelige å innpasse i noe mønster.

Dette gjelder f.eks. fjerning av den siste e'en i Henery. For å gi en ide om problemenes omfang har jeg laget en tabell over forekomsten av de ulike typer standardisering (vertikal variabel) og hvor vellykkede de ble (horisontal variabel). Verdiene vertikalt tilsvarer standardiseringstypene ovenfor, blank er ingen standardisering.

Horisontalt står "-" for mislykket standardisering, "." for manuell standardisering og "&" for ukurant term ("udøpt", "barn" o.l. kan knapt kalles fornavn).

		. MANUE	- MISLY	& UKURA	INGEN	SUM
1	DOBBEL	7	6	2	48	63
2	H BORT	15	3	0	58	76
3	C>S/K	1	0	0	10	11
4	AA> Å	2	0	0	5	7
5	OU>AU	0	1	0	1	2
6	W > V	1	0	0	2	3
7	PH> F	0	1	0	1	2
8	-NA>E	10	8	0	81	99
9	-RD>R	3	0	0	26	29
10	Æ > E	1	0	0	3	4
	INGEN	68	0	5	272	345
SUM		108	19	7	507	641

Vi ser at de 2419 menneskene i 1875-tellinga for Alta hadde 641 ulike fornavn. (Doble navn er delt opp, ett-tegns forkortelser er kuttet ut.) Når det gjelder opptelling av de ulike typer standardisering må jeg understreke en viktig begrensning. For hvert navn er det bare plass til en kode. Mange navn blir standardisert etter flere kriterier, og da er siste tilslag avmerket.

En form for maskinell standardisering er foretatt i nesten halvparten av tilfellene. Som regel dreier det seg om forandring av endelser etter regel 8 og 9, fjerning av h eller forenkling av doble tegn (regel 1 og 2). Langs den vertikale variabelen ser vi at det varierer hvor vellykket programmet arbeidet, men i under 10% av tilfellene gjorde det vondt verre. Dette er rettet med editor.

Verdien "manuelt" gjelder noe annet. Dette er tilfeller hvor navna ikke var standardisert maskinelt, men hvor vi allikevel kunne ha nytte av et redusert antall skrivemåter. I mange tilfeller gjelder det også navn som allerede var forandret i programmet, men ikke nok. Dette er gjort med editor og gjelder ca. 17% av navna. Med tillegg for de mislykkede, måtte 20% endres manuelt. Det viktigste er allikevel at de 641 navneformene er redusert med vel en en tredjedel til 420.

Resultatene blir noe anderledes for andre kilder eller regioner. Men det kan neppe rokke ved den konklusjon at mange navn kan standardiseres ved hjelp av forholdsvis enkle algoritmer. Disse vil imidlertid lett gripe inn der det er uønsket, samtidig som de ikke tar seg av en rekke navn som vi ønsker å standardisere. Problemene kan møtes med mer

kompliserte regler for konvertering, men antagelig vil en annen strategi være mer rasjonell: Søking i navnelister som nevnt innledningsvis.

Ei slik liste må inneholde to former av navna, den originale og den standardiserte. Og det var jo ei slik liste vi nettopp produserte! Denne lista inneholder da alle fornavn i den kilda vi tok utgangspunkt i, slik at navna der nå kan standardiseres ut fra den. Søking i slike lister tar nødvendigvis en del maskintid siden vi får et oppslag pr. navn, men ved såkalt binær søking går det forholdsvis raskt.

Et annet problem er at når vi går løs på nye kilder vil det dukke opp navn som ikke finnes i lista. Vi må derfor ha en prosedyre som standardiserer dem, først maskinelt og deretter manuelt. Når de er kontrollert kan de inkorporeres i navnelista slik at den gradvis bygges opp. For at den ikke skal bli for stor og uhåndterlig kan vi sette grenser for hvilke navn som kommer med, f.eks. kutte ut alle med forekomst lik en.

Ved å trekke inn folketellinga fra år 1900 for Harstad har jeg prøvd ut et slikt opplegg. Her bodde 2109 personer med tilsammen 3306 fornavn. 811 var ulike, og av dem var det 511 som ikke fantes i 1875-tellinga for Alta. Altså en betydelig restkategori, men hele 344 av disse forekom bare en gang. Dette tyder på at det vil gå forholdsvis raskt å bygge opp et register over frekvente fornavn med standarder.

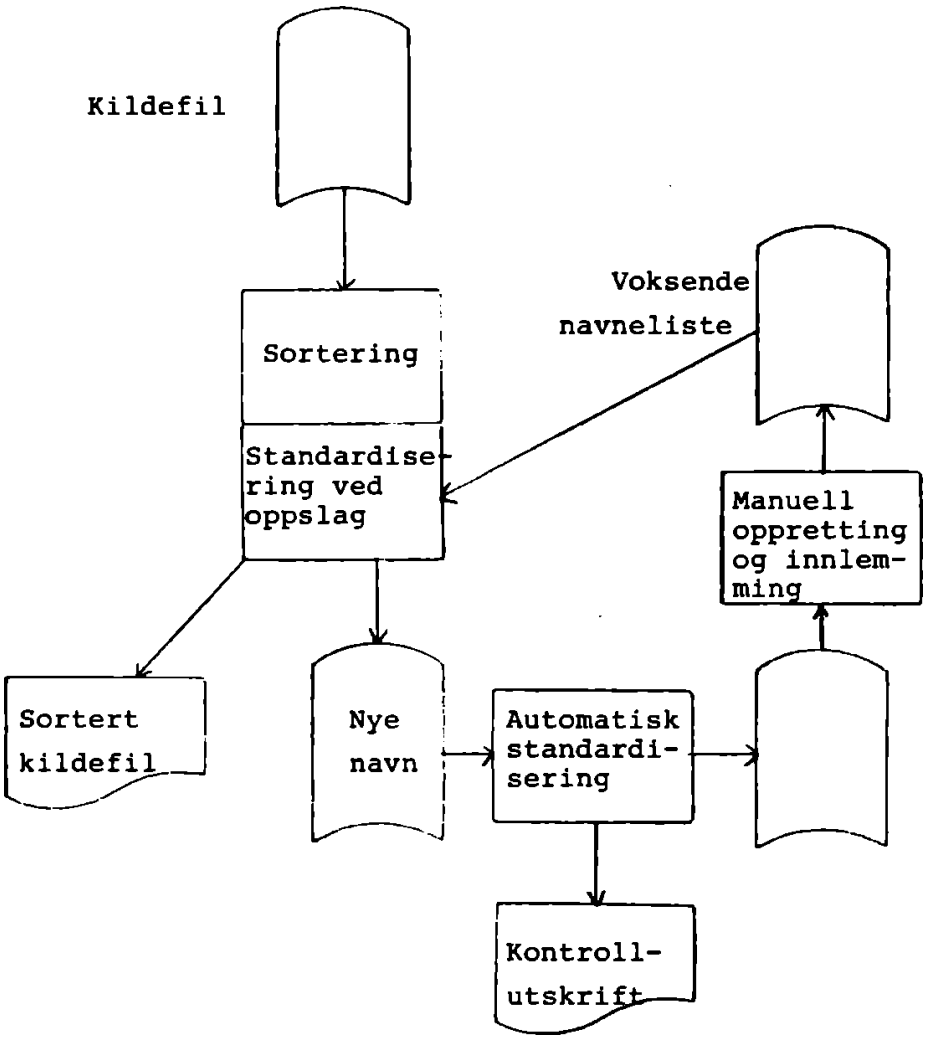
Det kan gjøres med utgangspunkt i følgende flytskjema:

Hver gang vi sorterer ei kilde på navn og ønsker normalisering, slår programmet opp i lista over navneformer med standarder. Når et navn ikke fins her, kommer det med i lista over nye navn. Denne krympes slik at hvert navn bare forekommer en gang sammen med en frekvens. Lista gjennomgår deretter en automatisk normalisering, og på grunnlag av kontrollutskriften foretas en manuell oppretting. Eventuelt kan man utelate de minst vanlige navna før den nye lista innlemmes i den gamle, voksende navnelista.

Hvis det var mange nye navn, vil det lønne seg å kjøre sorteringsprogrammet om igjen for den aktuelle kildefila slik at alle navn i utskriften er sortert etter sin normalversjon.

Slik det er skissert her tar systemet seg bare av fornavn. Det vil imidlertid være en enkel utvidelse å ta med etternavn. Og når det gjelder patronymika, kan man basere seg på fornavnlista.

FLYTSKJEMA FOR NORMALISERING AV PERSONNAVN



ALLMENN NYTTE AV DATA FRA RHD I NAVNEGRANSKING

Registreringssentral for historiske data ved Universitetet i Tromsø har som oppgave å databehandle folketellinger og kirkebøker fra 1800-tallet for utvalgte deler av Norge. Siden kildene skrives mest mulig bokstavrett av, mener vi EDB-utgavene også vil være til nytte i navnegransking.

Navnetilfanget i materialet omfatter både person- og stedsnavn. Når det gjelder personnavn skal de nevnte kildene inneholde hele befolkninga og dermed utgjøre en bortimot fullstendig navnesamling. (Et viktig unntak gjelder etniske minoriteter).

Når det gjelder stedsnavn gir nok de nevnte kildene bare en grovere oversikt. De aller fleste gårds- og bruksnavn er med, ofte med ulike skrivemåter i de ulike kildene. Det er noe mer tilfeldig i hvilken grad tellerne har fått med navn på mindre (husmanns) plasser og jordlapper.

Imidlertid avspeiler kildene godt hvor mange måter man kunne skrive (og uttale) et navn på. Dessuten åpner de store muligheter for forskere som vil sette navn i sammenheng med egenskaper som yrke, fødested o.l.

Hvordan kan vi så presentere data for brukerne? For det ene kan vi levere maskinskrevne kopier av kildene. Videre kan vi alfabetisere personer og bosteder etter de ulike navnetyper før utskrift. Og endelig kan vi levere magnetbånd med navnetilfanget til forskere som selv ønsker å bearbeide kildene med EDB.

Registreringssentralen vil i første omgang konsentrere seg om kilder fra utvalgte deler av landet. En foreløpig oversikt er vedlagt. Både Nord-, Midt-, Vest- og Øst-Norge blir representert. Vi tilstreber også en næringsgeografisk representativitet. Helt avgjørende er allikevel forskernes behov. Meld derfor fra om konkrete forskningsprosjekt hvor det er behov for kildematerialet.

På lenger sikt er det aktuelt å starte registrering av materiale som vil være mer detaljert mht. stedsnavn. Vi tenker da f.eks. på matrikkelforarbeide og pantebøker.