Rasmus Munksgaard og Oskar Enghoff

Køn og metodevalg blandt samfundsvidenskabelige specialeskrivende

Feministisk teori og forskning har argumenteret for to sammenhænge mellem køn og forskningsmetoder: Kvinder benytter oftere kvalitative metoder, og køn påvirker valget af forskningsområder. Tidligere forskning baseret på fagfællebedømte publikationer understøtter disse foreslåede sammenhænge, men anerkender bias som følge af homogeniserende mekanismer såsom akademisk professionalisering og fagfællebedømmelse. Vi komplementerer disse studier gennem en analyse af de »nedre lag af akademisk produktion«, specifikt 1.103 socialvidenskabelige specialer, hvilket giver en alternativ vinkel på studiet af køn og forskningsdesign.

Vi benytter nylige innovationer indenfor digital tekstanalyse og estimerer en *structural topic model* for at modellere korpussets latente tematiske struktur. Ud fra denne model tester vi empirisk de foreslåede sammenhænge mellem køn, forskningsmetoder og forskningsområder. Vi finder, at de kvindelige specialestuderende er mere tilbøjelige til at benytte kvalitative metoder, og at nogle forskningsområder er kønnede. Topic modelling bliver demonstreret som et effektivt redskab til at analysere akademiske tekster.

Søgeord: digitale metoder, akademisk produktion, topic modelling, køn

1. Introduktion

Danske universitetsstuderende på det Samfundsvidenskabelige Fakultet ved Københavns Universitet skriver og indleverer et speciale på sidste semester af deres kandidatstudie. I specialet kan de studerende indenfor visse rammer vælge et emne, som de herefter underkaster samfundsvidenskabelig analyse på baggrund af de kompetencer, de har tilegnet sig igennem 4,5 års studier. Den studerende nyder en høj grad af frihed i sit valg af emne, og det endelige produkt må formodes at være et produkt af blandt andet akademiske interesser, kompetencer, vejledning, aktuelle samfundsproblematikker og karrieremæssige overvejelser. Specialet afleveres fysisk og indsendes til et online arkiv, hvorefter det eksamineres og bedømmes.

I dette studie benytter vi os af et arkiv af millioner af ord produceret af danske specialeskrivende og udsætter det for en *fjern* analyse inspireret af, hvad digitale humanister har kaldt *distant readings* eller *makroanalyser* (Jockers, 2013). 1.103 specialer skrevet af 1.246 studerende på Københavns Universitets samfundsvidenskabelige fakultet indsamles gennem web-crawling og -scraping og analyseres herefter via digitale tekstanalytiske metoder – specifikt en *structural topic-model*.

Med udgangspunkt i dette korpus studerer vi sammenhængen mellem køn og produktionen af videnskab, som er et emne, der har fulgt sociologien og samfundsvidenskaberne i det hele taget siden 1960'erne (Brewer, 1989). Vi belyser emnet fra en alternativ vinkel, gennem både en innovativ metodologi og et atypisk datasæt.

Vi introducerer først kort diskussionen om sammenhængen mellem køn og metodologi og gennemgår herefter den empiriske litteratur, som har bidraget til denne. På baggrund af denne præsenterer vi vores hypoteser og vidensbidrag. Vi introducerer hernæst vores metode – *topic modelling* – heuristisk, præsenterer vores data og fund og afslutter med en kortfattet diskussion samt perspektiver for videre forskning.

2. Kønnede metoder

Diskussionen om skellet mellem kvalitative og kvantitative metoder har en underliggende kønnet dimension, og satirisk foreslår Gheradi & Turner (2002) at »*real men don't collect soft data*«. Figur 1 viser, at økonomi (som er den samfundsvidenskabelige disciplin, der er mest dedikeret til kvantitativ analyse) er den eneste af de fem studieretninger i vores datasæt, som har flere mandlige end kvindelige specialeskrivere, hvorimod antropologi, en disciplin, som i udpræget grad benytter sig af kvalitative metoder, er domineret af kvinder. Betragtes populationen i samfundsvidenskabelige discipliner på Københavns Universitet således isoleret, så fremstår Gherardi & Turners forslag umiddelbart som hvilende på en korrekt intuition (om end deres ordvalg kan diskuteres). Samtidig er det dog ikke muligt at vurdere de mange potentielle

Rasmus Munksgaard

Cand.scient.soc, ph.d.-studerende, School of Criminology, University of Montreal

E-mail: rasmus.munksgaard.andersen@gmail.com

Oskar Enghoff

Ph.d.-studerende, School of Law, University of Manchester

E-mail: oskar.enghoff@gmail.com

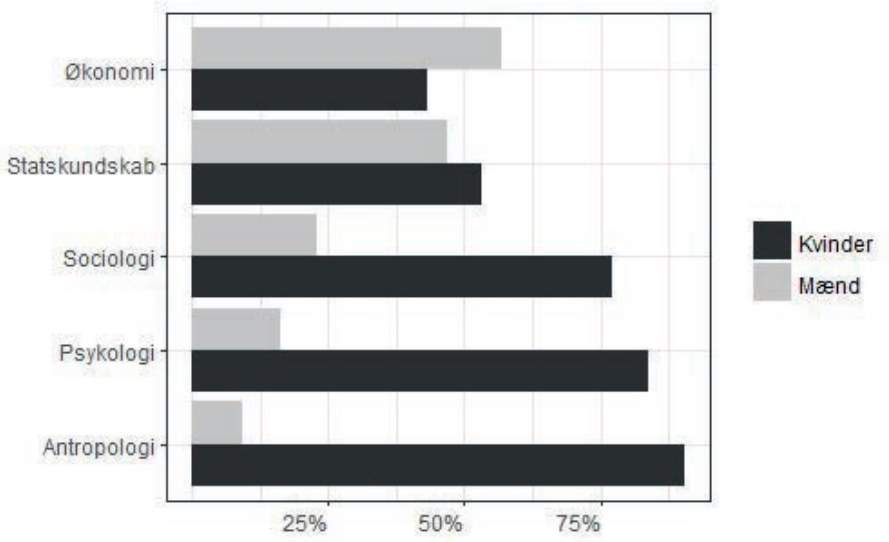
faktorer og mekanismer, der kan tænkes at producere disse udfald, på baggrund af denne figur.

Den foreslåede sammenhæng mellem metodologi og køn er særligt blevet behandlet i visse grene af feministisk teori og kritik (Hughes & Cohen, 2010; Ward & Grant, 1985), herunder særligt med henvisning til de såkaldte standpunktsteorier, der var udbredte i 1980'erne og 1990'erne (Harding, 1992). Omend den samlede litteratur ikke kan reducereres til enkelte positioner eller analyser, så er der en gennemgående tendens til at analysere applikationen af videnskabelige metoder i et kønsperspektiv. Gheradi & Turner (2002: 82) skriver eksempelvis, at der er en underliggende mening i distinktionen mellem hård og blød, kvantitativ og kvalitativ, data: Hård data er maskulin og sølet ind i maskuline praksisser, hvorimod blød data er »*soft-hearted*« og svær at tage seriøst. Følgelig bliver den kvantitative metodiske tilgang, ifølge disse forfattere, ofte taget for værende mere objektiv inden for samfundsvidenskab og videnskab generelt (Grant, Ward, & Rong, 1987: 857; Mcnamee, Willis, & Rotchford, 1990: 109; Gheradi & Turner, 2002).

Generelt udspecificeres detaljerne i dette argument om et prestigehierarki ikke i litteraturen, og en generalisering på tværs af de enkelte discipliner er problematisk. De forfattere, som har præsenteret dette argument, befinder sig i og omkring sociologien, fremfor eksempelvis økonomi og antropologi, og deres argument må derfor betragtes som vedrørende sociologi. Den prestige og validitet, som tilskrives metoder, må formodes at være anderledes indenfor eksempelvis økonomi og antropologi, som i højere grad end sociologi benytter én specifik metodologisk tilgang. Endvidere må man være opmærksom på, at prestigehierarkiet ikke nødvendigvis gør sig gældende, eller gør sig gældende på forskellige måder, i forskellige nationale sociologitraditioner (se f.eks. Abend 2006 for en sammenligning af mexicanske og amerikanske sociologer).

Oakley (1998) sidestiller dikotomien mellem kvalitative og kvantitative metoder med ideologisk kønnede repræsentationer i andre dikotomier såsom *naturlig* kontra *social*, *rationel* kontra *intuitiv* og så fremdeles. Den mest udbredte form for feministisk kritik kan ifølge Oakley (1998) her opsummeres som en kritik af validiteten af en 'maskulin' kvantitativ videnskabelig metode,

Figur 1. Fordelingen af mandlige og kvindelige specialeforfattere på tværs af institutter, baseret på forfatterskab til 1.103 specialer uploadet til Københavns Universitets database DISKURS. Figuren er baseret på det benyttede datasæt og er derfor ikke nødvendigvis repræsentativ for den samlede population af studerende



som er blind over for magtrelationer og er besat af positivisme, objektivitet og *p*-værdier (se også Cook & Fonow, 1986; Ramazanoglu, 1992; Sprague & Zimmerman, 1989). Den feministiske videnskabsteori og -kritik er dog mangfoldig og kan ikke reduceres udelukkende til en kritik af kvantitative metoder (Fonow & Cook, 2005).

Dele af den feministiske litteratur foreslår, at nogle metodologier er bedre egnede til feministiske videnskabelige dagsordner, og kvalitative metoder er blevet fremsat som »*the most appropriate way of enabling researchers to listen to and give voice to women*« (Hughes & Cohen, 2010: 190). Griffin & Phoenix (1994) foreslår endvidere, at kvalitativ metodologi kan informere forskningsagendaer med en dybdegående og fortolkende tilgang til sociale fænomener, særligt hvis den har et feministisk udgangspunkt. Ligeledes foreslår andre, at kvinders sociale positioner kan være et bedre udgangspunkt for kvalitativ forskning, eller at kvinder blot har en særlig affinitet for kvalitativ forskning, der er i overensstemmelse med stereotypiske feminine kvaliteter eller træk (Grant et al., 1987; Riger & Stephanie, 1992: 733; Grant & Ward, 1991: 211).

Der er således baggrund i litteraturen for, at der eksisterer en sammenhæng mellem køn og metodologi – uanset om man accepterer de påståede kasualforklaringer om kønnede affiniteter for metodernes applikation, metodologiernes emancipatoriske potentialer eller kønspolitiske ideologier, som

kommer til udtryk i metodologi (e.g. Oakley, 1998). I det følgende afsnit gennemgår vi den empiriske litteratur, som har beskæftiget sig med køn og metodevalg i samfundsvidenskab.

3. Køn og akademisk produktion

Ifølge Grant et al. (1987) foreslås der i den feministiske litteratur to grundlæggende sammenhænge mellem køn og metodologi indenfor samfundsvidenskab: Kvindelige forskere er mere tilbøjelige til at vælge kvalitative metoder, og kvalitative metoder er bedre egnede til studiet af kønsrelaterede forskningsagendaer. Disse påstande er ikke nødvendigvis empirisk begrundede, og litteraturen peger på, at emnet er mere komplekst.

Grant et al. (1987) tester påstanden om kvinders applikation af forskningsdesign i et studie af forfatterkøn i en stikprøve på 856 artikler fra 10 af de ledende amerikanske sociologiske tidsskrifter og finder her, at kvinder *er* mere tilbøjelige til at benytte kvalitative metoder og mindre tilbøjelige til at benytte kvantitative metoder, både i artikler om køn og om andre emner. Specifikt finder Grant et al. (1987: 862), at i artikler, som omhandler køn, benytter 79 % af kvinder og 91 % af mænd kvantitative metoder, mens i artikler, der ikke omhandler køn, benytter 71 % af mænd og 63 % af kvinder kvantitative metoder, og konkluderer, at »*findings support the existence of systematic links between gender and methods*«. McNamee et al. (1990) finder et lignende mønster fra 1960 til 1985 på tværs af fire førende amerikanske sociologiske tidsskrifter, hvor 52 % af artikler med en kvindelig forfatter og 65 % med en mandlig forfatter, benytter kvantitative metoder. Endeligt observerer Erola et al. (2015) ligeledes signifikante assoociationer mellem kvindeligt forfatterskab og metode i et komparativt studie af tre skandinaviske tidsskrifter i sociologi. Igen er kvinder mere tilbøjelige til at benytte kvalitative metoder (samt publicere nationalt) og mindre tilbøjelige til at benytte kvantitative metoder (samt publicere internationalt).

I sociologi er der således en tendens til kønsforskelle i metoder benyttet i videnskabelige publikationer, og denne kan spores både nationalt, internationalt og historisk. Uden for sociologiske tidsskrifter finder Plowman & Smith (2011), at der i organisationsforskning er en signifikant tilbøjelighed hos kvinder til at producere kvalitative studier.

Dunn & Waller (2000) tester påstanden om kvalitative metoders dominans i kønsforskning (særligt den, som har et feministisk udgangspunkt) på et datasæt bestående af 1.826 kønsrelaterede forskningsartikler i 15 amerikanske sociologiske tidsskrifter mellem 1984 og 1993 og finder, at der er et positivt (omend svagt) forhold mellem feministisk orienteret forskning og brugen af kvalitative metodologier. Omvendt så finder Dunn & Waller (2000), at 71 % af feministisk orienterede artikler i disse tidsskrifter benytter kvantitative metodologier, og bekræfter således Hughes & Cohens (2010) argument: *Feminists*

really do count, og feministiske agendaer forudsætter ikke rendyrket kvalitativ metode, tværtimod.

Både Dunn & Waller (2000) og Grant et al. (1987) observerer endvidere, at kvantitative analyser underbygger de fleste forskningsartikler i prestigefyldte amerikanske sociologitidsskrifter. Dette kan ses i forlængelse af den tidligere omtalte diskussion om metodologiske meritter, hvori kvantitativ metode – ifølge visse forskere, herunder visse feminister – generelt betragtes som værende mere prestigefyldt og troværdig indenfor samfundsvidenskaberne (Gheradi & Turner, 2002; Grant et al., 1987; Griffin & Phoenix, 1994). I forbindelse med de undersøgelser, som er foretaget af både Dunn & Waller (2000), Grant et al. (1987) og Plowman & Smith (2011), må der dog tages det forbehold, at de tidsskrifter, som udgør datagrundlaget, netop er særligt prestigefyldte (ikke mindst i amerikansk sammenhæng) og dermed indebærer et særligt selektionsbias. Således noterer de pågældende studier, at deres fund sandsynligvis ville være anderledes, hvis datagrundlaget bestod af interdisciplinære eller mindre tidsskrifter, igen fordi de studerede højtrangerede journaler i overvejende grad er kvantitativt orienterede.

Selve dette forhold kan dog siges at underbygge det argument, som præsenteres i dele af litteraturen. Altså at kvantitativ forskning generelt, i hvert fald i amerikansk sociologi, betragtes som havende bedre krav på objektivitet og højere videnskabelig status, eftersom en gennemgående tendens i studier af publikationer i førende amerikanske sociologiske tidsskrifter viser, at kvantitative studier er mere prævalente end kvalitative (Dunn & Waller, 2000; Grant et al. 1987; McNamee et al., 1990; Abend 2006). I forbindelse med dette argument så må det understreges, at disse studier netop er historisk og ikke mindst nationalt indlejrede. Vi er ikke bekendt med nyere empiriske studier udover Erola et al. (2015), som på baggrund af et specifikt skandinavisk datasæt har efterprøvet og fundet frem til beslægtede kønnede tendenser. Det siger sig selv, at opfattelser af videnskabelige metoders objektivitet og værdi, såvel i sociologien som i samfundsvidenskaberne generelt, har udviklet sig siden 1960'erne og også til stadighed udviser en del national variation (Abend 2006).

Et andet aspekt af denne diskussion er eventuelle kønsforskelle i forskeres epistemologiske standpunkter, hvilket har været et omdrejningspunkt i dele af den feministiske kritik (Riger & Stephanie, 1992). Et survey af Goldenberg & Grigel (1991) fandt dog, at køn havde meget lidt forklaringskraft i forhold til forskeres forståelse af 'god videnskab', samt at variationen mellem køn var større i naturvidenskab end samfundsvidenskab. I et andet studie af *personlig epistemologi* målte Unger, Draper, & Pendergrass (1986) universitetsstuderendes epistemologi på en skala fra konstruktivisme til logisk positivisme og fandt, at når der kontrolleredes for faktorer såsom kursusdeltagelse (dvs. hvilke uddannelsesretninger de studerende havde valgt), så mistede køn forklaringskraft.

I den empiriske litteratur finder vi således belæg for en sammenhæng mellem køn og metodevalg, omend der er tale om gradsforskelle fremfor absolutte afvigelser eller skel: I akademiske publikationer, såvel amerikanske som skandinaviske, er mænd mere tilbøjelige til at benytte kvantitative metoder, mens kvinder er mere tilbøjelige til at benytte kvalitative metoder. I studier af epistemologiske standpunkter observeres ligeledes kønsforskelle, men disse forskelle varierer på tværs af videnskabelige domæner eller forsvinder, når passende kontrolvariable inkluderes.

4. Køn i de nedre lag af akademisk produktion

Kvantitative studier af videnskabelig produktion, køn og metodevalg viser, at der er en sammenhæng mellem køn og metodevalg – men denne er kun studeret i begrænsede sfærer såsom prestigiøse tidsskrifter. I dette studie bygger vi på denne eksisterende empiriske litteratur omkring forskning, metodologi og køn, men vi gør dette gennem et studie af et specifikt domæne i produktionen af samfundsvidenskab, som vi kalder de 'nedre lag' af akademisk produktion. Med dette mener vi akademisk produktion i de tidligste stadier af en eventuel videnskabelige karriere, før udvælgelsen til ph.d.-studier og før udgivelsen af de tidsskriftartikler, som tidligere studier har benyttet som empirisk grundlag. Vores undersøgelse belyser desuden det hidtil uudforskede spørgsmål om, hvorvidt resultaterne fra den eksisterende internationale forskning kan genfindes i en dansk kontekst.

Vi benytter topic modelling, en metodisk tilgang til tekstanalyse, som vi uddyber nedenfor, hvilket lader os estimere *proportionen* af specifikke metodiske og tematiske elementer i specialerne. Således måler vi ikke blot, om fx kvinder oftere bruger kvalitative metoder, men om deres specialer indeholder *mere* kvalitativ metode. Yderligere så lader topic modelling os studere indholdet af specialerne induktivt og uden forudindtagelser baseret på fx titler, abstracts eller institutionelle tilhørsforhold. Dermed kan vi bl.a. tage højde for, at et speciale kan indeholde flere forskellige metoder og eventuelt kan misrepræsentere sig selv (fx ved at en specifik metode er mere central, end abstractet giver udtryk for). Brugen af topic modelling lader os endeligt nuancere eksisterende tilgange, som har afhængt af binære opdelinger i kvalitativ og kvantitativ metode, hvilket ikke nødvendigvis kan imødegå kompleksiteten i feministiske kald for kvalitativ forskning (se særligt Cook & Fonow, 1986). Vi kommer nærmere ind på dette nedenfor.

Begrænset af den tilgængelige data og valgte metode kan dette studie ikke identificere eller problematisere de bagvedliggende kausaliteter eller faktorer, som den mere teoretisk orienterede del af literaturen foreslår som forklaringer på kønsforskelle i metodevalg. Eksempelvis kan vi hverken be- eller afkræfte, at kvalitative metoder i sig selv er i overensstemmelse med kvinders præferencer og kvaliteter (Grant et al. 1987), eller at kvantitative metoder er udtryk for maskulin dominans (Oakley, 1998). Vores studie er således empi-

risk og deskriptivt: En replikation af en sammenhæng observeret i tidligere forskning, men indenfor et nyt domæne og med en alternativ metodologi. Med udgangspunkt i den eksisterende litteratur fremsætter vi de følgende hypoteser.

H1 Kvinders specialer indeholder mere kvalitativ metodologi end mænds specialer
H2 Mænds specialer indeholder mere kvantitativ metodologi end kvinders specialer
H3 Kvinders specialer indeholder mere indhold relateret til kønsproblematikker end mænds specialer

Endvidere forventer vi, at eventuelle korrelationer mellem akademisk produktion og køn vil variere på tværs af akademiske discipliner, og at visse forskningsområder (f.eks. kønsforskning) vil være mere prævalente inden for nogle discipliner end andre (f.eks. sociologi kontra økonomi). Indirekte tager vi derfor også højde for disciplinernes interne orden og prioritering af metodologier og tematikker. Vi sammenligner således tematisk indhold indenfor discipliner og ikke på tværs af discipliner.

5. Topic modelling

Vi bruger digital tekstanalyse til at estimere og analysere det tematiske indhold af vores korpus af kandidatspecialer. Via *topic modelling* kortlægger og udleder vi både det tematiske indhold og den konkrete fordeling af dette på tværs af specialerne. Denne metode er en statistisk tilgang til tekstanalyse, som er blevet succesfuldt anvendt i analyser af en lang række forskellige datasæt såsom romaner (Jockers, 2013; Jockers & Mimno, 2013), videnskabelige artikler (Blei & Lafferty, 2007; Blei et al., 2003; Hall, Jurafsky, & Manning, 2008), fritekstsvar fra spørgeskemaundersøgelser (Roberts et al., 2014), online politiske kommentarer (Levy & Franklin, 2014), indhold på online debatfora (Munksgaard & Demant, 2016; Törnberg & Törnberg, 2016), mikroblogs (Ramage, Dumais, & Liebling, 2010) og nyhedsartikler (DiMaggio, Nag, & Blei, 2013). I disse og andre undersøgelser har topic modelling vist sig effektivt i at udlede en række tematikker fra datasættet og i at 'kode' dokumenterne ud fra disse – på en måde, der ofte stemmer overens med manuelle, kvalitative kodninger af det samme datasæt (Baumer et al., 2017; Chang et al., 2009).

Topic modelling åbner op for en analytisk tilgang, der kan beskrives som *mixed-methods* eller 'kvali-kvantitativ' (Venturini & Latour, 2010). Det skyldes, at topic modelling både indebærer en bred vifte af muligheder for kvantificering og samtidig kan bidrage med en tematisk kodning samt identifikation af relevante dokumenter – eksempelvis ved at producere en liste over dokumenter, som indeholder en høj proportion af et særligt tema (topic) og derfor kan være relevante at underkaste en dybere læsning. Således kan forskeren

uhindret skifte mellem *fjerne* (hovedsageligt kvantitative) og *nære* (hovedsageligt kvalitative) læsninger af korpusset og derved potentielt undgå nogle af de metodiske faldgruber, som er forbundet med kun at vælge ét af disse to perspektiver (Jockers, 2013). Denne fleksible analytiske tilgang understøttes yderligere af de vidtrækkende muligheder for datavisualisering, som er indbygget i de fleste software-implementeringer af topic modelling – eksempelvis ordskyer, netværksdiagrammer og diverse visualiseringer af effekten af baggrundsvariable (se eksempelvis Freeman et al., 2015; Roberts, Stewart, & Tingley, 2014).

I det følgende afsnit introducerer vi topic modelling heuristisk og forsøger at kontekstualisere metoden, i forhold til hvordan tekster behandles kvalitativt indenfor samfundsvidenskaberne. En videre redegørelse for de matematiske aspekter af topic modelling er uden for denne artikels rækkevidde, men for en detaljeret indførsel i det statistiske grundlag anbefaler vi Blei & Lafferty (2007); Blei et al. (2003) og Roberts et al. (2014). For andre konceptuelle introduktioner anbefaler vi Grimmer & Stewart (2013); Jockers (2013); Mohr & Bogdanov (2013) og særligt Blei (2012).

5.1. Topic modelling – en heuristisk introduktion

Mohr & Bogdanov (2013) opsummerer topic modelling som »*an automated procedure for coding the content of a corpus of texts (including very large corpora) into a set of substantively meaningful coding categories called 'topics'*«. Dermed kan topic modelling til dels forstås som en automatiseret version af en kvalitativ analysestrategi (tematisk kodning af tekst), men der er imidlertid en række vigtige forskelle. Den mest åbenlyse forskel er, at topic modelling er automatiseret og derfor gør det muligt at arbejde med langt større datasæt end sædvanligt. I dette studie analyserer vi 1.103 specialer, hvilket svarer til ca. 46 millioner ord, eller 117.000 sider á 2.400 tegn.[1] En anden vigtig forskel er, at topic modelling er en *ikke-superviseret* modellering, hvilket betyder, at de analytiske kategorier eller koder ikke er defineret *a priori* (Grimmer & Stewart, 2013: 281). De bliver i stedet udledt direkte af korpussets indhold. Dermed repræsenterer topic modelling en induktiv tilgang til kodning af tekst, som begrænser mulighederne for, at eventuelle forskerbias kommer til udtryk i kodningen og selve kodningsapparatet. Baumer et al. (2017) har påpeget, at topic modelling har flere fællestræk med den iterative og datanære tilgang til kodning, der kendertegner analysestrategier baseret på *grounded theory*, og dermed repræsenterer topic modelling ikke nødvendigvis et radikalt brud med eksisterende sociologiske tilgange.

Vi benytter en *structural topic model* (STM) – en videreudvikling af *correlated topic modelling* (CTM) og *latent dirichlet allocation* (LDA) (Blei & Lafferty, 2007; Blei et al., 2003; Roberts, Stewart, Tingley, Airoldi, & others, 2013) – som alle falder under genren *unsupervised mixed-membership models* (se Grimmer & Stewart, 2013 for en bred introduktion til kvantitativ tekstanalyse). Under

disse modeller formoder vi, at der eksisterer et antal latente temaer (herefter 'topics') i et korpus. Disse topics er *sandsynlighedsfordelinger over et vokabular*. Med andre ord formoder vi, at der inden for hvert topic er nogle termer, der er mere sandsynlige end andre. Et topic, hvorunder termer som økonomi, *gæld*, *nationalprodukt* og *BNP* optræder ofte, kunne eksempelvis formodes at handle om nationaløkonomi. Fordi topics betragtes som sandsynlighedsfordelinger (distributioner) og ikke gensidigt udelukkende kategorier, kan termer optræde med høj sandsynlighed i flere topics. Vi kunne eksempelvis formode, at *gæld* også optrådte i et andet topic sammen med termer som *husholdning*, *kreditkort* og lignende, som ville være typiske for et topic, der handler om privatøkonomi. DiMaggio et al. (2013) argumenterer på baggrund af dette, at topic models opererer under antagelsen om *polysemi* – at en term kan have multiple betydninger, der varierer i forhold til den kontekst, det bruges i. Et andet centralt træk ved topic modelling er, at dokumenter er *distributioner over topics*. Altså indeholder hvert dokument en særlig distribution over alle de topics, som findes i hele korpusset. I en diskursanalytisk forståelsesramme argumenterer Munksgaard & Demant (2016) for, at topic-modeller derfor antager et af diskursanalysens kernepræmisser: Interdiskursivitet, dvs. anerkendelsen af, at tekster udgøres af flere diskurser eller temaer (Fairclough, 1992).

Topic modelling er en automatiseret procedure, som er 'sprogblind' i den forstand, at hver term i korpussets samlede vokabular blot optræder som et identificerende tal og ikke som en lingvistisk enhed med grammatiske og semantiske egenskaber. Desuden er det vigtigt at bemærke, at topic-modeller baserer sig på *bag-of-words*-antagelsen – at dokumenter betragtes som 'poser af ord', hvori rækkefølgen af ordene er irrelevant. Topic-modeller forkaster dermed sætningskonstruktioner og syntaktiske relationer. Konkret giver dette sig udtryk i, at korpusset repræsenteres som en dokument-term-matrice, hvor hvert dokument er en vektor af frekvenser for hver term i korpussets vokabular. På trods af at disse simplificeringer kan betragtes som kontroversielle, har de vist sig at kunne danne grundlag for effektive modeller inden for tematisk tekstanalyse, og således har topic modelling gjort sit indtog i en række discipliner ved at udlede kvantificerbare topics, der er genkendelige som kohærente temaer, og som ofte stemmer overens med menneskelige tolkninger (Mohr & Bogdanov, 2013).

5.2. Topic modelling – detaljerne

Topic modelling tilhører en bredere genre af *generative probabilistiske modeller*. Disse er modeller, hvor datamaterialet modelleres som resultatet af en hypotetisk 'generativ proces', og hvor formålet er at bruge data til at arbejde sig baglæns gennem denne proces og dermed estimere dens parametre (Blei, 2012). Den generative proces er en hypotetisk simplificering af den reelle proces, hvorved data er blevet genereret (i dette tilfælde en specialeskrivendes ar-

bejdsproces). I topic modelling udspringer den generative proces af følgende grundlæggende antagelser:

1. Et korpus indeholder et fastlagt antal topics (dvs. temaer)
2. Hvert dokument i korpusset indeholder en blanding af disse topics
3. Hvert ord i et dokument tilhører ét bestemt topic

På baggrund af disse antagelser opstilles den følgende generative proces, som i denne undersøgelse repræsenterer en hypotetisk simplificering af produktionen af et speciale:

1. For hvert dokument i korpusset vælges en distribution over korpussets topics
2. Hvert dokument genereres ved at tilføje ét ord ad gangen således:
   1. Vælg et topic fra dokumentets distribution over topics
   2. Vælg en term fra det valgte topics distribution over vokabulariet

Denne proces gentages for alle dokumenter i korpusset. De løbende 'valg' af topics og termer er ikke tilfældige, men er baseret på sandsynlighedsfordelinger, som estimeres af modellen. Via den generative proces konceptualiseres korpusset som genereret ud fra en række latente topics, som vi ikke kan observere direkte. Formålet med topic modelling er at arbejde sig baglæns gennem den generative proces, således at vi kan udlede det uobserverede/latente (topics) fra det observerede/manifeste (ordene i dokumenterne) (Blei, 2012). Det sker gennem en iterativ proces, hvor modellen estimerer de sandsynlighedsfordelinger over termer og topics på baggrund af den observerede forekomst af ord inden for dokumenterne (bemærk, at vi bruger 'term' i en leksikal forstand – dvs. en term i et vokabular – mens vi bruger 'ord' til at betegne specifikke forekomster af termer – dvs. et ord i dokument). Disse sandsynlighedsfordelinger kan tolkes som proportioner af individuelle termer og topics, og dermed giver modellens endelige output et estimat af, i hvor høj grad hvert dokument indeholder hvert topic, og i hvor høj grad hvert topic indeholder hver term.

5.3. K og fortolkninger

Topic-modellens automatiserede og ikke-superviserede karakter indskrænker som nævnt forskerens muligheder for at påvirke resultatet med deres eventuelle forudgående antagelser om data (Norris, 1997). Der er imidlertid en række justerbare parametre, som kan påvirke resultatet, ikke mindst *K* – antallet af topics, hvilket skal vælges forud for estimeringen af modellen. Det er typisk ikke meningsfuldt at tale om en 'sand' værdi af *K* for et givent korpus, da topic-modellens antagelse om et fast antal stabile, latente emner er en forsimpling i de fleste korpora. Derfor er det almen praksis at udregne

adskillige modeller med forskellige værdier af K og evaluere dem post-hoc på baggrund af enten kvantitative eller kvalitative kriterier. Idet eksisterende kvantitative mål for 'topic-kvalitet' som regel ikke er ensbetydende med topics, som er velegnede til den tiltænkte analyse (Chang et al., 2009), udvælges *K* typisk via en heuristisk og kvalitativ *trial and error*-proces (DiMaggio et al., 2013; Lucas et al., 2015). Her er en god model ensbetydende med en model, der kan bruges i den tiltænkte analyse, og der tages således højde for den ønskede tematiske granularitet, dvs. at de resulterende topics hverken er for specifikke eller for generelle, samtidig med at der tages højde for, at topics er analytisk brugbare (Jockers, 2013: kap. 8). Dermed er valget af *K* udtryk for den enkelte forskers erkendelsesinteresse og er således potentielt genstand for bias. I vores erfaring er de fortolkningsmæssige forskelle mellem modeller med forskellige værdier af *K* dog som regel minimale – forskellene består primært i, hvor specifikke eller generelle de enkelte topics er.

Topic modelling indebærer ikke en *a priori* angivelse af indholdet af topics. Således er den første opgave efter estimeringen af en model at fortolke de topics, som modellen har produceret (Grimmer & Stewart, 2013, 286). Med denne tolkning følger ligeledes en kvalitativ validering af modellen, som oftest vil lede til identifikationen af et antal tematisk inkohærente eller decideret *nonsensical* topics, mens andre topics vil have en mere klar tematisk struktur (DiMaggio et al., 2013: 582; Jockers, 2013: 129). Denne kvalitative validering og tolkning foregår ad to veje, som er gensidigt understøttende: Analytikeren studerer de termer, som er associeret med det pågældende topic, og læser samtidig *eksemplariske dokumenter* (Jockers, 2013: kap. 8; Lucas et al., 2015; Roberts et al., 2014). Eksemplariske dokumenter er de dokumenter, som indeholder en høj proportion af det pågældende topic, og analytikeren søger på baggrund af disse at etablere en konsistent og kohærent tematisk fortolkning af hvert topic.

Topic modelling er oprindeligt designet med henblik på informationssøgning i digitale arkiver (Blei, 2012). Selvom metoden efterfølgende har fundet anvendelse inden for samfundsvidenskaberne, er der endnu ikke en etableret teoretisk definition af, hvad et topic er udtryk for (Törnberg & Törnberg, 2016). Empirisk er et topic en distribution over et vokabular; et udtryk for en form for sprogligt *cluster*. Hvad de er udtryk for teoretisk er imidlertid åbent for diskussion: Digitale humanister og historikere har behandlet dem som *temaer* (Jockers, 2013; Jockers & Mimno, 2013), diskursanalytikere har argumenteret for, at de kan repræsentere *diskurser* i et Fairclough'sk begrebsapparat (Munksgaard & Demant, 2016; Törnberg & Törnberg, 2016), kulturel sociologi har forsøgt at bygge bro til begreber inden for *frame analysis* (DiMaggio et al., 2013), og topics er ligeledes søgt valideret gennem *grounded theory* (Baumer et al., 2017). Vi vælger en simpel tilgang til kodning af topics og forholder os til dem som temaer, hvilket der er tradition for blandt digitale humanister (se særligt Jockers, 2013; Jockers & Mimno, 2013). Vi vurderer, at denne tilgang

er mest passende, idet den litteratur, vi bygger videre på, arbejder med en relativt simpel og ateoretisk kategorisering, og fordi vi ikke har en ekspertise i forhold til indholdet af undervisningen på de enkelte institutter (Quinn et al., 2010). Vi betragter således topics som *temaer, der udtrykker et fælles koherent og stabilt indhold på tværs af tekster*. I vores datamateriale kan disse temaer forventes at repræsentere samfundsvidenskabelige forskningsområder, metoder og teorier.

6. Data og metode

I de følgende afsnit detaljerer vi vores dataindsamling og -bearbejdning, samt estimeringen af vores topic-model. Al databehandling blev udført i R (R Core Team, 2017) og R-pakkerne *tm*, *textcat*, *plyr*, *stm*, *ggplot2*, *magrittr*, *stringr*, *dplyr*, *stmBrowser*, *reshape2*, *httr*, *tibble* og *rvest* benyttedes til forskellige dele af dataindsamling, -bearbejdning og -analyse (Bache & Wickham, 2014; Feinerer et al., 2013; Feinerer, Hornik, & Meyer, 2008; Freeman et al., 2015; Roberts et al., 2014; Wickham, 2007, 2009, 2016a, 2016b, 2017; Wickham & Francois, 2016; Wickham, Francois, & Müller, 2016).

6.1. Data

Vores data stammer fra det åbne online arkiv DISKURS under Det Kongelige Bibliotek,[2] hvortil specialer fra Det Samfundsvidenskabelige Fakultet på Københavns Universitet uploades. Vi konstruerede en specifik URL, som returnerede links til alle uploadede specialer fra følgende samfundsvidenskabelige institutter: økonomi, sociologi, antropologi, statskundskab og psykologi. For hvert speciale blev den vedhæftede PDF-fil samt bibliografisk metadata (forfatternavn, institut, dato, titel, nøgleord) indsamlet. Denne hurtige og økonomiske indsamling af data falder under genren *web-o-metrics*, specifikt *web-crawling* (Thelwall, Vaughan, & Björneborn, 2005).

Tabel 1: Specialer pr. institut og køn

Institut	Kvinder	Mænd)	Total
Antropologi	72	8	80
Psykologi	243	48	291
Sociologi	103	32	135
Statskundskab	163	156	319
Økonomi	120	158	278
Total	701	402	1103

1.890 specialer inklusive metadata blev downloadet den 13. marts 2017. PDF-filerne blev indlæst ved hjælp af pakken *tm* (Feinerer et al., 2008). Vi frasorterede 165 specialer, som manglede metadata eller indeholdt fejl i PDF-filerne. Specialer på andre sprog end dansk,[3] klassificeret ved hjælp af pakken *textcat*

(Feinerer et al., 2013), blev ligeledes frasorteret. *textcat* var dog ikke i stand til korrekt at klassificere specialer på norsk, hvorfor vi benyttede en tentativ topic-model til at identificere og fjerne disse. I alt frasorterede vi 621 specialer på andre sprog end dansk. Endeligt frasorterede vi et fåtal af specialer, hvortil vi ikke kunne tilskrive køn (se nedenfor), samt specialer forfattet af grupper med blandet kønsfordeling, henholdsvis 9 og 58 specialer. Sidstnævnte blev frasorteret, på trods af at de kunne levere et komparativt perspektiv, da antallet vurderedes for småt til at kunne komme med stabile resultater. Det endelige datasæts fordeling på institutter og køn er vist i tabel 1, mens fordelingen på årstal er vist i tabel 2.

Tabel 2: Specialer pr. årstal

År	Antal specialer
2001	3
2002	0
2003	2
2004	1
2005	0
2006	3
2007	1
2008	1
2009	1
2010	5
2011	44
2012	179
2013	262
2014	268
2015	246
2016	87
Total	1103

En fordel ved at benytte specialer fremfor forskningsartikler (som i tidligere studier) er, at disse hverken er udsat for den samme grad af selektionsbias som akademiske publikationer, hvor redaktører og fagfæller har indflydelse på publikationen, eller den samme grad af professionalisering/homogenisering via etablerede praksisser i professionelle forskningsmiljøer (Dunn & Waller, 2000; McNamee et al., 1990). Omvendt omfatter vores datasæt kun ét institut per disciplin, hvor tidligere studier af tidskriftartikler favner adskillige institutter fra flere lande. Således får vi et alternativt empirisk grundlag for en analyse af relationen mellem køn og tematisk indhold. Vores datasæt kan betragtes som mere repræsentativt for eventuelle kønnede prædispositioner i akademisk produktion, idet kvinder er underrepræsenterede i de højere lag af akademisk produktion både med hensyn til ansættelse og publikation,

hvorimod vores datasæt består af ét produkt per bestået studerende uanset køn (Grant & Ward, 1991; Plowman & Smith, 2011). Det anvendte datasæt er endvidere kendetegnet ved at være 'naturligt' forekommende, omfangsrigt, forskelligartet og let tilgængeligt, hvilket gør det attraktivt i forhold til andre datakilder, der kan belyse køn i de nedre lag af akademisk produktion – fx eksamensopgaver eller løbende afleveringer før specialeskrivningen.

En udfordring, som de tidligere beskrevne studier af køn og akademisk produktion har haft, er klassificeringen af forfatternes køn. Dunn & Waller (2000) beskriver hvorledes de kodede baseret på fornavnets udbredelse blandt kønnene, og Plowman & Smith (2011) benyttede internetsøgninger i tilfælde af tvivl. For at klassificere forfatternes køn benyttede vi Nordisk Forskningsinstituts database over frekvensen af danske fornavne (Nordisk Forskningsinstitut, 2017), og kodede køn på baggrund af navnenes rang på listerne over hhv. mandlige og kvindelige navne. Vi anerkender, at fornavne *ikke* er en optimal indikator for køn, eftersom nogle navne gives til begge køn. Kim, for eksempel, koder vi som et mandligt navn baseret på sandsynlighed, omend 308 af 32.519 personer med fornavnet Kim i databasen er kvinder. De tidligere empiriske studier, vi bygger videre på, har alle benyttet en binær opdeling i mænd og kvinder. Dette har som konsekvens, at ikke-binære kønsidentiteter usynliggøres (Westbrook & Saperstein, 2015), og vi anerkender dette som en metodisk begrænsning, hvori visse kønsidentiteter fejlkodes. Vi diskuterer senere, hvorledes videre forskning kan tage højde for denne udfordring (se afsnit 8).

6.2. Estimering af en structural topic-model

I dette afsnit redegør vi yderligere for forbehandlingen af datasættet og specificerer derefter topic-modellen. Vi filtrerede og behandlede indholdet af de 1.103 udvalgte specialer ad flere omgange. Først fjernede vi alle tal og al tegnsætning, konverterede den tilbageværende tekst til små bogstaver og reducerede alle ord til deres ordstammer – det vil sige, at f.eks. *danse*, *danser* og *dansen* alle bliver til *dans*. Dermed forkaster vi grammatiske og syntaktiske nuancer, men øger til gengæld muligheden for at spore termer på tværs af forskellige bøjninger og brugsscenarier. Derefter fjernede vi alle termer på under tre bogstaver, dels for at tage højde for fejl i tekstkonversionen og dels for at reducere det samlede datasæt ved at fjerne meget korte termer uden tematisk vægt. Af sidstnævnte årsag fjernede vi også alle forekomster af såkaldte 'stopord', dvs. funktionelle termer uden tematisk vægt såsom *denne*, *hvad* og *før*. Denne 'preprocessing' er typisk for brugen af topic models (Denny & Spirling, 2017).

Efter denne filtrering af specialernes indhold konverterede vi hvert speciale til en vektor af frekvenser for hver term i korpussets samlede vokabular og kombinerede disse vektorer i en dokument-term-matrice. Som en afsluttende filtrering af dokumenternes indhold fjernede vi alle termer (kolonner i matricen), som optræder i færre end 30 individuelle specialer. Dette resulterede

i en reduktion af det samlede vokabular fra 429.694 til 18.482 termer, hvilket betyder, at modellen bliver væsentligt mindre tids- og ressourcekrævende at estimere, og at sporadiske stavefejl og tekstkonversionsfejl frasorteres yderligere. I forhold til modellens kvalitative indhold er konsekvenserne af denne afsluttende filtrering mere komplekse. En højere tærskel (antal specialer) betyder, at man fjerner mere 'støj', og at modellens topics dermed bliver mere overordnede, til en grad, hvor de kan blive ubrugelige i analysen. En lavere tærskel betyder omvendt, at modellens topics bliver mere specialiserede og nuancerede, med fare for at de bliver svære at tolke eller decideret inkohærente. Derfor baserede vi valget af tærsklen på en iterativ trial-and-error-proces, hvor vi tilstræbte en fornuftig balance mellem disse ekstremer. Alle skridtene i den beskrevne forbehandling af teksterne er typiske inden for brugen af topic modelling i socialvidenskabelige eller humanistiske analyser, herunder en trial-and-error-tilgang hvor modellens analytiske brugbarhed vægtes højest (Grimmer & Stewart, 2013; Jockers, 2013; Lucas et al., 2015).

Dokument-term-matricen og det filtrerede vokabular blev endelig brugt som input til selve estimeringen af modellen. Vi estimerede og sammenlignede adskillige modeller med forskellige værdier af K og valgte til sidst en model med $K = 50$, da denne parametrisering producerede topics med en passende granularitet, kohærens og relevans for den tiltænkte analyse (se afsnit 5.3).

7. Analyse

I de følgende afsnit præsenterer vi vores fund. Resultater fra topic-modeller er ofte omfangsrige, og med 50 topics fordelt over fem institutter er det udenfor artiklens omfang at beskrive disse i detalje. Vi nøjes derfor med en kort redegørelse for den overordnede tematiske struktur i datasættet og tester derefter de tidligere præsenterede hypoteser.

7.1. En latent tematisk struktur i samfundvidenskabelige specialer

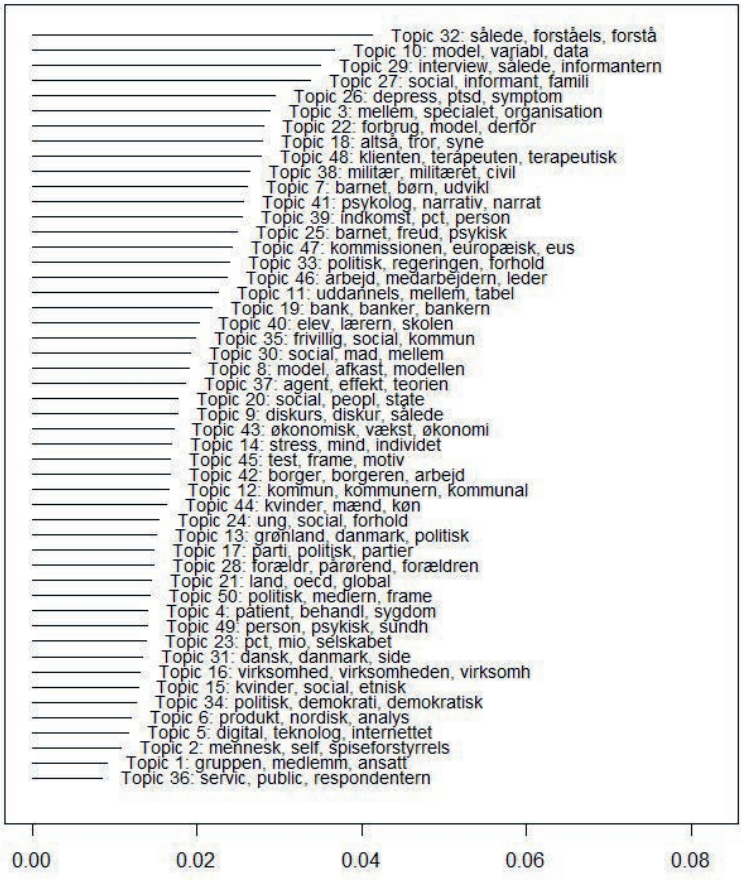
Vores model indeholder 50 topics, og hvert dokument indeholder en proportion (θ) af hvert topic. Udregnes den gennemsnitlige værdi af θ for hvert topic, får vi således den gennemsnitlige proportion af hvert topic i hele korpusset. Figur 2 viser den tematiske struktur i korpusset ved at plotte disse gennemsnitlige proportioner sammen med de tre mest sandsynlige termer for hvert topic.

Med udgangspunkt i litteraturen om anvendelse af topic modelling inden for humaniora og socialvidenskab har vi valgt en kvalitativ og heuristisk tilgang til fortolkning og validering (se afsnit 5.3) fremfor at anvende diverse kvantitative valideringsmål (fx Mimno et al., 2011), som har været genstand for kritik på grund af deres manglende evner til at indfange en models 'tolkbarhed' og analytiske brugbarhed (Chang et al., 2009; Jockers, 2013). Med udgangspunkt i den samme litteratur tillader vi os desuden at acceptere og ig-

norere enkelte inkohærente topics, da disse er et tilbagevendende fænomen inden for topic modelling, som kan være et udtryk for, at modellen samler tematisk 'støj' (dvs. diverse sporadisk indhold, som ikke kan konsolideres til individuelle topics med den valgte granularitet) og ikke har konsekvenser for validiteten af de resterende topics (DiMaggio et al., 2013: 582; Jockers, 2013: 128-30). Vi har som nævnt anvendt samme heuristiske tilgang til valget af K (og den resulterende tematiske granularitet), og valget af *K = 50* er således baseret på en *trial and error*-proces, hvor vi løbende evaluerede modellens tolkbarhed og analytiske anvendelighed ved forskellige værdier af K.

Forventeligt er en række temaer relateret til samfundsvidenskabelige metoder, herunder **10** (*model, variabl, data*) og **29** (*interview, sålede, informantern*). Ligeledes finder vi også en række samfundsvidenskabelige forskningsområder såsom sundhed og sygdom (**26**, **49** og **4**), kommunalpolitik (**12**), uddannelse (**40**, **11**) og europæisk politik (**47**). Disse er traditionelle emner for sam-

Figur 2. Forventet proportion af topics



fundsvidenskabelige specialer, men vi ser også nutidige emner såsom teknologi og internet (**5**).

Vi undlader en komplet udredning af den tematiske struktur, da dette ikke er relevant for vores analyse. Vi har imidlertid gjort en interaktiv visualisering af alle topics tilgængelig online,[4] ved hjælp af *stmBrowser*-pakken til R (Freeman et al., 2015). Læseren har her mulighed for at udforske den latente tematiske struktur i korpusset på tværs af institutter og køn med mulighed for at browse specialetitler. Det tematiske indhold i specialer er i høj grad præget af det institut, hvor specialet er skrevet, som vi detaljerer senere. For at give et indblik i, hvilke topics der er mest udbredte for hver disciplin, og lade læseren vurderes modellens 'face validity' (Quinn et al., 2010: 216) har vi vedlagt en tabel over de 10 mest prævalente topics per institut i artiklens appendiks.

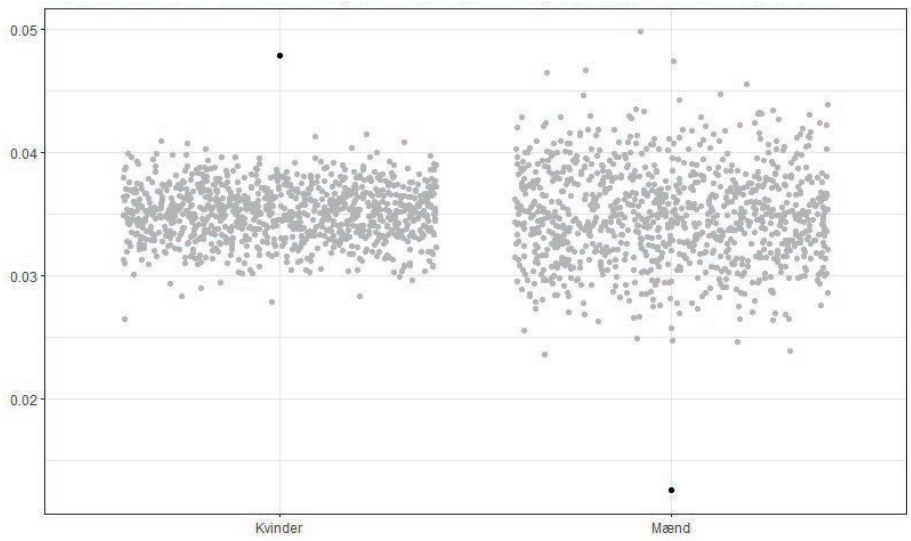
7.2. Køn og kvalitativ metode

Den første hypotese, vi tester, er hvorvidt kvinders specialer indeholder mere kvalitativ metodologi end mænds specialer. Som diskuteret tidligere baserer vi denne antagelse på tidligere forskning, som har vist signifikante kønsforskelle i publikationen af kvalitative studier (Grant et al., 1987; Plowman & Smith, 2011), samt feministiske argumenter for brugen af kvalitativ forskning (Griffin & Phoenix, 1994; Hughes & Cohen, 2010). Ud af de 50 topics fra modellen kan særligt ét kobles direkte til kvalitativ metodologi – topic **29**, hvis 100 mest sandsynlige termer er visualiserede som ordsky i figur 3. Som beskrevet tidligere er *ordskyer* (*wordclouds*) et ofte benyttet redskab til visualisering og tolkning af en topic-model (Jockers, 2013; Jockers & Mimno, 2013). Mens en typisk ordsky blot er baseret på ordfrekvenser, så kan en ordsky baseret på et topic være mere informativ, fordi et topic er en distribution over et vokabular. Sandsynligheder for de enkelte termer inden for et topic repræsenterer den betingede sandsynlighed for, at denne term indgår i et dokument, der indeholder det givne topic. Ordskyen visualiserer derfor et sandsynlighedsvægtet vokabular indenfor et topic. Metaforisk så vil den specialeskrivende i sin arbejdsproces putte hånden ned i en pose med termer i varierende størrelser efter sandsynlighed, tage den første term hun finder, og skrive det. Figur 3 viser denne 'pose'. Termer som *interview* og *informantern* indikerer, at dette topic handler om *interviewmetodologi*. Denne tolkning harmonerer med de øvrige termer med høj sandsynlighed i ordskyen og bekræftes gennem læsning af eksemplariske dokumenter. Blandt de mindre sandsynlige termer findes mere generelle termer såsom *mennesk, social, fællesskab, individ* og *identitet*, hvilket indikerer, at disse socialvidenskabeligt klingende emner ofte er knyttet til kvalitative studier (men også kan optræde sammen med andre metodologier, jf. diskussionen af polysemi i topic modelling i afsnit 5.1). Dermed kan topic 29 siges at indfange en bredere samfundsvidenskabelig genre, der dog først og fremmest er kendetegnet ved brugen af kvalitativ metodologi.

Figur 3. Topic 29 – ordsky vægtet efter sandsynlighed

Eftersom modellen modellerer hvert dokument som en distribution over topics, så vil alle dokumenter have en proportion af dette topic mellem 0 og 1. Vi bruger således ikke dette topic til at foretage en binær opdeling, men til at give udtryk for varierende grader af tilstedeværelse af (eller fokus på) en bestemt metodologi. Ved at udregne den gennemsnitlige proportion af topic 29 for kvindelige og mandlige forfattere kan vi danne os et billede af sammenhængen mellem køn og interviewmetodologi i korpusset. Der er imidlertid et problem ved denne tilgang, idet disse gennemsnit er estimater af de 'sande' værdier, og som alle estimater er de derfor forbundet med en vis statistisk usikkerhed. Vi kender ikke denne usikkerhed, eller *spredning*, og for at imødegå denne problematik benytter vi et *permutationsplot*, inspireret af Jockers & Mimno (2013). Fra modellen og den indsamlede data udtrækker vi to variable med 1.103 observationer: Proportionen af topic 29 og forfatternes køn – en simpel tabel med to kolonner. Vi udregner den gennemsnitlige proportion af dette topic for hvert køn og konstruerer herefter en række *simulerede* datasæt. Simulationen består i, at vi blander (permuterer) kolonnen med køn tilfældigt og udregner den gennemsnitlige topic-proportion for hvert køn under denne simulerede kønsfordeling. Vi simulerer således en nulhypotese, hvor der ingen sammenhæng er mellem køn og indholdet af interviewmetodologi i specialer. Denne procedure gentages 1.000 gange med henblik på at lave en stykvis approksimering af topic-proportionens spredning omkring det estimerede gennemsnit. For at parafrasere Jockers & Mimno (2013) *simulerer vi*

Figur 4. Topic 29 – gennemsnitlig proportion pr. køn

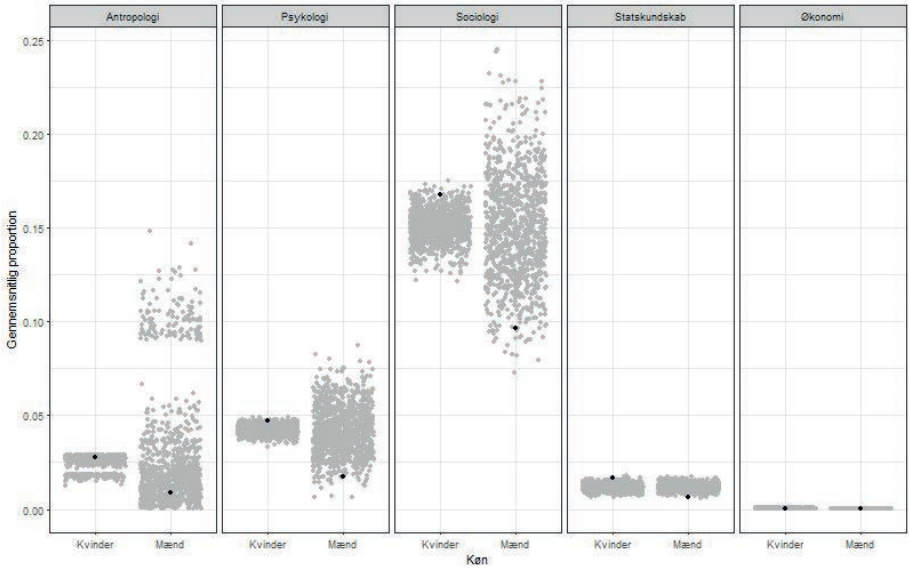


1.000 verdener, hvori forfatterskab af specialer, som indeholder interviewmetodologi, tilskrives køn tilfældigt, imens vi holder den overordnede kønsfordeling fast.

Figur 4 viser den gennemsnitlige proportion af interviewmetodologi-temaet på tværs af køn i 1.000 simulationer. Gennemsnittene for de reelt observerede data er markeret med sort, mens de simulerede gennemsnit under nulhypotesen er markeret med gråt. Eftersom der er flere kvinder i vores datasæt, er spredningen indenfor gruppen af mænd højere. Figuren indikerer umiddelbart, at den gennemsnitlige proportion af interviewmetodologi er ca. 3,5 % (procentpoint) større i specialer skrevet af kvindelige forfattere. Som illustreret tidligere i figur 1 så er der imidlertid ikke en ligelig fordeling af køn på tværs af de fem institutter. Økonomi har som det eneste en overvægt af mænd og er samtidig formentligt den mest kvantitativt orienterede disciplin, mens det formentligt mest kvalitativt orienterede studie, antropologi, har en klar overvægt af kvindelige studerende. For at teste, hvorvidt sammenhængen mellem topic 29 og køn kan forklares af forskellige kønsfordelinger på institutterne, udregner vi de gennemsnitlige proportioner for hvert køn på hvert institut og laver et permutationsplot af denne fordeling.

Denne test er vist i figur 5, som viser, at kønsforskellene varierer betydeligt på tværs af institutter. Den skiftende spredning i simulationerne skyldes den varierende kønsfordeling på institutterne. De observerede gennemsnit understøtter indirekte vores tolkning af topic 29 som repræsenterende for interviewmetodologi, idet specialer på økonomi stort set ikke indeholder dette tema. I sammenligning med sociologi er proportionerne på både antropologi

Figur 5. Topic 29 – gennemsnitlig proportion pr. køn pr. institut



og psykologi lavere, mens proportionerne på økonomi nærmer sig nul for begge køn. Vi bemærker, at der er en meget lav proportion af temaet interviewmetodologi på antropologi, hvilket kunne give indtrykket af, at modellen ikke udleder metoder korrekt. Skiftevis tætte og fjerne læsninger af topictermer, permutationsplot og specialer viser dog, at der er et separat topic omkring antropologisk metode, hvilket forklarer denne diskrepans. Det betyder, at vores analyse af topic 29 ikke kan sige noget om eventuelle kønsforskelle på dette institut.

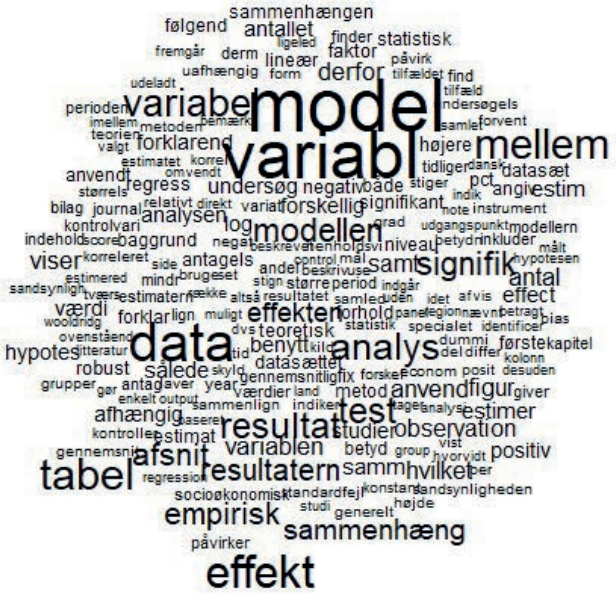
Man kan konstatere en lille kønsforskel på statskundskab, og en lidt større på psykologi, men den mest bemærkelsesværdige forskel er på sociologi, hvor den gennemsnitlige proportion i specialer med kvindelige forfattere er ca. 7 % (procentpoint) højere end i specialer med mandlige forfattere. Da disse gennemsnit ikke ligger fuldstændigt uden for den simulerede spredning (som i figur 4), er det nødvendigt at estimere, om afvigelsen er signifikant. Igen med inspiration fra Jockers & Mimno (2013) konstruerer vi en simuleret *p*-værdi ved at måle den proportion af den simulerede spredning, som ligger henholdsvis over og under gennemsnittene for kvindelige og mandlige forfattere. På Sociologisk Institut er den simulerede $p = 0{,}021$ for denne kønsforskel, hvorfor den kan betragtes som signifikant ved et signifikansniveau på 5 %. Der er således en observerbar sammenhæng mellem køn og proportionen af temaet interviewmetodologi indenfor enkelte institutter, og herunder særligt på sociologi.

7.3. Køn og kvantitativ metode

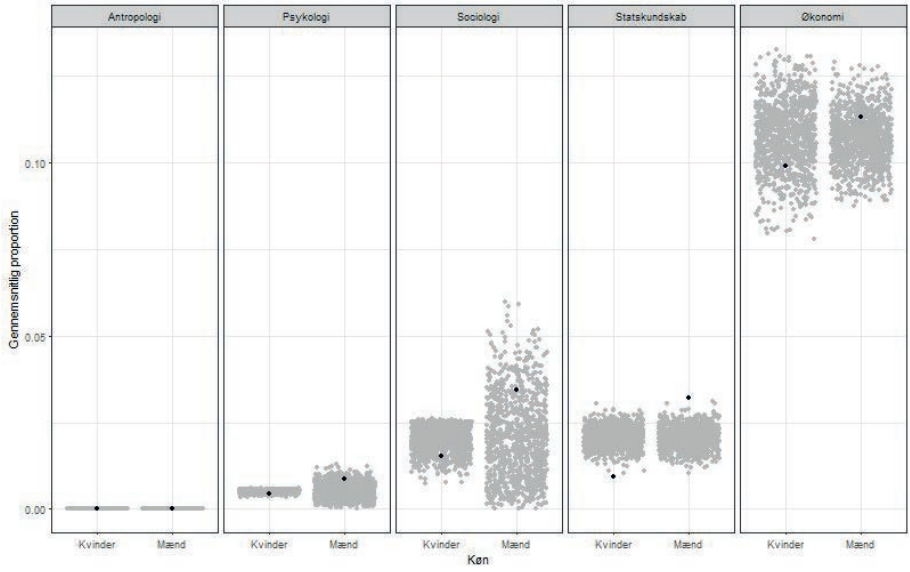
Som beskrevet tidligere har den feministiske litteratur foreslået en sammenhæng mellem køn og brugen af kvantitativ metode (Cook & Fonow, 1986; Oakley, 1998; Ramazanoglu, 1992), og der er empirisk belæg for signifikante forskelle i brugen af kvantitativ metode mellem kønnene (Dunn & Waller, 2000). Fordelingen af interviewmetodologi kan tolkes som værende indirekte understøttende for denne påståede sammenhæng, men for at undersøge hypotesen direkte analyserer vi et topic, som klart afspejler kvantitativt indhold – topic 10. Figur 6 viser de mest sandsynlige termer i dette topic som en ordsky. Ordskyen illustrerer et tematisk vokabular, som er præget af termer knyttet til kvantitativ metode såsom *model*, *variabl*, *data*, *tabel*, *effekt* og *signifik* (ordstammen af f.eks. *signifikant* og *signifikans*).

Igen benytter vi et permutationsplot for at undersøge, hvorvidt specialer med mandlige forfattere har et højere indhold af dette topic. Figur 7 viser gennemsnittene fra 1.000 simulationer under nulhypotesen (at forfatterens køn og institut ikke har indflydelse på proportionen af det kvantitative topic). Som det fremgår, fordeler disciplinernes brug af kvantitativ metode sig forventeligt, idet antropologi har mindre kvantitativt indhold for både mænd og kvinder, og økonomi har mere. På statskundskab ser vi, at den gennemsnitlige proportion for hvert køn falder helt uden for spredningen i simulationerne, og der er en signifikant kønsforskel ($p \approx 0$), hvor specialer skrevet af mandlige forfattere i gennemsnit indeholder 2,3 % mere af topic 10.

Figur 6. Topic 10 – gennemsnitlig proportion pr. køn pr. institut



Figur 7. Topic 29 – gennemsnitlig proportion pr. køn pr. institut



7.4. Køn og forskningsagendaer

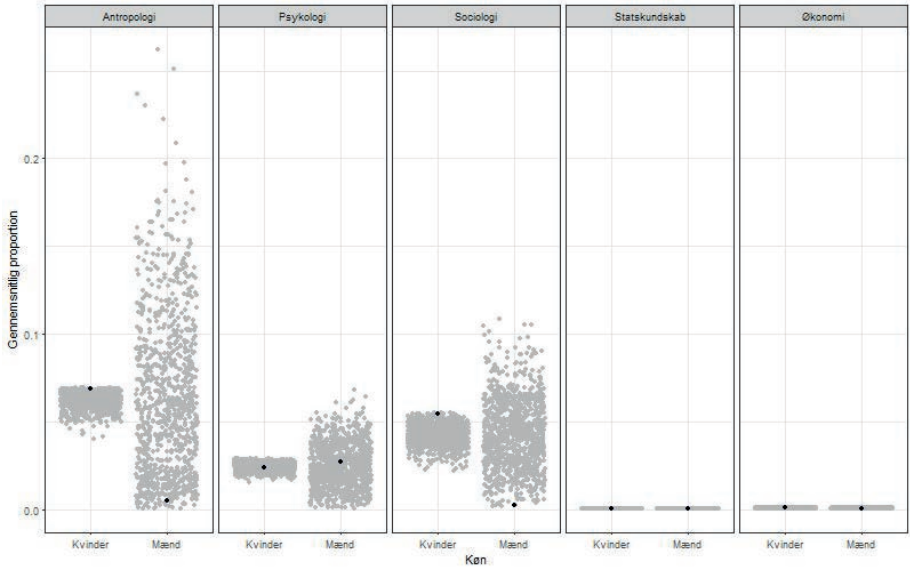
Den sidste hypotese drejer sig om sammenhængen mellem køn og forskningsagendaer. Vi har identificeret en række topics med signifikante kønsforskelle inden for enkelte institutter, men i forlængelse af den eksisterende litteratur vælger vi at fokusere på kønsforskelle inden for forskning i temaet køn. Modellen producerede ét topic med eksplicitte referencer til køn (topic 44), som er visualiseret som ordsky i figur 8.

Dette topic fremgår næsten udelukkende i antropologiske, sociologiske og psykologiske specialer og har proportioner, der nærmer sig 0 på økonomi og statskundskab. Eksemplariske specialetitler som 'En diskursanalyse af Sygeplejersken i et kønsperspektiv', 'Prostitution, køn og seksualitet' og 'Kejsersnit på moders ønske' illustrerer, at dette topic indfanger en bred vifte af problematikker relateret til køn og seksualitet. Figur 9 viser et permutationsplot af den gennemsnitlige proportion af topic 44 for hvert køn på hvert institut.

Permutationsplottet indikerer, at der er en kønsforskel på de to institutter, hvor flest specialer beskæftiger sig med emnet – antropologi og sociologi. Overrepræsentationen af dette topic blandt kvindelige forfattere er signifikant på begge institutter, med henholdsvis $p = 0{,}025$ og $p = 0{,}004$ (dvs. signifikante ved signifikansniveauer på henholdsvis 5 % og 1 %). På antropologi indeholder specialer skrevet af kvindelige forfattere i gennemsnit 6,41 % mere af topic 44 end specialer skrevet af mandlige forfattere, mens forskellen på sociologi er på 5,22 %.

Figur 8. Topic 44 – Ordsky vægtet efter sandsynlighed

Figur 9. Topic 10 – gennemsnitlig proportion pr. køn pr. institut



8. Diskussion

Vi finder, at kvinders specialer fra Sociologisk Institut indeholder en højere grad af interviewmetodologi, at kvinders specialer fra Institut for Statskundskab indeholder en lavere grad af kvantitativ metodologi, og at kvinders specialer fra Sociologisk og Antropologisk Institut indeholder mere kønsforskning. Alle forskelle er små (mellem 1 % og 7 %), men statistisk signifikante. Dette er en kønnet tendens (om end relativt svag) som både er i overensstemmelse med feministiske kald for kritiske kvalitative metodologier (Brewer, 1989; Cook & Fonow, 1986; Ward & Grant, 1985) og med den eksisterende forskning i køn og metodevalg (Grant et al., 1987: McNamee et al., 1990; Erola et al., 2015). At de observerede forskelle er relativt små kan skyldes, at metodevalg blandt specialeskrivere på Københavns Universitet ikke er kønnet i samme grad som blandt publicerede forskere. Omvendt kan det også skyldes, at eksempelvis det topic, vi bruger til at operationalisere kvalitativ metode, ikke er fuldstændig entydigt (jf. at termer som 'individ' og 'fælleskab' er blandt de 100 mest sandsynlige). Det har ikke været muligt for os at teste disse hypoteser, men det bemærkes, at der i de tidligere empiriske studier af køn og metodevalg er tale om forskelle, der, omend de er signifikante, er begrænsede i deres styrke (se afsnit 3).

At der imidlertid *kan* observeres en signifikant kønsforskel (uanset størrelsen) mellem mandlige og kvindelige studerende, er særligt relevant for sociologien, som i modsætning til eksempelvis økonomi og antropologi benytter både kvalitative og kvantitative metoder i vid udstrækning, og fordi vores resultater peger på, at tendensen er særligt stærk på sociologi. Kønsforskellen kan være særligt relevant med henblik på de studerendes videre karrierer efter kandidatstudiet. Hvis kvantitativ metode vægtes højere i professionelle miljøer, indenfor og/eller udenfor universitetet – sådan som dele af litteraturen har argumenteret for er en generel tendens i (amerikansk) sociologi og samfundsvidenskab (Gherardi & Turner, 2002; Grant & Ward, 1991; Grant et al., 1987) – så kan kvinders lavere brug af denne metode påvirke beskæftigelse og karriere efter specialeskrivning. I så fald kan danske samfundsvidenskabelige uddannelser være med til at reproducere eksisterende uligheder på arbejdsmarkedet (Larsen, Holt, & Rode, 2016).

Endvidere kan vores fund med fordel søges vurderet i relation til anbefalingen fra Forskningsrådet for Samfund og Erhverv (2006), gående på, at dansk sociologi i øget grad opkvalificerer og fokuserer på kvantitative metoder. Vores fund peger i den sammenhæng på en kønsforskel mellem de hermed efterspurgte kompetencer i forskningen, på den ene side, og produktionen af samfundsvidenskab før eventuelle forskningskarrierer, på den anden. I forbindelse med dette fund så bemærker vi, at en række topics fra vores model, som ikke er behandlet i denne artikel, tyder på, at eksisterende kønsforskelle i arbejdsområder (f.eks. den offentlige sektor kontra det private erhvervsliv; se Larsen et al., 2016) afspejles i specialeskrivendes emnevalg. Af

disse grunde ser vi en analyse af sammenhængen mellem kandidatspecialet og arbejdsmarkedet som særlig relevant og gangbar (omend uden for dette studies rækkevidde), navnligt hvis køn inkluderes i analysen.

Med henblik på videre forskning ser vi perspektiver for at kombinere specialer og metadata i DISKURS med dimittendundersøgelser. Vi har tidligere nævnt problemerne i kønsklassifikation, og med spørgeskemasvar fra en dimittendundersøgelse kunne denne valideres. Den kunne ligeledes inkludere kønsidentiteter udenfor det binære spektrum (Westbrook & Saperstein, 2015). Dimittendundersøgelser kunne også bruges til at inkludere andre baggrundvariable såsom forældres sociale klasse i tråd med diskussioner i den feministiske videnskabskritik om sociale positioner og produktionen af videnskab (Harding & Norberg, 2005, 2010).

Et andet relevant perspektiv er betydningen af de individuelle samfundsvidenskabelige discipliner og tilhørende institutter. Vi har observeret de tydeligste kønsforskelle i brug af metode inden for de samfundsvidenskabelige discipliner, der typisk udviser den største metodiske diversitet (sociologi og statskundskab). Det er værd at undersøge, om det er denne diversitet, der ligger til grund for kønsforskellene, eller om forskningstraditioner inden for specifikke institutter spiller en større rolle – hvilket kunne undersøges via et komparativt studie på tværs af universiteter og/eller lande.

Endeligt er der variable i vores indsamlede datasæt, som på trods af kodning ikke blev taget i brug grundet deres kompleksitet eller irrelevans for vores specifikke hypoteser, herunder vejleders køn. På baggrund af sammenhængen mellem køn og brug af metode er det værd at overveje samarbejdet med specialevejlederen som en yderligere faktor. Grant et al. (1987) viser, hvorledes samarbejdsmønstre i akademia varierer mellem kønnene, og vores data lader til at indikere, at de specialeskrivende har en tendens til at vælge vejledere af samme køn. Et nærmere studie af disse mekanismer kunne være af interesse – eksempelvis en undersøgelse af, hvorvidt vejledervalg og forfatterskaber påvirker både tematisk indhold og metodologiske tilbøjeligheder, og hvorvidt vejledningen i sig selv kan være med til at reproducere kønsforskelle.

Udover de ovenstående betragtninger om køn og akademisk produktion har vores undersøgelse bibragt en række indsigter om anvendelsen af topic modelling i socialvidenskabelig tekstanalyse. Vores undersøgelse er endnu et eksempel på denne metodes anvendelighed i at arbejde eksplorativt med større tekstkorpora – i forhold til at opdage temaer og at 'kode'/klassificere dokumenter. Vi har imidlertid taget metoden et skridt videre, idet vi har målt statistiske forskelle på tværs af dokumenter og dokumentkategorier (som fx DiMaggio et al., 2013 el. Jockers, 2013). Her har det vist sig muligt at anvende topic modelling til at teste og nuancere allerede etablerede hypoteser, men der har ligeledes vist sig en metodologisk udfordring i at tolke signifikansen af resultaterne. Eksempelvis er det ikke nødvendigvis åbenlyst, hvad det be-

tyder, at ét speciale indeholder 7 % mere kvalitativ metode end et andet; en udfordring, der styrkes af, at det pågældende topic ikke er fuldkommen entydigt og ikke indfanger alle typer af kvalitativ metodologi. Derfor kan vi med vores undersøgelse understrege vigtigheden i at koble en analyse baseret på topic modelling med andre tilgange til den undersøgte problematik for at validere eller blot sandsynliggøre resultaterne – særligt når de målte forskelle er relativt svage. I vores tilfælde har vi baseret os på den eksisterende forskning i køn og akademisk produktion for at vurdere betydningen af vores resultater.

9. Konklusion

Vi har i dette studie benyttet et alternativt datasæt og en alternativ metodologi til en replikation af tidligere empiriske studier, ansporet af diskussionen om sammenhængen mellem køn og metodologi. Denne diskussion har eksisteret siden 1960'erne indenfor sociologien og samfundsvidenskaberne generelt (Brewer, 1989), og tidligere forskning har fundet tendenser til, at kvinder i samfundsvidenskab (særligt sociologi) i højere grad benytter sig af kvalitative metodologier end mænd (Grant et al., 1987; McNamee et al. 1990; Erola et al. 2015; Plowman & Smith, 2011). Vi observerer disse tendenser på et tidligere akademisk stadie i en dansk population af kandidatstuderende. På trods af at datamaterialet ikke er udsat for den samme selektionsbias som tidligere forskning i tidsskriftsartikler, og at metodologien adskiller sig, så er vores fund i overensstemmelse med litteraturen: Kvinder er mere tilbøjelige til at bruge kvalitativ metode, og mænd er mere tilbøjelige til at bruge kvantitativ metode. Førstnævnte er særligt tilfældet blandt sociologistuderende, hvor kvinders specialer gennemsnitligt indeholder 7 % (procentpoint) mere af et topic, som repræsenterer kvalitativ metodologi. Endeligt finder vi også en bekræftelse af, at kvindelige forfattere i højere grad skriver om køn og seksualitet.

Disse fund må overvejes i lyset af de studerendes videre karrierer. Dele af litteraturen (Gheradi & Turner, 2002; Grant et al., 1987; Griffin & Phoenix, 1994) argumenterer for, at kvantitative metoder i samfundsvidenskaben generelt ses som overlegne, og tidligere studier har vist, at prestigiøse journaler i særligt den amerikanske sociologi overvejende publicerer kvantitativ forskning (Grant & Ward, 1991). Hvis sådanne tendenser er de samme på andre arbejdsmarkeder, og herunder også i en europæisk kontekst som den danske, så kan kønsforskelle i metodevalg reproducere ulighed. Endvidere har Forskningsrådet for Samfund og Erhverv (2006) argumenteret for en opkvalificering og øget fokus på kvantitative metoder i dansk sociologi, og kønsforskelle i applikationen af metodologier kan derfor også have relevans for de studerendes potentielle forskningskarrierer. Disse forhold kalder på yderligere studier.

Noter

1. Typiske retningslinjer for specialelængder angiver 2.400 tegn inkl. mellemrum som én sidelængde.

2. Tilgængeligt på https://diskurs.kb.dk/

3. Vi observerede særligt dette i specialer fra Psykologisk Institut.

4. Visualiseringen er tilgængelig på rasmusmunksgaard.net/viz/metodevalg

Litteraturliste

Abend, G. (2006). Styles of sociological thought: Sociologies, epistemologies, and the Mexican and U.S. quests for truth. *Sociological Theory* 24(1), 1-41. https://doi.org/10.1111/j.0735-2751.2006.00262.x

Bache, S.M., & Wickham, H. (2014). *magrittr: A Forward-Pipe Operator for R.* Tilgængelig på https://CRAN.R-project.org/package=magrittr

Baumer, E.P.S., Mimno, D., Guha, S., Quan, E., & Gay, G.K. (2017, 4). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*. In press.

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84. https://doi.org/10.1145/2133806.2133826

Blei, D.M., & Lafferty, J.D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics*, *1*(1), 17-35. https://doi.org/10.1214/07-AOAS114

Blei, D.M., Ng, A.Y., Jordan, M.I., & Lafferty, J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4/5), 993-1022.

Brewer, R.M. (1989, 3). Black women and feminist sociology: The emerging perspective. *The American Sociologist*, *20*(1), 57-70. https://doi.org/10.1007/BF02697787

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., & Blei, D.M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems, NIPS '09,* 288-296.

Cook, J.A., & Fonow, M.M. (1986). Knowledge and Women's Interests: Issues of Epistemology and Methodology in Feminist Sociological Research. *Sociological Inquiry*, *56*(1), 2-29. https://doi.org/10.1111/j.1475-682X.1986.tb00073.x

Denny, M. & Spirling, A. (2017). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. Tilgængelig på https://ssrn.com/abstract=2849145. https://doi.org/10.2139/ssrn.2849145

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, *41*(6), 570-606. https://doi.org/10.1016/j.poetic.2013.08.004

Dunn, D., & Waller, D. (2000, 4). THE METHODOLOGICAL INCLINATIONS OF GENDER SCHOLARSHIP IN MAINSTREAM SOCIOLOGY JOURNALS. *Sociological Spectrum*, *20*(2), 239-257. https://doi.org/10.1080/027321700279974

Erola, J., Reimer, D., Räsänen, P., & Kropp, K. (2015). No crisis but methodological separatism: a comparative study of Finnish and Danish publication trends between 1990 and 2009. Sociology, 49(2), 374-394. https://doi.org/10.1177/0038038514542495

Fairclough, N. (1992). Intertextuality in critical discourse analysis. *Linguistics and Education*, *4*(3), 269-293. https://doi.org/10.1016/0898-5898(92)90004-G

Feinerer, I., Buchta, C., Geiger, W., & Rauch, J. (2013). The textcat package for n-gram based text categorization in R. *Journal of Statistical Software*, 52(6), 1-17.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in {R}. *Journal of Statistical Software*, *25*(5), 1-54. https://doi.org/10.18637/jss.v025.i05

Freeman, M.K., Chuang, J., Roberts, M.E., Stewart, B.M., & Tingley, D. (2015). *stmBrowser: Structural Topic Model Browser*. Tilgængelig på http://cran.r-project.org/package=stmBrowser

Forskningsrådet for Samfund og Erhverv (2006) Dansk sociologis muligheder. København: Forskningsrådet for Samfund og Erhverv, Arbejdsgruppe under Forskningsrådet for Samfund og Erhverv.

Gheradi, S., & Turner, B. (2002). Real Men Don't Collect Soft Data. I M. Huberman & M.B. Miles (Eds), *The qualitative researcher's companion.* United Kingdom. London: Sage.

Goldenberg, S., & Grigel, F. (1991, 9). Gender, science and methodological preferences. *Social Science Information*, *30*(3), 429-443. https://doi.org/10.1177/053901891030003003

Grant, L., & Ward, K.B. (1991). Gender and Publishing in Sociology. *Source: Gender and Society*, *5*(2), 207-223. https://doi.org/10.1177/089124391005002005

Grant, L., Ward, K.B., & Rong, X.L. (1987, 12). Is There An Association between Gender and Methods in Sociological Research? *American Sociological Review*, *52*(6), 856-862. https://doi.org/10.2307/2095839

Griffin, C., & Phoenix, A. (1994). The Relationship between Qualitative and Quantitative Research: Lessons from Feminist Psychology. *Journal of Community & Applied Social Psychology*, *4*(4), 287-298. https://doi.org/10.1002/casp.2450040408

Grimmer, J., & Stewart, B.M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267-297. https://doi.org/10.1093/pan/mps028

Hall, D., Jurafsky, D., & Manning, C.D. (2008). *Studying the history of ideas using topic models.* Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613763

Harding, S., & Norberg, K. (2005). New Feminist Approaches to Social Science Methodologies: An Introduction. *Signs: Journal of Women in Culture & Society*, *30*(4), 2009-2015. https://doi.org/10.1086/428420

Harding, S. (1992). Rethinking standpoint epistemology: What is »strong objectivity?«. *The Centennial Review*, *36*(3), 437-470.

Hughes, C., & Cohen, R.L. (2010, 7). Feminists really do count: the complexity of feminist methodologies. *International Journal of Social Research Methodology*, *13*(3), 189-196. https://doi.org/10.1080/13645579.2010.482249

Jockers, M.L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Jockers, M.L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, *41*(6), 750-769. https://doi.org/10.1016/j.poetic.2013.08.005

Larsen, M., Holt, H., & Rode, M. (2016). *Et kønsopdelt arbejdsmarked: Udviklingstræk, konsekvenser og forklaringer*. København: SFI – Det Nationale Forskningscenter for Velfærd.

Levy, K.E.C., & Franklin, M. (2014). Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. *Social Science Computer Review*, *32*(2), 182-194. https://doi.org/10.1177/0894439313506847

Lucas, C., Nielsen, R.A., Roberts, M.E., Stewart, B.M., Storer, A., &

Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, *23*(2), 254-277. https://doi.org/10.1086/428417

Margaret Fonow, M., & Cook, J.A. (2005). Feminist methodology: New applications in the academy and public policy. *Signs: Journal of Women in Culture and Society*, *30*(4), 2211-2236.

McNamee, S.J., Willis, C.L., & Rotchford, A.M. (1990). Gender Differences in Patterns of Publication in Leading Sociology. *The American Sociologist*, *21* (2), 99-115.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, & Andrew McCallum (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (2),* 262-72.

Mohr, J.W., & Bogdanov, P. (2013). Introduction: Topic models: What they are and why they matter. *Poetics*, *41*(6), 545-569. https://doi.org/10.1016/j.poetic.2013.10.001

Munksgaard, R., & Demant, J. (2016). Mixing politics and crime The prevalence and decline of political discourse on the cryptomarket. *International Journal of Drug Policy*, *35*, 77-83. https://doi.org/10.1016/j.drugpo.2016.04.021

Nordisk Forskningsinstitut. (2017). *Danskernes Navne.* Tilgængelig på http://www.danskernesnavne.navneforskning.ku.dk/Personnavne.asp

Norris, N. (1997, 3). Error, bias and validity in qualitative research. *Educational Action Research*, *5*(1), 172-176. https://doi.org/10.1080/09650799700200020

Oakley, A. (1998). Gender, Methodology and People's Ways of Knowing: Some Problems With Feminism and the Paradigm Debate in Social Science. *Sociology*, *32*(4), 707-731. https://doi.org/10.1177/0038038598032004005

Plowman, A.D., & Smith, A.D. (2011, 5). The gendering of organizational research methods. *Qualitative Research in Organizations and Management: An International Journal*, *6*(1), 64-82. https://doi.org/10.1108/17465641111129399

Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., & Radev, D.R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209-228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria.

Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing microblogs with topic models. *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010.*

Ramazanoglu, C. (1992). ON FEMINIST METHODOLOGY: MALE REASON VERSUS FEMALE EMPOWERMENT. *Sociology*, *26*(2), 207-212. https://doi.org/10.1177/0038038592026002003

Riger, S., & Stephanie. (1992). Epistemological debates, feminist voices: Science, social values, and the study of women. *American Psychologist*, *47*(6), 730-740. https://doi.org/10.1037/0003-066X.47.6.730

Roberts, M.E., Stewart, B.M., & Tingley, D. (2014). Stm: R package for structural topic models. *R package version 0.6*, *1*.

Roberts, M.E., Stewart, B.M., Tingley, D., Airoldi, E.M., & others. (2013). The structural topic model and applied social science. In *Advances*

in neural information processing systems workshop on topic models: Computation, application, and evaluation.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B. & Rand, D.G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064-1082. https://doi.org/10.1111/ajps.12103

Sprague, J., & Zimmerman, M.K. (1989). Quality and Quantity: Reconstructing Feminist Methodology. *The American Sociologist*, *20*(1), 71-86. https://doi.org/10.1007/BF02697788

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, *39*(1), 81-135. https://doi.org/10.1002/aris.1440390110

Törnberg, A., & Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society*, *27*(4). https://doi.org/10.1177/0957926516634546

Unger, R.K., Draper, R.D., & Pendergrass, M.L. (1986). Personal Epistemology and Personal Experience. *Journal of Social Issues*, *42*(2), 67-79. https://doi.org/10.1111/j.1540-4560.1986.tb00225.x

Venturini, T., & Latour, B. (2010). The social fabric: Digital traces and quali-quantitative methods. *Proceedings of Futur en Seine 2009*. Tilgængelig på http://www.academia.edu/download/38150764/

Ward, K.B., & Grant, L. (1985). The Feminist Critique and a Decade of Published Research in Sociology Journals. *The Sociological Quarterly*, *26*(2), 139-157. https://doi.org/10.1111/j.1533-8525.1985.tb00220.x

Westbrook, L., & Saperstein, A. (2015, 8). New Categories Are Not Enough. *Gender & Society*, *29*(4), 534-560. https://doi.org/10.1177/0891243215584758

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12), 1-20. https://doi.org/10.18637/jss.v021.i12

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Verlag New York. Tilgængelig på http://ggplot2.org

Wickham, H. (2016a). *httr: Tools for Working with URLs and HTTP.*

Tilgængelig på https://CRAN.R-project.org/package=httr

Wickham, H. (2016b). *rvest: Easily Harvest (Scrape) Web Pages.* Tilgængelig på https://CRAN.R-project.org/package=rvest

Wickham, H. (2017). *stringr: Simple, Consistent Wrappers for Common String Operations.* Tilgængelig på https://CRAN.R-project.org/ package=stringr

Wickham, H., & Francois, R. (2016). *dplyr: A Grammar of Data Manipulation.* Tilgængelig på https://CRAN.R-project.org/package= dplyr

Wickham, H., Francois, R., & Müller, K. (2016). *tibble: Simple Data Frames.* Tilgængelig på https://CRAN.R-project.org/package= tibble

Appendix

Appendix: 10 mest prævalente topics per institut

Psykologi	Sociologi	Økonomi	Statskundskab	Antropologi
depress, ptsd, symptom, lidels, kognit, mellem, patient, forskellig, dsm, person	interview, sålede, informantern, oplev, social, beskriv, oplevels, forhold, hvilket, fortæller	model, variabl, data, effekt, tabel, mellem, variabel, analys, test, resultat	militær, militæret, civil, politisk, afghanistan, usa, mellem, stater, intern, forhold	social, informant, famili, liv, kapitel, arbejd, gennem, blandt, ibid, mellem
klienten, terapeuten, terapeutisk, terapi, behandl, terapien, terapeut, kognitiv, allianc, forhold	social, mad, mellem, billed, mennesk, rum, del, sociolog, gør, forskellig	forbrug, model, derfor, modellen, energi, afsnit, samled, produkt, forbruget, afgift	mellem, specialet, organisation, mål, interview, praksi, anvend, organisationen, grad, sålede	social, mad, mellem, billed, mennesk, rum, del, sociolog, gør, forskellig
sålede, forståels, forstå, forhold, måde, social, mellem, foucault, verden, perspektiv	arbejd, medarbejdern, leder, medarbejder, ledels, arbejdet, forhold, ledelsen, mellem, opgav	indkomst, pct, person, arbejdsmarkedet, arbejd, pension, offentlig, modellen, derm, mellem	kommissionen, europæisk, eus, national, european, rådet, forhold, europa, politisk, sålede	kvinder, mænd, køn, seksualitet, seksuell, sex, overgreb, kvindelig, mandlig, kvind
psykolog, narrativ, narrat, individet, sålede, white, mennesk, forhold, liv, praksi	borger, borgeren, arbejd, ledig, social, borgern, aktiv, københavn, forhold, mellem	bank, banker, bankern, finansiell, banken, period, aktiv, modellen, model, derfor	politisk, regeringen, forhold, dansk, politik, folketinget, derfor, specialet, reger, polici	social, peopl, state, will, usa, new, can, like, http, mellem
barnet, freud, psykisk, udvikl, ubevidst, mellem, oplevels, forhold, følelser, forstå	ung, social, forhold, kriminalitet, anbragt, derfor, tidliger, mellem, del, adfærd	model, afkast, modellen, derfor, test, period, værdi, figur, forskellig, investor	test, frame, motiv, effekt, mellem, hypotes, national, respondentern, holdn, respondent	altså, tror, syne, ting, måske, gang, rigtig, siger, måde, gør
barnet, børn, udvikl, barn, adfærd, tilknytn, mellem, børnene, forældr, bowlbi	kvinder, social, etnisk, kvindern, mænd, sålede, forskellig, beboer, forhold, boligområd	uddannels, mellem, tabel, videregående, variabl, indkomst, person, model, pct, alder	grønland, danmark, politisk, rusland, dansk, grønlandsk, mellem, kina, økonomisk, forhold	sålede, forståels, forstå, forhold, måde, social, mellem, foucault, verden, perspektiv
stress, mind, individet, social, mellem, forhold, oplevels, psykisk, psykologisk, resilien	sålede, forståels, forstå, forhold, måde, social, mellem, foucault, verden, perspektiv	pct, mio, selskabet, http, www, kild, figur, derfor, sas, hvilket	frivillig, social, kommun, kommunen, mellem, arbejd, københavn, offentlig, aktører, samarbejd	kvinder, social, etnisk, kvindern, mænd, sålede, forskellig, beboer, forhold, boligområd

Psykologi	Sociologi	Økonomi	Statskundskab	Antropologi
altså, tror, syne, ting, måske, gang, rigtig, siger, måde, gør	diskurs, diskur, sålede, diskursen, mellem, identitet, analys, social, derm, analysen	agent, effekt, teorien, mellem, model, adfærd, præferenc, inform, agenten, princip	politisk, mediern, frame, facebook, politik, artikl, kampagn, negat, kommunik, medier	patient, behandl, sygdom, patienten, patientern, praksi, læger, medicinsk, region, praktiserend
interview, sålede, informantern, oplev, social, beskriv, oplevels, forhold, hvilket, fortæller	kvinder, mænd, køn, seksualitet, seksuell, sex, overgreb, kvindelig, mandlig, kvind	produkt, nordisk, analys, hvilket, vækst, iss, derfor, side, strategisk, virksomheden	parti, politisk, partier, vælgere, vælgern, mellem, venstr, partiern, partiet, blok	interview, sålede, informantern, oplev, social, beskriv, oplevels, forhold, hvilket, fortæller
mennesk, self, spiseforstyrrels, bed, psykologisk, fedm, eat, person, overvægtig, selvværd	altså, tror, syne, ting, måske, gang, rigtig, siger, måde, gør	økonomisk, vækst, økonomi, offentlig, bnp, land, krisen, gæld, ern, politik	politisk, demokrati, demokratisk, tillid, institution, land, social, korrupt, værdier, polit	gruppen, medlemm, ansatt, pair, både, humor, blir, forhold, ble, informanten