Kvantificering af gruppe-diskussioner

– metodetriangulering via edb?

Af Marcus Schmidt og Hans Solgaard

Resumé

Artiklen beskriver den kvalitative analyseproces, og ser på hvorledes fremgangsmåden kan støttes ved anvendelse af computere. Dette efterfølges af præsentationen af et nyt avanceret programmel til tekst-analyse, der bygger på et neuralt netværk. Artiklen giver anvisninger på hvordan en tekst bør prepareres, førend en kvantitativ analyse kan finde sted. Programmets funktionsmåde søges "valideret" ved hjælp af et lille empirisk eksempel. Afslutningsvis fortæller forfatterne om erfaringer med at bruge metoden på komplette gruppediskussioner.

Indledning

Kvalitative data og analysemetoder har altid været en væsentlig del af metodegrundlaget i de forskellige samfundsvidenskaber, herunder også i erhvervsorienterede forskningsområder som marketing og organisation. I markeds- og kommunikationsforskningen benytter man således i vidt omfang fokusgrupper, dybdeinterview, case studier, åbne spørgsmål, projektive teknikker, ansættelsesinterviews, personalesamtaler og lignende. Disse metoder resulterer i data i form af ord (tekst), billeder og/eller film, der ikke umiddelbart kan underkastes statistisk analyse på samme måde som kvantitative data (tal).

Kvalitative data vil typisk være beskrivelser og/eller forklaringer af hændelsesforløb eller uddybning af holdninger i en konkret kontekst. Takket være de kvalitative teknikker kan analytikeren granske og afdække et bredt spektrum af fænomener. Derigennem opnås et indblik i komplicerede problemområder. Tekst (og billeder) er nu engang lettere forståelige og dermed mere overbevisende for mange beslutningstagere, end side op og side ned med tal, tabeller og grafer.

Men der er naturligvis en skyggeside. Kvalitative data er dels besværlige og tidskrævende at indsamle, dels vanskelige at analysere. Det sidste fordi, der ikke findes klart formulerede og stringente analysemetoder til formålet. Miles (1979, 517) bemærker således, *"...the analyst faced with a bank of qualitative data has very few guidelines for protection against self-delusion, let alone the presentation of unreliable and invalid conclusions to scientific or policy making audiences"*. Dette var i 1979. Siden da er forholdene heldigvis ændret på dette punkt, således at analyse af kvalitative data i dag kan siges at være mindre kunst og mere håndværk end tidligere. Det er især den stigende anvendelse af computere til tekstanalyse, der har befordret denne udvikling og som har gjort det muligt at behandle større mængder af kvalitative data på en stringent og replicerbar måde[1]. Se Pfaffenberger (1988), Tesch (1990), Fielding and Lee (1991), Kelle (1995) og især Weitzmann and Miles (1995).

Set ud fra en poppersk videnskabsopfattelse kan man sige, at den kvalitative analyse med computerens hjælp har overskredet demarkationslinien mellem mytologi og videnskab.

I en vis grad kan kvalitative analyser nu om stunder endog leve op til det såkaldte falisificeringskriterium, idet der kan opstilles hypoteser med hensyn til sprogbrug, relationer mellem ord mm., der kan efterprøves ved hjælp af statistiske tests.

Brug af computere til kvalitative analyser støder derimod også på kritik: 1. Fokusering på kvantitative forhold, statistik osv. resulterer i at programmet styrer processen. Konsekvensen bliver at metoden styrer problemet fremfor det omvendte. 2. De kreative og intuitive tilgange sløres, nedtones og træder i baggrunden. 3. Der kommer for meget fokus på små, isolerede fænomener. 4. Samtidigt med at detaljer opprioriteres, går det holistiske overblik tabt, idet fokuseringen foregår på mikroniveau. Se hertil Evans (1989) og Robson and Hedges (1993).

Vi skal ikke her komme nærmere ind på de anførte kritikpunkter. Men vi medgiver ud fra egen erfaring, at det bestemt ikke er ukompliceret at bruge kvantitative teknikker til analyse af tekst, hvilket formentligt vil fremgå tydeligt under læsningen af denne artikel. Vi har organiseret resten af artiklen således, at vi først kort beskriver den kvalitative analyseproces, og ser på hvorledes fremgangsmåden kan støttes ved anvendelse af computere. Herunder anfører vi kort de vigtigste computerprogrammer, der benyttes i dag. Dette efterfølges af præsentationen af et nyt avanceret tekst-analyse programmel, CATPAC, baseret på et neuralt netværk. I den forbindelse gør vi rede for, hvordan en tekst skal prepareres, førend den er klar til computeranalysen. Derpå beskriver vi programmets virkemåde og væsentligste karakteristika. Programmets funktionsmå-

de søges "valideret" ved hjælp af et lille empirisk eksempel. Afslutningsvis fortæller vi kort om vores erfaringer med at bruge metoden på komplette gruppediskussioner.

Den kvalitative analyseproces og fremkomsten af computer-støttet analyse

En af de væsentligste kritikpunkter ved den traditionelle kvalitative analyse er dens ateoretiske og usystematiske fremgangsmåde. Et klassisk eksempel herpå findes hos Popper. Han diskuterede engang en konkret opførsel hos et barn med Adler.

> *"Once, in 1919, I reported to him a case, which to me did not seem particularly Adlerian, but which he found no difficulty in analyzing in terms of his theory of inferiority feelings, although he had not even seen the child. Slightly shocked, I asked him how he could be so sure. "Because of my thousandfold experience", he replied; whereupon I could not help saying: "And with this new case, I suppose, your experience has become thousand-and-one-fold." (Popper, 1963, 35).*

Det hænder, at man i faglige tidsskrifter eller ved foredrag under konferencer støder på en kvalitativ forsker, der på entusiastisk vis beretter om eureka-oplevelsen, det magiske øjeblik eller om "det illuminerende citat", der medførte at brikkerne pludselig faldt på plads og hvor det med ét blev lysende klart, hvordan tingene er skruet sammen. Nærgående spørgsmål om de tekniske detaljer i forbindelse med undersøgelsen bliver ved samme lejlighed ofte mødt med undvigende svar eller bliver betragtet som upassende og "uvedkommende indblanding i indre forhold".

Heldigvis er sådanne tilfælde af "selvindlysende" forskningsresultater sjældnere i dag end for 10-20 år siden. Forklaringen på dette er givetvis at den metode-opfattelse, der er fremherskende i de hårde videnskaber (matematik, fysik, biologi osv.), stille og roligt er ved at udvikle sig til en slags standard indenfor de bløde videnskaber (sociologi, psykologi etc.).

Det er nok også i det lys man skal se fremkomsten af computer-programmer til analyse af kvalitative informationer. I tabel 1 vises en oversigt over software til behandling af kvalitative data. Oversigten er ikke udtømmende, men rummer de væsentligste aktører. Ud over de 27 anførte programmer findes der specielle tiltag fra Kina (Leung and Yeh 1997) og Norge (Helmersson 1997), der ikke er nævnt. Også Grunert and Bader (1986) har arbejdet seriøst med en kvantificering af fokusgruppe-fragmenter på baggrund af en tidlig version af *TEXTPAC*. Af oversigten i tabel 1 fremgår ikke direkte, at en del programmer foreligger i forskellige versioner og med diverse tillægsmoduler, der reelt er selvstændige enheder. Så samlet kan man konstatere, at det i dag findes omkring 50 forskellige programmer på markedet til analyse af kvalitative oplysninger.

Inddelingen i tabel 1 kan diskuteres, idet der forekommer adskillige overlapninger med hensyn til, hvilke features de enkelte programmer indeholder. Man kunne derfor med lige så god ret placere en del af programmerne under flere af kategorierne. Inddelingen skal derfor kun forstås som en grov tilnærmelse[2]. Men groft set gælder (især for så vidt angår 1-4), at den efterfølgende kategori kan det, som

Tabel 1: Software til analyse af kvalitative data.

	Kategori	*Væsentlige faciliteter*	*Udbydere*
1	Text Retrievers	Avancerede søgefunktioner, opsporing af synonymer samt ord med vis lighed, etablering af bookmarks i og hyperlinks til dokumenter og filer etc.	WordCruncher, Metamorph, ZyINDEX, Sonar Professional, Orbis, Text Collector
2	Textbase Managers	Database-indeksering, markeringer af ord med "pop-ups" til memos, noter, kommentarer og lignende	askSam, FolioVIEWS, MAX, Tabletop
3	Code-and-Retrieve Programs	Kodning, strukturing og sortering af tekstfragmenter, konstruktion af hierarkiske relationer etc.	Ethnograph, QUALPRO, HyperQual2, Kwalitan, Martin
4	Code-Based Theory-Builders	Opbygning af teorier om teksten, primitiv kausalanalyse og hypotesetest	NUD.IST, ATLAS/ti, AQUAD, HyperRESEARCH, QCA
5	Conceptual Network-Builders	Grafisk modellering, konstruktion af relationer, etablering af ord-matricer og lister, visse statistiske analyser	Inspiration, MECA, MetaDesign, SemNet
6	Content Analysis Programs	Indholdsanalyse assisteret af omfattende encyklopædier over værdier og socialpsykologiske begreber	TEXPAC, General Enquirer III
7	ANN-Based Programs	Neural netværks-analyse	CATPAC (Metamorph)

programmerne i den foregående kategori kunne, men samtidigt indeholder noget nyt, da de er opstået på et senere tidspunkt.

Programmerne under 5-7 skiller sig ud. Kategori 5 minder en del om grafisk præsentationssoftware, medens softwaren i kategori 6 har en del træk tilfælles med elektroniske leksika (CATPAC i kategori 7 behandles forneden).

De fleste af programmerne i tabel 1 rummer booleanske søge-faciliteter, indeksering og kan opstille lister med ordfrekvenser. De bedste features og brugervenlige interfaces, der ved programmernes fremkomst for ca 10-15 år siden var spredt ud over et stort antal forskellige producenter, er i tidens løb blevet overtaget af de fleste analyse-pakker. Følgen er at programmerne i dag kan mange af de samme ting og derfor er begyndt at ligne hinanden ret meget[3]. De fås normalt i varianter til flere forskellige styresystemer og prisen svinger mellem halvtreds til hundrede dollars for en students-lab udgave til omkring tusind dollar for en fuld udbygget version.

Det er i øvrigt slet ikke sikkert, at en anskaffelse er nødvendig, idet adskillige af faciliteterne (søgning, sortering, indeksering, visse statistiske oplysninger osv) er integreret i de gængse kontorpakkers standardmoduler. Så, hvis man i forvejen råder over noget sådant og til og med har adgang til relationsdatabase-software svarende til Access eller Dbase IV, da er man på forhånd ganske godt dækket ind. Man skal også være opmærksom på, at en del af programmerne enten kun fungerer når teksten, der skal indlæses, er på engelsk eller amerikansk (fx General Enqirer III) eller at visse centrale features formentlig ikke virker efter hensigten, når sproget er et andet. Selv hvis alt lader til at fungere, vil en række systemfiler være sat op til en-

gelsk hvilket medfører en forkert læsning (programmet vil fx tro at "possess" - og ikke "park" - er synonym for "have", når det behandler en dansk tekst).

Etnograph, NUD.IST og General Enquirer er formentlig de tre, der anvendes mest. Når ingen af programmerne har opnået betydelig udbredelse uden for de akademiske cirkler er årsagen givetvis den, at de er tidskrævende at anvende. Bearbejdning og analyse består typisk i en detaljeret gennemgang og nærlæsning af teksten, hvorefter der skal indlægges koder, etableres kategorier mm. I følge Catterall and Maclaren (1998, 216) tager det i den første tid omkring 1 minut for at kode ½ tekstlinie i NUD.IST. Med det tempo bliver det en opgave på ca 3 arbejdsdage for blot at kode en enkelt gruppediskussion. Når der oveni dette påløber 1-2 dage for at få et bånd skrevet ud, ja så ender man med at bruge en hel uge blot på det forberedende arbejde. Det vil helt sikkert sprænge de ressourcemæssige rammer for det budget, der er afsat til en kommerciel fokusgruppe.

En andet væsentlig svaghed er, at selv de bedste af de traditionelle programmer (kategori 1-6) ikke kommer ud over det subjektive element, når man vil opbygge en konceptuel model over tekstens betydningsindhold ("Aussagekraft"). Man er faktisk henvist til at konstruere sin model "ud af den blå luft". Bortset fra sit forhåndskendskab til teksten og de deraf følgende apriori-formodninger har man nemlig ingen hjælp; det skulle da lige være kollegers input. Det er altså analytikeren selv, der fortæller programmet, hvordan modellen skal se ud, hvilke nøgleord, der skal have links til andre osv. Men for at et program kan komme ud over dette aldeles subjektive element, bør arbejdsdelingen mellem software og analytiker fungere omvendt, nemlig sådan at det er *programmet*, der giver *analytikeren* nogle vink mht. centrale relationer i teksten. Og det er præcis her, programmer som CATPAC, der bygger på kunstig intelligens, kommer ind i billedet.

Tekstanalyse ved hjælp af neurale netværk

Selv om et program som CATPAC stadigvæk kræver en vis form for forhåndskodning samt fastlæggelse af nøgleord for at virke optimalt, ja så kommer det selvorganiserende neurale netværk faktisk selv med et bud på relationsmønsteret nøgleordene imellem. Og det må vel siges at være et konkret fremskridt, sammenlignet med de øvrige traditionelle programmer (Moore, Burbach, and Heeler 1995).

Brugen af neurale netværk til tekstanalyse er først og fremmest blevet introduceret af folkene bag tekstanalyseprogrammet CATPAC (CATegory PACkage) fra Terra Research and Computing. Den oprindelige udvikling påbegyndtes sidst i 1970'erne; i 1989 blev programmet så udbygget til at inkludere et neural netværk. Se Woelfel and Fink (1980), Woelfel (1993), Woelfel and Stoyanoff, (undated). Den foreliggende version er CATPAC™ 4-Windows, Version 1.0.

For en generel introduktion til neurale netværk se for eksempel Masson and Wang (1990) samt Rasmussen (1995), mens en mere detaljeret beskrivelse kan findes hos White et al. (1992). Man skelner mellem to hovedtyper af netværk, - (a) *supervised* netværk, som typisk anvendes

til klassifikations-formål. Her *trænes* netværket på historiske data til at kunne genkende eller forudsige et af flere mulige udfald - og (b) *unsupervised* netværk, der anvendes til identifikation og kvantificering af mønstre, som måtte findes i datasæt eller tekstfragmenter. Den type netværk, der anvendes til tekstanalyse, er således af typen *unsupervised,* også kaldet *selforganizing.*

Det princip, som neurale netværk er baseret på, er forholdsvis enkelt at beskrive. I et kunstigt neuralt netværk forsøger man at efterligne den måde, hvorpå den menneskelige hjerne fungerer. En biologisk hjerne består af en række *neuroner* (nerveceller), det vil sige en slags "kontakter", som i det simpleste tilfælde kan befinde sig i en af to tilstande nemlig at være enten "tændt" eller "slukket" (aktiveret/ikke-aktiveret). I komplicerede tilfælde kan flere mellemtilstande forekomme. En neuron tændes, når den påvirkes eller stimuleres i passende grad, og vil som følge af påvirkningen udsende et signal (man siger, at den *fyrer*). Da neuronerne er indbyrdes knyttet til hinanden via synapser (nervetråde), udgør de samlet et netværk af forbundne kar, der kan overføre signaler fra en neuron til en eller flere andre neuroner.

Et neuralt netværk starter med et sæt neuroner, i tekstanalyse-tilfældet med én neuron for hvert ord, svarende til et "sanseindtryk" fra den tekst, som det neurale netværk læser. Det neurale netværk "sanser" n ord ad gangen. I CATPAC sættes n til værdien 7, svarende til det antal ord, som et menneske i følge den kognitive psykologi menes at kunne overskue i et enkelt "view". Jfr. Miller (1956, 81-97) Man husker i snit højst 7 parametre ved et produkt. Forbindelserne mellem neuronerne skabes ved at et *scanning window* ruller henover teksten i læseretningen. Vinduet indeholder altså først ordene nr. 1 til nr. 7 i teksten, dernæst ordene nr. 2 til nr. 8, osv. Når et ord optræder i vinduet *aktiveres* dets *neuron*, og når to eller flere ord optræder i samme vindue etablerer netværket en forbindelse eller association mellem ordene. For hvert vindue, hvori et ord ikke optræder, mister det lidt aktivering, omvendt forstærkes aktiveringen desto oftere et ord forekommer. Neuronernes associationsvægte kan være negative eller positive. Positive når ordene ofte forekommer sammen i teksten, og negative mellem ord, som sjældent eller aldrig forekommer sammen i teksten.

Resultatet af netværkets læsning af teksten udmøntes i en symmetrisk associationstabel, hvor hvert ord optræder engang i hver søjle og hver række. Tabellen er nærmest at sammenligne med en korrelationsmatrice (en varians/kovariansmatrice) eller mere generelt en dis-similarity/similarity matrice - velkendte begreber fra markedsanalysen. Denne tabel kan så underkastes sædvanlig statistisk analyse. I CATPAC anvendes klyngeanalyse og multidimensional skalering til at illustrere hvad associations-matricen kan bruges til. Men den kan sagtens importeres til de gængse statistiske programpakker og fungere som input til andre multivariate analysemetoder.

Forberedelse af tekst til kvantitativ analyse: Filglatnings-processen

Før man overhovedet giver sig i lag med CATPAC må det helt sikkert anbefales, at man har gjort et grundigt forarbejde. Det kommer nemlig ikke noget godt ud af blot

at indlæse en tekst (endsige en med et avanceret filformat). Det anbefales, at man gennemløber en forberedelsesfase, svarende til det, som fremover kaldes *fil-glatnings-processen*. Det tilrådes også at man i første omgang holder sig til tekster af en overskuelig længde, dvs. allerhøjst to sider, men helst ikke mere end en side. Ved en gruppediskussion, hvor mødelederen skifter spørgsmål eller emne fx ti gange i løbet af seancen anbefales, at man nøjes med et enkelt mere eller minde sammenhængende emne som "rådata" til sin analyse. Man kan så om nødvendigt gentage processen flere gange og foretage en særkørsel for hvert emne-afsnit.

Kort om fil-glatnings processens faser: Allerførst skal man transformere dvs. global-erstatte de danske tegn (æ til ae ø til oe og å til aa), eliminere orddelinger, holde sig til en record-længde på maksimalt 80 karakterer per linie samt fjerne andre interpunktionstegn end komma og punktum. Da programmet ignorerer tal er man i givet fald nødsaget til at verbalisere dem (fx omskrive Q 10 til QTI). Herefter kan teksten konverteres til ASCII-tekst med extension *.txt*.

1. Man skal nu først og fremmest have defineret et antal keywords, som man er interesseret i. Eksempler på keywords kunne være mærkenavne, relevante stednavne, holdnings-betonede ytringer og ord, der udtrykker præference.
2. Dernæst skal man have "ensrettet" vigtige ord, der minder meget om hinanden. Et typisk eksempel er samme ord i ental, flertal, bøjninger og evt. som verbum (barn, boerns, barnlig etc.). Her skal man så vælge en version, der anvendes som gennemgående standard, fx {boern}[4]. Det foreslås at ord, der i teksten ikke forefindes i standardformen, markeres ved at skrive det sidste bogstav stort. Stod der oprindeligt {barns} eller {boernene}, da indgår ordet som {boerN} i analysen[5].
3. Teksten skal renses for *homonymer*. Det kunne være en relevant oplysning, når et besøg i Tivoli opfattes som en{dyr} sag. Nu kan det imidlertid forekomme, at der i samme diskussion om børne-ferier falder en bemærkning om {dyr} i Zoologisk Have.
4. Der skal checkes for synonym-relationer. {boern}, {unger}, {mine døtre} o.lign. anvendes ofte i flæng. I sådanne situationer skal man enes om en gennemgående standard. Det foreslås, at man i de tilfælde, hvor keywords substituerer et synonym, markerer dette ved at man staver ordet ved skiftevis at anvende store og små bogstaver. Står der fx {unger}, skrives dette som {BoErN}.
5. Man skal kontrollere for negationer som {ikke}, {aldrig} og {hader}. Sådanne ord er typisk uinteressante i sig selv, hvorimod de kan "ompolere" betydningsindholdet af de keywords, i hvis tekstuelle omverden de befinder sig. I sådanne tilfælde er det nødvendigt at neutralisere de implicerede keywords.
6. Til sidst skal man være klar over betydningen af pronominer. Skal man substituere {boern} i alle de tilfælde, hvor {de} grammatisk set kan føres tilbage til {boern}? Det er indtil videre valgt at se bort fra en sådan form for substituering.

For en langt mere detaljeret redegørelse for filglatingsprocessens faser henvises til

Schmidt (1998a, siderne 21-30 og 46-59).

Case eksempel med uddrag af fokus gruppe for rugbrød

I tabel 2 ses en lille og relativt overskuelig tekst. Det drejer sig om et uddrag fra en gruppediskussion blandt 8 husmødre i alderen 20-49 år med børn i husstanden, der alle var kunder hos en af de store detailhandelskæder. Interviewet foregik for nogle år siden i en dansk provinsby[6].

Den lille gruppediskussion består af præcist 173 ord. Af disse ord staves 108 på forskellig vis (Dvs. at der forekommer 108 "unique words"). Først skal man nu have foretaget en række smårettelser: " f.eks." skal udskrives, idet det ellers læses som to ord, {f} og {eks} hvilket øger forvirringen. {mave-/tarmfunktioner} kan sammenskrives til et ord{mavetarmfunktioner}. Tankestregen i [03], omdannes til et komma og spørgsmålstegnet i [07] bliver til et punktum. Citationstegnene i [07] udgår. - De sidstnævnte ændringer mht. tegn er ikke alle strengt nødvendige. Men det har i løbet af arbejdet med softwaren vist sig, at sådanne tiltag letter kørsler såvel som tolkning betydeligt. Tal kan godt stå fx i kantede parenteser til identifikation af den pågældende tekststreng eller record. Dem ignorerer programmet. I øvrigt gås der slavisk frem efter de ovenfor skitserede skridt i fil-glatnings-processen.

I det foreliggende tilfælde kunne det interessere på hvilken måde franskbrød og rugbrød vurderes ud fra et sundhedsmæssigt synspunkt, hvad man synes om smagen, og hvornår man spiser hvad. Det foreslås derfor, at der på nuværende tidspunkt oprettes en slags log-bog eller protokol,

Tabel 2: Uddrag af en gruppediskussion mht. brød.

[01]	Min bedre halvdel køber næsten altid brød med birkes.
[02]	Vi får kun lyst brød og rundstykker søndag morgen. Det er vist ikke sundt.
[03]	Vi har næsten altid seks forskellige brød hjemme - mest på grund af familiens kræsenhed.
[04]	Franskbrødet, det usunde, spiser vi om morgenen. Ellers kun groft til madpakken og senere på dagen.
[05]	Vi spiser kun det grove, alene på grund af smagen.
[06]	Kernerne giver en god smag til brødet.
[07]	Rugbrød? Det står for mig som "noget godt", f.eks. sammen med ost. Og så er det sundt.
[08]	Den rigtige rugsmag er vigtig.
[09]	Udviklingen går mod grovere og grovere sorter. Man er blevet mere bevidst om, at fuldkornsbrød er godt for mave-/tarmfunktioner.
[10]	Det er sværere at få den ældre generation til at spise groft.
[11]	Jeg er helt tryg, når rugbrødet er fra Brugsen. Så er det gennemprøvet og gennemanalyseret. Det er vigtigt for mig.
[12]	Børnene er langt mere kostbevidste. Rugbrødet opleves af dem som sundere.
[13]	Man kan overhovedet ikke undvære det grove med kernerne.
[14]	Bagerne kan slet ikke bage den slags. Det smuldrer.

Tabel 3: Logbog over ord-varianter, der indgår som keywords i CATPAC inputfilen.

Oprindelige forekomster (umiddelbart efter ASCII-konvertering)	Identifikation	Glattet ord, (- = uændret)
[soendag] morgen morgenen	02 04	morgen morgeN
lyst broed franskbroedet	02 04	FrAnSkBrOeD franskbroeD
rugbroed rugbroedet fuldkornsbroed	07 11,12 09	- rugbroeD RuGbRoEd
groft, det grove (med kernerne) broed (med birkes) grovere og grovere sorter	04,10 05,13 01,03 09	 # udgår!
smagen smag rugsmag	05 06 08	smaG - smaG
god godt	06 07,09	godT -
sundt sundere ikke sundt det usunde	07 12 02 04	sunD sunD usunD usunD

hvori det på detaljeret vis noteres, hvilke keywords plus afarter (inkl. synonymer) man er interesseret i, samt *hvordan* de kommer til at indgå i den endelige version af teksten, der udgør input-filen til CATPAC. Se i den forbindelse tabel 3, der udgør koblingen mellem tabel 2 og tabel 4 forneden. Tabel 4 indeholder teksten i præcis den version, i hvilken den indlæses i CATPAC.

Tabel 3 sammenfatter 16 forskellige sekvenser og "glatter" dem til i alt 7 forskellige keywords {morgen, franskbroed, rugbroed, smag, godt, sund, usund}. Efter moden overvejelse er det beslutet ikke at substituere {groft}, {broed med birkes} og lignende med {RuGbRoEd}.

Det skyldes, at koblingen til {rugbroed} ikke forekom som "hævet over enhver tvivl". Derimod er vurderingen, at det lige akkurat var acceptabelt at substituere {lyst broed} med {FrAnSkBrOeD} og klart forsvarligt at erstatte {fuldkornsbroed} med {RuGbRoEd). Erfaringen tilråder her at anlægge et konservativt kriterium mht. nærhed af sproglige varianter. Men analytikeren må fra sag til sag afgøre, om der bør anlægges et mere liberalt skelnemærke.

Når man har gennemført disse ændringer er man fremme ved tabel 4, en version af teksten, der er parat til at blive indlæst i CATPAC. Den indeholder nu 170 ord (100 unique words). Af disse samles interessen om præcis 7 unique keywords, der samlet forekommer 18 gange i teksten. De resterende 152 ord (herunder 93 unique words)

Tabel 4: Glattet version af gruppe-diskussion: Input-fil til CATPAC.

[01] Min bedre halvdel koeber naesten altid broed med birkes.
[02] Vi faar kun FrAnSkBrOeD og rundstykker soendag morgen.
Det er vist usunD.
[03] Vi har naesten altid seks forskellige broed hjemme, mest paa grund af familiens kraesenhed.
[04] FranskbroeD, usunD, spiser vi om morgeN.
Ellers kun groft til madpakken og senere paa dagen.
[05] Vi spiser kun det grove, alene paa grund af smaG.
[06] Kernerne giver en godT smag til broedet.
[07] Rugbroed. Det staar for mig som noget godt, for eksempel sammen med ost.
Og saa er det sunD.
[08] Den rigtige smaG er vigtig.
[09] Udviklingen gaar mod grovere og grovere sorter. Man er blevet mere bevidst om, at RuGbRoEd er godt for mavetarmfunktioner.
[10] Det er svaerere at faa den aeldre generation til at spise groft.
[11] Jeg er helt tryg, naar rugbroeD er fra Brugsen. Saa er det gennemproevet ´ og gennemanalyseret. Det er vigtigt for mig.
[12] Boernene er langt mere kostbevidste. rugbroeD opleves af dem som sunD.
[13] Man kan overhovedet ikke undvaere det grove med kernerne.
[14] Bagerne kan slet ikke bage den slags. Det smuldrer.

er derimod uden interesse. Dette meddeles CATPAC ved at indtaste eller indkopiere disse 93 ord i en såkaldt excludefil (med extension .exc)[7]. På den måde sikres, at CATPAC ignorerer disse ord[8].

Resultater

Tabel 5 viser resultatet af ordtællingen, akkompagneret af en clusteranalyse foretaget med udgangspunkt i de syv definerede keywords.

I venstre kolonne er ordene opført på en måde, der minder om det fra de hierarkiske clusteranalysemodeller så velkendte dendrogram. Det fremgår tydeligt af figuren, at der tale om to klynger, hvoraf den ene {franskbroed, morgen, usund, smag} forekommer mere fasttømret eller udkrystalliseret end den anden {godt, rugbroed, sund}.

I tabel 6 vises en tabel, der svarer til en kovarians- eller korrelationsmatrice, som den kendes fra traditionelle analyser. Der noteres en klar sammenhæng mellem {franskbroed}, {morgen} og {usund}. Den anden kobling som sås i tabel 5 genfindes ikke her. Men det skyldes formentligt netop at koblingen er svagere. Et næste na-

Tabel 5: Frekvens- og clusteranalyse af teksten i tabel 4.

Centroid method	Frekvens	Pct.
franskbroed . . . . .<<<<<	2	11%
morgen . . . . . .<<<<<<	2	11%
usund . . . . . .<<<<<<	2	11%
smag . . . . . .<<<<	3	17%
<		
godt . . . . . .<<<	3	17%
rugbroed . . . . . .<<<	4	22%
sund . . . . . .<<	2	11%
I alt	18	100%

Tabel 6: Associationsmarice over de syv keywords.

	franskbroed	godt	morgen	rugbroed	smag	sund
godt	-.37					
morgen	.97	-.35				
rugbroed	-.41	-.22	-.40			
smag	.30	-.20	.32	-.25		
sund	-.40	-.31	-.40	-.28	-.33	
usund	.98	-.36	.99	-.41	.32	-.41

turligt led i analyseprocessen ville være at anvende tabel 6 som input til en faktoranalyse (ikke vist her).

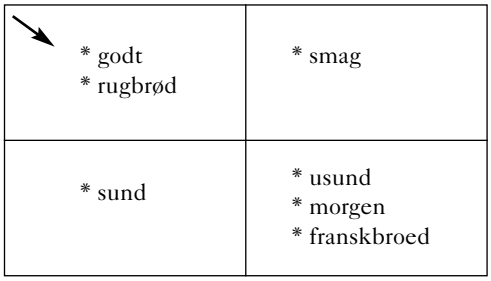
Tabel 7 viser resultatet af en såkaldt ORESME-interactive clustering, der er en del af CATPAC.

Tabel 7: Interaktiv clusteranalyse baseret på neuralt netværk (Threshold level =.15).

Keywords	franskbroed	rugbroed
godt		X
morgen	X	
smag	X	X
sund		X
usund	X	

Det noteres, at neuronet {franskbroed} aktiverer neuronerne {morgen}, {smag} og {usund}, hvorimod {rugbroed} aktiverer {godt},{smag} og {sund}. Nu skal de tekniske koefficienter vælges med omhu og man skal være meget forsigtig og kritisk i sin tolkning. Men det skal ikke meget fantasi til at forestille sig potentialet af denne særanalyse. Man kunne tænke sig genusprodukterne {franskbroed} og {rugbroed} erstattet med 2-3 mærkevarer. Det kunne være virksomhedens egne og de nærmeste konkurrenters mærker - og så kunne de øvrige keywords være begreber (værdier, vurderinger, opfattelser, ytringer o. lign.), som man er interesseret i at få relateret til mærkevarerne i analysen. Der findes gode eksempler på (under udgivelse andetsteds), at således etablerede koblinger mellem mærker og begreber, hvis de tolkes kreativt og med omtanke, på direkte vis kan udmøntes i form af forbedret skræddersyet reklamestrategi, for nu bare at tage et eksempel.

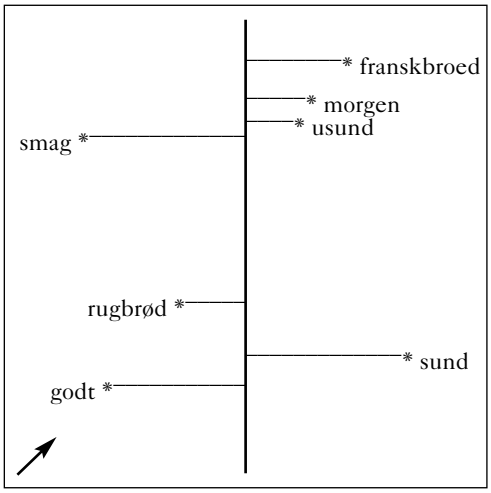
Figur 1: Positioneringsanalyse (brødcasen).



Figur 1 gengiver det todimensionale MDS-plot med udgangspunkt i brød-eksemplet.

Påny ses en klar association mellem {usund}, {morgen} og {franskbrød} medens forbindelsen mellem {rugbroed}, {godt} og især {sund} forekommer mindre klar. Dette understreges ved at zoome ind på 3. dimensionen. Se Figur 2. Koblingen mellem anskuelsesvinklen for de to figurer kan etableres på følgende måde: Hvis man bevæger sig ned på horisontalniveau og derpå skuer fra pilen i det øvre venstre hjørne af Figur 1 og ud i rummet, så ligger ordene nogenlunde fordelt på en højre venstreskala som når man gentager samme øvelse, men dengang startende fra pilen i det nedre venstre hjørne af Figur 2.

Figur 2: Ordenes beliggenhed på 3. aksen.



Figur 2 understreger atter den tætte association mellem {franskbroed}, {morgen} og {usund}. Det noteres, at den tætte association mellem {rugbroed} og {godt} svækkes lidt, når man ser på tredjeaksen. Det bliver endvidere klart, at {sund} ligger langt fra disse to ord i denne dimension[9].

Konklusion

Den lille analyse byggede på 18 forekomster af 7 keywords. Det er let og overskueligt. Men en real-life gruppediskussion bygger på ca 500 forekomster af 30-40 keywords, fordelt over 10-15 sider. I et sådant tilfælde er det særdeles kompliceret at få et holistisk overblik over problemet.

Den ene af forfatterne har anvendt programmet til en omfattende analyse af to gruppediskussioner omhandlende et turistkatalog. Analysen viste en række forhold, der intuitivt kan forekomme nærliggende, men som ikke så let lader sig underbygge: Kvinderne nævner i snit ord, der relaterer til børn som {barnets},{unger}, {datter}, {børneferie} dobbelt så ofte som mændene. De i alt 5 mænd med småbørn (0-6 år) anvender ikke et eneste af disse børnerelaterede ord i den ca $2^1/_2$ timers lange diskussion, medens de 6 kvinder med børn i alderen 7-13 år nævner dem 25 gange (pronominer undtaget). En sådan forskel er højsignifikant i en χ^2-test. Kvinder anfører også dobbelt så mange attraktioner {Legoland}, {BonBonLand}, {Kattegatmuseet} osv. som mænd. I modsætning til mænd nævner de mange dyr (fx {hajer}, {søløver} og {heste}). Den multivariate analyser viste tillige, at det standardiserede keyword {boern} i analysen af specielt kvindernes ytringer klyngede sig til en række ord, der betyder noget for børn som {Tivoli} og {Sommerland}. Mere derom i Schmidt (1998b).

Generelt set kan man forestille sig mange anvendelsesområder for den form for analyse, som her introduceres. Input-brødteksten kunne være udskriften af den årlige medarbejdersamtale, analysens fokus kunne være navngivne ansatte og blandt keywords kunne indgå termer, der fortæller noget om samarbejdsrelationer i afdelingen. Et managementfirma kunne som input-fil anvende udskrift af en jobsamtale, en politolog kunne tage udgangspunkt i indlæggene under behandlingen af et punkt i folketingssalen (den kan nemt downloades fra Folketingets web) osv. CATPAC kan anvendes til analyse af svar fra åbne spørgsmål (se Wassmann 1992). Men da en fil med svar fra åbne spørgsmål indeholder records, der er uafhængige af hinanden er følgen, at en af ANN-algoritmens klare forcer, nemlig at kunne identificere et mønster ud fra en sammenhængende informationsstrøm, ikke udnyttes fuldt ud.

Summary

The article describes a qualitative analysis process and how this process can be supported by the use of computers. New advanced software for text analysis, based on a neural network, is presented. The article gives instructions on how a text should be prepared prior to being subjected to a quantitative analysis. How the program works is being validated by means of an empirical example. In conclusion, the authors describe their experience in the use of the method for complete group discussions.

Noter

[1] EDB-programmer til analyse af kvalitative data går under fællesbetegnelsen CAQDAS (Computer Assisted Qualitative Data Analysis Software)

[2] For en grundig gennemgang af programmerne henvises læseren til Weitzmann and Miles (1995). Forfatterne diskuterer de fleste af programmerne i tabel 1 på den måde, at hvert program helliges et detaljeret review-kapitel. Det er selvfølgelig et problem, at deres bog efterhånden er næsten fem år gammel, og derfor ikke har de 2-3 nyeste versioner af programmerne med.

[3] Det er en udvikling ikke ulig den, som man har kunnet se med hensyn til de programmer, der indgår i de kendte forretingspakker.

[4] Fremover sættes et ord i {sådanne} parenteser, i de tilfælde, hvor de opfattes som et "datapoint".

[5] Dette trick er nødvendig, idet ASCII-formatet ikke tillader egentlige typografiske virkemidler til en lettere identifikation som kursiv, understregning, skyggeskrift o.lign. Erfaringen viser nemlig det praktiske ved, at man hele tiden har et overblik over, om det pågældende keyword optræder i den form, hvori det oprindeligt figurerede i teksten, eller om det som led i processen er erstattet med en anden grammatisk form respektive med et synonym. En sådan fremgangsmåde kan forekomme pertentlig. Men når man analyserer tekst er det bydende nødvendigt at man anlægger en ambitiøs præcisions-tærskel. Der tolereres ingen som helst slinger i valsen! For hvis man ikke opretholder en entydige kobling mellem den oprindelige tekst samt den, der anvendes til analysen, risikerer man alt for let at miste overblikket. Derfor anbefales, at man opretter en "logbog", der fører nøjagtig kontrol og registrerer hver eneste manipulative ændring, som foretages i forhold til den oprindelige tekst. Nærmere derom forneden og i Schmidt (1998a).

[6] Da der i den tilgrundliggende tekniske rapport ikke er registreret en personkode, har det ikke været muligt at koble ytringer (tallet i kantet parentes) med subjekter. Der deltog kun 8 respondenter, og derfor vil nogle af koderne vedrøre samme respondent. Man ved bare ikke hvem der hører hvortil. Men det er også ganske ligegyldigt i det foreliggende tilfælde. Sagen er i let modificeret form optryk i Hollensen og Schmidt (1998).

[7] Bemærk, at programmet, hvis den ikke for anden besked, indlæser den engelske default-fil. Den diskvalificerer så automatisk en række ord, der er uinteressante på engelsk, og som har et andet, men ikke nødvendigvis uinteressant betydningsindhold på dansk som {and}, {by}, {gave} o. lign.

[8] Helt konkret bygger CatPac's analyse på præcis 18 forekomster af følgende 7 ord {FrAnSkBrOeD morgen usunD fransbroeD usunD morgeN smaG godT smag rugbroed godt sunD smaG RuGbRoEd godt rugbroeD rugbroeD sunD}. Hverken mere eller mindre.

[9] Man kunne ganske vist have valgt en logaritmisk skalering, som i det foreliggende tilfælde i tre og især i to dimensioner på dramatisk vis trækker {godt}, {rugbroed} og {sund} sammen. Optisk set opnås da to "rene klynger", der ligger i hver sin ende af to modstående kvadranter og med {smag} omkring origo. Men en sådan projektion snyder, rent bortset fra at man i dette tilfælde kan være i tvivl om en todimensional projektion er udtryk for det rigtige valg af dimensionalitet.

Litteratur

Catterall, Miriam and Pauline Maclaran: Using Computer Software for the Analysis of Qualitative Market Research Data. *Journal of the Market Research Society*, 40(3): 207-22, 1998.

Evans, T: Analysis and interpretation in S. Robson and A. Foster (Eds.) *Qualitative Research in Action*. London: Sage, 1989.

Fielding. N .G., and R. M. Lee: *Using Computers in Qualitative Research*. London: Sage, 1991.

Grunert, Klaus, and Margarete Bader: "A Systematic Way to Analyse Focus Group Data", *EMAC Proceedings* (Helsinki): 825-40, 1986.

Helmersson, Helge: "Metodestudier av konsumentpreferenser - Intuitiv textanalys jämförd med Pertex-analys." *Working paper*. Lund, Sweden: University of Lund, 1997.

Hollensen, Svend og Marcus Schmidt: *Scener fra dansk erhvervsliv*. 3. ed. Nyt Nordisk Forlag Arnold Busck, 1998.

Kelle, Udo: *Computer-Aided Qualitative Data Analysis*. London: Sage, 1995.

Leung, Josef, and Ching-Long Yeh: "Natural Language Processing - Verbatim Text Coding and Data Mining Report Generation." *ESOMAR Proceedings* (Edinburgh): 393-409, 1997.

Masson, Egill and Yih-Jeou Wang: Introduction to Computation and Learning in Artificial Neural Networks. *European Journal of Operations Research*, 47:1-28, 1990.

Miles, Matthew B: Qualitative Data as an attractive Nuisance: The Problem of Analysis. *Administrative Science Quarterly*, 24:590-601, 1979.

Miles, Matthew B., and A. Michael Huberman: *Qualitative Data Analysis*. Thousand Oaks, Cal.: Sage, 1994.

Miller, G. A.: The Magic Number Seven, Plus and Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review*, 63: 81-97, 1956.

Moore, Karl, Robert Burbach, and Roger Heeler: Using Neural Nets to Analyze Qualitative Data. *Marketing Research: A Magazine of Management and Applications*, 7(1): 35-9, 1995.

Pfaffenberger, B.: *Microcomputer Applications in Qualitative Research*. Newbury Park: Sage, 1988.

Popper, Karl R. *Conjectures and Refutations*. London: Routledge and Kegan Paul 1963.

Rasmussen, Knud Erik.: "Neurale netværk som beslutningsstøtteværktøj". *Ledelse og Erhvervsøkonomi* 59 (Januar): 57-68, 1995.

Robson, S., and A. Hedges: Analysis and Interpretation of Qualitative Findings. *Journal of the Market Research Society*, 35 (1):22-7, 1993.

Schmidt, Marcus: *Kvantitativ analyse af kvalitative data (herunder især gruppediskussioner)*. Handelshøjskole Syd 1998a.

Schmidt, Marcus: Quantitative Analysis of Qualitative Interviews: Theoretical Considerations and Empirical Analysis, *AMA Educator's Proceedings:* 168-77. 1998b.

Schmidt, Marcus: Kvantificering af tekst. *21. Sumposium i anvendt Statistik*, København: SFI, 247-158. 1999a. (www3.hhs.dk/~marcus/DKpapers/sympos99/sympos99.htm).

Schmidt, Marcus: Multivariate Analysis of Focus Group Interviews. *EMAC Proceedings* (Berlin). 1999b. (www3.hhs.dk/~marcus/GBPapers/EMAC99/EMAC99.htm).

Tesch, R.: *Qualitative Research: Analysis Types and Software Tools*. London: Falmer Press, 1990.

Wassmann, David. A.: "Using Catpac to read Qualitative data," Paper presented at *The Advanced Research Techniques Forum*, Lake Tahoe, NV: American Marketing Association, 1992.

Weitzman, E. A., and M. B. Miles.: *Computer Programs for Qualitative Data Analysis*. Thousand Oakes, Cal.: Sage, 1995.

White et al.: *Artificial Neural Networks: Approximation and learning Theory*. Cambridge, MA: Blackwell Publishers, 1992.

Woelfel, J.: "Artificial Neural Networks in Policy Research: A Current Assessment." *Journal of Communication* 43(1): 63-80, 1993.

Woelfel, J., and E. L. Fink.: *The Measurement of Communication Processes: Galileo Theory and Method*. New York: Academic Press, 1980.

Woelfel, Joseph and Nick Stoyanoff: "CATPAC: A Neural Network for Qualitative Analysis of Text." *Working Paper*. Buffalo, NY: University of New York at Buffalo, undated.

Bemærk, at tre af de i litteraturlisten nævnte kilder er tilgængelige i fuld tekst og kan downloades fra www3.hhs.dk/~marcus. Den præcise adresse, hvorfra de pågældende artikler kan hentes, er anført foroven i litteraturlisten sammen med (umiddelbart efter) den pågældende kilde.