Statistics Norway
Statistisk sentralbyrå

Ole Villund

Notater

Klassifisering ved hjelp av tekst
- noen resultater fra yrkeskodingen i Arbeidskraftundersøkelsen

Seksjon for arbeidsmarked

1 Innhold

2 Innledning

2.1 Formål

Dette notatet beskriver noen egenskaper ved tekst og bruk av tekst til klassifisering. Formålet er å dokumentere noen metoder og resultater fra analyse av tekst og klassifisering ved hjelp av tekst. Den tekstbaserte yrkeskodingen i Arbeidskraftundersøkelsen (AKU) har vært viktig i utviklingen av metoder for yrke i registerbasert statistikk, og alle eksemplene i dette notatet er om yrke i AKU. Programsystemet for behandling av yrke i Arbeidstakerregisteret omfatter metoder for automatisk tekstbasert yrkeskoding. Utviklingen av dette systemet har i stor grad vært støttet av analyser av forholdet mellom tekst og yrke i AKU, som også forutsatt oppbygging av kompetansen innenfor tekstbaserte metoder.

Notatet gir også en nærmere omtale av kjennemerket **yrke** i forhold til viktige egenskaper. Dette blant annet for å vise sammenhengen mellom yrke og andre kjennemerker enn teksten som brukes til klassifisering. Yrke er også et viktig kjennemerke i seg selv, og i register brukes det til detaljert statistikk og levering av mikrodata til forskning. Notatet er beregnet på intern dokumentasjon og kompetanseutvikling, og kan ha interesse for oppdragsgivere som bruker av yrkesstatistikk eller mikrodata med yrke. Endel av stoffet kan kanskje være interessant også for utenforstående som jobber med tekstdata.

2.2 Oppsummering

- Yrke er et viktig kjennemerke, og kan ikke direkte avledes av andre kodede variabler.
- Yrkesklassifiseringen i AKU er helt avhengig av teksten som oppgis som yrkestittel og arbeidsoppgaver, selv om det kan brukes noen andre tilleggopplysninger.
- Tekst er en spesiell datatype med egenskaper som ikke kan sammenliknes med nummeriske og kategorielle variabler, og som krever egne metoder og kompetanse.
- Når man har samlet opp større mengder tekst som er klassifisert og kontrollert, kan dette materialet utnyttes til:
    - Kvalitetskontroller: som retrospektiv analyse, eller overvåkning av inndata.
    - Automatisk koding: helautomatisk, eller datastøttet manuell koding.
    - Generering av kodeindekser til bruk ved søking (webkatalog), manuelle eller automatiske systemer.
    - Grunnlagsanalyser: basiskunnskap om tekstdata, som f.eks. språklige trender.
- I dette notatet vises endel egenskaper ved yrke, og i hvilken grad yrke kan klassifiseres automatisk utfra tekst.

2.3 Bruk av yrkesdata

2.3.1 Statistikk

Yrke har vært med i **Arbeidskraftundersøkelsen** (AKU) siden starten i 1972, og det publiseres årlig statistikk etter yrke for alle sysselsatte. AKU er en meget stor spørreundersøkelse og det lages detaljerte yrkesfordeling etter kjønn og aggregerte yrkesfordelinger på flere kjennemerker som arbeidstid, fylke og alder. AKU-tall for yrke er årsgjennomsnitt og publiseres sammen med 4.kvartal. Yrke er kodet etter Standard for yrkesklassifisering (NOS C521) "STYRK" fra 1994, som bygger på ISCO-88 COM (EUs versjon av ILO-yrkesstandard). Norsk offisiell statistikk etter yrke er derfor i stor grad sammenliknbar med andre innen EU-området. Andre viktige egenskaper ved AKU er lange tidsserier og kompletthet. Praktisk talt alle sysselsatte i utvalget får yrkeskode på det mest detaljerte nivå.

Den registerbaserte sysselsettingsstatistikken inneholder mer detaljerte tabeller og er særlig aktuell for tall på kommunenivå. Klassifisering og imputering av yrke i register er også beregnet på levering av mikrodata til forskning. Fra 2001 har det vært det krav til innrapportering av yrke på innmeldinger til **Arbeidstakerregisteret** (AA), som administreres av Rikstrygdeverket (RTV). Hovedbegrunnelsen for dette var ønske om sykefraværsstatistikk etter yrke, men det er klart at yrke er interessant i mange andre sammenhenger. Eksempler på bruk av yrke i register som kan være yrkesfordeling i kommuner, detaljert yrke etter utdanning, næring, osv. Videre vil yrke i mikrodata fra register være interessant for forskning på arbeidsmiljø, sykdom, lønn, osv.

Vi regner med at yrke i AKU er av god kvalitet, noe som også blir belyst i dette notatet. Årsaken til at AKU-data brukes i arbeidet med yrke i register er to hovedoppgaver:

- Kontroll og sammenlikning. Dette gjelder både på mikro (jobber) og makro (tabeller).
- Estimering av yrke på sysselsatte i register som ikke har yrkeskode. Dette gjelder både de som mangler yrkeskode i AA, og sysselsatte hvor data ikke hentes fra AA.

Yrke ble tatt med som nytt kjennemerke i den registerbaserte sysselsettingsstatistikken fra og med statistikkåret 2003, noe som er dokumentert i diverse notater nevnt i referanseliste. Endel av dette notatet ble til med tanke på videreutvikling av metodene for klassifisering og estimering av yrke i register og registerbasert statistikk.

2.3.2 Analyser

Kjennemerket yrke er definert i STYRK som gruppering av arbeidsoppgaver, klassifisert etter kompetansenivå og spesialisering. Kompetansenivå er gruppert etter formell utdanning, men skal også omfatte tilsvarende reell kompetanse som erfaring, internopplæring, o.l. Spesialisering grupperes etter ferdigheter, verktøy, maskiner, materialer, produkter, m.m. Andre kriterier er grad av selvstendighet, rutinearbeid og manuelle oppgaver.

Arbeidsoppgaver og derved yrke vil være avhengig av egenskaper både ved arbeidsplassen og den som utfører arbeidet. Bedriftens aktivitet altså hva som produseres, klassifiseres ved næringskode. Andre viktige egenskaper ved bedriften er størrelsen, organisasjonsform, m.v. Den sysselsattes formelle kompetanse framkommer av utdanningskode, andre bestemmende egenskaper er som regel mer indirekte målbare. En må kunne si at yrke er viktig fordi det skal beskrive arbeidsoppgaver direkte, og det er en interessant egenskap ved arbeidsmarkedet på en måte som ikke kan avledes fra andre kjennemerker.

Etterspørselen av arbeidskraft innen bestemte yrker bestemmes av arbeidsgivers behov for å få utført visse arbeidsoppgaver og mulighet til å betale for dette. Tilbudet vil avhenge av bl.a. arbeidssøkeres kompetanse altså både formelle og uformelle kvalifikasjoner og mobilitet. Det kan nok variere i hvilken grad arbeidsoppgavene i en gitt stilling er faste eller kan tilpasses noe til den som skal utføre dem. Videre kan arbeidsoppgavene kan endres som følge av omlegging av produksjon, f.eks. pga. konkurransesituasjonen, omstillinger i bedriften, omorganisering, allokering av kompetanse, m.m. Slike endringer kan ha betydning for sysselsetting, lønnsdannelse, kompetansebehov og arbeidsmiljø. Ved hjelp av **yrke** kan man analysere dette på tvers av f.eks. næringsstruktur og formell utdanning.

3 Yrke i AKU-data

Her beskrives yrkesklassifiseringen og noen viktige egenskaper ved yrke i AKU. Visse data kobles på fra register f.eks. arbeidsmarkedsstatus i register og utdanningskode. I notatet brukes ordinære AKU-filer uten ytterligere påkoblede datakilder, slik at vi f.eks. ikke tar stilling til yrkeskoden i register. Dette for å gi et mest mulig riktig bilde av yrkesklassifisering utfra de variabler som er tilgjengelig på klassifiseringstidspunktet.

3.1 Klassifisering av yrke i AKU

De som blir definert som sysselsatte i AKU blir spurt om yrkestittel og arbeidsoppgaver. Dette blir notert som vanlig tekst på inntil 30+30=60 tegn. Teksten blir lagret for senere klassifisering. Dette til forskjell fra f.eks. svensk AKU, der klassifiseringen skjer ved selve intervjuet. Yrkeskodingen skjer i sin helhet "manuelt", som er en datastøttet skjønnsmessig klassifisering ved Seksjon for databearbeiding. Foruten hovedopplysningene yrkestittel og arbeidsoppgaver har koderne nødvendige data som bedriftens navn, næring, størrelse, sektor, osv.; personens utdanning, arbeidsmarkedsstatus, o.s.v. Det er svært få som ikke kan klassifiseres, slik at praktisk talt alle sysselsatte yrkeskode på det mest detaljerte nivå (4-siffer yrkeskode). Retningslinjer for kodingen er gitt i *Standard for yrkesklassifisering* (NOS C521) og interne instrukser.

Til bl.a. bruk i kodingen finnes *Yrkeskatalogen* som inneholder et varierende antall yrkestitler til hvert yrke og som oppdateres løpende. Innenfor et yrke kan yrkestitlene være rene synonymer, mindre variasjoner i arbeidsoppgaver eller skille mellom f.eks. formann, fagarbeider og lærling. Yrkeskatalogen er ikke en kodingsindeks i vanlig forstand da den har flere bruksområder, som oppslagsverk for arbeidsgivere som skal levere yrkeskode til Arbeidstakerregisteret. Titlene er som regel ganske formelle og lange i forhold til det som brukes i bedriftene.

I de tilfeller det er oppgitt en egnet yrkestittel og konkrete arbeidsoppgaver, regner vi med at yrkeskodingen i AKU er av god og jevn kvalitet. Dette fordi klassifiseringen utføres av en stabil gruppe av personer med lang erfaring. Viktige hjelpemidler er lett tilgjengelige data, instruks- og oppslagsverk, intern drøfting og rådgivning fra Seksjon for arbeidsmarked. Noen feilkilder i kodingsarbeidet er tastefeil, misforståelse av instruksen og uvaner, som f.eks. at man koder alle tvilstilfeller til en restgruppe. Det er et stort potensial i å bevisstgjøre IT-miljøer på betydning av brukergrensesnitt i forhold til datakvalitet.

3.2 Datakilde til forsøk

AKU er en løpende PC-assistert intervjuundersøkelse med omlag 24 000 personer pr. kvartal. For å undersøke tekst og yrkeskoder er det aktuelt å bruke data for **sysselsatte**, dette utgjør omlag 15 000 personer pr. kvartal. Selv om dette er et meget stort utvalg, er over halvparten av de eksisterende yrkeskoder så små at det f.eks. ikke estimeres nivåtall i den vanlige publiseringen. Og når vi skal analysere tekst, trenger vi enda flere observasjoner. Vi kan jo ikke øke utvalgstørrelsen direkte siden vi undersøker eksisterende data. Så for å minske usikkerheten tar vi med data fra flere år og lager et utvidet utvalg. Vi må vurdere noen forhold som kan gi andre feilkilder enn utvalgsusikkerhet for et punktestimat. Viktigst i denne sammenhengen er:

- Endringer i selve yrkesstrukturen.
- Endringer i sammenhengen mellom yrke og forklaringsvariablene.
- Paneleffekter pga. at hver person er med 2 år.
- Klyngeeffekter pga. at trekkeenheten er familie mens statistikkenheten er person.

Det har vært en langsom økning av sysselsatte kvinner og fortsatt skjer en utjevning mellom kjønnene i sysselsettingsnivå. Men innenfor de fleste yrker er det store kjønnsforkjeller. Av langsiktige endringer kan vi nevne at yrker som krever høy kompetanse øker jevnt, mens lavere kontoryrker og yrker i primærnæringen reduseres. Det skjer også en endring av aldersfordelingen av de sysselsatte, som mer skyldes endringer i befolkningsstrukturen enn selve sysselsettingen. Det skjer liten endring av forskjellen mellom menn og kvinner i yrkesstruktur. Hovedtendensen for denne perioden er stabilitet og vi velger å se bort fra eventuelle trendeffekter i undersøkelsene.

Tabell 3-1: Sysselsatte pr. kvartal etter kjønn, alder og yrke. AKU 2000-2004. Prosent.

	2000k1	2000k2	2000k3	2000k4	2001k1	2001k2	2001k3	2001k4	2002k1	2002k2	2002k3	2002k4
I alt	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Kvinner	46.7	46.5	46.6	46.6	46.6	46.6	46.6	47.0	47.2	47.0	47.0	47.1
Menn	53.3	53.5	53.4	53.4	53.4	53.4	53.4	53.0	52.9	53.0	53.0	52.9
16-19 år	3.5	4.1	4.4	4.3	3.2	3.9	4.3	4.2	3.6	3.8	4.5	4.3
20-24 år	8.1	8.4	8.6	8.3	8.0	8.4	8.2	7.9	7.9	8.3	8.3	8.0
25-39 år	37.4	37.4	37.7	38.0	37.1	36.9	37.3	37.6	36.4	36.5	36.6	37.0
40-54 år	35.9	35.4	35.2	35.4	35.6	35.1	34.9	34.9	35.2	34.8	34.4	34.7
55-66 år	14.2	13.8	13.4	13.5	15.1	14.8	14.5	14.4	15.8	15.6	15.3	15.2
67-74 år	0.9	1.0	0.7	0.6	1.0	0.9	0.8	0.9	1.1	1.0	0.8	0.8
1 Lederyrker	6.5	6.8	6.7	6.9	7.0	7.0	7.0	7.4	7.7	7.6	7.6	7.6
2 Akademiske yrker	10.6	10.4	10.6	10.7	11.2	11.4	11.7	11.8	11.5	11.0	11.0	11.1
3 Høyskoleyrker	22.8	22.5	22.2	22.5	23.3	22.9	22.9	22.8	22.6	22.8	22.8	23.5
4 Kontoryrker	8.8	8.8	9.0	8.6	8.4	8.4	8.7	8.2	8.0	8.0	8.2	7.9
5 Salg/serviceyrker	21.7	21.7	21.5	21.6	21.3	21.3	21.0	21.8	21.6	21.9	21.8	22.1
6 Bønder,fiskere ol	3.7	3.8	3.7	3.6	3.5	3.7	3.4	3.3	3.4	3.5	3.4	3.1
7 Håndverkere	11.4	11.1	11.5	11.6	11.3	11.0	11.0	10.6	10.8	11.1	11.2	10.9
8 Operatør./sjåfør.	7.9	8.0	8.0	7.8	7.7	7.8	7.9	7.9	8.0	7.6	7.7	7.7
9 Andre yrker	6.7	7.0	6.8	6.7	6.4	6.5	6.5	6.1	6.4	6.6	6.3	6.1

	2003k1	2003k2	2003k3	2003k4	2004k1	2004k2	2004k3	2004k4	Std.av.	Rel. Feil
I alt	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0 %	0 %
Kvinner	47.1	47.1	47.4	47.3	47.2	47.3	47.2	47.2	1 %	1 %
Menn	52.9	52.9	52.6	52.7	52.8	52.7	52.8	52.8	1 %	1 %
16-19 år	3.4	3.8	4.3	4.2	3.4	3.8	4.4	4.3	10 %	19 %
20-24 år	7.8	8.3	8.1	8.0	7.9	8.1	8.1	7.8	3 %	10 %
25-39 år	35.7	35.7	35.8	35.9	35.2	35.5	35.4	35.7	2 %	2 %
40-54 år	35.1	34.8	34.6	34.9	35.3	34.8	34.8	35.1	1 %	2 %
55-66 år	16.8	16.3	16.2	16.1	17.1	16.8	16.4	16.2	7 %	5 %
67-74 år	1.1	1.1	1.0	0.9	1.1	1.1	1.0	0.9	15 %	84 %
1 Lederyrker	7.4	7.5	7.5	7.5	7.4	7.3	7.0	6.9	5 %	11 %
2 Akademiske yrker	11.1	10.8	10.9	11.3	11.6	11.6	11.7	12.0	4 %	7 %
3 Høyskoleyrker	23.1	22.9	23.0	23.3	22.9	23.1	23.4	23.8	2 %	3 %
4 Kontoryrker	7.9	7.8	7.9	7.5	7.5	7.5	7.6	7.4	6 %	10 %
5 Salg/serviceyrker	22.1	22.5	22.3	22.4	22.8	22.8	23.0	23.0	3 %	3 %
6 Bønder,fiskere ol	3.3	3.4	3.3	3.1	3.2	3.1	3.0	2.8	8 %	21 %
7 Håndverkere	10.9	11.1	11.0	11.1	11.1	11.2	11.1	11.2	2 %	7 %
8 Operatør./sjåfør.	8.0	7.6	7.7	7.5	7.2	7.1	7.4	7.4	3 %	10 %
9 Andre yrker	6.2	6.3	6.3	6.1	6.2	6.2	5.9	5.5	5 %	12 %

Til sammenlikning er vist standardavviket til andelene over tid og standardfeilen til estimatet (utvalgsusikkerheten). Utover et visst sesongmønster er fordelingen meget stabil. For å balansere utvalgstørrelse med de andre effektene velger vi å konstruere et datasett med alle sysselsatte fra 2000-2004, men med den siste records pr. person. I tabellen sammenliknes tre ulike datasett med hensyn til viktige variabler:

- AKU samlet data 2000-2004. Alle records, hver person kan være med inntil 8 ganger.
- AKU samlet data 2000-2004. 1 record pr. person, valgt nyeste.
- AKU årsgjennomsnitt 2004. Vektet som for vanlig statistikk.

Tabell 3-2: Alternative delutvalg. AKU 2000-2004, etter kjønn alder og yrke. Antall, prosent og -poeng.

	Samlet (alle)		Person (siste)		År (vektet)		Skjevheter		
	Antall	Prosent	Antall	Prosent	Antall	Prosent	Samlet-År	Person-År	Samlet-Person
I alt	304 003	100.0	61 346	100.0	2 275 534	100.0	0.0	0.0	0.0
Kvinner	143 773	47.3	29 233	47.7	1 074 464	47.2	0.1	0.4	-0.4
Menn	160 230	52.7	32 113	52.3	1 201 070	52.8	-0.1	-0.4	0.4
16-19 år	12 740	4.2	3 278	5.3	90 115	4.0	0.2	1.4	-1.2
20-24 år	24 085	7.9	5 797	9.5	181 510	8.0	-0.1	1.5	-1.5
25-39 år	107 339	35.3	21 164	34.5	806 244	35.4	-0.1	-0.9	0.8
40-54 år	108 435	35.7	20 361	33.2	796 252	35.0	0.7	-1.8	2.5
55-66 år	48 487	16.0	9 879	16.1	378 245	16.6	-0.7	-0.5	-0.2
67-74 år	2 917	1.0	867	1.4	23 168	1.0	-0.1	0.4	-0.5
1 Lederyrker	21 610	7.1	4 026	6.6	162 980	7.2	-0.1	-0.6	0.5
2 Akademiske yrker	32 598	10.7	6 229	10.2	267 158	11.7	-1.0	-1.6	0.6
3 Høyskoleyrker	68 831	22.6	13 367	21.8	529 759	23.3	-0.6	-1.5	0.9
4 Kontoryrker	24 433	8.0	4 822	7.9	170 892	7.5	0.5	0.4	0.2
5 Salg/serviceyrker	67 427	22.2	14 455	23.6	521 756	22.9	-0.7	0.6	-1.4
6 Bønder,fiskere ol	11 268	3.7	2 326	3.8	68 766	3.0	0.7	0.8	-0.1
7 Håndverkere	34 439	11.3	6 791	11.1	253 189	11.1	0.2	-0.1	0.3
8 Operatør./sjåfør.	23 838	7.8	4 868	7.9	165 938	7.3	0.5	0.6	-0.1
9 Andre yrker	19 559	6.4	4 462	7.3	135 096	5.9	0.5	1.3	-0.8

Det er ikke betydelige skjevheter i forhold til de nyeste data, så vi går i det videre utfra at gevinsten ved økt utvalgsstørrelse vil oppveie eventuelle ulemper.

Som nevnt er trekkeenheten familie, og man kjenner til at det for enkelte yrker er sammenheng mellom ektefellers yrker (f.eks. som følge av utdanningssted) og mellom foreldre og barn ("arvelige" yrker). Hvis denne *intraklynge-korrelasjonen* er sterk for yrke, vil dette gi skjevheter i estimering av yrkestall fra AKU-data. Det kan uansett være interessant å belyse som et fenomen i det norske arbeidsmarked. Vi betrakter yrke som en nominell kategorivariabel og parameteriserer denne ved å lage binære indikatorvariabler. Korrelasjonen kan da regnes ut for hver gruppering, og vi begrenser dette til yrkesfelt (1-siffer kode). For enkelhets skyld tar vi med de to eldste i hver familie uavhengig av sivilstatus.

Tabell 3-3: Korrelasjon mellom to eldste familiemedlemmers yrke. AKU 2000-2004.

	Nesteldste	1 Lederyrker	2 Akademiske yrker	3 Høyskoleyrker	4 Kontoryrker	5 Salg/serviceyrker	6 Bønder,fiskere ol	7 Håndverkere	8 Operatør./sjåfør.	9 Andre yrker
Eldste person	1 Lederyrker	0.049	0.036	0.039	0.042	-0.016	-0.041	-0.056	-0.046	-0.042
	2 Akademiske yrker	0.034	0.202	0.073	-0.028	-0.096	-0.040	-0.047	-0.065	-0.067
	3 Høyskoleyrker	0.015	0.043	0.109	-0.011	-0.039	-0.054	-0.022	-0.053	-0.060
	4 Kontoryrker	0.025	-0.018	-0.020	0.004	-0.010	-0.027	0.030	0.024	0.009
	5 Salg/serviceyrker	0.013	-0.061	-0.060	-0.031	0.023	-0.017	0.094	0.061	0.020
	6 Bønder,fiskere ol	-0.036	-0.047	-0.032	-0.031	-0.034	0.369	-0.026	0.006	-0.010
	7 Håndverkere	-0.048	-0.082	-0.052	0.051	0.111	-0.047	-0.024	-0.016	0.048
	8 Operatør./sjåfør.	-0.048	-0.080	-0.070	0.020	0.065	-0.020	0.004	0.069	0.081
	9 Andre yrker	-0.029	-0.044	-0.055	-0.023	-0.009	0.001	0.072	0.078	0.074

I delutvalget er det 15 617 personer som er i familier med minst 2 sysselsatte. Feilmarginen er da omlag 0.016 for de markerte tallene. Vi kan derfor si at finnes en sammenheng for yrkene innen bønder og fiskere og for akademiske yrker, selv om det på dette nivå ikke er en spesielt sterk sammenheng. Vi finner heller ingen sterk sammenheng i de største yrkene på detaljert nivå, selv om det nok kan være sterkere sammenhenger i spesielle (små) yrker.

3.3 Aktuelle kjennemerker

Det er betydelige forskjeller i yrkesfordelingene utfra demografiske variabler som kjønn og alder, og økonomiske som lønn, inntekt og arbeidsmarkedsstatus. Dette er interessante egenskaper i mange tabeller og analyser på aggregert nivå, men har også andre anvendelser. En kan ikke bruke disse direkte for å *klassifisere* yrke i AKU, men de inngår i en sannsynlighetsmodell for estimering av yrke i registerbasert sysselsettingsstatistikk for grupper uten oppgitt yrke.

Tabell 3-4: Antall i utvalget etter kjønn, alder og yrkesfelt. Personutvalg AKU 2000-2004.

		I alt	1 Leder yrker	2 Akadem. yrker	3 Høyskole yrker	4 Kontor yrker	5 Salg/serv. yrker	6 Bønder, fiskere ol	7 Hånd-verkere	8 Operatør /sjåfør.	9 Andre yrker
I alt		61 346	4 026	6 229	13 367	4 822	14 455	2 326	6 791	4 868	4 462
	16-30	16 776	272	855	2 678	1 120	5 954	551	2 071	1 220	2 055
	31-40	14 872	1 074	1 858	3 932	1 147	2 798	387	1 746	1 206	724
	41-50	13 793	1 278	1 611	3 308	1 156	2 634	501	1 456	1 154	695
	51-74	15 905	1 402	1 905	3 449	1 399	3 069	887	1 518	1 288	988
Kvinner		29 233	1 119	2 679	7 108	3 299	10 467	607	533	807	2 614
	16-30	8 130	115	352	1 519	634	4 093	155	152	195	915
	31-40	7 081	319	868	2 185	782	2 045	90	153	192	447
	41-50	6 701	361	723	1 776	855	2 051	126	119	195	495
	51-74	7 321	324	736	1 628	1 028	2 278	236	109	225	757
Menn		32 113	2 907	3 550	6 259	1 523	3 988	1 719	6 258	4 061	1 848
	16-30	8 646	157	503	1 159	486	1 861	396	1 919	1 025	1 140
	31-40	7 791	755	990	1 747	365	753	297	1 593	1 014	277
	41-50	7 092	917	888	1 532	301	583	375	1 337	959	200
	51-74	8 584	1 078	1 169	1 821	371	791	651	1 409	1 063	231

Nivåtallene er et slags gjennomsnitt for en fem år, og gir et visst bilde av hovedtrekkene. Det er enda mer interessant å se på de strukturelle forskjellene altså ulikheter i prosentvis fordeling, som vist i neste tabell. Enheten for fordelingen er prosent og for avvikene i prosentpoeng.

Tabell 3-5: Forskjeller i yrkesfordeling etter alder og kjønn. Personutvalg AKU 2000-2004. Prosent og -poeng.

		I alt	1 Leder	2 Akademiske	3 Høyskole	4 Kontor	5 Salg/service	6 Bønder,fiskere ol	7 Håndverkere	8 Operatør./sjåfør.	9 Andre
I alt		100.0	6.6	10.2	21.8	7.9	23.6	3.8	11.1	7.9	7.3
	16-30	100.0	1.6	5.1	16.0	6.7	35.5	3.3	12.3	7.3	12.2
	31-40	100.0	7.2	12.5	26.4	7.7	18.8	2.6	11.7	8.1	4.9
	41-50	100.0	9.3	11.7	24.0	8.4	19.1	3.6	10.6	8.4	5.0
	51-74	100.0	8.8	12.0	21.7	8.8	19.3	5.6	9.5	8.1	6.2
Kvinner		100.0	3.8	9.2	24.3	11.3	35.8	2.1	1.8	2.8	8.9
	16-30	100.0	1.4	4.3	18.7	7.8	50.3	1.9	1.9	2.4	11.3
	31-40	100.0	4.5	12.3	30.9	11.0	28.9	1.3	2.2	2.7	6.3
	41-50	100.0	5.4	10.8	26.5	12.8	30.6	1.9	1.8	2.9	7.4
	51-74	100.0	4.4	10.1	22.2	14.0	31.1	3.2	1.5	3.1	10.3
Menn		100.0	9.1	11.1	19.5	4.7	12.4	5.4	19.5	12.6	5.8
	16-30	100.0	1.8	5.8	13.4	5.6	21.5	4.6	22.2	11.9	13.2
	31-40	100.0	9.7	12.7	22.4	4.7	9.7	3.8	20.4	13.0	3.6
	41-50	100.0	12.9	12.5	21.6	4.2	8.2	5.3	18.9	13.5	2.8
	51-74	100.0	12.6	13.6	21.2	4.3	9.2	7.6	16.4	12.4	2.7
		I alt	1 Leder	2 Akademiske	3 Høyskole	4 Kontor	5 Salg/service	6 Bønder,fiskere ol	7 Håndverkere	8 Operatør./sjåfør.	9 Andre
I alt		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	16-30	0.0	-4.9	-5.1	-5.8	-1.2	11.9	-0.5	1.3	-0.7	5.0
	31-40	0.0	0.7	2.3	4.6	-0.1	-4.7	-1.2	0.7	0.2	-2.4
	41-50	0.0	2.7	1.5	2.2	0.5	-4.5	-0.2	-0.5	0.4	-2.2
	51-74	0.0	2.3	1.8	-0.1	0.9	-4.3	1.8	-1.5	0.2	-1.1
Kvinner		0.0	-2.7	-1.0	2.5	3.4	12.2	-1.7	-9.2	-5.2	1.7
	16-30	0.0	-5.1	-5.8	-3.1	-0.1	26.8	-1.9	-9.2	-5.5	4.0
	31-40	0.0	-2.1	2.1	9.1	3.2	5.3	-2.5	-8.9	-5.2	-1.0
	41-50	0.0	-1.2	0.6	4.7	4.9	7.0	-1.9	-9.3	-5.0	0.1
	51-74	0.0	-2.1	-0.1	0.4	6.2	7.6	-0.6	-9.6	-4.9	3.1
Menn		0.0	2.5	0.9	-2.3	-3.1	-11.1	1.6	8.4	4.7	-1.5
	16-30	0.0	-4.7	-4.3	-8.4	-2.2	-2.0	0.8	11.1	3.9	5.9
	31-40	0.0	3.1	2.6	0.6	-3.2	-13.9	0.0	9.4	5.1	-3.7
	41-50	0.0	6.4	2.4	-0.2	-3.6	-15.3	1.5	7.8	5.6	-4.5
	51-74	0.0	6.0	3.5	-0.6	-3.5	-14.3	3.8	5.3	4.4	-4.6

Vi kan også bruke tallene til å studere oddsforholdene for de enkelte yrker, f.eks. sammenhengen mellom yrkesfelt 5, kjønn og alder:

	5 Salg/serv.	Øvrige yrker
16-30 år	5 954	10 822
31-74 år	8 501	36 069
Odds:	2.33	

	5 Salg/serv.	Øvrige yrker
Kvinner	10 467	18 766
Menn	3 988	28 125
Odds:	3.93	

	5 Salg/serv.	Øvrige yrker
Kvinner 16-30 år	4 093	4 037
Øvrige sysselsatte	10 362	42 854
Odds:	4.19	

Ser vi på 'menn' og ' håndverkere', er oddsforholdet over 13! Hvis vi regner på de publiserte tallene finner vi liknende resultater. På detaljert nivå (4 siffer) finner vi at i en god del yrker er andelen nær 100% av det ene kjønn. Over 20% jobber i yrker hvor det er mer enn 80% kvinner, nærmere 30% jobber i yrker med mindre enn 20% kvinner. Sagt på en annen måte er rundt halvparten av sysselsatte i yrker som er sterkt kjønnsdelte.

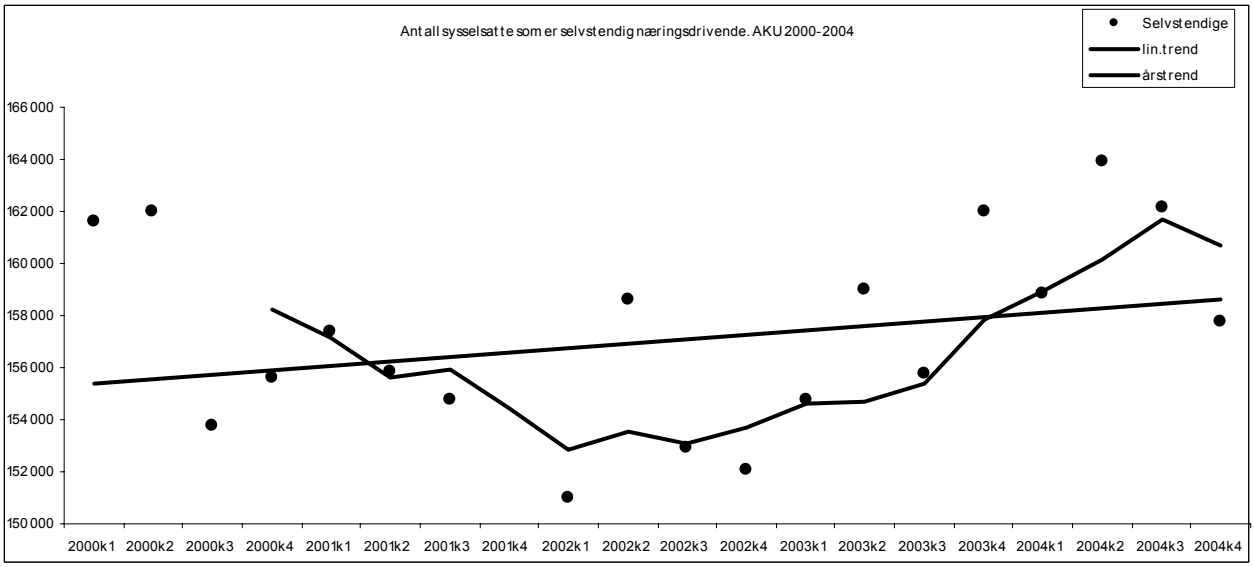
I f.eks. lederyrkene finner vi at også alder er en viktig faktor. Lønn og inntekt har nok også en stor sammenheng med yrke, men slik informasjon finnes ikke i dagens AKU. Det er gjennomført et pilotprosjekt for å estimere dette på bakgrunn av kobling med registerdata på mikronivå. Et alternativ til dette er imputering på et mer aggregert nivå. Ingen av metodene er implementert i løpende AKU. Her har vi ikke forsøkt å koble til lønns- eller inntektsdata.

3.4 Yrkesstatus

Med yrkesstatus skiller vi her mellom **selvstendig næringsdrivende** og øvrige sysselsatte. Den siste gruppen kalles her **ansatte** siden de utgjør flest i denne gruppa. Det er flere grunner til å studere yrkesfordelingen etter yrkesstatus nærmere, siden det er ikke publisert yrkesstatistikk etter yrkesstatus. Det estimeres yrkesnivåer for selvstendige i registerbasert sysselsettingsstatistikk, og de mangler yrkesdata i register. AKU-data viser at det er klare forskjeller i yrkesfordelingen til ansatte og selvstendige, noe som har gitt opphav til egne estimeringmetoder.

Vi viser først noen generelle trekk i omfanget av selvstendig gruppa, siden vi vil studere denne spesielt. Sett i sammenheng med stabiliteten i sysselsettingen totalt kan se ut som det er en tendens til at mengden selvstendige øker noe den siste tiden.

Figur 3-6: Antall sysselsatte selvstendig næringsdrivende pr. kvartal. AKU 2000-2004.



Det er vanskelig å si om dette skyldes generell økning i sysselsettingen eller spesielle forhold for denne gruppa. Diagrammet viser også punktestimat pr. kvartal, for å illustrere at svingningene er større enn det som ser ut som en

trend. Komponentene i svingningene kan være reelle sesongvariasjoner og tilfeldigheter pga. utvalgsusikkerhet. Legg merke til at skalaen er strukket ut, avstanden fra laveste til høyeste punkt er under 10%.

En oversikt over demografi og yrkesfordelingen etter yrkesstatus viser at selvstendig næringsdrivende er typisk menn, gamle og bønder. De er også noen flere håndverkere og mange færre innen høyskoleyrker. På grunn av svingningene og muligens usikkerhet, undersøker vi ikke endringer over tid i denne strukturen.

Tabell 3-7: Kjønn- alders- og yrkesfordeling etter yrkesstatus. Personutvalg AKU 2000-2004. Antall og prosent.

	I alt	Ansatte	Selvstendige	I alt	Ansatte	Selvstendige
I alt	61 346	57 134	4 212	100.0	100.0	100.0
Kvinner	29 233	28 012	1 221	47.7	49.0	29.0
Menn	32 113	29 122	2 991	52.3	51.0	71.0
16-30	16 776	16 407	369	27.3	28.7	8.8
31-40	14 872	13 875	997	24.2	24.3	23.7
41-50	13 793	12 641	1 152	22.5	22.1	27.4
51-74	15 905	14 211	1 694	25.9	24.9	40.2
1 Lederyrker	4 026	3 884	142	6.6	6.8	3.4
2 Akademiske yrker	6 229	5 750	479	10.2	10.1	11.4
3 Høyskoleyrker	13 367	12 828	539	21.8	22.5	12.8
4 Kontoryrker	4 822	4 780	42	7.9	8.4	1.0
5 Salg/serviceyrker	14 455	13 877	578	23.6	24.3	13.7
6 Bønder,fiskere ol	2 326	1 031	1 295	3.8	1.8	30.7
7 Håndverkere	6 791	6 074	717	11.1	10.6	17.0
8 Operatør./sjåfør.	4 868	4 521	347	7.9	7.9	8.2
9 Andre yrker	4 462	4 389	73	7.3	7.7	1.7

Fordi det også er store både kjønnsforskjeller i yrke og i yrkesstatus, kan det være interessant å se på sammenhengen mellom disse tre. For eksempel er det overvekt av bønder og håndverkere, og spørsmålet blir da hvor mye av dette som skyldes at det allerede er overvekt av menn som jo er overrepresentert i de yrkene. For å klarlegge dette sammenlikner vi oddsforholdet for yrkesfeltene, sammenliknet med oddsforholdene kontrollert for kjønn.

Tabell 3-8: Sammenheng mellom yrke, kjønn og yrkesstatus. Personutvalg AKU 2000-2004. Oddsforhold.

	I alt	Ansatte	Selvst.									
I alt	61 346	57 134	4 212		Ansatte		Selvst.		Odds kontrollert for			
Kvinner	29 233	28 012	1 221	Odds	Kvinner	Menn	Kvinner	Menn	Selvst.		Mann	
Menn	32 113	29 122	2 991	2.4	28 012	29 122	1 221	2 991	Kvinner	Menn	Ansatte	Selvst.
1 Lederyrker	4 026	3 884	142	0.5	1 083	2 801	36	106	0.6	0.3	2.6	1.2
2 Akademiske yrker	6 229	5 750	479	1.1	2 537	3 213	142	337	2.1	1.0	1.2	1.0
3 Høyskoleyrker	13 367	12 828	539	0.5	6 923	5 905	185	354	3.2	0.5	0.8	0.8
4 Kontoryrker	4 822	4 780	42	0.1	3 272	1 508	27	15	0.2	0.1	0.4	0.2
5 Salg/serviceyrker	14 455	13 877	578	0.5	10 066	3 811	401	177	0.0	0.4	0.3	0.1
6 Bønder,fiskere ol	2 326	1 031	1 295	24.2	297	734	310	985	9.7	19.0	2.4	1.4
7 Håndverkere	6 791	6 074	717	1.7	481	5 593	52	665	2.6	1.2	13.6	6.4
8 Operatør./sjåfør.	4 868	4 521	347	1.0	775	3 746	32	315	3.0	0.8	5.2	4.4
9 Andre yrker	4 462	4 389	73	0.2	2 578	1 811	36	37	0.3	0.2	0.7	0.4

Hvis man bare vet at en person er selvstendig er det 24 ganger større sjanse at vedkommende er bonde eller fisker, enn for ansatte. Merk at det er i utgangspunktet en lav sjanse for alle, under 4%. Når vi kontrollerer for kjønn, blir forholdet kraftig redusert. Dette sier altså noen om hvilke kjennemerker som har betydning for yrke. Dette vil få konsekvenser for estimering og imputering av yrke, ved f.eks. hvis vi har en gruppe som er for liten til å estimere yrke stratifisert etter yrkesstatus, kan vi i kompensere for dette ved å ta hensyn til kjønn og alder.

3.5 Utdanning

Med utdanning mener vi utdanningskode etter Norsk standard for utdanningsgruppering "NUS 2000" (NOS C 617). Dette er en 6-sifret kode på det mest detaljerte nivå, enkeltutdanning. 1.siffer angir utdanningsnivå, 2.siffer angir fagfelt. Fagfeltet er gjennomgående i kodehierarkiet i NUS, en struktur som ikke har noe motstykke i STYRK.

I AKU blir utdanningskode koblet på fra register over befolkningens høyeste utdanning. Personer som oppgir at de har deltatt i undervisning, får ny utdanningskode for dette (2 siffer).

Det er interessant å studere sammenhengen mellom utdanning og yrke, enten man skal bruke utdanning i estimering av yrke eller andre formål. Det publiseres vanligvis ikke krysstabeller med disse to variabler. Selv med samlede data fra flere kvartaler og selv med høyeste aggregeringsnivå på begge variabler, er de fleste grupper for små til å kunne si noe sikkert om nivået.

Tabell 3-9: Antall i utvalget etter utdanningsnivå og yrkesfelt. Personutvalg AKU 2000-2004.

	I alt	1 Leder yrker	2 Akademiske yrker	3 Høyskole yrker	4 Kontor yrker	5 Salg/service yrker	6 Bønder, fiskere ol	7 Hånd-verkere	8 Operatør. /sjåfør.	9 Andre yrker
I alt	61 364	4 026	6 229	13 367	4 822	14 455	2 326	6 791	4 868	4 480
0 Ingen/førskole	241	1	3	9	9	63	32	11	11	102
1 Barneskole	55	1	2	.	2	18	1	2	8	21
2 Ungdomsskole	8 408	240	74	493	587	2 599	633	1 046	1 245	1 491
3 Videreg., grunnk.	15 384	738	340	1 775	1 822	5 042	894	1 820	1 590	1 363
4 Videreg., avslut.	17 409	935	604	2 694	1 428	4 932	570	3 447	1 715	1 084
5 Påbygging videreg.	2 386	249	238	947	270	323	29	170	95	65
6 Høyskole ol lavere	13 630	1 403	2 450	6 907	645	1 368	130	272	183	272
7 Universitet høyere	3 572	421	2 296	530	58	109	37	21	20	80
8 Forskerutdanning	279	38	222	12	1	1	.	2	1	2

Men selv om vi ikke kan gi et brukbart estimat på *nivå*, er det meningsfylt å se på *sammenhengen.* Vi viser derfor henholdsvis andeler og korrelasjoner. Man skal i den sammenheng være klar over at korrelasjonskoeffisientene er laget ved at variablene er parameterisert. Når vi betrakter dette som nominelle variabler, lages altså ni binære indikatorvariabler for hver av utdanningsnivå og yrkesfelt. Hvis vi gjør om begge kjennemerker til nummeriske variabler finner vi en samlet korrelasjon på –0.504. En slik omgjøring er rimelig for utdanningsnivå, men ikke så rimelig for yrkesfelt. Utdanningsnivå er stort sett ordinal med hensyn til lengden på utdanning, selv om det ikke er en lineær sammenheng. Yrkesfelt kan vi si er delvis ordinal.

Tabell 3-10: Yrkesfordeling etter utdanningsnivå. Personutvalg AKU 2000-2004. Prosent.

	I alt	1 Leder yrker	2 Akademiske yrker	3 Høyskole yrker	4 Kontor yrker	5 Salg/service yrker	6 Bønder, fiskere ol	7 Hånd-verkere	8 Operatør. /sjåfør.	9 Andre yrker
I alt	100	7	10	22	8	24	4	11	8	7
0 Ingen/førskole	100	0	1	4	4	26	13	5	5	42
1 Barneskole	100	2	4	0	4	33	2	4	15	38
2 Ungdomsskole	100	3	1	6	7	31	8	12	15	18
3 Videreg., grunnk.	100	5	2	12	12	33	6	12	10	9
4 Videreg., avslut.	100	5	3	15	8	28	3	20	10	6
5 Påbygging videreg.	100	10	10	40	11	14	1	7	4	3
6 Høyskole ol lavere	100	10	18	51	5	10	1	2	1	2
7 Universitet høyere	100	12	64	15	2	3	1	1	1	2
8 Forskerutdanning	100	14	80	4	0	0	0	1	0	1

Tabell 3-11: Korrelasjon mellom utdanningsnivå og yrkesfelt. Personutvalg AKU 2000-2004. Pearson.

	1 Leder yrker	2 Akademiske yrker	3 Høyskole yrker	4 Kontor yrker	5 Salg/service yrker	6 Bønder, fiskere ol	7 Hånd-verkere	8 Operatør. /sjåfør.	9 Andre yrker
0 Ingen/førskole	-0.016	-0.019	-0.027	-0.010	0.004	0.031	-0.013	-0.008	0.085
1 Barneskole	-0.006	-0.006	-0.016	-0.005	0.006	-0.003	-0.007	0.007	0.036
2 Ungdomsskole	-0.060	-0.122	-0.154	-0.013	0.069	0.078	0.017	0.101	0.160
3 Videreg., grunnk.	-0.041	-0.152	-0.144	0.086	0.126	0.061	0.014	0.051	0.035
4 Videreg., avslut.	-0.030	-0.139	-0.096	0.008	0.071	-0.017	0.175	0.045	-0.026
5 Påbygging videreg.	0.031	-0.001	0.087	0.026	-0.047	-0.027	-0.025	-0.029	-0.035
6 Høyskole ol lavere	0.081	0.138	0.374	-0.062	-0.170	-0.079	-0.155	-0.130	-0.109
7 Universitet høyere	0.052	0.446	-0.042	-0.058	-0.120	-0.036	-0.083	-0.068	-0.048
8 Forskerutdanning	0.019	0.155	-0.029	-0.019	-0.037	-0.013	-0.022	-0.019	-0.017

Det er en tydelig sammenheng mellom de lengre utdanningene og yrker som etter definisjonen forutsetter høyere kompetanse. Det illustrerer også en av begrensningene ved å bruke korrelasjon på denne måten. F.eks. de aller fleste med forskerutdanning har akademiske yrker, allikevel er ikke koeffisienten spesielt høy pga. at de utgjør svært få i dette yrkesfeltet. Det vi måler her er symmetrisk assosiasjon, ikke kausalitet. Sagt på en annen måte tar ikke metoden stilling til hva som kom først av høna eller egget, selv om vi i modellene regner yrket som avhengig variabel (virkning).

Som eksempel krever visse yrker en formell utdanning og autorisasjon, som endel helsepersonell. Utdanning er da en viktig forklaringsvariabel. I andre tilfeller kan videreutdanning komme som en følge av at man har først begynt i et yrke. Økt interesse eller stigende kompetansekrav kan da gjøre at yrke blir årsaken og utdanning virkningen.

For å øke samsvaret mellom indikatorene, kan vi ta utgangspunkt i *kompetansenivå i yrkesstandarden.* En må uansett ha i mente at dette betegner *realkompetanse* ikke nødvendigvis skole/ formell utdanning, men vi kan vise en oversikt:

Yrkesfelt	Tilsvarende kompetanse	Yrkesklasse
1 Lederyrker	Ingen krav	C
2 Akademiske yrker	Minst 4 år universitet, e.l.	D
3 Høyskoleyrker	1-3 høyskole, e.l. (høyere utdanning enn vidr.sk.)	C
4 Kontoryrker	Videregående skole (1-3 år etter grunnskole)	B
5 Salg/serviceyrker		
6 Bønder,fiskere ol		
7 Håndverkere		
8 Operatør./sjåfør.		
9 Andre yrker	Ingen krav	A

Selv om felt 1 og 9 synes å ha samme kompetansekrav (nemlig ingen!), velger vi utfra ledernes reelle utdanningsnivå å gruppere lederyrkene sammen med høyskoleyrkene. Felt 9 er i hovedsak hjelpearbeidere og manuelle yrker. Vi velger også å slå sammen endel utdanningsnivåer som er små og ikke har særlig betydning for yrke.

	Off. utdanningsnivå	Utdanningsklasse
9	Uoppgitt	A
0	Ingen utdanning og førskoleutdanning	
1	Barneskoleutdanning	
2	Ungdomsskoleutdanning	
3	Videregående, grunnutdanning	B
4	Videregående, avsluttende utdanning	
5	Påbygging til videregående utdanning	
6	Universitets- og høgskoleutdanning, lavere nivå	C
7	Universitets- og høgskoleutdanning, høyere nivå	D
8	Forskerutdanning	

Vi får nå 16 grupper istedenfor 81, og viser sammenhengen på dette nivå. Det er tydelig at selv om utdanningsnivå vil være en viktig variabel i modellering av yrkesfelt, er det betydelig virkning av andre faktorer som f.eks. uformell kompetanse. Slik kompetanse kan ikke måles direkte men må avledes av f.eks. ansiennitet, alder og lønn.

Tabell 3-12: Fordeling i utvalget etter utdannings- og yrkesgruppering. Personutvalg AKU 2000-2004.

antall	I alt	A Andre yrker	B Yrkene 4-8	C Høyskoleyrke	D Akademiker
I alt	61 364	4 480	33 262	17 393	6 229
A Ikke vgs	8 704	1 614	6 267	744	79
B VGS nivå	35 179	2 512	24 147	7 338	1 182
C Høyskole	13 630	272	2 598	8 310	2 450
D Univ./dr	3 851	82	250	1 001	2 518

andel	I alt	A Andre yrker	B Yrkene 4-8	C Høyskoleyrke	D Akademiker
I alt	100 %	7 %	54 %	28 %	10 %
A Ikke vgs	100 %	19 %	72 %	9 %	1 %
B VGS nivå	100 %	7 %	69 %	21 %	3 %
C Høyskole	100 %	2 %	19 %	61 %	18 %
D Univ./dr	100 %	2 %	6 %	26 %	65 %

korrelasjon	A Andre yrker	B Yrkene 4-8	C Høyskoleyrke	D Akademiker
A Ikke vgs	0.176	0.145	-0.179	-0.124
B VGS nivå	-0.007	0.336	-0.193	-0.261
C Høyskole	-0.109	-0.377	0.387	0.138
D Univ./dr	-0.051	-0.248	-0.013	0.473

Dette illustrerer også at *graden av sammenheng varierer* med utdanningsnivå. Det er m.a.o. ikke bare interessant å måle sammenhengen mellom yrke og andre variabler, men også de enkelte kategorier. Vi skal gi et aktuelt eksempel på detaljert nivå med utvalgt helsepersonell. Yrkeskodene som inngår er: 2230 SPESIALSYKEPLEIERE OG JORDMØDRE, 3231 SYKEPLEIERE og 2221 LEGER. Helsefaglig utdanning tilsvarer fagfelt 6. Høyere helseutdanning defineres som nivå 7 og 8 (embetsstudium, doktorgrad og høyere) og den andre gruppa betegner nivå 6

(høyskole eller kortere universitetsutd.). Kodene blir da 66, 76 og 86. Dette er en ganske grov inndeling, fordi vi bruker utdanningskoden i AKU, som er fra 2-5 siffer.

Tabellen viser antall i utvalget, og forventet antall gitt tilfeldig fordeling. Det siste er altså ikke fordi noen forventer en tilfeldig fordeling her, men fordi forskjellen fra det reelle tallet illustrerer sammenhengen mellom kategoriene. En observerer en tydelig sammenheng, men det er også klart at for å predikere disse yrker må vi ha ytterligere data, spesielt å koble til detaljert utdanningskode fra register, men også andre variabler.

Tabell 3-13: Utvalgte helseutdanninger og -yrker. AKU-data 2000-2004. Antall og prosent.

	ANNET YRKE	LEGE	SPES.SYKEPL .	SYKEPL.	I ALT
ANNEN UTD.	57 329	40	124	398	57891
forventet	*55 821*	*367*	*451*	*1 252*	
Rad%	99.0	0.1	0.2	0.7	94.3
Kol.%	96.9	10.3	25.9	30.0	
HELSEUTD. HØYERE	692	343	3	7	1045
forventet	*1 008*	*7*	*8*	*23*	
Rad%	66.2	32.8	0.3	0.7	1.7
Kol.%	1.2	88.2	0.6	0.5	
HELSEUTD.	1 149	6	351	922	2428
forventet	*2 341*	*15*	*19*	*53*	
Rad%	47.3	0.3	14.5	38.0	4.0
Kol.%	1.9	1.5	73.4	69.5	
I ALT	59 170	389	478	1 327	61 364
	96.4	0.6	0.8	2.2	100.0

Foruten å beregne korrelasjonene som i forrige eksempel kan vi også vurdere et samlet mål istedenfor å få 12 ulike tall. Tre andre mål forteller om den samlede sammenhengen uavhengig av antallet, og man kan betrakte dem som Chi-kvadrat baserte men justert for utvalgsstørrelsen. Phi- og C-koeffisientene er imidlertid avhengig av *tabellstørrelsen* og det på litt ulik måte. Felles er at jo høyere tall, jo sterkere sammenheng.

Tabell 3-14: Utvalgte helseutdanninger og -yrker. AKU-data 2000-2004. Assosiasjonsmål.

Phi Coefficient	0.7991
Contingency Coefficient	0.6243
Cramer's V	0.5651

Siden vi skal sammenlikne nominelle kategorivariabler med ulike antall kategorier velger vi Cramer's V for å måle styrken på assosiasjonen. For sammenhengen mellom utdanningsnivå og yrkesfelt generelt er Cramer's V = 0.2668. Hver av disse har 9 kategorier, som gir 81 mulige celler. Hvis vi bruker 4 utdanningsgrupper og 4 yrkesgrupper, øker det til 0.4085. I enkelte yrker finner vi høyere samsvar, som f.eks. noen helseyrker.

3.6 Næring

Bedriftens næringskode skal avspeile aktiviteten på arbeidsplassen og er derfor høyst aktuell som variabel i modellering av yrke. Det er generelt en interessant å studere yrkesfordeling i næringer, for bl.a. å kunne si noe om kompetanse, lønn, mobilitet og omstillingsevne. Det vil være store forskjeller mellom yrker med hensyn til spesialisering og generell (næringsuavhengig) kompetanse. Forholdet mellom tilbud og etterspørsel av kompetanse vil være en viktig når man skal studere utviklingen på arbeidsmarkedet.

Når vi tar utgangspunkt i AKU-data, finner vi raskt at det er mange sider ved yrke/næringsrelasjonene som ikke kan studeres. Selv i dette store utvalget blir det fort for små grupper til å lage publiserbar statistikk. For å analysere dette på et detaljert nivå er vi nok avhengig av å få yrke i registrene komplett. En registerbasert analyse har på sin side nytte av å sammenlikne de aggregerte tallene med AKU, f.eks. for å avdekke systematiske forskjeller i kodingen.

Her vises noen tabeller for å illustrere ulikheter i yrkesstruktur etter næring. I denne sammenheng er det greit å ha en ide om feilmarginer pga. utvalgsusikkerhet. Vi tar utgangspunkt i analyse av yrke *innen* en liten gruppe. I en næring med 1000 personer vil feilmarginen være opptil 3.2%-poeng (ved α=.05 og p=.5). Ved en gruppe på 100 personer, så blir den tilsvarende feilmarginen 10%-poeng. Altså hvis et yrke utgjør 50% av gruppa i utvalget, kan man bare anslå at det utgjør mellom 40 og 60 prosent i den tilsvarende gruppa i populasjonen. Generelt er den absolutte presisjonen større, jo mindre andelen er. Typisk for hvert enkelt yrke er svært *små* andeler. Den relative feilen er da ofte stor.

I forhold til næring ser vi at yrkesfelt 1 (adm. ledere) og 9 (hjelpearb., manuelt arbeid, m.v.) er generelle, mens feltene 5,6,7 er tydelig mer spesifikke. Det største håndverksyrket, "7125 Tømrere", utgjør ikke overraskende en stor del av næringen "45 Bygg og anlegg" med over 20% av de sysselsatte. På den annen side finner vi 7% tømrere i helt andre næringer. Mens "5221 Butikkmedarbeidere" er næringsspesifikt er "3415 Salgskonsulenter" (høyskolekompetanse) et mer generelt yrke. For å kunne studere slike forhold mer i detalj, må vi utnytte mulighetene i registerdata.

Tabell 3-15: Yrkesfordeling etter næring. Personutvalg AKU 2000-2004. Antall, prosent og -poeng.

Næring	I alt	1 Leder	2 Akadem.	3 Høyskole	4 Kontor	5 Salg/serv.	6 Bønder. ol	7 Håndv.	8 Opr../sjåfør.	9 Andre
I alt	61 313	4 026	6 229	13 367	4 822	14 455	2 326	6 791	4 868	4 429
01-05 Jord-/skogbruk,fiske	2 632	37	6	81	28	32	2 269	98	29	52
10-14 Bergverk,olje/gass	891	93	143	211	65	16	1	114	220	28
15-16 Nærings/nyt.industri	1 469	108	26	128	106	97	6	257	618	123
17-19 Tekst./klær/skoind.	177	13	4	12	12	4	0	44	82	6
20-22 Tre-/graf. industri	1 554	119	72	298	111	59	0	322	422	151
23-37 Annen industri	4 311	360	237	667	287	48	1	1 463	1 093	155
40-41 Kraft/vannforsyning	447	35	28	105	56	13	3	135	61	11
45 Bygg/anleggsvirks.	4 153	216	71	237	177	61	4	2 840	408	139
50-52 Handel/rep.av varer	9 232	909	133	1 274	860	4 799	10	846	168	233
55 Hotell/restaurant	2 119	159	7	41	161	1 186	0	7	21	537
60-64 Transport/komm.	4 190	210	147	771	1 008	219	1	156	1 402	276
65-67 Bank,finans,utleie	1 215	185	85	807	99	12	0	0	3	24
70-74 Forr. tjenster m.m.	5 691	441	1 450	1 733	644	523	2	233	76	589
75 Offentlig forvaltn.	3 828	416	942	967	379	259	1	88	29	747
80 Undervisning	4 956	181	1 280	2 428	177	495	6	20	7	362
85 Helse/sosialtj.	11 711	349	1 206	2 999	381	5 894	10	111	109	652
90-99 Div. tjenester, m.m	2 737	195	392	608	271	738	12	57	120	344

Næring	I alt	1 Leder	2 Akadem.	3 Høyskole	4 Kontor	5 Salg/serv.	6 Bønder. ol	7 Håndv.	8 Opr../sjåfør.	9 Andre
I alt	100 %	7 %	10 %	22 %	8 %	24 %	4 %	11 %	8 %	7 %
01-05 Jord-/skogbruk,fiske	100 %	1 %	0 %	3 %	1 %	1 %	86 %	4 %	1 %	2 %
10-14 Bergverk,olje/gass	100 %	10 %	16 %	24 %	7 %	2 %	0 %	13 %	25 %	3 %
15-16 Nærings/nyt.industri	100 %	7 %	2 %	9 %	7 %	7 %	0 %	17 %	42 %	8 %
17-19 Tekst./klær/skoind.	100 %	7 %	2 %	7 %	7 %	2 %	0 %	25 %	46 %	3 %
20-22 Tre-/graf. industri	100 %	8 %	5 %	19 %	7 %	4 %	0 %	21 %	27 %	10 %
23-37 Annen industri	100 %	8 %	5 %	15 %	7 %	1 %	0 %	34 %	25 %	4 %
40-41 Kraft/vannforsyning	100 %	8 %	6 %	23 %	13 %	3 %	1 %	30 %	14 %	2 %
45 Bygg/anleggsvirks.	100 %	5 %	2 %	6 %	4 %	1 %	0 %	68 %	10 %	3 %
50-52 Handel/rep.av varer	100 %	10 %	1 %	14 %	9 %	52 %	0 %	9 %	2 %	3 %
55 Hotell/restaurant	100 %	8 %	0 %	2 %	8 %	56 %	0 %	0 %	1 %	25 %
60-64 Transport/komm.	100 %	5 %	4 %	18 %	24 %	5 %	0 %	4 %	33 %	7 %
65-67 Bank,finans,utleie	100 %	15 %	7 %	66 %	8 %	1 %	0 %	0 %	0 %	2 %
70-74 Forr. tjenster m.m.	100 %	8 %	25 %	30 %	11 %	9 %	0 %	4 %	1 %	10 %
75 Offentlig forvaltn.	100 %	11 %	25 %	25 %	10 %	7 %	0 %	2 %	1 %	20 %
80 Undervisning	100 %	4 %	26 %	49 %	4 %	10 %	0 %	0 %	0 %	7 %
85 Helse/sosialtj.	100 %	3 %	10 %	26 %	3 %	50 %	0 %	1 %	1 %	6 %
90-99 Div. tjenester, m.m	100 %	7 %	14 %	22 %	10 %	27 %	0 %	2 %	4 %	13 %

Næring	I alt	1 Leder	2 Akadem.	3 Høyskole	4 Kontor	5 Salg/serv.	6 Bønder. ol	7 Håndv.	8 Opr../sjåfør.	9 Andre
I alt	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
01-05 Jord-/skogbruk,fiske	0 %	-5 %	-10 %	-19 %	-7 %	-22 %	82 %	-7 %	-7 %	-5 %
10-14 Bergverk,olje/gass	0 %	4 %	6 %	2 %	-1 %	-22 %	-4 %	2 %	17 %	-4 %
15-16 Nærings/nyt.industri	0 %	1 %	-8 %	-13 %	-1 %	-17 %	-3 %	6 %	34 %	1 %
17-19 Tekst./klær/skoind.	0 %	1 %	-8 %	-15 %	-1 %	-21 %	-4 %	14 %	38 %	-4 %
20-22 Tre-/graf. industri	0 %	1 %	-6 %	-3 %	-1 %	-20 %	-4 %	10 %	19 %	2 %
23-37 Annen industri	0 %	2 %	-5 %	-6 %	-1 %	-22 %	-4 %	23 %	17 %	-4 %
40-41 Kraft/vannforsyning	0 %	1 %	-4 %	2 %	5 %	-21 %	-3 %	19 %	6 %	-5 %
45 Bygg/anleggsvirks.	0 %	-1 %	-8 %	-16 %	-4 %	-22 %	-4 %	57 %	2 %	-4 %
50-52 Handel/rep.av varer	0 %	3 %	-9 %	-8 %	1 %	28 %	-4 %	-2 %	-6 %	-5 %
55 Hotell/restaurant	0 %	1 %	-10 %	-20 %	0 %	32 %	-4 %	-11 %	-7 %	18 %
60-64 Transport/komm.	0 %	-2 %	-7 %	-3 %	16 %	-18 %	-4 %	-7 %	26 %	-1 %
65-67 Bank,finans,utleie	0 %	9 %	-3 %	45 %	0 %	-23 %	-4 %	-11 %	-8 %	-5 %
70-74 Forr. tjenster m.m.	0 %	1 %	15 %	9 %	3 %	-14 %	-4 %	-7 %	-7 %	3 %
75 Offentlig forvaltn.	0 %	4 %	14 %	3 %	2 %	-17 %	-4 %	-9 %	-7 %	12 %
80 Undervisning	0 %	-3 %	16 %	27 %	-4 %	-14 %	-4 %	-11 %	-8 %	0 %
85 Helse/sosialtj.	0 %	-4 %	0 %	4 %	-5 %	27 %	-4 %	-10 %	-7 %	-2 %
90-99 Div. tjenester, m.m	0 %	1 %	4 %	0 %	2 %	3 %	-3 %	-9 %	-4 %	5 %

Tabell 3-16: Yrkesfordeling i utvalgte næringer. Personutvalg AKU 2000-2004. Antall, prosent og -poeng.

Næringsundergruppe	I alt	1 Leder	2 Akadem.	3 Høyskole	4 Kontor	5 Salg/serv.	6 Bønder. ol	7 Håndv.	8 Opr../sjåfør.	9 Andre
I alt	61 313	4 026	6 229	13 367	4 822	14 455	2 326	6 791	4 868	4 429
45211 OPPFØRING AV BYGNINGER	1 277	48	25	74	38	4	0	1 012	29	47
45310 ELEKTRISK INSTALLASJONSARBEID	705	40	13	72	40	14	1	513	1	11
50200 VEDLIKEHOLD OG REP. AV MOTORV.	552	22	3	16	41	27	0	399	18	26
52110 BUTIKKHANDEL MED BREDT VAREUTV.	1 664	133	2	16	40	1 420	0	9	7	37
52420 BUTIKKHANDEL MED KLÆR	611	50	1	27	13	510	0	8	1	1
55301 DRIFT AV RESTAURANTER OG KAFEER	955	79	1	6	7	640	0	4	9	209
60240 GODSTRANSPORT PÅ VEI	687	31	0	24	62	3	0	11	527	29
64110 POSTTJENESTER	597	16	16	12	506	1	0	0	36	10
65120 BANKVIRKSOMHET ELLERS	660	103	35	435	58	9	0	0	3	17
75110 GENERELL OFF. ADM.	792	135	259	221	133	13	0	1	4	26
75220 FORSVAR	601	44	92	87	46	28	0	61	16	227
80102 GRUNNSKOLEUNDERVISNING	2 624	88	57	1 868	60	373	0	1	1	176
80220 UNDERVISNING I YRKESRETTEDE FAG	637	33	267	203	24	50	5	3	3	49
85111 ALMINNELIGE SOMATISKE SYKEHUS	2 177	53	498	832	150	465	4	27	9	139
85118 SOMATISKE SYKEHJEM	759	16	30	120	7	526	0	0	8	52
85141 HJEMMESYKEPLEIE	1 196	17	22	207	17	821	0	0	17	95
85321 HJEMMEHJELP	1 081	65	7	257	3	711	0	0	0	38
85327 BARNEHAGER	715	47	7	204	3	429	0	0	0	25

Næringsundergruppe	I alt	1 Leder	2 Akadem.	3 Høyskole	4 Kontor	5 Salg/serv.	6 Bønder. ol	7 Håndv.	8 Opr../sjåfør.	9 Andre
I alt	100	7	10	22	8	24	4	11	8	7
45211 OPPFØRING AV BYGNINGER	100	4	2	6	3	0	0	79	2	4
45310 ELEKTRISK INSTALLASJONSARBEID	100	6	2	10	6	2	0	73	0	2
50200 VEDLIKEHOLD OG REP. AV MOTORV.	100	4	1	3	7	5	0	72	3	5
52110 BUTIKKHANDEL MED BREDT VAREUTV.	100	8	0	1	2	85	0	1	0	2
52420 BUTIKKHANDEL MED KLÆR	100	8	0	4	2	83	0	1	0	0
55301 DRIFT AV RESTAURANTER OG KAFEER	100	8	0	1	1	67	0	0	1	22
60240 GODSTRANSPORT PÅ VEI	100	5	0	3	9	0	0	2	77	4
64110 POSTTJENESTER	100	3	3	2	85	0	0	0	6	2
65120 BANKVIRKSOMHET ELLERS	100	16	5	66	9	1	0	0	0	3
75110 GENERELL OFF. ADM.	100	17	33	28	17	2	0	0	1	3
75220 FORSVAR	100	7	15	14	8	5	0	10	3	38
80102 GRUNNSKOLEUNDERVISNING	100	3	2	71	2	14	0	0	0	7
80220 UNDERVISNING I YRKESRETTEDE FAG	100	5	42	32	4	8	1	0	0	8
85111 ALMINNELIGE SOMATISKE SYKEHUS	100	2	23	38	7	21	0	1	0	6
85118 SOMATISKE SYKEHJEM	100	2	4	16	1	69	0	0	1	7
85141 HJEMMESYKEPLEIE	100	1	2	17	1	69	0	0	1	8
85321 HJEMMEHJELP	100	6	1	24	0	66	0	0	0	4
85327 BARNEHAGER	100	7	1	29	0	60	0	0	0	3

Næringsundergruppe	I alt	1 Leder	2 Akadem.	3 Høyskole	4 Kontor	5 Salg/serv.	6 Bønder. ol	7 Håndv.	8 Opr../sjåfør.	9 Andre
I alt	0	0	0	0	0	0	0	0	0	0
45211 OPPFØRING AV BYGNINGER	0	-3	-8	-16	-5	-23	-4	68	-6	-4
45310 ELEKTRISK INSTALLASJONSARBEID	0	-1	-8	-12	-2	-22	-4	62	-8	-6
50200 VEDLIKEHOLD OG REP. AV MOTORV.	0	-3	-10	-19	0	-19	-4	61	-5	-3
52110 BUTIKKHANDEL MED BREDT VAREUTV.	0	1	-10	-21	-5	62	-4	-11	-8	-5
52420 BUTIKKHANDEL MED KLÆR	0	2	-10	-17	-6	60	-4	-10	-8	-7
55301 DRIFT AV RESTAURANTER OG KAFEER	0	2	-10	-21	-7	43	-4	-11	-7	15
60240 GODSTRANSPORT PÅ VEI	0	-2	-10	-18	1	-23	-4	-9	69	-3
64110 POSTTJENESTER	0	-4	-7	-20	77	-23	-4	-11	-2	-6
65120 BANKVIRKSOMHET ELLERS	0	9	-5	44	1	-22	-4	-11	-7	-5
75110 GENERELL OFF. ADM.	0	10	23	6	9	-22	-4	-11	-7	-4
75220 FORSVAR	0	1	5	-7	0	-19	-4	-1	-5	31
80102 GRUNNSKOLEUNDERVISNING	0	-3	-8	49	-6	-9	-4	-11	-8	-1
80220 UNDERVISNING I YRKESRETTEDE FAG	0	-1	32	10	-4	-16	-3	-11	-7	0
85111 ALMINNELIGE SOMATISKE SYKEHUS	0	-4	13	16	-1	-2	-4	-10	-8	-1
85118 SOMATISKE SYKEHJEM	0	-4	-6	-6	-7	46	-4	-11	-7	0
85141 HJEMMESYKEPLEIE	0	-5	-8	-4	-6	45	-4	-11	-7	1
85321 HJEMMEHJELP	0	-1	-10	2	-8	42	-4	-11	-8	-4
85327 BARNEHAGER	0	0	-9	7	-7	36	-4	-11	-8	-4

Tabell 3-17: Fordeling av utvalgte yrker etter næring. Personutvalg AKU 2000-2004. Antall, prosent og -poeng.

	I alt	3231 SYKEPL.	3310 GR.SK.-LÆRER	3415 SALGS-KONS.	4114 KONTOR-MED.	5131 BARN/UNG.-ARB.	5132 HJELPE-PL.	5139 PLEIE-PERS.	5221 BUTIKK-MED.	7125 TØMRER	9132 RENGJ.-PERS.	9999 ANDRE
I alt	61 313	1 327	1 864	1 142	1 088	1 755	1 957	1 519	4 828	1 035	1 870	42 928
01-05 Jord-/skogbruk,fiske	2 632	0	0	1	15	0	0	0	15	6	8	2 587
10-14 Bergverk,olje/gass	891	6	0	8	10	0	0	0	0	0	16	851
15-16 Nærings/nyt.industri	1 469	1	0	59	35	0	0	0	70	0	34	1 270
17-19 Tekst./klær/skoind.	177	0	0	4	3	0	0	0	4	0	4	162
20-22 Tre-/graf. industri	1 554	1	0	49	31	0	0	1	7	19	19	1 427
23-37 Annen industri	4 311	1	0	105	69	0	0	0	21	5	34	4 076
40-41 Kraft/vannforsyning	447	0	0	11	14	0	0	0	0	0	5	417
45 Bygg/anleggsvirks.	4 153	0	0	13	71	0	0	0	29	960	52	3 028
50-52 Handel/rep.av varer	9 232	1	0	711	207	1	0	0	4 472	11	86	3 743
55 Hotell/restaurant	2 119	0	0	2	8	2	0	0	89	0	191	1 827
60-64 Transport/komm.	4 190	2	0	57	92	0	0	1	17	3	90	3 928
65-67 Bank,finans,utleie	1 215	0	0	6	23	0	0	0	0	0	15	1 171
70-74 Forr. tjenster m.m.	5 691	9	0	102	123	8	3	1	70	16	431	4 928
75 Offentlig forvaltn.	3 828	5	3	1	106	3	3	0	0	10	172	3 525
80 Undervisning	4 956	1	1 853	1	74	361	2	3	1	0	282	2 378
85 Helse/sosialtj.	11 711	1 300	7	1	147	1 369	1 945	1 511	9	4	335	5 083
90-99 Div. tjenester, m.m	2 737	0	1	11	60	11	4	2	24	1	96	2 527

	I alt	3231 SYKEPL.	3310 GR.SK.-LÆRER	3415 SALGS-KONS.	4114 KONTOR-MED.	5131 BARN/UNG.-ARB.	5132 HJELPE-PL.	5139 PLEIE-PERS.	5221 BUTIKK-MED.	7125 TØMRER	9132 RENGJ.-PERS.	9999 ANDRE
I alt	100 %	2 %	3 %	2 %	2 %	3 %	3 %	2 %	8 %	2 %	3 %	70 %
01-05 Jord-/skogbruk,fiske	100 %	0 %	0 %	0 %	1 %	0 %	0 %	0 %	1 %	0 %	0 %	98 %
10-14 Bergverk,olje/gass	100 %	1 %	0 %	1 %	1 %	0 %	0 %	0 %	0 %	0 %	2 %	96 %
15-16 Nærings/nyt.industri	100 %	0 %	0 %	4 %	2 %	0 %	0 %	0 %	5 %	0 %	2 %	86 %
17-19 Tekst./klær/skoind.	100 %	0 %	0 %	2 %	2 %	0 %	0 %	0 %	2 %	0 %	2 %	92 %
20-22 Tre-/graf. industri	100 %	0 %	0 %	3 %	2 %	0 %	0 %	0 %	0 %	1 %	1 %	92 %
23-37 Annen industri	100 %	0 %	0 %	2 %	2 %	0 %	0 %	0 %	0 %	0 %	1 %	95 %
40-41 Kraft/vannforsyning	100 %	0 %	0 %	2 %	3 %	0 %	0 %	0 %	0 %	0 %	1 %	93 %
45 Bygg/anleggsvirks.	100 %	0 %	0 %	0 %	2 %	0 %	0 %	0 %	1 %	23 %	1 %	73 %
50-52 Handel/rep.av varer	100 %	0 %	0 %	8 %	2 %	0 %	0 %	0 %	48 %	0 %	1 %	41 %
55 Hotell/restaurant	100 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	4 %	0 %	9 %	86 %
60-64 Transport/komm.	100 %	0 %	0 %	1 %	2 %	0 %	0 %	0 %	0 %	0 %	2 %	94 %
65-67 Bank,finans,utleie	100 %	0 %	0 %	0 %	2 %	0 %	0 %	0 %	0 %	0 %	1 %	96 %
70-74 Forr. tjenster m.m.	100 %	0 %	0 %	2 %	2 %	0 %	0 %	0 %	1 %	0 %	8 %	87 %
75 Offentlig forvaltn.	100 %	0 %	0 %	0 %	3 %	0 %	0 %	0 %	0 %	0 %	4 %	92 %
80 Undervisning	100 %	0 %	37 %	0 %	1 %	7 %	0 %	0 %	0 %	0 %	6 %	48 %
85 Helse/sosialtj.	100 %	11 %	0 %	0 %	1 %	12 %	17 %	13 %	0 %	0 %	3 %	43 %
90-99 Div. tjenester, m.m	100 %	0 %	0 %	0 %	2 %	0 %	0 %	0 %	1 %	0 %	4 %	92 %

	I alt	3231 SYKEPL.	3310 GR.SK.-LÆRER	3415 SALGS-KONS.	4114 KONTOR-MED.	5131 BARN/UNG.-ARB.	5132 HJELPE-PL.	5139 PLEIE-PERS.	5221 BUTIKK-MED.	7125 TØMRER	9132 RENGJ.-PERS.	9999 ANDRE
I alt	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
01-05 Jord-/skogbruk,fiske	0 %	-2 %	-3 %	-2 %	-1 %	-3 %	-3 %	-2 %	-7 %	-1 %	-3 %	28 %
10-14 Bergverk,olje/gass	0 %	-1 %	-3 %	-1 %	-1 %	-3 %	-3 %	-2 %	-8 %	-2 %	-1 %	25 %
15-16 Nærings/nyt.industri	0 %	-2 %	-3 %	2 %	1 %	-3 %	-3 %	-2 %	-3 %	-2 %	-1 %	16 %
17-19 Tekst./klær/skoind.	0 %	-2 %	-3 %	0 %	0 %	-3 %	-3 %	-2 %	-6 %	-2 %	-1 %	22 %
20-22 Tre-/graf. industri	0 %	-2 %	-3 %	1 %	0 %	-3 %	-3 %	-2 %	-7 %	0 %	-2 %	22 %
23-37 Annen industri	0 %	-2 %	-3 %	1 %	0 %	-3 %	-3 %	-2 %	-7 %	-2 %	-2 %	25 %
40-41 Kraft/vannforsyning	0 %	-2 %	-3 %	1 %	1 %	-3 %	-3 %	-2 %	-8 %	-2 %	-2 %	23 %
45 Bygg/anleggsvirks.	0 %	-2 %	-3 %	-2 %	0 %	-3 %	-3 %	-2 %	-7 %	21 %	-2 %	3 %
50-52 Handel/rep.av varer	0 %	-2 %	-3 %	6 %	0 %	-3 %	-3 %	-2 %	41 %	-2 %	-2 %	-29 %
55 Hotell/restaurant	0 %	-2 %	-3 %	-2 %	-1 %	-3 %	-3 %	-2 %	-4 %	-2 %	6 %	16 %
60-64 Transport/komm.	0 %	-2 %	-3 %	-1 %	0 %	-3 %	-3 %	-2 %	-7 %	-2 %	-1 %	24 %
65-67 Bank,finans,utleie	0 %	-2 %	-3 %	-1 %	0 %	-3 %	-3 %	-2 %	-8 %	-2 %	-2 %	26 %
70-74 Forr. tjenster m.m.	0 %	-2 %	-3 %	0 %	0 %	-3 %	-3 %	-2 %	-7 %	-1 %	5 %	17 %
75 Offentlig forvaltn.	0 %	-2 %	-3 %	-2 %	1 %	-3 %	-3 %	-2 %	-8 %	-1 %	1 %	22 %
80 Undervisning	0 %	-2 %	34 %	-2 %	0 %	4 %	-3 %	-2 %	-8 %	-2 %	3 %	-22 %
85 Helse/sosialtj.	0 %	9 %	-3 %	-2 %	-1 %	9 %	13 %	10 %	-8 %	-2 %	0 %	-27 %
90-99 Div. tjenester, m.m	0 %	-2 %	-3 %	-1 %	0 %	-2 %	-3 %	-2 %	-7 %	-2 %	0 %	22 %

For en samlet vurdering av næring tar vi igjen bruk assosiasjonsmålet Cramer's V. Den totale sammenhengen mellom næringshovedområde og yrkesfelt gir et 0.462, som altså er en betydelig sammenheng. For aggregeringen som er benyttet i tabellene over, målet omtrent likt (0.467). Hvis vi lager en tilsvarende gruppering av 10 største næringer som for de utvalgte yrkene i de siste tabellene, får vi igjen omtrent samme mål (0.468). Holder vi restgrupper utenfor får vi en delpopulasjon på 9 288 hvor vi vurderer detaljert næring mot detaljert yrke. Her stiger Cramer's V til 0.669. Vi kan med en viss rett si at næring har større betydning enn utdanning, når vi betrakter variablene som helhet.

Det er et kjent problem fra yrke i register at man må benytte næringskode på det meste detaljerte nivå, og at de hyppige endringene får store konsekvenser for kodeopplegget. Vi har nå vist en enkel måte å evaluere ulike kategorielle variabler i forhold til yrkesklassifisering. Det er i denne sammenhengen viktig å ta hensyn til det reelle innholdet i kategoriene og nøye vurdere de aggregeringer som benyttes, samt å velge metoder som passer til dette.

4 Bruk av tekstdata til klassifisering

4.1 Innledning

Dette kapittelet tar for seg bruk av tekst i forhold til klassifisering av yrke, men belyser noen sider ved tekst generelt. En beskrivelse og undersøkelse av tekst i yrkesrelaterte kjennemerker har ulike formål:

- Utviklingen av automatisk tekstbasert yrkesklassifisering i Arbeidstakerregisteret (AA).
- Dokumentere kvaliteten av den manuelle yrkeskodingen i AKU.
- Evaluering og kvalitetsmålinger på manuell og automatisk klassifisering. Dette både eksisterende klassifiseringsrutiner og eventuelt utvikling av nye metoder.
- Utvikling av indekser og kataloger til bruk ved manuell eller automatisk klassifisering.
- Demonstrere karakteristiske egenskaper ved tekst til forskjell fra andre datatyper. Dette for å kanskje skape interesse for å øke kompetansen på dette felt.

Tekstdata brukes til klassifisering av yrke og andre variabler i AKU og i mange andre undersøkelser. Det leveres også i noen grad ved rapportering av yrke til AA. Yrkestittel og arbeidsoppgaver i AKU er tekstvariabler, og det er mulig å oppgi yrke som tekst i meldinger på papirform til Arbeidstakerregisteret. Både i forbindelse med manuell koding i AKU og utvikling av automatiske kodemetoder i Arbeidstakerregisteret, er det interessant å analysere bruk av tekst nærmere. Automatisk klassifisering har hatt stor betydning for yrke i Arbeidstakerregisteret, og kunne være interessant for andre seksjoner og eksterne etater som driver med liknende registrering og klassifisering. Automatiske klassifiseringsmetoder er neppe aktuelt for AKU i overskuelig framtid, men også den manuelle kodingen kan dra nytte av slike undersøkelser.

Det kan også være at det kan være noen generelle resultater som kan være interessante for andre prosjekter, siden tekst ligger til grunn for mange andre klassifiseringer. Generelt kan vi vel si at tekstvariabler foreløpig har hatt liten plass i kurs og håndbøker i statistikk.

Mye av litteraturen om tekstdata dreier seg om å analysere store tekstmasser som bøker, aviser eller internettsider ved å trekke ut nøkkelord for å klassifisere og utføre søk. En slik analyse gir et ukjent antall kategorier. Klassifisering til statistikk, f.eks. av yrke innebærer et begrenset antall standardiserte kategorier som funksjon av relativt korte tekster. Det skiller seg altså fra de generelle metodene både ved at klassene er fastlagt på forhånd, og at tekstmassene er mye mindre. Dette kapittelet nevner noen systematiske tilnærminger og resultater fra AKU-data.

4.2 Egenskaper ved tekst

Tekstvariabler skiller seg fra tradisjonelle nummeriske og kategorielle variabler, altså "vanlige variabler" i statistikksammenheng. Tekst kan omtales som en kategorivariabel, men man kan også med en viss rett beskrives som kontinuerlig – ikke nummerisk men fordi det nærmest er uendelig antall mulige kombinasjoner. Tekst som kategori er nominell fordi det ikke er gitt en betydningsmessig rekkefølge, slik som med f.eks. månedsnavn (ordinal).

Det er vanlig å dele opp en tekst i **setninger** og **ord**. Det er imidlertid ikke enkelt å definere disse to begreper. Vi observerer spesielle egenskaper ved tekstdata som f.eks. at mange ord brukes svært sjelden, og noen svært få brukes ofte. Ord med lik skrivemåte kan bety ulike ting (homonymer), andre ganger kan ord som skrives helt forskjellig kan bety det samme (synonymer). Små forskjeller i skrivemåte kan noen ganger ha stor semantisk betydning. Andre ganger kan ortografisk variasjon (både tillatte former og skrivefeil) gi mange sammenfallende kategorier.

Til bruk for klassifisering må vi finne systematiske måter å karakterisere tekstbruken på. Eksempler på slike karakteristikker er rent deskriptive metoder og metoder for å måle sammenhengen mellom tekst og kategoriene i den aktuelle klassifisering.

4.3 Deskriptive metoder og resultater

Selv om det er levert betydelige mengde tekst til Arbeidstakerregisteret, brukes kun AKU-data i analysene i dette notatet. Dette fordi AKU er en mer ensartet kilde både med hensyn til leveringsmetode og klassifikasjonsrutiner. For vurdering og resultater om tekst i register henvises til tidligere notater i denne serien. Det kan bli aktuelt å studere tekstmateriale fra register på bakgrunn av de metoder som utvikles på AKU-data. Utfra det vi vet om hyppighet og variasjon i tekst vil det være gunstig å ha en betydelig utvalgsstørrelse. I mange av forsøkene inngår derfor data fra mange utvalg, f.eks. et samlet utvalg av 20 kvartalsfiler 2000-2004, samt konstruksjon av årsfiler 1996-2005.

En forutsetter da at strukturen (forholdet mellom tekst og klassifikasjon) ikke endrer seg særlig over tid. Mulige feilkilder i denne antagelsen er f.eks. justering av instruksen til koderne, og språklige trender. Først benyttes endel rent deskriptive undersøkelser av tekstmaterialet som belyser noen av de statistiske egenskaper til tekst.

4.4 Frekvensanalyser

Tabellen viser frekvensfordelingen for tekst og en tekstgruppering. Grupperingsmetoden er bruk av den såkalte *soundex*-verdien av de 30 første tegn. Dette er en funksjonen som beskrives nærmere lenger ut. *Soundex* har vært nyttet noe i automatiske yrkeskoding i register.

Tabell 4-1: Frekvensanalyse av tekst. AKU. Vektet gjennomsnitt 2000-2004.

	Frekvensfordeling	
	Tekster	Tekstgrupper
N	53 785	48 295
Mean	43	47
Std Deviation	376	421
100% Max	36 844	38 484
99 %	325	404
95 %	76	84
90 %	58	61
75% Q3	35	37
50% Median	18	19
25% Q1	9	9
10 %	6	6
5 %	5	5
1 %	4	4
0% Min	1	1

I dette materialet med 304 049 records er det brukt over 50.000 forskjellige tekster. Fordelingen er karakteristisk for tekstsdata: stor variasjon, og svært få tekster er brukt særlig mange ganger. Dette illustrerer utfordringene både ved å klassifisere utfra tekst og å måle kvaliteten på anvendelsen av tekst.

4.4.1 Ord som enhet

Noe av det mest opplagte vi gjør for å systematisere tekst, er å dele den opp i ord. Dette er en fundamental egenskap ved tekst som ikke kan sammenliknes med andre tradisjonelle datatyper. For å vise hvordan teksten som leveres til yrkesklassifisering skiller seg fra tekst generelt, kan vi ta utgangpunkt i frekvens vs betydningsmessig signifikans. I sammenheng med ekstrahering av nøkkelord fra dokumenter antar H.P. Luhn (1958) m.fl. at både ord som brukes svært ofte eller svært sjelden vil være *minst nyttige* i klassifiseringen. Vi får en illustrasjon på dette ved de hyppigst forekommende ord i norsk tekst, gjengitt i tabellen under.

Tabell 4-2: Vanligste ord i norsk tekst generelt.

1	I	6	TIL	11	AV	16	IKKE
2	OG	7	SOM	12	MED	17	DEN
3	DET	8	EN	13	AT	18	OM
4	ER	9	FOR	14	HAR	19	ET
5	PÅ	10	Å	15	DE	20	HAN

Dette er ord som nok i liten grad vil karakterisere teksten de finnes i, selv om de kan ha stor betydning i en gitt sammenheng. Når vi studerer vanlige ord som oppgis som yrke og arbeidsoppgaver, finner vi et tildels helt annet bilde.

Tabell 4-3: Vanligste ord oppgitt i yrke og arbeidsoppgaver. AKU 2000-2004.

1	LEDER	6	SALG	11	SEKRETÆR	16	UNDERVISNING
2	OG	7	HJELPEPLEIER	12	SJÅFØR	17	FOR
3	AV	8	SYKEPLEIER	13	RENHOLDER	18	BUTIKKMEDARBEIDER
4	DAGLIG	9	KONSULENT	14	EKSPEDITØR	19	I
5	LÆRER	10	ASSISTENT	15	SELGER	20	ADM

Selv om noen av ordene er generelle og intetsigende i forhold til formålet, må vi si at hvertfall halvparten er ganske spesifikke mht. arbeidsoppgaver. Dette gjør at tekstmaterialet må sies å ha en annen karakteristikk enn norsk tekst generelt. Det er også slik at i dette materialet tilsier ikke høy frekvens generalitet. Altså at et ord forkommer hyppig kan

skyldes at det brukes i mange ulike sammenhenger og er uspesifikt med hensyn til yrke. Men i andre tilfeller beskriver det spesifikke og vanlige arbeidsoppgaver. Dette gjør at vi i det aktuelle tekstmaterialet ikke uten videre kan anvende resultater fra generell tekst. Men det er naturligvis viktig å kjenne til slike metoder for å bygge videre på til vårt bruk.

4.4.2 Frekvensfunksjon

Zipf (1949) med senere forbedring av Mandelbrot (1954) formulerte en lovmessighet mellom rangeringen av frekvensen til et ord og den målte frekvensen selv (altså hvor ofte det forekommer).
I den enkleste modellen er frekvensen omvendt proporsjonalt med rangen for en konstant *k:*

$$f = \frac{k}{r}$$

Dette stemmer som regel unntatt for spesielt høye og lave ranger. Den forbedrete tilpasningen, for parametere *P, B, q:*
$f = P(r+q)^{-B}$ som i engelsk tekst kan ha verdiene $P=10^{5.4}$, $B=1.15$, $q=100$

Den opprinnelige funksjonen kan betraktes som en eksponentialfunksjon og hvis man plotter den i et dobbeltlogaritmisk system tilsvarer den rett linje. Vi utnytter dette til visuell vurdering av regulariteten i vårt materiale. Som forventet er det endel avvik i de høyeste og laveste rangeringene. Utfra denne enkle analysen kan vi hvertfall skille ut 3 klasser av ord, uten å undersøke selve innholdet (betydning av hvert ord):

- Meget sjeldne ord: kan være enkle å klassifisere, men har liten effekt i automatiske metoder.
- Ord som har en regelmessig frekvensfordeling, og derfor er tilgjengelige for analytiske metoder. Disse kan derfor grupperes etter betydning og spesifisitet.
- Meget vanlige ord: enkelte kan være for generelle til klassifisering, men spesifikke kan ha stor nytte til kataloger, omkoding osv.

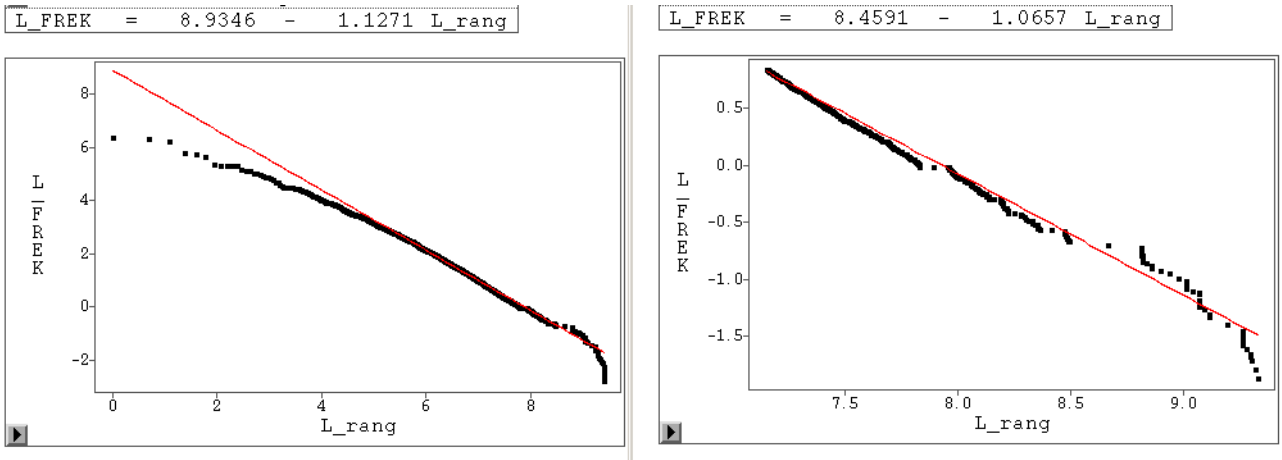
Videre vil en slik metode anvendt på et ukjent tekstmateriale kunne indikere i hvilken grad man har med regulære tekstdata å gjøre. Vi ser her litt nærmere på forholdet mellom frekvens og rangering av ord. Dette viser seg å være hyperbel-liknende funksjon, men følger ikke helt formelen $f = \frac{k}{r}$

For å illustrere dette plotter vi log(f) mot log(r), og viser en lineær modell utfra $\log(f) = \alpha \cdot \log(r) + \beta$

Altså finnes frekvensen ved: $f = e^{\alpha \cdot \log(r) + \beta}$

Som antatt passer dette best for midlere verdier. Forsøker også en robust tilpasning ved å fjerne de 10% av ordene med høyeste og laveste frekvenser. Vi får da kuttet langt mer i de høyere frekvensene og vi ser at det er særlig de sjeldne ordene hvor det ikke er en regulær sammenheng. Diagrammene viser hvordan dette arter seg for den komplette ordmengden, og for den trimmede listen (til høyre).

Figur 4-4: Frekvens og rangering av ord. AKU 2004.



4.5 Tekstlengde

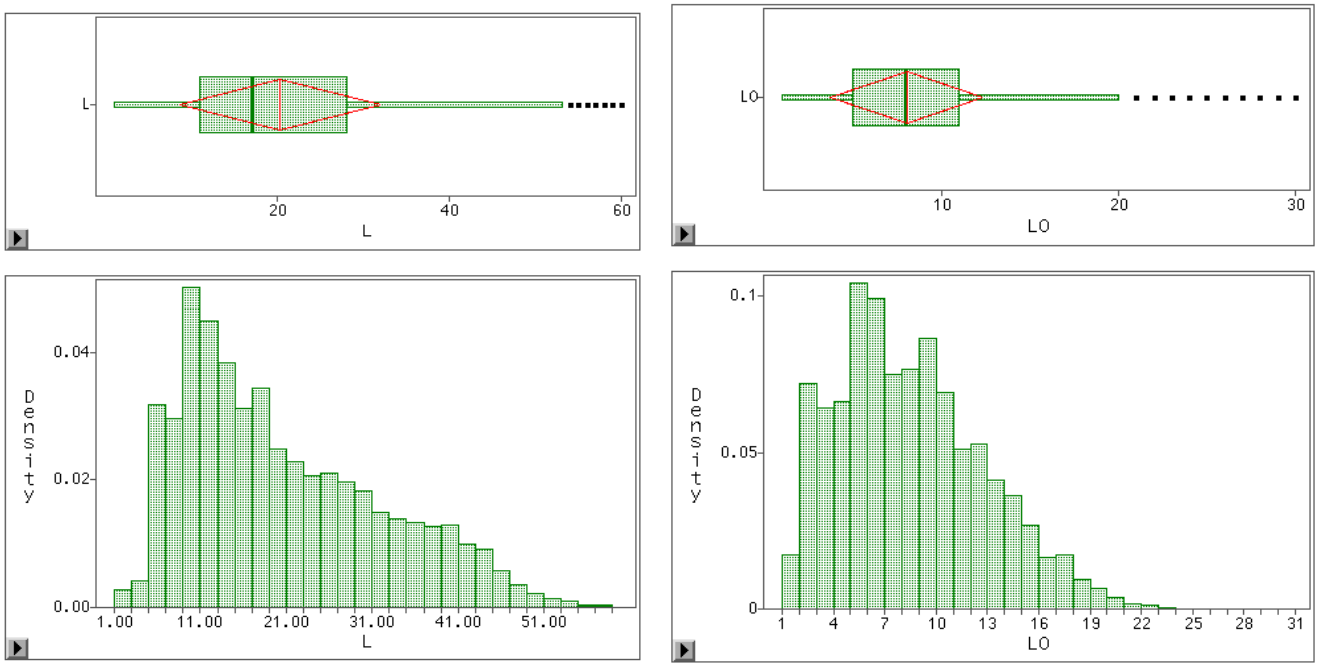
Viser noen målinger av lengden på tekster og ord, målt i antall tegn. Størrelsen er interessant i forhold til effektivitet, hvor godt man utnytter tilgjengelig plass, og for å undersøke eventuell sammenheng mellom lengde og egnethet. Det er ikke nødvendigvis gitt at mye mer tekst gjør en enhet lettere å klassifisere.

Det er i Arbeidstakerregisteret satt av plass til 40 tegn, i Arbeidskraftundersøkelsen er det plass til 30 tegn for yrke og 30 for arbeidsoppgaver. I disse forsøkene er yrke og arbeidsoppgaver slått sammen, og den tilgjengelig plass er da 60 tegn. Diagrammet til venstre viser distribusjon av lengden av hele teksten. Det er stor variasjon i hvor mye man skriver, men de fleste er ganske korte. Typisk lengde er 17 bokstaver og halvparten har mellom 11 og 28 bokstaver. Det er svært

få tilfeller hvor man trenger mer plass enn 40 tegn, som er feltstørrelsen i Arbeidstakerregisteret. Det betyr at mange av de yrkestitteltekster som er vanskelig å yrkeskode, neppe skyldes mangel på plass.

Diagrammet til høyre viser at det for det meste brukes korte, greie ord. Lange, spesialiserte titler som gjerne finnes i eksempler og i yrkeskatalogen er ikke typisk. Dette har stor betydning for utvikling av lister for automatisk klassifisering. Det betyr også at en enkel kontroll mot en vanlig ordliste burde ha stor effekt når det gjelder kvaliteten på tekst til manuell koding. Hverken i AKU og AA kontrolleres teksten maskinelt ved inndata.

Figur 4-5: Antall tegn i samlet tekst og i enkeltord. AKU 2000-2004.



Moments			
N	305931.00	Sum Wgts	305931.00
Mean	20.30	Sum	6209454.00
Std Dev	11.71	Variance	137.09
Skewness	0.73	Kurtosis	-0.28
USS	167973532	CSS	41940802.5
CV	57.69	Std Mean	0.02

Quantiles			
100% Max	60	99.0%	50
75% Q3	28	97.5%	46
50% Med	17	95.0%	43
25% Q1	11	90.0%	38
0% Min	1	10.0%	7
Range	59	5.0%	6
Q3-Q1	17	2.5%	5

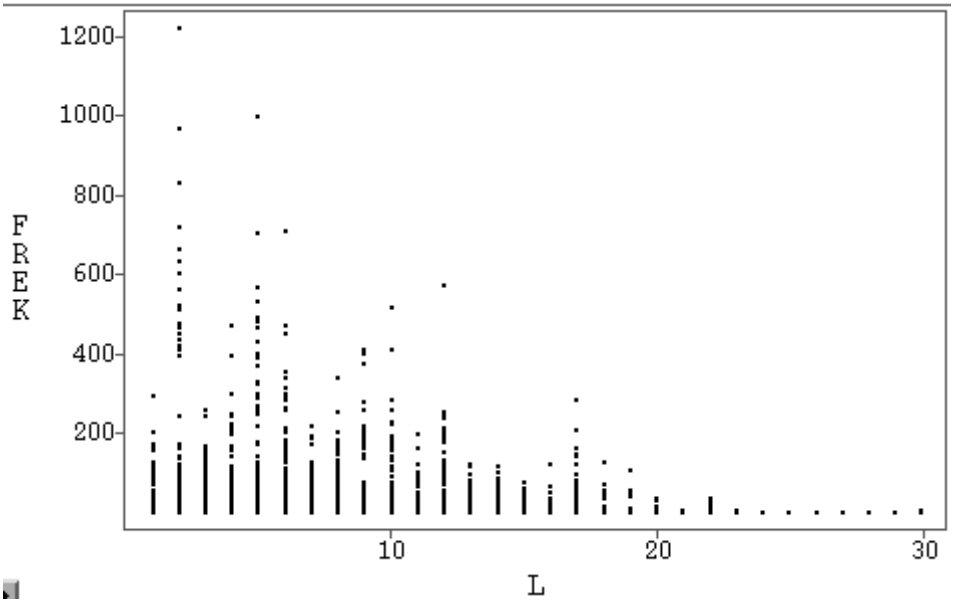
Moments			
N	720247	Sum Wgts	720247.0
Mean	8.0	Sum	5791328.0
Std Dev	4.3	Variance	18.5
Skewness	0.6	Kurtosis	-0.1
USS	59924406	CSS	13357770
CV	53.6	Std Mean	5.07E-03

Quantiles			
100% Max	30	99.0%	19
75% Q3	11	97.5%	17
50% Med	8	95.0%	16
25% Q1	5	90.0%	14
0% Min	1	10.0%	3
Range	29	5.0%	2
Q3-Q1	6	2.5%	2
Mode	5	1.0%	1

4.5.1 Lengde av enkeltord

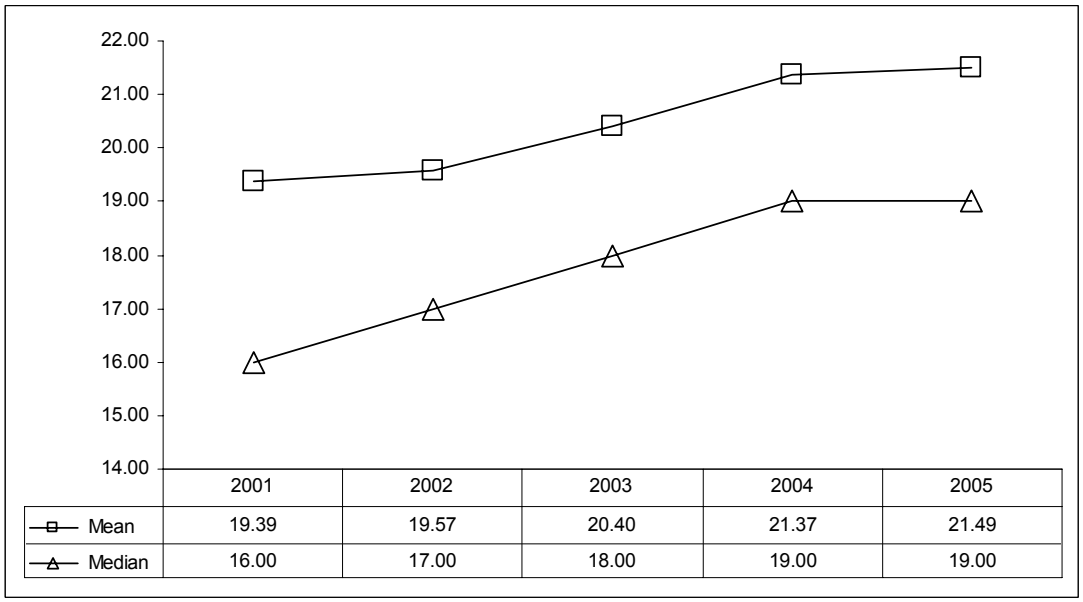
Et plot av forholdet mellom ordlengde og frekvens viser at det er en klar sammenheng, selv om den lineære korrelasjonen ikke er særlig høy. Ikke overraskende er det spesielt lange ord som er sjeldnest. I tråd med tidligere viste analyser av lengde, er de vanligste ordene ganske korte. Dette reduserer variasjonsmulighetene og tilfeldige feil, noe som kan gi bedre effektivitet ved automatiske metoder.

Figur 4-6: Frekvens etter ordlengde, AKU 1996-2004, justert for persondubletter.



Generelt ser det ut til at et brukes stadig lengre ord, som vist i neste figur. Det har vært litt forskjellig utvikling i de siste 10 år, men de siste 5 år viser en jevn vekst. Særlig det at medianverdien så vel som gjennomsnittet øker, indikerer en reell utvikling. Det kan være på sin plass å framheve betydningen av robuste metoder, når vi holder på med tekst og små grupper.

Figur 4-7: Gjennomsnittlig og median ordlengde, AKU 2001-2005, justert for persondubletter.



	2001	2002	2003	2004	2005
Mean	19.39	19.57	20.40	21.37	21.49
Median	16.00	17.00	18.00	19.00	19.00

4.5.2 Ordmengde

En opptelling av hvor mange ord som oppgis i hver tekst tilsier at det overveiende er få ord og at omlag 30% har en enkel 1-ords tittel. Det betyr at man kommer langt med svært enkle kodelister, men også at det er betydelig gevinst hvis man kan etablere en metode for å gruppere 2-ords titler, altså å identifisere ord som regelmessig opptrer i en bestemt rekkefølge.

Tabell 4-8: Antall ord oppgitt i yrke og arbeidsoppgaver. AKU 2000-2004.

	Distinkte ord		Ordbruk i alt	
Ord i teksten	Antall	Prosent	Antall	Prosent
1	8829	29.75	51533	40.58
2	6583	22.18	24299	19.14
3	4641	15.64	16279	12.82
4	3996	13.46	14518	11.43
5	2572	8.67	9218	7.26
6–12	3058	10.3	11138	8.78

4.5.3 Ordplassering

Den innbyrdes plasseringen av ord er viktig i forbindelse med søk og matching av ord i yrkestittel og arbeidsoppgaver. Viser noen eksempler på dette i tabellen nedenfor, hvor det går fram at det er store forskjeller i den typiske plasseringen av de enkelte ord. F.eks. "butikkmedarbeider" er både svært vanlig og forekommer som regel alene. Da dette ordet også nærmest er entydig yrkeskodet, er det lite å hente ved å søke i teksten eller å be om mer tekst. Ordet "ledelse" er aldri plassert som 1. ord, og derfor ingen vits å ha med i en indeks som bare sammenlikner tekst direkte (begynnelsen). Dette demonstrerer betydningen av tilpasset analyse og bruk av trinnvise metoder. For å bygge opp systemer og kataloger for manuell, datastøttet eller automatisk klassifisering av tekst er det viktig å vurdere de spesielle egenskaper som tekst har. Som en metafor på dataløsninger kan det være bedre å tenke på et tre med greiner og kvister, ikke en bredest mulig motorvei.

Tabell 4-9: Plassering av utvalgte ord. Yrke og arbeidsoppgaver. AKU 2000-2004. Prosent.

	I alt	*1.*	*2.*	*3.*	*4.*	*5.++*
I alt	100	70	20	6	3	1
ADJUNKT	100	94	6	0	0	0
ASSISTENT	100	85	13	1	1	0
BUTIKKMEDARBEIDER	100	99	1	0	0	0
DAGLIG	100	84	9	5	1	1
EKSPEDITØR	100	97	2	1	0	0
HJELPEPLEIER	100	96	4	0	0	0
INGENIØR	100	74	24	1	0	0
KONSULENT	100	62	34	2	1	0
LEDELSE	100	0	35	41	16	8
OMSORG	100	1	46	24	23	6
REGNSKAP	100	11	48	25	12	3
RENHOLD	100	17	65	10	6	3
RENHOLDER	100	97	2	0	0	0
RÅDGIVER	100	68	25	5	1	1
SEKRETÆR	100	81	17	2	0	0
SELGER	100	83	12	2	1	1
SJÅFØR	100	84	13	2	1	0
SNEKKER	100	88	9	2	1	1
SYKEPLEIER	100	88	11	1	0	0
UNDERVISNING	100	1	83	9	5	2
VEDLIKEHOLD	100	4	52	20	18	6

4.6 Gruppering og sammenlikning

Materialet viser at det er stor variasjon med svært mange ulike tekster. Både for å analysere tekstvariabler og å søke/klassifisere utfra tekst, fordres det metoder for å gruppere ulike tekstverdier. I forrige avsnitt ble det så vidt nevnt soundex-funksjonen i SAS. Dette er en funksjon som grupperer utfra lydlikhet, en algoritme som er laget for engelsk språk, se Knuth (1973).

Tabell 4-10: SOUNDEX-funksjonen

I. Alle doble forekomster slettes.
II. Behold første bokstav.
III. Dropp alle forekomster av: A E H I O U W Y
IV. Grupper øvrige bokstaver slik:
- 1: B F P V
- 2: C G J K Q S X Z
- 3: D T
- 4: L
- 5: M N
- 6: R

Tiltross for forskjellene i norsk og engelsk ortografi kan denne funksjonen gi gode resultater for å gruppere ord med ulike skrivemåter, skrivefeil, nynorsk, m.m. I automatisk klassifisering på registerdata brukes Soundex-funksjonen til gruppering med støtte av andre kjennemerker. Det er også laget en egen "SSB-soundex", som ble utviklet i BULL-assembler på 1970-tallet, og senere implementert i SAS som en felles makro. Selv om denne algoritmen har samme navn, er den spesialberegnet for navnesøk, og har mindre anvendelse i generell tekst. Tabellen viser eksempel fra AKU-data på gruppering ved SAS-soundex.

Tabell 4-11: Eksempel på gruppering av skrivemåter. AKU 2000-2004.

HJELPEPLEIERR, HJELPEPLEOER, HJELPEPLEIEER, HJELPEPELEIR, HJELPEPLØEIER, HJELPEPLEEIER, HJELPEPLEIRE, HJELPEPLEIER, HJELPEPLEEIERE, HJELPEPELIER, HJELPEPELEIER, HJELPEPLIER, HJELOPEPLEIER, HJELPEPLERI, HJELPEPLIEIER, HJELPEPLEIOER, HJELPEPLEI9ER, HJELPEPLEER, HJELPEPLEIR, HJELPEPLEIAR, HJELPEPLER, HJELPEPLEIER4, HJELPEPLEIERE.

Begge disse funksjoner er entydige ved at de kun tar verdien som argument, og leverer et fast resultat. Forholdet mellom effektivitet og spesifisitet er altså fastlagt. For å kunne avstemme dette parametrisk finnes en annen funksjon i SAS, nemlig "spedis" (spelling distance) som gir den *asymmetriske staveavstanden* mellom to ord fra søkeord til nøkkelord. Resultatet er altså ikke en gruppering direkte, men en eksakt verdi for likheten mellom to ord. Avstanden beregnes som vist under.

Tabell 4-12: SPEDIS-funksjonen

1. Beregn perturbasjoner med følgende verdier:
   a. 0 ingen endring
   b. 25 slett en dobbel bokstav
   c. 50 fordoble enkel bokstav
   d. 50 bytt om to etterfølgende bokstaver
   e. 50 slett siste bokstav
   f. 35 legg til bokstav tilslutt
   g. 50 slett indre bokstav
   h. 100 legg til indre bokstav
   i. 100 bytt ut indre bokstav
   j. 100 slett første bokstav
   k. 200 legg til første bokstav
   l. 200 endre første bokstav
2. Staveavstanden er lik sum av perturbasjoner dividert med lengden på søkeord.

Eksemplene i tabellen er igjen fra reelle AKU-data. Dette kan altså brukes for å samle ulike skrivemåter av samme ord, og hvor man kan **justere** grensene for likhet. For å illustrere at det er *asymmetrisk* mål oppgis også avstanden for det omvendte tilfelle.

Tabell 4-13: Eksempel på måling av likhet i skrivemåter. AKU 2000-2004.

Nøkkelord	Søkeord	Staveavstand	Omvendt
ARBEIDSLEDER	ARBEIDSLEDER	0	0
ARBEIDSLEDER	ARBEIDSSLEDER	2	3
ARBEIDSLEDER	ARBEIDESLEDER	4	7
ARBEIDSLEDER	ARBEIDSLEDERE	4	2
ARBEIDSLEDER	AREBEIDSLEDER	4	7
ARBEIDSLEDER	ARBEIDLEDER	8	4
ARBEIDSLEDER	ARBEIDSLADER	8	8
ARBEIDSLEDER	ARBEIDSLEDAR	8	8
ARBEIDSLEDER	ARBEIDSLEDEL	8	8
ARBEIDSLEDER	ARBEIDSLEIER	8	8
ARBEIDSLEDER	ARBEISLEDER	8	4
ARBEIDSLEDER	AREIDSLEDER	8	4
ARBEIDSLEDER	ARBEIDELDER	12	9
ARBEIDSLEDER	ARBEIDSLEDELS	12	10
ARBEIDSLEDER	ARBEIDLEDAR	16	13
ARBEIDSLEDER	ARBEIDSLEDELSE	16	12
ARBEIDSLEDER	ARBEIDSLEDENDE	16	12
ARBEIDSLEDER	ARBEIDSLEDESLE	16	12
ARBEIDSLEDER	ARBEIDSLEIAR	16	16
ARBEIDSLEDER	ARBEISLEDAR	16	13
ARBEIDSLEDER	ARBEIDSLDELSE	20	21

Ved enda større avstand enn vist i tabellen, begynner vi å få ord med annen betydning, f.eks. : "arbeidssøkende" har avstand=27.

4.7 Omfang og bruk av tekst

I dette avsnittet har vi beskrevet tekstdata og demonstrert metoder som benytter rent tekniske mål. Før vi går videre med metoder som analyserer innhold, kan vi vise virkningen av noen av de tekniske sider som er gjennomgått på aktuelle data fra AKU. Tabellene vises omfangen av levert tekst i antall tegn og ord, etter kvartal. Det er ser ut til å være ganske stabilt over tid selv om det altså er stor variasjon i hva som skrives.

Det er kanskje en langsiktig trend at yrkestitler og teksten blir noe lengre, og at det blir vanligere å bruke flere ord. Slike mål på tilstanden og utviklingen er eksempler på bakgrunnsinformasjon som ikke sier noe om selve kodingen, men som burde nyttiggjøres både ved opplæring av intervjuere og kodere, samt ved utvikling av kataloger og kodesystemer. Særlig bruk av lange og uspesifikke titler og engelsk språk, bør være gjenstand for kontroll og oppfølging. I intervjusituasjonen kan dette gjøres direkte, men det i registersammenheng krever andre mekanismer for kontroller og tilbakemelding til oppgavegiver.

Tabell 4-14: Lengde av tekst i AKU yrke og arbeidsoppgaver. Prosent.

	Ialt	*1–30*	*31–40*	*41–60*
2000-1	100	80	14	6
2000-2	100	79	14	7
2000-3	100	80	14	6
2000-4	100	80	13	7
2001-1	100	81	13	7
2001-2	100	81	13	7
2001-3	100	81	12	7
2001-4	100	81	12	6
2002-1	100	81	12	6
2002-2	100	81	12	7
2002-3	100	80	13	7
2002-4	100	80	13	7
2003-1	100	79	14	7
2003-2	100	79	14	7
2003-3	100	79	14	7
2003-4	100	77	15	8
2004-1	100	77	15	8
2004-2	100	77	15	8
2004-3	100	76	15	9
2004-4	100	75	16	9

Tabell 4-15: Antall ord i tekst i AKU yrke og arbeidsoppgaver. Prosent.

	Ialt	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7–12*
2000-1	100.0	0.6	37.7	26.6	14.2	10.8	5.6	3.0	1.5
2000-2	100.0	0.6	37.4	26.5	14.3	10.7	5.7	3.2	1.7
2000-3	100.0	0.6	38.4	26.2	13.7	10.5	5.5	3.3	1.7
2000-4	100.0	0.7	39.1	25.8	13.5	10.2	5.5	3.3	1.8
2001-1	100.0	0.7	39.9	25.6	13.5	10.1	5.2	3.2	1.8
2001-2	100.0	0.7	40.6	25.4	13.4	10.0	5.2	3.1	1.7
2001-3	100.0	0.6	42.2	24.5	13.0	9.8	5.0	3.0	1.7
2001-4	100.0	0.5	43.2	24.0	13.0	10.0	4.9	2.7	1.7
2002-1	100.0	0.5	43.7	23.5	12.8	9.8	5.1	2.9	1.8
2002-2	100.0	0.4	43.4	23.5	12.8	10.2	5.0	2.9	1.7
2002-3	100.0	0.5	42.3	23.8	12.7	10.3	5.3	3.1	2.0
2002-4	100.0	0.5	41.8	23.9	12.8	10.4	5.2	3.4	2.0
2003-1	100.0	0.5	40.2	24.2	13.3	10.6	5.5	3.6	2.1
2003-2	100.0	0.5	39.2	25.1	13.7	10.8	5.5	3.3	1.9
2003-3	100.0	0.5	39.2	25.2	13.6	10.8	5.6	3.2	1.9
2003-4	100.0	0.6	37.2	24.9	13.6	11.5	6.3	3.9	2.2
2004-1	100.0	0.5	35.5	25.4	13.7	12.0	6.4	4.1	2.3
2004-2	100.0	0.5	35.2	25.5	14.0	11.9	6.5	4.0	2.3
2004-3	100.0	0.4	34.9	25.5	14.0	12.2	6.8	4.1	2.2
2004-4	100.0	0.5	34.1	24.8	14.6	12.8	6.9	4.1	2.3

4.8 Forholdet mellom ord og yrke

Klassifisering utfra tekst foregår ved at det antas en større elle mindre sammenheng mellom ordene i teksten og kategorien (her: yrket). For å måle den samlede sammenhengen er det upraktisk å bruke korrelasjon eller Chi-kvadrat baserte metoder. Dette pga. at begge variabler er nominelle kategorielle og med svært mange verdier. Merk også at hvis vi så på hele teksten ville det blir enda mange flere ulike verdier for den uavhengige variabel.

Vi forsøker to metoder med utgangspunkt i at $p_{o,y} = P(yrke = y | ord = o)$ er andelen eller sannsynligheten for et yrke y gitt et ord o:

$$ent_o = \sum_{\forall y} (p_{o,y} \cdot \log(p_{o,y}))$$ Entropien for ord o

$$uss_o = \sum_{\forall y} (p_{o,y})^2$$ Konsistensen for ord o

Begge måler i hvilken grad hvert enkelt ord kodes til et eller mange yrker. Tabellen viser gjennomsnittet for alle ordene pr. år. Det er litt svingninger, men det ser ut til at det er en synkende tendens i perioden. Det betyr altså at enkeltord i mindre grad kan brukes for å klassifisere yrke direkte. At enkeltord alene betyr mindre kan skyldes to ting:

- Sammenhengen (ordkombinasjoner) får større betydning.
- Bruken av tilleggsvariabler får større betydning.

Det kan også være interessant å se dette i forhold til hvor vanlige ord er, i lys at jo de fleste ord forekommer svært sjelden. Viser derfor også sammenhengsmålene vektet med ordfrekvensen. Det vektede forholdet blir da mye lavere, og viser litt annen utvikling. Det viser tydelig at årsaken til at mange ord er entydige indikatorer, skyldes rett og slett at de er veldig sjeldne, ofte unike. Dette demonstrerer igjen behovet for en eller annen form for prekategorisering før selve klassifiseringen, f.eks. gruppering etter noen av de ortografiske funksjonene som er nevnt.

Tabell 4-16: Samlet sammenheng mellom ord og yrke. Gjennomsnitt pr. år. AKU 1996-2004.

	Uvektet		Gj.sn. vektet med frekvens	
	Entropi	USS	Entropi	USS
1996	-0.263	0.866	-1.598	0.449
1997	-0.265	0.864	-1.639	0.460
1998	-0.260	0.867	-1.636	0.462
1999	-0.270	0.862	-1.627	0.463
2000	-0.265	0.865	-1.605	0.465
2001	-0.263	0.866	-1.613	0.466
2002	-0.267	0.864	-1.623	0.468
2003	-0.268	0.863	-1.645	0.471
2004	-0.270	0.862	-1.693	0.471

Om vi ser på den rene entropien, altså spredningen av fordelingen på ulike ord uavhengig av forholdet til yrke finner vi et mer loivende bilde. Det ser ut til at måltallene øker i hele perioden, som tolkes som at ordbruken konsentrer seg om litt færre ord som brukes i større grad. Det totale antall ord har også gått ned, selvom det må bemerkes at dette isolert sett er veldig lite robust. Viser derfor også summen av andelene av de ti vanligste ordene. Resultatene indikerer nok at konsentrasjonen er jevnt økende, om enn ikke stor.

Tabell 4-17: Spredning av ordbruk, etter år. AKU 1996-2004.

	Entropi	USS	Antall ord	Andel av topp 10
1996	-7.467	0.00324	14524	13.5 %
1997	-7.505	0.00310	13696	12.9 %
1998	-7.492	0.00304	13496	12.8 %
1999	-7.433	0.00305	12628	12.8 %
2000	-7.430	0.00320	12734	13.3 %
2001	-7.397	0.00333	12262	13.6 %
2002	-7.378	0.00339	12022	13.8 %
2003	-7.396	0.00338	12369	13.7 %
2004	-7.392	0.00359	12753	14.2 %

Tilslutt noe om bruken av avanserte titler, fremmedord og engelsk. Vi har ikke mulighet for å kontrollere mot ordlister, slik at dette må vises indirekte. Metoden i tabellen er ganske enkelt å telle opp forekomsten av bokstavene c,q,w,x,z, som ikke forekommer i tradisjonell norsk. Resultatene kan indikere at bruken har økt de senere årene, noe som ville samsvare med en vanlig oppfatning blant de som jobber med yrkestitler.

Tabell 4-18: Bruk av visse bokstaver, etter år. AKU 1996-2004.

	c	q	w	x	z
1996	414	5	22	415	10
1997	219	2	13	103	6
1998	171	4	11	32	3
1999	183	4	13	13	5
2000	165	7	13	13	7
2001	171	3	13	16	6
2002	180	3	16	14	5
2003	192	3	21	14	4
2004	224	4	18	21	4

5 Nytte av tekst og tekstanalyse

Vi gjør forsøk og viser resultater av praktisk nytte av tekstanalyse. Først omtales klassifisering av yrke utfra tekstene som foreligger i datamateriale fra Arbeidskraftundersøkelsen. Alle sysselsatte spørres om yrkestittel og arbeidsoppgaver og i utvalget utgjør disse ca. 15 000 personer hvert kvartal. Yrkeskodingen i AKU foregår "manuelt", i praksis betyr det en skjønnsmessig klassifisering av personell som har lang erfaring med kodearbeid, støttet av instruksverk og

datatekniske hjelpemidler. Vi tar det for gitt at validiteten i yrkeskodingen er meget god, og derfor kan brukes som referanseverdi for en analyse. Som det kan gå fram av avsnittet foran betrakter vi ikke tekst som en helt vanlig kategorivariabel og vi gjør ikke tekstvariabler til gjenstand for tradisjonelle metoder som regresjonsanalyser, osv.

En kan trygt si at teksten som oppgis som yrkestittel og arbeidsoppgaver er hovedkilden til yrkesklassifiseringen. Det nyttes også annen informasjon, vesentlig kjennemerker fra register som bedriftens næring og størrelse, til en viss grad personens utdanning, og i enkelte tilfeller diverse opplysninger som bedriftens navn, størrelse og organisering. I dette notatet analyseres ikke virkningen av de enkelte tilleggsvariabler. Man kan anslå den samlede effekten av tilleggsvariabler indirekte fra sammenhengen vi påviser mellom tekst og yrkeskode. I tilfeller med liten sammenheng, forutsettes da en større effekt av tilleggsvariabler. Det kan selvsagt i noen tilfeller tilskrives uspesifikke inndata, eller teoretisk også ustabilitet i selve kodingen. Selv om det ikke er noe realistisk mål å bli helt uavhengig av tilleggsinformasjon, kan det være mye å hente på å forbedre tekstene som oppgis i AKU og kanskje særlig titlene som leveres til Arbeidstakerregisteret.

Når vi betrakter tekst til yrkesklassifisering, vil dette skille seg fra generelle tekstanalyser som leksikalske klassifiseringer, osv. Det er først og fremst fordi vi her har tilknyttet et entydig kjennemerke til innhold/mening, nemlig yrkeskoden. Noen lingvistiske begrep kan da defineres operasjonelt, uten å ta stilling til innholdet f.eks.:

- "Synonymer" = ulike ord som gir samme yrkeskode.
- "Homonym" = samme ordet gir ulike yrkeskoder.

I denne sammenhengen ser vi altså bort fra hva som er "god norsk", formell ortografi, grammatikk, osv.

5.1 Forsøk og resultater

5.1.1 Bruk av enkeltord

Istedenfor å operasjonalisere det generelle lingvistiske begrepsapparat, kan vi betrakte hvert enkelt *ord* som en mer eller mindre egnet *indikator* for et yrke. En kan tenke på det som at et ord i en tekst forteller noe om hva teksten handler om. I denne klassifiseringen er det optimale at teksten peker på et bestemt yrke. Vi skal undersøke om enkelte ord alene kan brukes for å klassifisere. For å beregne dette defineres:

Antallet med ord *t*

$$X_t$$

Antallet med ord *t* og yrke *u*

$$Y_{t,u}$$

Andelen med yrke *u* av de med ord *t*

$$P_{t,u} = \frac{Y_{t,u}}{X_t}$$

Denne andelen kan tolkes som sjansen for å ha yrket *u* gitt at ordet *t* er oppgitt. Vi skal nå ikke bruke dette til å estimere yrkesstatistikker, men gjøre forsøk med automatisk koding. En kunne nå si at bare "entydige" ord kunne brukes, altså de med p=100%. Det kan bli upraktisk av flere grunner. Mange har p=100% fordi de forekommer kun 1 gang, og derfor liten effekt. Mange har nær 100%, hvor den største andelen er korrekt og med en liten mengde tilfeldige feil. I andre tilfeller er det en mer distinkte andeler fordelt til flere yrker, bl.a. utfra andre tilleggskjennemerker. En enkel raffinering av dette var å velge det største yrket for alle med en andel over en viss størrelse. Denne størrelsen måtte bestemmes vilkårlig. Vi ønsker et mål som er uavhengig i forhold til om spredningen skyldes tilfeldige feil eller tilleggsinformasjon og uten å fastsette en vilkårlig størrelse. Vi velger her å ta med *alle* yrkesandelene til hvert ord i den *uvektede, ukorrigerte kvadratsummen*:

$$uss_t = \sum_{u=1}^{U} (P_{t,u})^2$$

Dette er ganske robust i forhold til mange små andeler. Jamfør også med entropi og andre beregninger på side 26 i notat 2005/43.

Denne kan tolkes som sjansen for å klassifisere riktig yrkeskode utfra kun ordet, gitt tilfeldig fordeling innen de som har ordet. Legg merke til at i de tilfeller spredningen skyldes tilfeldige feil, ville det blitt mer korrekt å anvende bare den største andelen. I de tilfeller hvor spredningen på flere yrker skyldes bruk av tilleggsopplysninger, blir tilfeldig fordeling til alle andelene mer korrekt. Siden vi ikke vet forholdet mellom tilleggsdata og mengden tilfeldig feil, gjør vi forsøket enklest mulig ved å kode med det mest sannsynlige yrket. I tillegg stiller vi krav til at ordet må ha en viss frekvens, altså kutte ut de mest sjeldne. Grensen kutt kan da balanseres mellom effekten altså utbytte, hvor stor del vi får kodet; og kvaliteten målt ved den forventede mikrokonsistensen.

5.1.2 Data

Også i disse forsøkene benyttes et samlet datasett fra flere kvartalsfiler fra AKU, som inneholder alle sysselsatte som har oppgitt yrke/arbeidsoppgaver og fått yrkeskode. Vi gjengir først noen generelle undersøkelse av kvaliteten. Det partielle frafallet av disse data er svært lite, med en svak trend i retning av enda mindre frafall.

Tabell 5-1: Partielt frafall av yrkesdata og -kode. AKU 2000-2004.

		Har tekst		Mangler tekst	
Kvartal	I alt	Har kode	Ukodet	Har kode	Ukodet
2000-1	100	99.32	0.11	0.56	0.01
-2	100	99.32	0.07	0.56	0.04
-3	100	99.25	0.07	0.64	0.04
-4	100	99.10	0.11	0.74	0.05
2001-1	100	99.06	0.09	0.75	0.10
-2	100	99.16	0.11	0.67	0.06
-3	100	99.15	0.12	0.64	0.09
-4	100	99.26	0.13	0.53	0.08
2002-1	100	99.22	0.14	0.51	0.14
-2	100	99.36	0.15	0.43	0.07
-3	100	99.25	0.15	0.56	0.04
-4	100	99.29	0.16	0.50	0.05
2003-1	100	99.30	0.11	0.54	0.05
-2	100	99.31	0.15	0.49	0.05
-3	100	99.30	0.15	0.52	0.03
-4	100	99.24	0.14	0.55	0.07
2004-1	100	99.31	0.12	0.54	0.03
-2	100	99.29	0.13	0.54	0.03
-3	100	99.46	0.12	0.38	0.04
-4	100	99.42	0.09	0.46	0.03

På bakgrunn av undersøkelsene om ordfrekvens, skiller vi mellom vanlige og sjeldne ord. Særlig sjeldne ord vil kunne være lett klassifiserbare, men ha liten effekt i systematiske metoder. Først undersøkes mengden ord som har frekvens over 100. Resultatet kan vi si er at selv om de fleste ord er sjeldne, inneholder de fleste tekster vanlige ord. Tabellen viser hvordan dette slår ut i de samlede AKU-data, i forhold til demografi og yrke. Andelen som overhodet ikke inneholder noen vanlige ord er i størrelsesorden 15%, med noe større andeler blant yngre menn og gruppen "andre yrker". Vi kan da anta at det blir brukt mer spesielle yrkestitler her.

Tabell 5-2: Andeler av records med og uten vanlige ord, etter demografi og yrke. AKU 2000-2004. Prosent.

		I alt	Inneholder et vanlig ord	Ingen vanlige ord
I alt	I alt	100.0	86.4	13.6
	16-19 år	100.0	83.1	16.9
	20-24 år	100.0	82.2	17.8
	25-39 år	100.0	86.7	13.3
	40-54 år	100.0	87.3	12.7
	55-66 år	100.0	86.9	13.1
	67-74 år	100.0	83.9	16.1
Menn	I alt	100.0	84.4	15.6
	16-19 år	100.0	79.3	20.7
	20-24 år	100.0	76.8	23.2
	25-39 år	100.0	85.3	14.7
	40-54 år	100.0	85.4	14.6
	55-66 år	100.0	85.3	14.7
	67-74 år	100.0	82.5	17.5
Kvinner	I alt	100.0	88.6	11.4
	16-19 år	100.0	86.9	13.1
	20-24 år	100.0	88.6	11.4
	25-39 år	100.0	88.2	11.8
	40-54 år	100.0	89.3	10.7
	55-66 år	100.0	88.8	11.2
	67-74 år	100.0	86.1	13.9
9 Andre yrker		100.0	78.3	21.7
1 Lederyrker		100.0	87.5	12.5
2 Akademiske yrker		100.0	85.7	14.3
3 Høyskoleyrker		100.0	87.0	13.0
4 Kontoryrker		100.0	83.3	16.7
5 Salgs- og serviceyrker		100.0	91.1	8.9
6 Bønder, fiskere ol.		100.0	90.2	9.8
7 Håndverkere		100.0	83.4	16.6
8 Operatører, sjåfører ol		100.0	84.1	15.9

5.1.3 Forsøk

Her undersøkes i hvilken grad enkeltord kan brukes alene for kode et bestemt yrke. Vi forsøker med krav til *uss>.8* og *n>50*. Med denne kvaliteten kunne vi kodet ca. 30% automatisk. Det er vel et ganske stort utbytte når man koder bare

ved hjelp av ett eneste ord i teksten. Den neste tabellen viser resultater av forsøk ved å måle konsistensen (samsvaret) med den opprinnelige yrkeskoden. I alt må vi kunne si at samsvaret er høyt og stabilt. Det er ikke overraskende at *lederyrker* både har spesielt lavt utbytte og laveste samsvar. Dette er yrker som oftest er avhengige av tilleggskjennemerker, hvis de i det hele tatt lar seg klassifisere. Hvis en strammer inn testkriteriet, synker utbyttet – uten at kvaliteten blir så mye bedre. Derimot kan man øke utbytte ved å anvende de grupperingsfunksjoner som er nevnt. Tabellen viser også resultatet ved bruk av soundex-funksjonen.

Tabell 5-3: Effekt av automatisk yrkeskoding, etter yrkesfelt og kvartal. AKU 2000-2004. Prosent.

	KUN 1 ORD		SOUNDEX	
	Utbytte	Samsvar	Utbytte	Samsvar
I alt	30.3	96.7	35.5	86.6
9 Andre yrker	44.6	95.5	49.2	88.1
1 Lederyrker	7.4	89.5	9.0	74.5
2 Akademiske yrker	25.1	97.2	31.6	62.2
3 Høyskoleyrker	16.2	95.5	27.0	90.3
4 Kontoryrker	21.5	92.5	23.8	83.8
5 Salgs- og serviceyrker	56.9	97.9	59.6	94.2
6 Bønder, fiskere ol.	20.5	96.7	23.3	80.1
7 Håndverkere	36.2	97.8	40.2	87.6
8 Operatører, sjåfører ol	15.5	94.1	21.1	68.1
2000-1	29.7	96.4	34.7	86.3
2000-2	29.5	96.6	34.3	86.5
2000-3	29.5	96.7	34.4	86.6
2000-4	29.6	97.3	34.7	87.1
2001-1	29.4	97.2	34.4	87.0
2001-2	29.9	97.0	34.8	87.0
2001-3	30.1	97.1	35.1	87.6
2001-4	29.7	97.1	34.9	87.2
2002-1	29.8	97.1	35.2	87.4
2002-2	30.2	97.0	35.6	87.6
2002-3	30.1	97.0	35.7	86.4
2002-4	30.3	96.9	35.9	86.1
2003-1	30.9	96.4	36.4	86.0
2003-2	30.9	96.5	36.4	86.4
2003-3	31.0	96.6	36.5	86.4
2003-4	31.1	96.3	36.6	86.2
2004-1	31.2	96.4	36.6	86.4
2004-2	30.9	96.0	36.3	86.0
2004-3	30.7	95.8	35.9	86.0
2004-4	30.7	95.8	35.9	85.9

Det er ikke slik at det økte utbytte bare fører til mindre samsvar for de enkelt jobber som blir kodet i tillegg med metode 2. Som vist i tabellen under, er det litt bedre samsvar blant de flere som man får kodet. Det er også noen som får *bedre* kode med metode 2.

Tabell 5-4: Effekt av to metoder for automatisk yrkeskoding. AKU 2000-2004. Koblet delutvalg.

Ord 1 \ soundex	I alt	Ikke	Samsvar
I alt	111 151	14 902	96 249
Mangler	18 150	8 678	9 472
Ikke	3 111	2 907	204
Samsvar	89 890	3 317	86 573

Ord 1 \ soundex	I alt	Ikke	Samsvar
I alt	100 %	13 %	87 %
Mangler	100 %	48 %	52 %
Ikke	100 %	93 %	7 %
Samsvar	100 %	4 %	96 %

5.2 Kombinasjon av ord

Ved å utnytte mer informasjon som ligger i teksten, enn bare isolerte ord og teksten som helhet, kunne man anta at en i større grad kunne nyttiggjøre seg det semantisk innholdet i teksten og derved får bedre klassifisering. Det neste trinnet

kan da være å studere kombinasjoner av ord. Det er da enklest å først se på kombinasjoner av 2 ord. Vi får nå to nye kriterier å ta stilling til:

- Rekkefølge, som noen ganger kan være betydningsfull: "ASS.DIR." er ikke lik "DIR.ASS.", andre ganger er det ikke: "elektrikerlærling" gir samme kode som "lærling, elektriker".
- Avstand, altså om man tillater andre ord mellom ordene i kombinasjonen.

For å starte litt enkelt samtidig som vi får større delutvalg, velger vi fri rekkefølge og fri avstand. For å få et håndterlig antall kombinasjoner og motvirke tilfeldige feil, velges ord med en viss grunnfrekvens. Videre må selve kombinasjonen også være over et visst nivå. Det viser seg at det er ganske få kombinasjoner som både er vanlige og særlig spesifikke utover de som inneholder ord som allerede er spesifikke alene. For å kunne ha utbytte av dette bør vi undersøke nærmere i en enda større datamengde.

Tabell 5-5: Effekt av forsøk med 2-ords kombinasjoner. AKU 2000-2004.

Ordet alene	Ord 2	Spes. Alene	Spes. Komb.
BARNEHAGEASSISTENT	BARN	1.000	0.998
BUSS	SJÅFØR	0.992	0.836
BUTIKKMEDARBEIDER	SALG	0.984	0.978
EKSPEDITØR	EKSPEDERE	0.851	0.894
EKSPEDITØR	KUNDER	0.859	0.894
EKSPEDITØR	SALG	0.892	0.894
FRISØR	FRISØR	0.969	0.980
HJELPEPLEIER	ELDRE	0.984	0.968
HJELPEPLEIER	OG	0.983	0.968
HJELPEPLEIER	OMSORG	0.966	0.968
HJELPEPLEIER	PASIENTER	1.000	0.968
HJELPEPLEIER	PLEIE	0.998	0.968
HJELPEPLEIER	STELL	0.985	0.968
HJELPEPLEIER	AV	0.982	0.968
KOKK	KOKK	1.000	0.926
KOKK	LAGE	0.847	0.926
KOKK	MAT	0.904	0.926
LEGE	LEGE	1.000	0.902
OMSORGSARBEIDER	PLEIE	0.927	0.816
REGNSKAPSFØRER	REGNSKAP	0.809	0.859
RENHOLDER	RENHOLD	0.998	0.976
RØRLEGGER	RØRLEGGER	1.000	0.969
SERVITØR	SERVERE	0.896	0.868
SERVITØR	SERVERING	0.841	0.868
UNGDOMSARBEIDER	BARNE	0.891	0.876
UNGDOMSARBEIDER	OG	0.904	0.876
VAKTMESTER	VAKTMESTER	1.000	0.967

5.3 Konklusjon

Det er vist at tekstdata her skiller seg både fra vanlige kategorivariabler og fra generell tekst i andre sammenhenger. Videre analyser av dette kan ha interesse utover yrkesklassifisering, siden det i mange undersøkelser oppgis tekst som klassifiseres til statistiske grupperinger. Det kan fortsatt være rom for forbedringer i manuelle og automatiske metoder innen klassifisering og i evalueringen av denne.

Det er demonstrert noen metoder for å analysere tekst både som grunnlagsviten, i forhold til klassifiseringer, og mulighetene for å utvikle automatiske kodemetoder. Arbeidskraftundersøkelsen og andre utvalgsundersøkelser inneholder betydelig mengde tekstdata som bør utnyttes videre for å øke kompetansen på området, og på sikt skaffe ny kunnskap.

6 Referanser

Luhn, H. P.: The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2):159-165, 1958

Zipf, George K.: Human Behaviour and the Principle of Least-Effort, Addison-Wesley, Cambridge MA, 1949

Knuth, D.E.: The Art of Computer Programming, Volume 3. Sorting and Searching, Reading, MA: Addison-Wesley. 1973

Bø, Tor Petter og Inger Håland: Statistisk sentralbyrå Notat 2002/24 "Dokumentasjon av arbeidskraftundersøkelsen"

Villund, Ole: Statistisk sentralbyrå notater: 2005/43, 2005/14, 2004/46, 2003/80, 2003/79, 2001/70

De sist utgitte publikasjonene i serien Notater

2005/31 T. Hægeland, L.J. Kirkebøen og O. Raaum: Skoleresultater 2004. En kartlegging av karakterer fra grunn- og videregående skoler i Norge. 89s.

2005/32 A.Rolland: Brukertilfredshetsmålinger i offentlig sektor. Utredning for Moderniseringsdepartementet og regjeringens handlingsplan for modernisering. 96s.

2005/33 K. Aasestad, A. Finstad o g K. Loe Hansen: Bruk av helsefarlige produkter i grafisk industri. 27s.

2005/34 S.W. Bogen, K. Digre, A. Hedum, T. Hægeland, T.K. Schjerven og B. Vold: Et system for statistikk omstatlig virksomhet. Forprosjektnotat. 44s.

2005/35 Kostra. Arbeidsgrupperapporter 2005. 230s.

2005/36 D. Rafat: Produksjonsopplegg for foreløpige tall i industristatistikken. 46s.

2005/37 T. Dale og B. Hole: Evaluering av elektroniske skjemaer i KOSTRA. Case: Skjema 20 - Fysisk planlegging, kulturminner, natur og nærmiljø. 55s.

2005/38 A. Sundvoll: Kirkelig tjenestestatistikk i KOSTRA-drakt. Et pilotprosjekt. 48s.

2005/39 G.I. Gundersen, B. Hoem, P. Løkkevik og D. Splide. Gjennomgang av metoder og datakilder i energiregnskapet. 50s.

2005/40 K. Loe Hansen: Bruk av helsefarlige produkter i båtbyggerbransjen. 27s.

2005/41 S. Skaare: Undersøkelsen om samvær og bidrag 2004. 67s.

2005/42 A. Haglund, A. Hedum, T. Schjerven og K.Ø. Sørensen: Offentlig sektor og BoF. 63s.

2005/43 O. Villund: Yrkesdata for selvstendig næringsdrivende. Dokumentasjonsnotat. 44s.

2005/44 O. Villund: Alder i AKU endring av definisjoner og trekkgrunnlag. 27s.

2005/45 J.I. Hamre: Estimering av fylkesfordelte og sektorfordelte tall for egenmeldt sykefravær. Dokumentasjon av metode og system, og resultater. 67s.

2005/46 A-K. Mevik: Revisjon av Strukturstatistikk for industrien. Et forslag til selektiv revisjon. 43s.

2005/47 A. Sundvoll: Utvikling av webskjema i UT-prosjektet. Dokumentasjonsrapport. 75s.

2005/48 E. Frilseth og P. Ø. Andreassen: Brukerundersøkelsen 2004. Brukernes. 64s. tilfredshet med SSBs produkter og tjenester. 64s.

2005/49 E. Rauan: Undersøking om foreldrebetaling i barnehagar, august 2005. 45s.

2005/50 A. Rolland: Brukertilfredshets-undersøkelser som offentlig styringsverktøy. 27s.

2005/51 S. Blom: Holdninger til innvandrere og innvandring 2005. 50s.

2005/52 A. Sundvoll, B. Thomassen og K. Thorsen: Balansert målstyring i Avdeling for IT og datafangst. Dokumentasjonsrapport. 35s.

2005/53 B. Castberg, P.O. Haugen, E. Knutsen og S. Myro: Økt tilgang på regnskapsdata: Konsekvenser for revisjon, tekniske løsninger og ny regnskapsstatistikk. 45s.

2005/54 A. Holmøy: Forbruksundersøkelsen 2004. Dokumentasjonsrapport. 95s.

2005/55 A. Schjalm: Flagging - Koder for dokumentasjon av revisjon. 23s.

2005/56 H. Haanæs, A. Kløvstad og J.E Wålberg: Dokumentasjon av statistikk for skogavvirkning til salg. 63s.

2006/1 S. Abonyo og T. Hagen: Tidsbruksundersøkelse - hvor lang tid bruker oppgavegiver på rapportering til kvartalsvis lønnsstatistikk. 24s.