artikkelit

ESA REUNANEN & RISTO SUIKKANEN

Latent Class Analysis: samoilua latentissa tilassa

Latent Class Analysis (LCA) on määrällinen metodi, joka on kehitetty laatueroasteikollisia muuttujia käyttävän tutkimuksen tarpeisiin. Paul F Lazarsfeld kehitti sen peruspiirteet jo 1950-luvulla, mutta menetelmän tehokkaan käytön edellyttämät algoritmit kehitettiin vasta paria vuosikymmentä myöhemmin. Sisällönanalyysin välineenä LCA:ta on ryhdytty käyttämään 1990-luvun mittaan.

Artikkeli pyrkii johdattamaan LCA:n perusperiaatteisiin sekä antamaan käytännön esimerkkejä sen suorittamisesta. Samalla tehdään myös rakennekuvausmenetelmiä koskevia alustavia metodologisia huomautuksia. Vaikkakin vaikuttaa siltä, että LCA on mielenkiintoisella tavalla yhdistettävissä laadulliseen sisällönanalyysiin sen suhde laadulliseen typologisointiin tai etäisyysindikaattorien käyttöön perustuvaan ryhmittelyyn on jännitteinen.

Kuten joukkoviestintätutkimuksen historiaan kohdistuvissa tai sitä sivuavissa esityksissä usein todetaan, Bernard Berelsonin 1952 julkaistu *Content analysis in communication research* sai klassisen aseman sisällönanalyysina tunnetun tutkimustekniikan itseymmärryksen kirjaajana. Berelson (1952, 18) määritteli tekniikan siten, että se on "kommunikaation (eli viestinnän) ilmisisällön objektiivista, systemaattista ja määrällistä kuvailua varten soveltuva tutkimustekniikka". Termi 'määrällinen sisällönanalyysi' on kiteytynyt kuvaamaan mainittuihin määreisiin kiinnittyvää tutkimusperinnettä.

Berelson oli tietysti hyvin tietoinen myös 'laadullisen' sisällönanalyysin olemassaolosta. Mainitussa teoksessa on asiaa koskettelevia luku (1952, 114 ja alkaen). Käyttäessään termiä 'laadullinen sisällönanalyysi' Berelson ympäröi sanaparin ensimmäisen osan aina johdonmukaisesti lainausmerkeillä (kyseessä oli *"laadullinen" sisällönanalyysi).* Berelsonin suhde tällaiseen tutkimustapaan ei silti ollut erityisen vihamielinen. Sillä oli käyttöä – kuten Pietilä (1997, 152) toteaa – vähintäänkin hypoteesien muodostamisessa. Berelsonia ilmeisesti vaivasi vähäiseen formalisaatioon ja kvantifiointiin liittyvä argumentatiivinen epämääräisyys mutta toisaalta hän oli kyllä valmis sanomaan yhtä ja toista myönteistä tästä kaikenlaiseen vanhaan mutta osin arvovaltaiseenkin ryönään – kirjallisuuskritiikkiin, filosofiaan, retoriikan tutkimuksen ja ties mihin – liittyvästä tulkintataiteesta. Lisäksi hän oli valmis (1952, 133-134) pohtimaan teitä lähestymistapojen yhdistämiseksi.

Määrällisen sisällönanalyysin herättämän keskustelun ja kritiikin avainmääreitä olivat "määrällisyys" ja "ilmisisältö". On ehkä paikallaan todeta, että mainitut määreet esittäessään Berelson samalla rajasi (emt., 13-20) kahdella tapaa metodin sovellusalaa. Tekniikkaa saattoi soveltaa vain kohteisiin, joissa "sisällön yksiköillä" oli sama painoarvo, ts. "määrällistä analyysiä ei voi käyttää, mikäli yksi sana tai lause on yhtä 'tärkeä' kuin koko muu sisältö yhteensä" (emt., 20). Sinänsä numeraalinen kuvailu ei kylläkään ollut välttämätöntä, esitys saattoi tukeutua myös ylimalkaisempiin määrällisiin ilmaisuihin.[1]

Vaatimus rajoittumisesta ilmisisältöön taas johti rajaamaan käyttöalueen kohteisiin, joista voitiin olettaa, että oli olemassa "common universe of discourse" yleisön ja viestijän välillä, mikä takaisi sen, että viestijän lataamat ja yleisön purkamat merkitykset olisivat kutakuinkin yhtenevät. Esimerkkinä hän mainitsi junaonnettomuudesta kertovan uutisen ja asetti toiseen ääripäähän tuolloisesta perspektiivistä 'modernina' pidetyn runouden, jota ei ollut syytä ottaa sisällönanalyysin kohteeksi ainakaan siihen sisältyvien merkitysten kannalta.[2] Berelson tuskin millään muotoa kielsi sitä mahdollisuutta, että uutinen junaonnettomuudesta saattaisi generoida tusinoittain muita merkityksiä ("rautatielaitos on huonosti hoidettu", "elämä on alati uhattua"). Hän lienee vain halunnut rajoittua siihen, mitä piti kaikille yhteisenä. Tässä tapauksessa se olisi viesti siitä, että on tapahtunut junaonnettomuus. Berelson (emt., 16) halusi rajauksella tehdä nimenomaista pesäeroa pragmatiikkaan eli rajata pois sen selvittelyn, mitä muuta jotain sanomalla mahdollisesti sanotaan.

Klassisen aseman saanut reaktio Berelsonin kirjaan oli Siegfried Kracauerin artikkeli *The Challenge of Qualitative Content Analysis*, (alkuperäinen julkaisupaikka oli Public Opinion Quarterly 52/53, nr. 4). Kracauer maalaili aluksi näkymän, jonka mukaan määrällinen analyysi oli tyypillisimmässä muodossaan toteutunut viestinnän "puolesta" tai "vastaan" suuntautuvan sisällön selvittämiseen. Tällöin kohteen sisältö oli ahdettu erilaisia jatkumoja pilkkoviin luokitusskaaloihin. Enemmän tai tai vähemmän monimutkaisten kokonaisuuksien pilkkominen elementteihin ei Kracauerin mukaan välttämättä ollenkaan taannut tarkkaa analyysia, sillä tutkijoiden tällä tavoin tuottama...

> sangen atomistinen data tuskin tekee mahdolliseksi saada selkoa datassa vallitsevista suhteista. On huomionarvoista, että usein juuri nämä keskinäissuhteet ovat koko tekstin suuntautuneisuuden kannalta ratkaisevia. Jokainen tulkintaa harjoittanut, olipa hän hahmopsykologian kannalla tai ei, tietää, että viestintä saa rakentumisensa perusteella usein 'suunnan', joka poikkeaa sen elementtien luetteloinnin ilmaisemasta suunnasta (1990, 340).

Artikkelinsa lopussa Kracauer (emt., 350) vielä toisti, että kohdetekstit ovat kokonaisuuksia eivätkä faktakasoja. Niiden elementit...

> menettävät sisältönsä, mikäli sisältö otetaan kirjaimellisesti ja revitään irti siitä viittausten ja kietoutumisten verkosta johon se kuuluu. Se on olemassa vain tässä kudoksessa ja tämän kudoksen myötä. Kudos on vielä sirpaleinen elämänilmaus. Se vaatii vastakaikua kehkeytyäkseen täyteen mittaansa. Viestinnässä ja

viesteissä ei niinkään ole kyse lukkoon lyödyistä entiteeteistä kuin monitulkintaisista (ambivalent) haasteista. Ne haastavat lukijan tai analysoijan ottamaan vastaan ja reagoimaan. Vain lähestymällä näitä kokonaisuuksia koko olennollaan kykenee analysoija keksimään ja määrittämään merkityksen – tai jonkin niihin sisältyvistä merkityksistä – ja auttamaan täten niitä toteuttamaan itsensä.

Kracauerin ajatus lienee tämän esityksen lukijoille perin tuttu: kohde koostuu elementeistä, joita voi havainnoida ja luetella, mutta näistä elementeistä syntyvä kokonaisuus ei tällä tavoin välttämättä paljastu.

Niin eri suuntaan kuin Paul F. Lazarsfeldin yhteiskuntatutkimusta koskevat ajatukset muuten 1950-luvulla mahdollisesti menivätkin, niillä oli tietyssä mielessä kosketuskohta Kracauerin esittämän kritiikin kanssa. Wienin piirin loogikoiden tuottamiin teksteihin viitannut Lazarsfeld katsoi yhteiskuntatieteellisen tutkimuksen olevan enimmiltään Hempelin ilmaisua käyttäen "tutkimuksen esiteoreettisessa vaiheessa" (Lazarsfeld 1959, 485). Niinpä nyt oli hiottava tutkimusmenetelmiä riittävän empiirisen tietoaineksen tuottamiseksi sekä käytävä niitä koskevaa keskustelua, jotta olisi tarkka selko siitä, miten tutkimus- ja mittausmenetelmät ovat olleet määrittämässä käytettyjä tai tutkittuja teoreettisia käsitteitä. Näin olisi olemassa empiiristen tutkimustulosten ja niiden tuottamisprosessia koskevan tiedon 'varasto', jota "jonain päivänä kehittyväksi toivomamme "teoria" voisi järjestelmällisesti käyttää hyväkseen" (ema. 485).

Varastoimiskelpoisen empiirisen tiedon tuotantoa esti kuitenkin metodisten taitojen ja kykyjen kehittymättömyys. Lazarsfeldin mukaan tutkijat olivat aseettomia ristiriitaisten tai eri suuntiin osoittavien empiiristen indikaattoreiden käsittelyssä ja tulkinnassa. Ne tulisi saattaa mielekkääksi kokonaisuudeksi mutta "on hämmästyttävää miten pahasti tutkijat voivat tässä epäonnistua kun he jättävät erityisasiantuntemuksensa piirin" (ema. 482).

Lazarsfeld itse tähtäsi faktorianalyysin kaltaiseen monimuuttujamenetelmään, joka kuitenkin perustuisi todennäköisyyslaskentaan ja olisi sovellettavissa laatueroasteikollisiin muuttujiin (Lazarsfeld 1950,469). Hän käytti menetelmästä nimeä *Latent Structure Analysis* (LSA), mutta nyttemmin siitä on alettu käyttää nimitystä *Latent Class Analysis* (LCA), mikä johtunee luokkien tietynlaisen laskentatavan omaksumisesta (Andersen 1982, 2).

Lazarsfeld visioi LSA:sta varsin yleistä sosiaalitutkimuksen menetelmää. Ajatuksena oli, että kyselyvastaukset tai muut havainnot muodostavat ns. manifestin heti nähtävän rakenteen. Tutkimuksen varsinainen kiinnostuksen kohde olisi kuitenkin manifestin rakenteen taustalla piilevissä latenteissa tekijöissä, kuten asenteissa tai luonteenpiirteissä, joita ei voi suoraan havaita. Tavoitteena oli LSA:n avulla määrittää manifestien vastausten perusteella vastaajien 'paikka' tällaisella latentilla asenne- tai luonneulottuvuudella. Mikäli manifestien vastausten taustalla on useita latentteja ulottuvuuksia, voidaan puhua 'latentista tilasta', johon vastaajat sijoittuvat eri tavoin. (Lazarsfeld 1950, 1959.)

Ongelma 1950-luvulla oli, ettei menetelmän tehokkaan soveltamisen edellyttämiä algoritmeja ollut käytettävissä. Matemaattiset ongelmat ratkaistiin 1970-luvulla, ja tietokoneiden suorituskyvyn kehittyminen 1980-luvulla olisi periaatteessa tehnyt mahdolliseksi LCA:n yleistymisen, mutta ainakaan toistaiseksi siitä ei ole tullut faktorianalyysin kaltaista laajalti tunnettua ja yleisesti käytettyä metodia. Syistä emme tietenkään voi olla varmoja. Yksi syy lienee menetelmän edellyttämän matemaattisen ja teknisen arsenaalin kehityksen hitaus ja asteittaisuus. Tuona aikana asemansa vakiinnuttanut faktorianalyysi ehti vallata LCA:lle soveliasta maastoa. Myös laadullisten menetelmien kokema noste 1970-luvulta alkaen on ollut omiaan supistamaan uusista määrällisistä menetelmistä kiinnostuneiden määrää.

Lazarsfeldin omista teksteistä ilmenee, että 1950-luvulla menetelmää käytet-

tiin asenne- ja mielipidetiedusteluihin, mm lukija- ja kuuntelijatutkimuksiin. 1980-luvulla LCA:ta on käytetty ainakin jonkin verran kyselyaineistojen analyysiin. 1990-luvulla menetelmää on pyritty käyttämään myös sisällönanalyysin välineenä. Pyrkimykset ruumiillistuvat pitkälti Konstanzin yliopistoon ja sen rauhantutkimusyksikön johtajaan Wilhelm Kempfiin, joka on viitannut mahdollisuuteen tehdä LCA:n avulla Kracauerin kritiikin kestävää tilastollista analyysia (Kempf 1994, 6). Suomessa Ari Heinonen (1997) on kokeillut LCA:n käyttöä kyselyaineiston analyysiin.

Tässä esityksessä yritämme tehdä selkoa LCA:n periaatteesta ja sen käyttämisestä. Käytämme tässä apuna mainitusta Heinosen tutkimuksesta nappaamaamme esimerkkiaineistoa. Yritämme tuoda LCA:n ominaispiirteitä esiin myös vertaamalla sitä muutamiin muihin monimuuttujamenetelmiin. Kirjoituksen lopuksi esittelemme lyhyesti tuoreimpia LCA:lla tehtyjä tutkimuksia ja pohdimme LCA:n käyttökelpoisuutta sisällönanalyysissä.

LCA ja 'paikallisen riippumattomuuden' periaate

LCA:n taustalla siintää Lazarsfeldin lanseeraama 'empiirinen sosiaalitutkimus'. 1920- ja -30-luvun Wienissä perusnäkemyksensä (psykologian harjoittajana) luonut Lazarsfeld oli omaksunut kannan, jonka mukaan sosiaalitutkimuksen tehtävänä – eikä sillä sitten ollut niin väliä, minkä tieteenalan piirissä tutkimusta harjoitettiin – oli tutkia ja ymmärtää inhimillistä toimintaa. Tarvittavat teoreettiset käsitteet (asenne, status, motivaatio...) olivat yleensä välittömän havainnon tavoittamattomissa. Lazarsfeldin intohimona olikin – yhtenäistieteen hengessä – pohdinta, jonka tavoitteena oli havaintomateriaalin tuottamisen, keräämisen ja tulkinnan mahdollisimman kontrolloitu suhde siihen, mistä materiaalin tuli kertoa. Sana 'latentti' nimessä 'Latent Structure Analysis' viitanneekin kahtaalle. Yhtäältä se viittaa teoreettisten käsitteiden kuvaamaan 'latenttiin tilaan' (latent space). Toisaalta se huitoo kohti empiirisessä aineistossa piileviä matemaattisia suhteita, joita aineisto sinällään ei ilman erityisiä laskentaoperaatioita paljasta. Ja lopulta se viittaa nimenomaan näiden kahden (latentin tilan ja matemaattisten suhteiden) kytkentään.

Latentin tilan ja siitä tuotetun aineiston kytkennän kannalta keskeinen käsite on niin sanottu *paikallisen riippumattomuuden periaate*. Periaate on eräänlainen 'jos..niin..' -aksiooma. Sen mukaan havaintoyksiköt ovat samankaltaisia suhteessaan tähän tai tuohon tutkittavaan ominaisuuteen (latenttiin ulottuvuuteen), mikäli ne tuottavat kyseistä ominaisuutta mittaavissa testeissä tilastollisesti riippumattoman jakauman. Asiaa voidaan havainnollistaa tarkastelemalla kahden eri ryhmän vastauksia kahteen kysymykseen.

TAULUKKO 1. Tilastollisesti riippumaton ja riippuvainen jakauma.

Tilastollisesti riippumaton jakauma

		Kys. 2		
		+	–	
Kys. 1	+	75	15	90
	–	15	3	18
		90	18	108

Tilastollisesti riippuvainen jakauma

		Kys. 2		
		+	–	
Kys. 1	+	35	19	54
	–	19	35	54
		54	54	108

Kysymysparin ristiintaulukointi ensimmäisessä ryhmässä kertoo, että 108 vastaajasta 90 on vastannut ensimmäiseen kysymykseen myönteisesti, 18 kielteisesti. Toiseen kysymykseen vastattaessa ovat ensimmäiseen kysymykseen myönteisesti vastanneet hajonneet siten, että 75 on vastannut myönteisesti myös toiseen kysymykseen. Ensimmäiseen kysymykseen myönteisesti vastanneista 15 on taas vastannut kielteisesti jälkimmäiseen kysymykseen. Suhde on siis 75:15. Tarkasteltaessa vastaavaa hajoamista ensimmäiseen kysymykseen kielteisesti vastanneiden keskuudessa, havaitaan, että ryhmä hajoaa samassa suhteessa eli suhteessa 15:3. Tällainen jakauma, jossa edelliseen kysymykseen vastattaessa syntyneet osaryhmät jakaantuvat aina uuteen kysymykseen vastattaessa samassa suhteessa on tilastollisesti riippumaton jakauma. Tällaisen jakauman synnyttävä havaintoyksikköjen joukko on paikallisen riippumattomuuden periaatteen nojalla homogeeninen suhteessa tutkittuun ominaisuuteen.

Jälkimmäisessä ryhmässä annettujen vastausten jakauma ei ole edellä kuvatulla tavalla tilastollisesti riippumaton. Siinä ensimmäiseen kysymykseen on vastannut myönteisesti 54 ja kielteisesti samoin 54. Myönteisesti vastanneiden vastaukset jakautuvat toisen kysymyksen osalta suhteessa 35:19, kun kielteisesti vastanneiden vastaukset jakautuvat suhteessa 19:35. Tässä ryhmässä siis vastaus ensimmäiseen kysymykseen 'vaikuttaa' selvästi siihen, miten toiseen kysymykseen vastataan, ja kyseessä on siten heterogeeninen havaintoyksikköjen joukko.

Jakauman tilastollinen riippuvuus on osoitus siitä, että kysymyksillä (muuttujilla) on jokin tai joitakin 'yhteisiä tekijöitä', jotka erottelevat havaintoyksikköjä toisistaan. Riippuvuus on siis yhtäältä osoitus siitä, että kysymyksillä on jotain yhteistä ja toiseksi siitä, että havaintoyksiköt ovat tämän yhteisen tekijän suhteen erilaisia. Taulukon 1 riippumattomassa jakaumassa kysymyksillä voi hyvinkin olla jotain yhteistä, koska varsin monet ovat vastanneet molempiin myönteisesti. Tämä yhteinen tekijä ei kuitenkaan erottele vastaajia, koska ensimmäiseen kysymykseen myönteisesti ja kielteisesti vastanneet vastaavat samassa suhteessa toiseen kysymykseen.

Tilastollisesti riippumattomassa jakaumassa vallitsee myös sellainen lainalaisuus, että laskennallinen todennäköisyys eri vastausvaihtoehtojen yhdistymiselle vastaa niiden todellista osuutta aineistossa. Niinpä laskennallinen todennäköisyys vastata taulukon 1 ensimmäisen kysymysparin molempiin kysymyksiin myöntävästi on 90/108*90/108 = 0,69 eli täsmälleen molempiin kysymyksiin annettujen myönteisten vastausten todellinen osuus aineistosta 75/108 = 0,69. Toisessa kysymysparissa 'plus-plus'-vastausten todellinen osuus (35/108 = 0,32) on selvästi laskennallista todennäköisyyttä (54/108*54/108 = 0,25) suurempi, mikä siis on osoitus tilastollisesta riippuvuudesta.

LCA:n tavoitteena on jakaa heterogeenisia joukkoja homogeenisiksi osajoukoiksi. Asian havainnollistamiseksi oletamme Lazarsfeldin (1959, 496-500) käyttämää esimerkkiä mukaillen, että taulukossa 1 esitetty heterogeeninen joukko olisi vastannut seuraaviin kahteen kysymykseen:

Kysymys 1: Hallitsevatko isot öljy-yhtiöt liian suurta osaa öljymarkkinoista?

Kysymys 2: Tuhlaako öljyteollisuus luonnonvaroja?

Esimerkissä siis vastaajat ovat havaintoyksiköitä, kysymykset muuttujia ja annetut vastaukset muuttujien saamia arvoja. Myönteisten vastausten voisi ajatella kuvaavan kielteistä suhtautumista ja kielteisten vastausten myönteistä suhtautumista öljyteollisuuteen. Vaikka siis kysymykset tarkkaan ottaen koskevat eri asioita, niiden molempien voi ajatella mittaavan myös yleistä asennoitumista öljyteollisuuteen. Juuri tämä 'yleinen asennoituminen' olisi siis se latentti tekijä, joka selittäisi havaittua riippuvuutta ja jonka suhteen heterogeenisen joukon vastaajat

ovat erilaisia.

Öljyteollisuutta koskeva asenne voidaan ajatella ulottuvuudeksi, jolla kukin vastaaja sijoittuu tiettyyn kohtaan. Asenneulottuvuuden kielteiseen päähän sijoittuvien todennäköisyys vastata kysymyksiin myönteisesti on suuri, kun asenneulottuvuuden myönteiseen päähän sijoittuvien todennäköisyys vastata myönteisesti on pieni. Osa vastaajista taas ei asennoidu öljyteollisuuteen juuri mitenkään, joten heidän vastauksensa ovat yhtä todennäköisesti myönteisiä tai kielteisiä. Aineisto on paikallisen riippumattomuuden periaatteen mukaisesti jaettavissa kolmeen homogeeniseen osajoukkoon taulukon 2 osoittamalla tavalla:

TAULUKKO 2. Heterogeenisen aineiston jakautuminen homogeenisiin osajoukkoihin.

Hallitsevatko isot öljy-yhtiöt liikaa öljymarkkinoita?

Koko aineisto

Tuhlaako öljyteollisuus luonnonvaroja?	+	−	
+	35	19	54
−	19	35	54
	54	54	108

Hallitsevatko isot öljy-yhtiöt liikaa öljymarkkinoita?

Tuhlaako öljyteollisuus luonnonvaroja?	*Joukko 1*			*Joukko 2*			*Joukko 3*		
	+	−		+	−		+	−	
+	1	5	6	9	9	18	25	5	30
−	5	25	30	9	9	18	5	1	6
	6	30	36	18	18	36	30	6	36

Koko aineiston heterogeenisuus johtuu LCA:n logiikan mukaan siitä, että siinä on sekoittuneena erilaisia sisäisesti homogeenisia osajoukkoja. Tämä sekoittuneisuus voidaan purkaa jakamalla aineisto tilastollisesti riippumattomiin osajoukkoihin. Paikallisen riippumattomuuden periaatteen nojalla näiden osajoukkojen pitäisi sitten olla homogeenisia niiden latenttien tekijöiden suhteen, jotka aiheuttivat muuttujien riippuvuuden koko aineistossa.

Taulukossa 2 esitetyt kolme 36 vastaajan joukkoa koostuisivat siten öljyteollisuuteen asennoitumiseltaan samankaltaisista vastaajista. Joukkoon 1 kuuluvien vastaajien asenne on myönteisin, mikä näkyy siitä että kyseisessä joukossa kysymyksiin vastataan todennäköisimmin kielteisesti. Joukko 2 taas koostuu vastaajista, joilla ei ole öljyteollisuuteen selvää asennetta: vastaus on yhtä todennäköisesti myönteinen kuin kielteinenkin. Kolmannessa joukossa puolestaan vastataan todennäköisesti myönteisesti, mikä viittaa kielteiseen asenteeseen.

Puhuttaessa osajoukkojen homogeenisuudesta tarkoitetaan homogeenisuutta tutkittavan latentin (muuttujille yhteisen) tekijän suhteen, ei sitä, että kukin osajoukko koostuisi samalla tavalla vastanneista. Esimerkiksi joukossa 2 on yhtä paljon kaikkia eri vastauskombinaatioita. Silti tämä on osoitus nimenomaan jou-

kon homogeenisuudesta: asenteeltaan epävarmojen tai välinpitämättömien henkilöiden vastausten pitäisikin jakautua tasaisesti kaikille eri vaihtoehdoille. LCA:n yhteydessä puhe homogeenisuudesta tarkoittaa siis homogeenisuutta nimenomaan muuttujia yhdistävien 'latenttien' tekijöiden suhteen, ei välttämättä homogeenisuutta siinä mielessä, että vastaukset olisivat kussakin joukossa mahdollisimman samanlaisia. Tutkitun ominaisuuden suhteen samankaltaiset vastaajat eivät aina tuota samankaltaisia vastauksia Tosin näinkin varsin usein on asianlaita. Esimerkiksi joukoissa 1 ja 3 havainnot kasautuvat hyvin vahvasti tiettyihin vastauskombinaatioihin.

Osajoukoista täytyy todeta vielä se, että LCA:n ryhmittelyn perusteella ei voida varmasti sanoa, mitkä yksittäiset havainnot kuuluvat mihinkin osajoukkoon. Esimerkiksi molempiin yllä oleviin kysymyksiin myönteisesti vastannut henkilö voi kuulua mihin tahansa kolmesta osajoukosta. Silti voidaan sanoa, että hän kuuluu todennäköisimmin joukkoon 3, toiseksi todennäköisimmin joukkoon 2 ja kaikkein epätodennäköisimmin joukkoon 1.

LCA-luokkien muodostaminen käytännössä: esimerkkianalyysi

Tarkastellaan seuraavaksi LCA:n etenemistä ja laskentaperiaatteita konkreettisen esimerkkiaineiston avulla, joksi olemme valinneet suomalaisten päätoimittajien Internet-asenteita koskevan kyselytutkimuksen (Heinonen 1997). LCA:n toiminnan kannalta ei ole väliä, onko kyseessä sisällönanalyysi- vai kyselyaineisto. Esimerkkikyselyssä päätoimittajille esitettiin väittämiä, joista heidän piti sanoa ovatko he samaa vai eri mieltä. Tähän esimerkkiin on valittu kolme väittämää, joista kukin siis muodostaa yhden muuttujan. Kunkin muuttujan arvo voi olla 0 (samaa mieltä), 1 (ei osaa sanoa) ja 2 (eri mieltä):

Väittämä 1: Teknologia kehittyy niin nopeasti, että sanomalehtien on varauduttava jo nyt verkkojulkaisemiseen.

Väittämä 2: Internetin käytön teknisyys ja kalleus rajoittavat vielä pitkään sen suosiota yleisön keskuudessa.

Väittämä 3: Tietokoneruudulta luettavat verkkolehdet eivät kiinnosta suurta yleisöä.

LCA:ssa käytettävä Lacord-ohjelma laskee ensin näiden muuttujien jakaumat koko aineistossa. Tätä kutsutaan LCA-menetelmässä 'yksiluokkaiseksi ratkaisuksi', ja siinä on kyse yksinkertaisesti vastausten suhteellisista jakaumista (taulukko 3). Erona tavallisiin prosenttitaulukoihin on vain se, että Lacord esittää tulokset prosenttien sijaan todennäköisyyksinä, jotka ilmaistaan desimaalilukuina. Todennäköisyys 1 tarkoittaa täyden sadan prosentin osuutta aineistosta, 0,5 tarkoittaa viittäkymmentä prosenttia jne.

TAULUKKO 3. Käsitykset Internet-tekniikan kehityksestä ja sen merkityksestä sanomalehdille.

			Verkkolehtiin varauduttava jo nyt	Teknisyys rajoittaa vielä pitkään	Verkkolehdet eivät kiinnosta
1.CLASS	1.000	*0* samaa mieltä	0.861	0.750	0.556
		1 ei osaa sanoa	0.056	0.056	0.056
		2 eri mieltä	0.083	0.194	0.389

Taulukosta näkyy, että 1-luokan koko on 1.000, mikä tarkoittaa että luokkaan kuuluu 100 prosenttia vastauksista. Koko aineisto siis kuuluu yksiluokkaisessa ratkaisussa samaan luokkaan. Numerot 0 - 2 viittaavat annettuihin vastauksiin eli muuttujien arvoihin. Taulukosta näkyy esimerkiksi, että 75 prosenttia vastaajista on samaa mieltä väittämän kanssa, että "Internetin käytön teknisyys ja kalleus rajoittavat vielä pitkään sen suosiota yleisön keskuudessa". Eri mieltä väittämän kanssa on ollut 19,4 prosenttia vastaajista.

Kokonaisjakaumaa laskettaessa lasketaan samalla tunnusluku, joka kuvaa sitä kuinka hyvä kuvaus kokonaisjakauma on aineistosta edellä esitetyn homogeenisuuden mielessä. Nykyisin käytettävässä laskutavassa tämä tunnusluku on LOG-Like, joka kuvaa koko ratkaisun todennäköisyyttä. LOG-Like puolestaan lasketaan ns. *koodauskuvioiden* todennäköisyyksien perusteella. Koodauskuvio tarkoittaa yksittäiselle havainnolle eri muuttujien arvoista muodostuvaa yhdistelmää. Esimerkiksi taulukkoon 3 sisältyvien väittämien arvot voivat yhdistyä keskenään 27 eri tavalla. Esimerkkejä koodauskuvioista olisivat 000 (samaa mieltä kaikkien väittämien kanssa), 020 (samaa mieltä ensimmäisen ja viimeisen, mutta eri mieltä keskimmäisen väittämän kanssa).

Kullekin koodauskuviolle lasketaan siis (yksittäisten muuttuja-arvojen todennäköisyyksien tulona syntyvä) todennäköisyysarvo. Koodauskuvion todennäköisyysarvo on sitä suurempi, mitä useammassa havainnossa on koodauskuvion muuttuja-arvoja. Esimerkiksi koodauskuvion 000 todennäköisyysarvoksi tulee 0,861 x 0,750 x 0,556 = 0,359037 ja koodauskuvion 020 todennäköisyysarvoksi 0,861 x 0,194 x 0,556 = 0,09287. Koko aineiston 'todennäköisyys' puolestaan saataisiin kertomalla jokaisen havainnon todennäköisyysarvot keskenään. Käytännössä näin ei kuitenkaan tehdä, sillä luvusta tulisi niin pieni, että sillä olisi vaikea operoida. Vastaavaan lopputulokseen päästään, kun jokaisen havainnon todennäköisyydestä otetaan logaritmi (ln) ja nämä logaritmit lasketaan yhteen. LOG-Like -tunnusluku lasketaan tällä tavalla. Mitä suurempia havaintojen todennäköisyydet ovat, sitä 'parempi' on niiden perusteella laskettu LOG-Like -tunnusluku ja sitä homogeenisemmasta joukosta on kysymys. Tämä perustuu siihen, että heterogeenisessa aineistossa koodauskuvioiden laskennallinen todennäköisyys jää yleensä niiden todellista osuutta pienemmäksi (toisinaan laskennallinen osuus myös ylittää todellisen frekvenssin). Jos aineistoa muutetaan homogeenisemmaksi, laskennalliset todennäköisyydet alkavat lähestyä koodauskuvioiden todellista osuutta ja myös LOG-Like -tunnusluku alkaa kasvaa.

Taulukossa 4 on esitetty esimerkkiaineiston koodauskuviot, niiden lukumäärä (N), osuus aineistosta (N/72), muuttuja-arvojen perusteella laskettu koodauskuvion todennäköisyys (p) sekä todennäköisyyksien summana muodostuva LOG-Like -tunnusluku. Koska kyse on yksiluokkaisesta ratkaisusta luokan koko (g) on täydet 1.

TAULUKKO 4. Esimerkkiaineiston koodauskuvioiden todennäköisyydet sekä niiden perusteella laskettu LOG-Like -tunnusluku: yksiluokkainen ratkaisu.

	Koodauskuvio	N	N/72	p	ln(p)	N*ln(p)
	0 0 0	27	0,375	0,359	-1,024	-27,657
Koko	0 0 1	1	0,014	0,036	-3,320	-3,320
aineisto	0 0 2	18	0,250	0,251	-1,382	-24,867
g=1,000	0 1 2	3	0,042	0,019	-3,976	-11,929
	0 2 0	4	0,056	0,093	-2,377	-9,506
	0 2 1	2	0,028	0,009	-4,672	-9,344
	0 2 2	7	0,097	0,065	-2,734	-19,136
	1 0 0	3	0,042	0,023	-3,757	-11,271
	1 1 1	1	0,014	0,000	-8,647	-8,647
	2 0 0	5	0,069	0,035	-3,364	-16,818
	2 2 0	1	0,014	0,009	-4,716	-4,716
	Yhteensä	72	1,000	0,899	-39,968	-147,211
					LOG-Like:	-147,211

Taulukosta havaitaan, että useimpien koodauskuvioiden osuus aineistosta (N/72) on suurempi kuin niiden todennäköisyys laskettuna muuttuja-arvojen tulona (p). Tämä tarkoittaa, että muuttujat ovat keskenään tilastollisesti riippuvaisia. Riippuvuus näkyy jollain tapaa koodauskuvioiden todennäköisyyksien summasta, koska se jää selvästi alle yhden. Oikeampi tunnusluku on kuitenkin LOG-Like, joka sekin on laskettu taulukossa 4.

LCA:n periaatteena on jakaa aineistoa osaryhmiin siten, että muuttujat ovat kussakin ryhmässä mahdollisimman riippumattomia. Käytännössä tämä tapahtuu niin, että Lacord-ohjelma ryhmittelee aineistoa ensin kahteen ryhmään (aloittaen satunnaisesta ryhmittelystä) ja iteroi ryhmitystä kunnes LOG-Like -tunnusluku on kahden LCA-luokan ratkaisussa paras mahdollinen.[3] Tämän jälkeen ohjelma etsii samalla tavoin parhaat mahdolliset kolme-, neljä-, viisi- ja kuusiluokkaiset ratkaisut. Yleensä luokkamäärän lisääminen parantaa tunnuslukua. Toisin sanoen kun luokkamäärää lisätään, luokat saadaan homogeenisemmiksi. Palaamme parhaan luokkamäärän valintaan tuonnempana.

TAULUKKO 5. Esimerkkiaineiston koodauskuvioiden todennäköisyydet sekä niiden perusteella laskettu LOG-Like -tunnusluku: kaksiluokkainen ratkaisu.

	Luokka 1: g_1=0,555				Luokka 2: g_2=0,445				Koko kaksiluokkainen ratkaisu: g=1				
Kood. kuvio	n_1	n_1/40	p_1	g_1*p_1	n_2	n_2/32	p_2	g_2*p_2	N	N/72	$g_1p_1+g_2p_2$	ln ($g_1p_1+g_2p_2$)	N*ln ($g_1p_1+g_2p_2$)
000	27	0,675	0,678	0,376	0	0	0,001	0,001	27	0,375	0,377	-0,976	-26,348
001	0	0	0,000	0,000	1	0,031	0,072	0,032	1	0,014	0,032	-3,441	-3,441
002	0	0	0,000	0,000	18	0,563	0,503	0,224	18	0,250	0,224	-1,497	-26,941
012	0	0	0,000	0,000	3	0,094	0,106	0,047	3	0,042	0,047	-3,055	-9,166
020	4	0,100	0,097	0,054	0	0	0,001	0,000	4	0,056	0,054	-2,919	-11,674
021	0	0	0,000	0,000	2	0,063	0,034	0,015	2	0,028	0,015	-4,190	-8,380
022	0	0	0,000	0,000	7	0,219	0,238	0,106	7	0,097	0,106	-2,245	-15,717
100	3	0,075	0,066	0,036	0	0	0,000	0,000	3	0,042	0,036	-3,312	-9,936
111	0	0	0,000	0,000	1	0,031	0,000	0,000	1	0,014	0,000	-8,442	-8,442
200	5	0,125	0,131	0,073	0	0	0,000	0,000	5	0,069	0,073	-2,619	-13,097
220	1	0,025	0,019	0,010	0	0	0,000	0,000	1	0,014	0,010	-4,565	-4,565
Yht.	40	1,000	0,991	0,550	32	1,000	0,954	0,425	72	1,000	0,975	-37,262	-137,709

LOG-Like: -137,709

Taulukossa 5 on esitetty miten Lacord-ohjelma jakaa esimerkkiaineiston koodauskuviot kahteen LCA-luokkaan.[4] Taulukosta havaitaan, että koodauskuvion laskennalliset todennäköisyydet (g1p1+g2p2) vastaavat nyt paremmin koodauskuvioiden todellisia osuuksia aineistossa (N/72). Samalla myös LOG-Like -tunnusluku on saatu selvästi paremmaksi. Olemme laatineet taulukot 4 ja 5 havainnollistaaksemme LCA:n laskentaperiaatetta. Tavallisessa LCA-analyysissä tällaisia koodauskuviokohtaisia taulukoita ei tarvita, vaan käytössä on ainoastaan kunkin ratkaisun LOG-Like -tunnusluku ja muuttuja-arvojen luokittaiset jakaumat. Tarkastellaan kaksiluokkaista ratkaisua vielä siinä muodossa, kuin Lacord ohjelma sen esittää (taulukko 6):

TAULUKKO 6. Käsitykset Internet-tekniikan kehityksestä ja sen merkityksestä sanomalehdille: kaksiluokkainen ratkaisu.

			Verkkolehtiin varauduttava jo nyt	Teknisyys rajoittaa vielä pitkään	Verkkolehdet eivät kiinnosta
1.CLASS	0.555	*0* samaa mieltä	**0.775**	**0.875**	**1.000**
		1 ei osaa sanoa	0.075	0.000	0.000
		2 eri mieltä	0.150	0.125	0.000
2.CLASS	0.445	*0* samaa mieltä	**0.969**	**0.594**	0.002
		1 ei osaa sanoa	0.031	0.125	0.125
		2 eri mieltä	0.000	0.281	**0.874**

Luokkien sisäinen homogeenisuus ja keskinäinen erilaisuus näkyvät selvästi myös taulukosta 6. Luokkia erottaa selvimmin tietokoneen ruudulta luettavien lehtien kiinnostavuutta koskeva väittämä. Kaikki ensimmäiseen luokkaan kuulu-

vat ovat väittämän kanssa samaa mieltä, kun taas toisessa luokassa melkein kaikki ovat sen kanssa eri mieltä. LCA:n tulosta voisi tulkita nimeämällä luokkia erottavan latentin ulottuvuuden vaikkapa 'teknologiaoptimismiksi'. Luokkaan 2 sijoitetut vastaajat olisivat näin selvästi 'teknologiaoptimistisempia' ja luokkaan 1 sijoitetut vastaavasti 'teknologiapessimistisempiä'. Tähän viittaa erityisesti verkkolehtien kiinnostuvuutta koskeva muuttuja. Kukaan teknologiapessimisteistä ei usko verkkolehtien kiinnostavan suurta yleisöä, kun taas optimisteista valtaosa on päinvastaisella kannalla. Myös kaksi muuta muuttujaa tukevat esitettyä tulkintaa, joskaan luokkien väliset erot eivät ole niiden kohdalla yhtä suuria.

TAULUKKO 7 Käsitykset Internet-tekniikan kehityksestä ja sen merkityksestä sanomalehdille: kolmiluokkainen ratkaisu.

			Verkkolehtiin varauduttava jo nyt	Teknisyys rajoittaa vielä pitkään	Verkkolehdet eivät kiinnosta
1.CLASS	0.134	*0* samaa mieltä	**0.896**	0.001	0.000
		1 ei osaa sanoa	0.104	0.416	0.290
		2 eri mieltä	0.000	**0.584**	**0.710**
2.CLASS	0.134	*0* samaa mieltä	0.067	**0.889**	**1.000**
		1 ei osaa sanoa	0.311	0.000	0.000
		2 eri mieltä	**0.622**	0.111	0.000
3.CLASS	0.732	*0* samaa mieltä	**1.000**	**0.861**	**0.576**
		1 ei osaa sanoa	0.000	0.000	0.023
		2 eri mieltä	0.000	0.139	**0.402**

Kolmiluokkaisessa ratkaisussa luokat ovat vielä hivenen homogeenisempia kuin kaksiluokkaisessa (LOG-Like = -135,415). Mikäli pidetään kiinni oletuksesta että muuttujien taustatekijänä on 'teknologiaoptimismi-teknologiapessimisimi' -ulottuvuus, näyttäisi ilmeiseltä, että kaikkein pessimistisimmät on nyt koottu 2-luokkaan. Heidän joukostaan on 'siivottu' pois ne, jotka pitivät verkkolehtiin varautumista tärkeänä. Kaikkein optimistisimmat puolestaan on koottu 1-luokkaan. Heidän joukostaan on poistettu ne, jotka arvelivat teknisyyden rajoittavan vielä pitkään internetin suosiota. Molemmat ääriryhmät ovat melko pieniä, kummassakin on 13 prosenttia vastaajista. Suurimmaksi ryhmäksi on jäänyt joukko, jonka vastaukset taustatekijöihin nähden vaikuttavat ensi katsomalta varsin satunnaisilta.

Kolmannen luokan voisi ajatella epävarmoista vastaajista koostuvaksi 'väliluokaksi', koska vastauksissa yhtäältä kannatetaan verkkolehtiin varautumista, mutta toisaalta uskotaan teknisyyden rajoittavan vielä pitkään internetin leviämistä. Parempi tulkinta kuitenkin on se, että optimismi-pessimismin lisäksi vastausten taustalla on myös toinen latentti ulottuvuus. Tämä voisi olla eräänlainen 'kaiken varalta' -ajattelutapa. Kolmannen luokan vastaajat haluavat varmistaa asemansa varmuuden vuoksi, vaikkeivät uskokaan internetin nopeaan leviämiseen. Kolmas luokka edustaisi siten 'varmistelijoita', kun taas ensimmäinen ja toinen luokka edustaisivat henkilöitä, jotka katsovat että vain tarpeellisilta näyttäviin toimenpiteisiin on syytä ryhtyä.

LCA:sta on syytä vielä huomata, että se pyrkii tuomaan esiin aineiston rakenteistumisen, ei ensisijaisesti sitä, miten yksittäiset havainnot sijoittuvat eri luok-

kiin. Kustakin havainnosta tiedetään ainoastaan millä todennäköisyydellä se kuuluu mihinkin luokkaan, ja yksi havainto voi eri todennäköisyyksin kuulua useisiinkin luokkiin.

Informatiivisimman luokkamäärän valitseminen

Luokkien homogeenisuus on täydellisintä silloin, kun muuttujat ovat kussakin luokassa täydellisesti riippumattomia. Homogeenisuutta voidaan parantaa lisäämällä luokkien määrää, ja homogeenisuus on täydellistä viimeistään silloin, kun jokaista erilaista koodauskuviota varten on oma luokkansa. Tällaista ratkaisua kutsutaan saturoituneeksi ratkaisuksi (saturated model). Esimerkkiaineistossa Internet-tekniikkaa koskeviin kysymyksiin vastattiin kaikkiaan 11 eri tavalla, joten tämän aineiston saturoituneessa ratkaisussa olisi siis 11 luokkaa. Luokkien määrää ei kuitenkaan kannata kasvattaa äärimmilleen, koska ratkaisusta tulee samalla monimutkaisempi ja epähavainnollisempi. Ongelmana onkin löytää 'taloudellisin' ratkaisu jostain yksiluokkaisen ja saturoituneen ratkaisun väliltä.

Taloudellisimman luokkamäärän valitsemista varten on olemassa useita erilaisia indeksejä. Ne nojaavat tiettyihin informaatioteoreettisiin perusteisiin, joiden avulla analyysin tarkkuuden tuottama hyöty ja ratkaisun monimutkaisuuden tuottama haitta tehdään vertailukelpoisiksi. Indeksit ovat siis teoreettisesti perusteltuja, mutta silti ne tuottavat erilaisia ratkaisuja eivätkä siten ole yleispäteviä. On myös huomattava, että tutkimustehtävän kannalta voi joskus olla tarkoituksenmukaista suosia yksinkertaisia ratkaisuja tarkkuuden kustannuksella (tai päinvastoin), vaikka analyysin laskennallinen informatiivisuus hiukan kärsisikin. Lacord-ohjelma laskee automaattisesti yhden tällaisen indeksin. Sitä kutsutaan keksijänsä mukaan AIC-indeksiksi (Akaike's Information Criterion). Indeksin kaava on seuraava (Kempf 1994, 14):

$$AIC = -2\ \ln(L(x)) + 2\ n(P)$$

Kaavan osa ln(L(x)) tarkoittaa yllä kuvattua LOG-Like -tunnuslukua. Kaavan osa n(P) puolestaan tarkoittaa niiden parametrien lukumäärää, jotka ratkaisua varten täytyy estimoida. Estimoitavia parametrejä ovat esimerkiksi LCA-luokkien koot ja muuttuja-arvojen osuudet kussakin LCA-luokassa. Estimoitavien parametrien määrä kasvaa jyrkästi, kun luokkien määrää lisätään. Luokkien määrää lisättäessä ln(L(x)) tulee lähemmäs nollaa, mikä pienentää AIC:tä koska ln(L(x)) on aina negatiivinen luku. Samalla kuitenkin estimoitavien parametrien määrä kasvaa, mikä puolestaan kasvattaa AIC-indeksiä. Paras ratkaisu on se, joka tuottaa pienimmän AIC-indeksin. AIC:n avulla siis etsitään luokkien määrä, jonka jälkeen luokkien määrän lisäämisen tuoma 'tarkkuusetu' ei enää riitä luokkien lisäämisen aiheuttamasta esityksen monimutkaisuudesta koituvaa haittaa.

Toinen tunnettu indeksi on nimeltään BIC (Best Information Criterion). Se antaa AIC:tä suuremman painon luokkien lisääntymisestä koituvalle haitalle. Lacord-ohjelma ei laske BIC:tä, mutta se on helposti laskettavissa kaavasta (Kempf 1994, 14):

$$BIC = -2\ \ln(L(x)) + \ln(n)\ n(P)$$

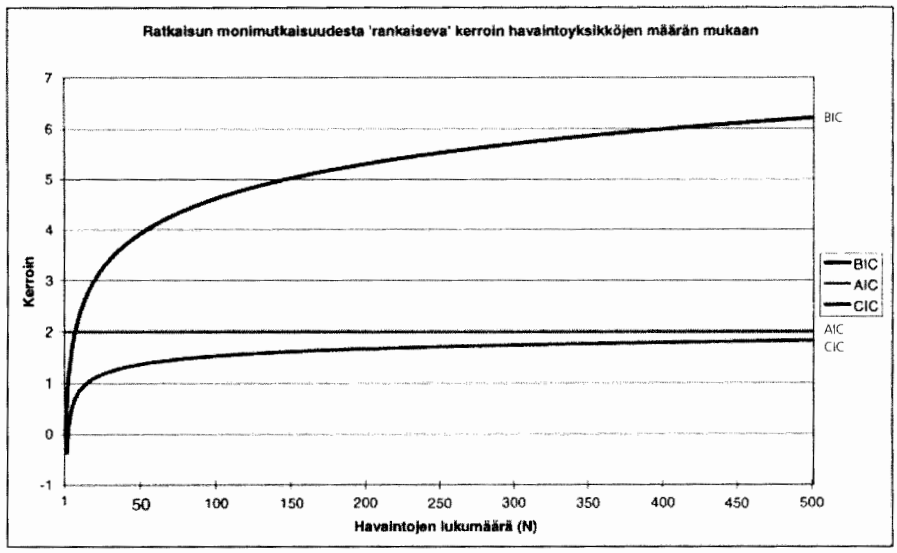
BIC eroaa AIC:stä vain siinä, että parametrien lukumäärän kerroin 2 on korvattu havaintoyksikköjen lukumäärän logaritmilla ln(n). Tämä luku on sitä suurempi, mitä enemmän on havaintoyksikköjä. Tämän vuoksi varsinkin suurilla aineistoilla BIC suosittaa usein pienempää luokkamäärää kuin AIC.

Stanley L. Scloven (1987, 337) mukaan eri indeksit eroavat toisistaan siinä,

kuinka nopeasti parametrien lukumäärästä 'rankaiseva' kerroin kasvaa havaintoyksikköjen lukumäärän kasvaessa. Nopeimmin kasvava kerroin on BIC:n ln(n), kun hitaimmin kasvava kerroin on ln(ln(n)). Hitaasti kasvava kerroin lähenee havaintoyksikköjen määrän kasvaessa AIC:ssä käytettyä vakiokerrointa 2. Nimeämme tässä hitaimmin kasvavaan kertoimeen perustuvan indeksin CIC:ksi:

$$CIC = -2\ ln(L(x)) + ln(ln(n))\ n(P)$$

KUVIO 1. Ratkaisun monimutkaisuudesta 'rankaisevat' kertoimet havaintoyksiköiden lukumäärän mukaan.



Scloven esittämän katsauksen perusteella vaikuttaa ilmeiseltä, että pienin teoreettisesti perusteltu luokkamäärä saadaan BIC-indeksin ja suurin CIC-indeksin avulla. Mikäli molemmat kertoimet suosittavat samaa luokkamäärää, luokkamäärä on yleispätevästi (globally) paras. Mikäli kertoimien suosittamat luokkamäärät eroavat toisistaan, luokkamääräksi voidaan valita jompi kumpi suositetuista tai jokin niiden väliltä.

Esimerkkianalyysissä havaintoyksikköjen lukumäärä on 72 ja sen logaritmi 4,28, tämän logaritmi puolestaan on 1,45. BIC:ssä siis kerrotaan estimoitavien parametrien lukumäärä 2:n sijasta 4,28:lla ja CIC:ssä 1,45:llä. Näin BIC siis painottaa CIC:tä ja AIC:tä enemmän luokkien lisäämisestä aiheutuvaa 'haittaa'.

TAULUKKO 8. Esimerkkianalyysin tunnuslukuja.

Luokkien määrä	LOG-Like [5]	Npar	LIK.ratio	DF	AIC	BIC	CIC
1.	-147,283	6	31,137	20	306,565	**320,226**	303,285
2.	-137,739	13	12,051	13	**301,479**	331,075	**294,369**
3.	-135,415	20	7,403	6	310,831	356,363	299,893
4.	-134,890	27	6,351	-1	323,779	385,250	309,016
5.	-132,154	34	0,88	-8	332,308	409,715	313,716
6.	-131,719	41	0,01	-15	345,438	438,781	323,018

Lacord esittää kustakin ratkaisusta LOG-Like- ja AIC-indeksien lisäksi estimoitujen parametrien lukumäärän (Npar), ratkaisun vapausasteen (DF) sekä LOG-Like -indeksin ja saturoituneen ratkaisun LOG-Liken erotuksen (LIK.ratio). BIC:tä ja CIC:tä se ei tulosta, mutta ne on helppo laskea yllä esitetyistä kaavoista. Jos vapausaste (DF) on pienempi kuin nolla, ratkaisu ei ole uniikki (uniquely defined) ja se on hylättävä. Esimerkkitapauksessa on siis valittava yksi-, kaksi- tai kolmeluokkainen ratkaisu. Monesti eri indeksien suosittamat luokkamäärät osuvat yksiin (esim. Kempf 1994, 14-15). Esimerkkianalyysissä BIC kuitenkin suositti yksiluokkaista ratkaisua, kun AIC:tä ja CIC:tä käyttäen olisi päädytty kaksiluokkaiseen ratkaisuun. Indeksit ovat kuitenkin vain apuneuvoja. Esimerkkitapauksessakin voi nähdäksemme valita yksi-, kaksi- tai kolmiluokkaisen ratkaisun tai käyttää niitä kaikkia tulkinnan apuna.

LCA verrattuna muihin monimuuttujamenetelmiin

LCA:n taustalla on Lazarsfeldin pyrkimys luoda faktorianalyysin kaltainen monimuuttujamenetelmä, joka olisi vapaa mitta-asteikollisista rajoituksista. Tästä samalla seuraa yksi LCA:n ja faktorianalyysin keskeinen ero: LCA:lle tarkasteltavat ilmiöt ovat laadullisia, faktorianalyysille määrällisiä. Toisin sanoen LCA:lle muuttujan arvot 1 ja 4 ovat samalla tavoin erilaisia kuin arvot 1 ja 2. Faktorianalyysi taas tulkitsee arvon 4 täsmälleen kolme yksikköä suuremmaksi kuin arvon 1. Tällä on tietenkin käytännön seurauksia sen suhteen, millaisiin tehtäviin menetelmiä kannattaa käyttää. Pyrimme havainnollistamaan asiaa tuonnempana.

Niin faktorianalyysi kuin LCA:kin tuovat esiin jotain latenttia. Faktorianalyysi etsii aineistoa strukturoivat latentit ulottuvuudet sekä kertoo sen, kuinka voimakasta tämä strukturointi on. Faktorin tulkinnassa kiinnitetään huomiota niihin muuttujiin, jotka 'latautuvat' voimakkaasti kyseisellä faktorilla. Tämä lataus voidaan tulkita faktorin ja kyseisen muuttujan väliseksi korrelaatioksi (Alkula ym. 1994, 270). LCA taas paljastaa LCA-luokkien avulla ennen muuta sen, miten aineisto 'makaa' tutkituilla latenteilla ulottuvuuksilla. Sekä LCA:n että faktorianalyysin tuloksia voi yrittää lukea ikään kuin toistensa näkökannalta, mutta tämä on aina vähän hankalaa. Ainakin on syytä huomata, että faktori ei ole sama asia kuin yksittäinen LCA-luokka. Pikemminkin voisi sanoa, että faktorit ovat ulottuvuuksia, joihin nähden LCA-luokat ovat sisäisesti mahdollisimman homogeenisia.

Faktorianalyysin paljastama tai analysoima faktori on siis kutakuinkin sama asia kuin latentti ulottuvuus, jota LCA tutkii siltä kannalta, miten aineisto sille sijoittuu. Faktorianalyysiä voidaan jatkaa laskemalla kullekin havainnolle ns. faktoripistemäärät ja 'sijoittamalla' havainnot pistemäärien avulla faktoreiden kuvaamille ulottuvuuksille. Tällöin aineistoa järjestetään periaatteessa hyvin samalla tavalla kuin LCA:ssakin. Näin ollen LCA:ta kannattaa verrata juuri faktorianalyysin avulla tuotettuun ryhmittelyyn.

Taulukossa 9 on kuvitteellinen aineisto, jossa on kymmenen opiskelijan saamat arvosanat ruotsin, englannin ja saksan kielissä. Tämä aineisto voidaan analysoida sekä faktorianalyysillä että LCA:lla mutta tulosten tulkinnassa on osattava ottaa huomioon menetelmien ominaispiirteet.

TAULUKKO 9. Kielten arvosanat esimerkkiaineistossa.

	Ruotsi	Englanti	Saksa
Aki	3	2	0
Arja	0	3	1
Auli	3	2	0
Axel	1	1	2
Asko	0	2	1
Adolf	1	1	2
Atte	1	0	3
Anna	1	0	2
Arto	1	0	3
Anne	1	1	2

Taulukosta huomataan, että ruotsin ja englannin numerot korreloivat keskenään hiukan positiivisesti, mutta saksan numerot korreloivat sekä ruotsin että englannin numeroihin selvästi negatiivisesti. Vaikuttaisi siis todennäköiseltä, että muuttujat voitaisiin 'tiivistää' yhteen tai kahteen faktoriin, joissa kielten numeroiden väliset riippuvuudet tulisivat esiin. Aineiston luonteen vuoksi kokeilemme tässä kuitenkin faktorianalyysiä muistuttavaa, ns. pääkomponenttianalyysiä, jonka "tavoitteena on sisällyttää maksimaalinen määrä alkuperäisten muuttujien sisältämää vaihtelua vain muutamaan - toisistaan riippumattomaan - pääkomponenttiin" (Ranta & Rita & Kouki 1991, 459).[6]

Analyysi tuottaa ensin yhden pääkomponentin (faktorin), joka selittää mahdollisimman hyvin vaihtelua, tämän jälkeen se tuottaa lisää pääkomponentteja, jotka selittävät 'jäljelle jäänyttä' vaihtelua. Analyysi tuotti esimerkkiaineistosta taulukossa 10 esitetyt kaksi pääkomponenttia, jotka yhdessä selittävät 97 % muuttujien vaihtelusta.

TAULUKKO 10. Arvosana-aineiston pääkomponentit

	PK 1	PK 2
Ruotsi	0,56	0,82
Englanti	0,84	-0,51
Saksa	-0,99	0,03
Selitysaste	66 %	31 %

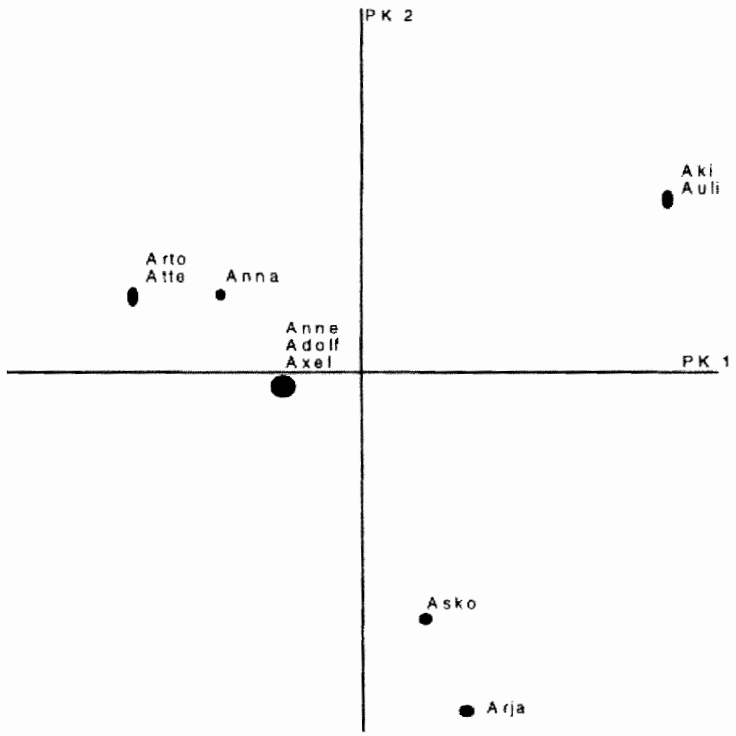
Ensimmäinen pääkomponentti on tärkein, ja se ilmaisee tekijää, joka saa ihmiset menestymään paremmin ruotsissa ja englannissa kuin saksassa tai vastaavasti paremmin saksassa kuin ruotsissa ja englannissa. Toinen taas selittää sitä, miksi jotkut vastaajat menestyvät paremmin ruotsissa kuin englannissa tai päinvastoin.

Analyysi näyttää siis varsin mainiosti tiivistävän taulukon sisältämää informaatiota. Hiukan mielikuvitusta käyttäen voisi esittää tulkinnan, että ensimmäisessä pääkomponentissa olisi kyse populaarikulttuurin vaikutuksesta. Paljon englanninkielistä ohjelmaa seuraavat saavat kohtalaisen hyviä arvosanoja englannissa, koska kuulevat sitä vapaa-aikanaan paljon. Sen sijaan aika ei riitä saksan epäsäännöllisten verbien opetteluun. Toisen pääkomponentin taas voisi ajatella kertovan

ruotsinkielisen ympäristön vaikutuksesta.

Kun havainnoille lasketaan faktoripistemäärät, voidaan tarkastella niiden sijoittumista pääkomponenttien kuvaamilla ulottuvuuksilla. Kuvion 2 vaaka-akseli kuvaa 'populaarikulttuuri' -komponenttia ja pystyakseli 'ruotsinkielisyys' -komponenttia. Havainnot on sijoitettu näiden ulottuvuuksien muodostamaan 'voimakenttään' SPSS-ohjelmalla laskettujen faktoripistemäärien (regression factor scores) avulla.

KUVIO 2. Arvosana-aineisto ryhmiteltynä faktoripistemäärien avulla.



Kuviossa henkilöt näyttävät sijoittuvan varsin selvästi kolmeen tai neljään eri ryhmään. Aki ja Auli olisivat siis hypoteesimme mukaan ruotsinkielisessä ympäristössä asuvia populaarikulttuurin kuluttajia. Asko ja Arja taas suomenkielisessä ympäristössä asuvia, jonkin verran populaarikulttuuria seuraavia. Atte, Arto ja Anna käyttäisivät aikansa populaarikulttuurin sijasta saksan opiskeluun, kuten myös Anne, Adolf ja Axel, joskaan eivät aivan yhtä suuressa määrin.

Jos faktoripistemäärien avulla tehty ryhmittely kolmeen esitetään LCA:lle ominaisessa muodossa, tulos näyttäisi seuraavalta:

TAULUKKO 11. Arvosana-aineisto ryhmiteltynä faktoripistemäärien avulla: kolmiluokkainen ratkaisu.

	Luokan koko	Arvosana	Ruotsi	Englanti	Saksa
1.CLASS	0,600	0	0,000	**0,500**	0,000
Arto, Atte		1	**1,000**	**0,500**	0,000
Anna, Anne		2	0,000	0,000	**0.667**
Adolf, Axel		3	0,000	0,000	**0,333**
2.CLASS	0,200	0	0,000	0,000	**1,000**
Aki, Auli		1	0,000	0,000	0,000
		2	0,000	**1,000**	0,000
		3	**1,000**	0,000	0,000
3.CLASS	0,200	0	**1,000**	0,000	0,000
Asko, Arja		1	0,000	0,000	**1,000**
		2	0,000	**0,500**	0,000
		3	0,000	**0,500**	0,000

LCA:n tavoitteena on kuvata sitä, miten aineisto ryhmittyy aineistoa jäsentäville latenteille ulottuvuuksille. Sikäli voisi olettaa, että LCA tuottaisi hyvin samantapaisen tuloksen kuin faktoripistemäärien avulla tehty ryhmittelykin. Menetelmien olennainen ero on kuitenkin siinä, että LCA:ssa tarkastellaan määrällisen vaihtelun sijasta vain eroja ja yhtäläisyyksiä. Tämä puolestaan voi johtaa erilaisiinkin ryhmittelyihin. LCA suositteli seuraavaa kaksiluokkaista ratkaisua:[7]

TAULUKKO 12. Arvosana-aineiston LCA-luokat: kaksiluokkainen ratkaisu.

	Luokan koko	Arvosana	Ruotsi	Englanti	Saksa
1,CLASS	0,400	0	**0,500**	0,000	**0,500**
		1	0,000	0,000	**0,500**
		2	0,000	**0,750**	0,000
		3	**0,500**	**0,250**	0,000
2,CLASS	0,600	0	0,000	**0,500**	0,000
		1	**1,000**	**0,500**	0,000
		2	0,000	0,000	**0,667**
		3	0,000	0,000	**0,333**

LCA-ratkaisu on muuten sama kuin faktoripistemäärillä tuotettu ratkaisukin, mutta siinä kaksi pienintä luokkaa on yhdistetty. Tulosta voi havainnollistaa kuvion 2 avulla. LCA on muodostanut kuvion vasemmasta puoliskosta oman luokkansa ja oikeasta puoliskosta omansa. Kuvion 'pystysuora' ulottuvuus ei erottele LCA-luokkia. Taulukossa 12 tämä näkyy siten, että luokassa 1 on sekä erittäin hyvin että erittäin huonosti ruotsissa menestyneitä, ei ketään siltä väliltä. Toisessa luokassa taas ovat välttävästi menestyneet. Juuri tässä tulee esiin LCA:n 'tunnottomuus' määrälliselle riippuvuudelle. LCA:n kannalta on aivan sama onko nume-

ro suuri tai pieni, koska LCA:n ryhmitykset perustuvat ainoastaan sille, ovatko numerot samoja vai erilaisia.

Esimerkkiaineiston kohdalla menetelmäksi ehkä kannattaisi valita pääkomponenttianalyysi, koska kyse on suhdelukuasteikollisista muuttujista. Pääkomponenttianalyysi 'osaa' tehdä eron suurten ja pienten arvosanojen välillä, mutta LCA:lla se menee 'yli ymmärryksen'. Toisaalta tässäkin aineistossa LCA tuo esiin sen epälineaarisen riippuvuuden, että neljä saksassa heikosti menestynyttä menestyvät ruotsissa joko erinomaisesti tai sitten eivät lainkaan, mutta eivät tältä väliltä. Mikäli tällainen tulos toistuisi suuremmassa aineistossa se voisi olla varsin mielenkiintoinen, mutta pääkomponentti- tai faktorianalyysissä se ei tulisi esiin.

Kuvitellaan seuraavaksi, että esimerkkiaineiston luvut eivät olisikaan arvosanoja, vaan vastauksia Matkailun edistämiskeskuksen kyselyyn. Oletetaan, että vastaajia olisi pyydetty kertomaan, mikä annetuista vaihtoehdoista heitä viehättää eniten Ruotsissa, Englannissa ja Saksassa. Vaihtoehtoina olisi ollut 0) ei mikään, 1) luonto, 2) kulttuuri, 3) elintaso. Muuttujina olisivat siis edelleen Ruotsi, Englanti ja Saksa, mutta ne olisivat nyt laatueroasteikollisia. Tällaisen asteikon yhteydessä pääkomponentit ja faktorit eivät kerro oikeastaan mitään[8], mutta LCA:n tulokset kertovat siitä mistä pitääkin. Ensimmäiseen luokkaan sijoitetut henkilöt arvostavat erityisesti englantilaista kulttuuria ja elintasoa, mutta eivät välttämättä näe Ruotsissa ja Saksassa mitään viehättävää. Jos Ruotsissa jokin viehättää, se on elintaso ja jos Saksassa jokin viehättää, se on luonto. Toiseen luokkaan kuuluvista taas kaikki pitävät ruotsalaisesta luonnosta ja useimmat saksalaisesta kulttuurista. Puolet ei näe Englannissa mitään viehättävää, puolet pitää luontoa Englanninkin viehättävimpänä piirteenä.

LCA:n perusteella voisi siis ajatella, että toinen LCA-luokka ilmentäisi tekijää, joka olisi sekä ruotsalaisen luonnon että saksalaisen kulttuurin arvostamisen taustalla. Yksi tällainen voisi olla saksalainen kulttuuri itse. Ehkä kakkosluokkaan kuuluvat ovat saksalaisia ja oppineet siellä arvostamaan skandinaavista luontoa ja samalla suhtautumaan hiukan epäluuloisesti Englantiin. Omaa kulttuuriperinnettä wagnereineen ja hegeleineen pidetään tietysti arvossa myös. Tällaista tulkintahypoteesia voi yrittää testata tutkimalla kakkosluokkaan kuuluvien henkilöiden taustatietoja, ja juuri näin LCA:ssa usein tehdäänkin tätä varten kehitetyn apuohjelman avulla. Tässä esimerkissä riittää, kun katsotaan alkuperäisestä aineistosta minkä nimiset henkilöt ovat vastannet Ruotsin kohdalla 1 (luonto) ja Saksan kohdalla 2 (kulttuuri): Axel, Adolf, Anna ja Anne. Hypoteesi tuntuu siis saavan tukea...

Tarkastelutavaltaan sekä LCA että faktorianalyysi ovat määrällisiä menetelmiä. Erona on kuitenkin se, että LCA:ssa tarkasteltavat ilmiöt ovat laadullisia, faktorianalyysissä määrällisiä (vrt. Suhonen 1994, 74-76). Sisällönanalyyseissä muuttujat ovat usein - vaikkakaan eivät aina - laatueroasteikollisia, ja siksi LCA tuntuisi soveltuvan niihin erityisen hyvin.

LCA ja etäisyysindeksit

LCA ja faktorianalyysi muistuttavat toisiaan siinä, niissä molemmissa analyysi perustuu muuttujien välisiin riippuvuuksiin. Niissä ei yksinkertaisesti panna samankaltaisia havaintoja samoihin luokkiin, vaan etsitään riippuvuutta selittäviä faktoreita tai jaotellaan havainnot mahdollisimman samankaltaisten havaintoyksikköjen joukoiksi. Tässä siis teemme eron havainnon ja havaintoyksikön välille. Havainto tarkoittaa esimerkiksi tiettyä koodauskuviota, havaintoyksikkö taas esimerkiksi asenteeltaan tietynlaista vastaajaa. Samankaltaiset havaintoyksiköt eivät siis välttämättä tuota samanlaisia havaintoja.

Havaintoja voidaan ryhmitellä myös vähemmän tilastollisesti, suoraan niiden manifestien erojen ja yhtäläisyyksien perusteella. Tähän on olemassa useita ns.

etäisyysmittoja, joista yleisimmin käytetty lienee ns. yleistetty etäisyysmitta (Pietilä 1970, 112), jota kutsutaan myös euklidiseksi etäisyydeksi (Everitt 1995, 46). Teemme tässä ryhmittelyn SPSS-ohjelman avulla käyttäen menetelmää, joka muodostaa ryhmät siten, että kussakin ryhmässä havaintojen etäisyys on mahdollisimman pieni. Ryhmittely etenee hierarkkisesti siten, että ensin etsitään keskenään täsmälleen samanlaiset havainnot ja muodostetaan näistä omat ryhmänsä. Esimerkkiaineistossa tällaisia ryhmiä on kuusi:

TAULUKKO 13. Arvosana-aineisto ryhmiteltynä 'koodauskuvioittain'.

	Ruotsi	**Englanti**	**Saksa**
Adolf, Axel, Anne	1	1	2
Aki, Auli	3	2	0
Atte, Arto	1	0	3
Arja	0	3	1
Asko	0	2	1
Anna	1	0	2

Seuraavaksi ohjelma laskee näiden 6 ryhmän keskinäisiä etäisyyksiä yleistetyn etäisyysmitan $D = \sqrt{\Sigma d^2}$ avulla. Esimerkiksi koodauskuvion 1 1 2 etäisyys koodauskuviosta 3 2 0 saadaan laskemalla ensin kuinka paljon kukin muuttuja-arvo eroaa koodauskuvioissa (d), nostamalla tämä toiseen potenssiin, laskemalla näin saadut erotusten neliöt yhteen ja ottamalla summasta neliöjuuri. Ensimmäisen muuttujan kohdalla erotus on 2, toisen kohdalla 1 ja kolmannen kohdalla 2. Kyseiset arvot nostetaan toiseen potenssiin ja lasketaan yhteen, eli tehdään laskuoperaatio 4+1+4 = 9. Tästä otetaan lopuksi neliöjuuri, joten koodauskuvioiden etäisyydeksi saadaan 3. Kaikkien ryhmien etäisyydet toisistaan lasketaan samalla tavalla ja lähimpänä olevat ryhmät yhdistetään toisiinsa. Näin jatketaan kunnes lopulta kaikki havainnot kuuluvat samaan ryhmään.

Ryhmittelyanalyysin ongelmana onkin 'oikean' ryhmäluvun määrittäminen eikä tähän ilmeisesti ole mitään yleispätevää ohjetta. Valitsimme esimerkeiksi kolme- ja kaksiryhmäiset ratkaisut, jotka esitämme vertailun helpottamiseksi LCA:lle ominaisessa muodossa.

TAULUKKO 14. Arvosana-aineisto ryhmitettynä etäisyysindeksin avulla: kolmiluokkainen ratkaisu.

	Luokan koko	Arvosana	Ruotsi	Englanti	Saksa
1.CLASS	0,600	0	0,000	**0,500**	0,000
Arto, Atte		1	**1,000**	**0,500**	0,000
Anna, Anne		2	0,000	0,000	**0.667**
Adolf, Axel		3	0,000	0,000	**0,333**
2.CLASS	0,200	0	0,000	0,000	**1,000**
Aki, Auli		1	0,000	0,000	0,000
		2	0,000	**1,000**	0,000
		3	**1,000**	0,000	0,000
3.CLASS	0,200	0	**1,000**	0,000	0,000
Asko, Arja		1	0,000	0,000	**1,000**
		2	0,000	**0,500**	0,000
		3	0,000	**0,500**	0,000

Etäisyysindeksin avulla saatu kolmiluokkainen ryhmittely on täsmälleen sama kuin faktoripistemäärilläkin. LCA taas esitti kaksiluokkaisen ratkaisun, jossa Aki, Auli, Asko ja Arja oli yhdistetty samaan luokkaan. Etäisyysindeksin avulla ryhmiteltäessä taas Asko ja Arja yhdistyvät kaksiluokkaisessa ratkaisussa suurimpaan ryhmään, jolloin tulos näyttäisi seuraavalta:

TAULUKKO 15. Arvosana-aineisto ryhmitettynä etäisyysindeksin avulla: kaksiluokkainen ratkaisu.

	Luokan koko	Arvosana	Ruotsi	Englanti	Saksa
1,CLASS	0,200	0	0,000	0,000	**1,000**
Aki, Auli		1	0,000	0,000	0,000
		2	0,000	**1,000**	0,000
		3	**1,000**	0,000	0,000
2,CLASS	0,800	0	**0,250**	**0,375**	0,000
Arto, Atte		1	**0,750**	**0,375**	**0,250**
Anna, Anne		2	0,000	**0,125**	**0,500**
Adolf, Axel		3	0,000	**0,125**	**0,250**
Asko, Arja					

Etäisyysindeksin käyttö näyttäisi johtavan ainakin esimerkkiaineistossa siihen, että havainnot vaihe vaiheelta yhdistyvät suurimpaan, 'keskimääräisten' ryhmään. Etäisyysindeksi tuottaa sikäli helposti ymmärrettäviä tuloksia, että ryhmittelyn perusteena käytetään muuttuja-arvojen erotuksia koodauskuvioiden välillä. Tällä tavoin vielä kaksiluokkaisessakin ratkaisussa Aki ja Auli erottuivat omana ryhmänään, koska he varsin selvästi erottuivat muista sekä ruotsin että englannin numeronsa suhteen.

Esimerkissä etäisyysindeksiä käytettiin ajatellen koulunumerojen suhdelukuasteikollisia muuttujia. LCA:n ja etäisyysindeksi tuottamien jaottelujen erot johtunevat paljolti tästä. Ilmeisesti numeroiden erojen tulkitseminen määrällisiksi on omiaan 'venyttämään' havaintojen välisiä etäisyyksiä ja siten loitontamaan erityisesti Akia ja Aulia omaksi ryhmäkseen. Jos taas erot tulkitaan laadullisiksi kuten LCA:ssa, Aki ja Auli ovat ikäänkuin 'lähempänä' Askoa ja Arjaa. Jos kyseessä olisi ollut Matkailun edistämiskeskuksen kysely, etäisyyksien venyttäminen numeroarvojen perusteella olisi tietysti ollut mieletöntä ja siten LCA tai jokin laadullista etäisyysindeksiä käyttävä ryhmittelymenetelmä (ks. Everitt 1995) olisi soveltunut analyysimenetelmäksi parhaiten.

Silmäys tehtyyn tutkimukseen: LCA sisällönanalyysin välineenä

Lazarsfeldin kokeilujen jälkeen LCA koki uuden tulemisensa tällä vuosikymmenellä, kun Konstanzin yliopiston rauhantutkimusyksikön johtaja Wilhelm Kempf kiinnostui menetelmän soveltamisesta sisällönanalyysiin. Kempfin johtamissa projekteissa menetelmää on käytetty ennen muuta sotaa ja konflikteja koskevan tiedonvälityksen tutkimiseen. Keskeisiä kohteita ovat olleet Persianlahden ja Bosnian sodat sekä niihin liittyvät tapahtumat. Meneillään on myös projekteja, joissa tarkastellaan journalismin, valtion ja kansalaisuuden suhteita toisen maailmansodan jälkeisessä Euroopassa.

Väiteanalyysi on toistaiseksi ollut LCA:n eräänlainen perusmuoto sisällönanalyysissä.

Journalism in the New World Order -projekti käytti väiteanalyysiä tarkastellessaan Persianlahden sodan osapuolten harjoittaman propagandan läpäisyä tiedotusvälineissä. Projektin tutkijat valitsivat propagandaa ja propagandatekstejä koskeva pohdinnan perusteella (ajattelun peruspiirteistä ks. Luostarinen 1986, 44 - 52) ensin yhdeksän teemaa, jotka he katsoivat sodan legitimoinnin ja siihen mobilisoinnin kannalta keskeisiksi.[9] Näistä teemoista sitten muotoiltiin sisällönanalyyttisiksi muuttujiksi väitteitä, joita käytettiin lehtiaineiston koodauksessa ja LCA:ssa.

Projekti tutki Yhdysvaltain hallinnon propagandapyrkimyksiin sisältyvän motivaatiologiikan tunkeutumista tiedonvälitykseen tarkastelemalla George Bushin lanseeraaman *New World Order* -formulan virittämien hahmotustapojen esiintymistä Yhdysvaltain, Saksan ja Pohjoismaiden lehdistössä. Erityisesti yritettiin selvittää sitä, millaisena lehtitekstit esittivät menneen ja tulevan leikkaukseksi ymmärretyn 'nykyhetken' (ks. Luostarinen 1986, 436-440). Seuraavassa esittelemme lyhyesti tätä analyysiä.

George Bush lanseerasi termin *New World Order* USA:n kongressissa pitämässään puheessa syyskuussa 1990. Kempfin ja Luostarisen mukaan (1997, 4-5) termi oli upotettu Bushin pitämissä puheissa ainakin kolmeen mielikuvayhteyteen. Ensinnäkin Bush maalaili näkyviin Hitlerin Saksan laajentumispolitiikan. Toiseksi termin taustana oli vuoden 1989 käännekohta: kylmän sodan ja ideologisen taistelun päättyminen oli luonut "nyt tai ei koskaan" -tilanteen päättäväiselle toiminnalle siintävän rauhantilan vakiinnuttamiseksi. Kolmas termiin liittyvä viritelmä oli lupaus "oikeudenmukaisesta maailmasta", jossa myös pienten kansojen elämä on turvattu. Näin Bushin puheeseen sisältyi tutkijoiden tyypillisenä pitämä menneisyyttä, nykyisyyttä ja tulevaisuutta koskeva motivaatiorakenne mainittujen historiallisten mielikuvaulottuvuuksien muodossa. Nämä ulottuvuudet muokattiin kolmeksi väitteeksi (muuttujaksi) seuraavalla tavalla:

1) **Menneisyyden opetuksia eli "Saksa" -ulottuvuus:** Diktaattoreja ei saa rohkaista osoittamalla heikkouden merkkejä sekä/tai teksti viittaa historiaan (liit-

tämispolitiikkaan/Hitleriin) tukeakseen tätä tulemaa.

2) **"Nyt on oikea hetki" – eli missä me olemme nyt -ulottuvuus:** Mahdollisuutta uuteen alkuun ei saa hukata sekä/tai teksti viittaa vallitsevaan historialliseen tilanteeseen (kommunismin häviö/demokratian voitto/YK:n aseman korostuminen) tukeakseen tätä ajatusta.

3) **Mihin olemme menossa eli "reilun pelin" -ulottuvuus:** Uuden politiikan väitetään tähtäävän oikeudenmukaisuuden ja eettisten periaatteiden kunnioittamiseen kansainvälisissä suhteissa sekä/tai teksti viittaa pienten kansojen oikeuksiin perusteena sille, ettei väkivaltaista anastusta voida hyväksyä.

Analyysin aineisto koostui norjalaisessa, ruotsalaisessa, saksalaisessa, suomalaisessa ja yhdysvaltalaisessa lehdistössä (2-3 lehteä kustakin maasta) tiettyinä ajankohtina julkaistuista Persianlahden sotaa koskevista uutisista ja pääkirjoituksista. Koodattaessa pyrittiin löytämään paitsi väitteiden esiintymät (+) myös niiden kieltäminen tai niihin tehdyt varaukset (-). Oli myös mahdollista, että juttu esitti sekä argumentin että siihen kohdistuvaa epäilyä.

Analyysin tulos uutismateriaalin osalta oli, että kyseiset argumentit eivät yleensä esiintyneet uutisteksteissä. Analysoiduista 4096 uutisjutusta vain 136:ssa kyseisiä väitteitä kosketeltiin tavalla tai toisella. Näihin juttuihin kohdistettu LCA tuotti seuraavan tuloksen (Kempf & Luostarinen 1997, 6):

TAULUKKO 16. Uusi Maailmanjärjestys -argumentteja sisältävien juttujen LCA-luokat ja kokonaisjakauma.

LCA-luokka	Luokan koko	Saksan opetukset (+)	Saksan opetukset (-)	Oikea hetki (+)	Oikea hetki (-)	Reilu peli (+)	Reilu peli (-)
1	0.433	**0.981**	0.051	0.000	0.000	0.086	0.017
2	0.279	0.163	0.000	**0.962**	0.079	0.206	0.000
3	0.244	0.000	0.000	0.013	0.000	**1.000**	0.000
4	0.044	0.000	0.000	0.000	0.000	0.000	**1.000**
Kokonaisj.	1.000	0.471	0.022	0.272	0.022	0.338	0.051

Osoittautui siis, että sikäli kuin väitteet esiintyivät, ne esiintyivät yleensä erikseen, sillä kolmessa ensimmäisessä luokassa kussakin korostuu yhden väitteen osuus. Samalla kuitenkin hahmottui myös kriittinen aineistoluokka, sillä neljäs (ja kaikkein pienin) analyysin tuottamista luokista näyttää koostuvan jutuista, jotka sisältävät vain kolmanteen väitteeseen kohdistuvaa kritiikkiä.

LCA ei ole kuitenkaan ollut pelkkä väiteanalyysin tai sellaista lähestyvän analyysin väline. Esimerkiksi mainittu *Journalism in the New World Order* -projekti käytti Bosnian sotaa koskevaa kirjoittelua tutkiessaan koodausapparaattia, joka oli viritelty ennen muuta psykologiassa käytettyjen käsitteiden varaan rakennetusta konfliktiteoreettisesta ajatusmallista ja sisälsi myös pragmaattisten elementtien tulkintaa (ks. Kempf, ilmestyy).Tutkimuksen ideana oli, että journalistiset tekstit saattavat avata horisonttia, joka sisältää konfliktin osapuolten vastavuoroiset näkökulmat itseensä ja toiseen tai sitten ne voivat rakentua tavalla, joka sulkee tällaista vastavuoroisten näkökulmien ottoa (jättäen jäljelle vain konfliktin jonkun osapuolen näkökulman itseensä ja muihin). Edellisen tavan ajateltiin tietysti olevan omiaan myötävaikuttamaan konfliktin purkamisessa, jälkim-

mäisen eskaloimaan sitä. Niinpä projekti pyrki suhteellisen monimutkaisen koodausapparaatin puitteissa kirjaamaan konfliktin kunkin osapuolen osalta muun muassa seuraavia seikkoja:

- Vastapuoleen tai vastapuoliin kohdistuvat uhkaukset ja painostus.
- Vastapuolen oikeuksien tai voimien vähättely, samoin omien voimien tai oikeuksien korostaminen.
- Osapuolten omien toimiensa puolustukseksi esittämät apologiat.
- Kaksisuuntaiset viestit ja kaksoissidokset.
- Omiin toimijoihin kohdistetut identifikaatiotarjoukset (samoin kuin vastaidentifikaatiot muiden osapuolten toimijoihin).

Niinikään koodatessa pyrittiin myös kirjaamaan edellä mainittuja seikkoja kyseenalaistavat teot. Koodaus ei siis edennyt niinkään konkreettisten väitteiden kuin tietyntyyppisten puhetekojen ja näistä syntyvien interaktiokuvioiden kirjaamisena.

Loppuhuomautuksia

Kracauerin esittämä kritiikki määrällisiä menetelmiä kohtaan perustui kohteen olennaisten rakenteellisten piirteiden pirstoutumiseen. Muiden tekstisisältöä numeeriseksi koodiksi muuntavien menetelmien tavoin myös LCA on alituisessa vaarassa altistua Kracauerin kritiikille atomisoivan datan tuottamisesta. Tähän Kracauerin osoittamaan ongelmaan ei liene muuta ratkaisua kuin muuttujien kehittäminen sellaisiksi, että ne saavat otteen tekstistä ja siinä kutoutuvasta "viittausten ja kietoutumisen verkosta". Sisällönanalyysin muuttuja-apparaatistoa on tietysti mahdollista kehittää paljonkin siitä, missä muodossa Kracauer oppi sen tuntemaan. Esimerkiksi edellä esitellyt konfliktiteoreettiset väitemuuttujat sekä puhetekojen luonnetta kuvaavat muuttujat lienevät tarkoituksiinsa varsin käypiä työkaluja. Lisäksi voisi ajatella, että tekstilingvistiikan, -semantiikan ja -pragmatiikan (yleensä kaikenlaisen diskurssianalyysinä itsensä esittelevän kieliteoreettisen tutkimuksen) kehittyminen voisi antaa tukea käytännön koodausta ja koodausapparaatin luomista koskeviin ongelmiin. Toki silti on aihetta painottaa, että tässä esityksessä olemme ennen muuta pyrkineet ymmärtämään menetelmän luonnetta ja toimintaa, emme sen ongelmia. Näin esityksestä tulee väistämättä hiukan fetisoivaa.

LCA:n vahvuudet tulevat esille yksittäisen havaintoyksikön sijasta koko aineiston tasolla. LCA on nimenomaan aineiston hahmoa tai struktuuria etsivä, ei niinkään havaintoja ryhmittelevä väline. Toisin kuin vaikkapa faktoripistemäärien avulla tuotetussa ryhmittelyssä LCA ei sijoita yksittäisiä havaintoyksiköitä yksiselitteisesti tiettyyn kohtaan latentteja ulottuvuuksia. Kustakin havaintoyksiköstä voidaan LCA:ssa sanoa vain, millä todennäköisyydellä se kuuluu mihinkin luokkaan. Faktorianalyysikin esittää latentteja ulottuvuuksia, mutta jos näiden lisäksi halutaan esittää muuttujien jakaumat, tullaan samalla kiinnittäneeksi yksittäiset havaintoyksikötkin tiettyihin kohtiin faktoreilla. Menetelmät lähestyvät asiaa eri suunnista ja saattavat tuottaa radikaalisti erilaisiin tulkintoihin johtavia tuloksia kuten tapahtui 1970- ja -80-luvuilla Englannissa suoritetuissa opetustyylejä koskevissa tutkimuksissa (Aitkin, Andersson, Hinde 1981). Tässä mielessä LCA:lla on jotain omaa annettavanaan.

Koska LCA perustuu muuttujaparien korrelaatioiden sijasta muuttuja-arvojen todennäköisyyksiin, sillä ei ole mitta-asteikollisia rajoituksia. Tämä on erityisen hyödyllinen ominaisuus sisällönanalyysissä, koska siinä käytetyt muuttujat ovat

usein laatueroasteikollisia. LCA soveltuu myös kyselyaineistojen analyysiin. Tosin kyselyjen asennemuuttujat voidaan tulkita järjestys- ja jopa suhdelukuasteikollisiksi, mikä voi puolustaa korrelaatiomenetelmien käyttöä. LCA ei tunnista vastauksista myöntyvyyden tai vastustuksen määrää, vaan tulkitsee vastausvaihtoehdot vain laadullisesti erilaisiksi.

Yksi LCA:n eduista on sen tuoma mahdollisuus yhdistää määrällistä ja laadullista analyysiä. Kun LCA:lla on saatu esiin aineiston piilevät luokat, voidaan näiden luokkien tyypillisiä tapauksia analysoida laadullisin menetelmin. LCA tavallaan huolehtii siitä, että laadullisen analyysin kohteet todellakin ovat aineistossa tyypillisiä. Toisin sanoen LCA antaa tietoa siitä, kuinka laajalle laadullisesta analyysistä saatuja tuloksia voidaan yleistää. Tällä tavoin LCA:ta ja laadullista analyysiä yhdistämällä voidaan siis hyödyntää yhtäaikaa sekä määrällisen analyysin tuomaa yleistettävyyttä että laadullisen analyysin tuomaa kohdeherkkyyttä. Lisäksi tyyppitapauksista tehdyt laadulliset analyysit voivat hyvinkin johtaa uusien muuttujien keksimiseen ja miksipä ei uuteen LCA:han näiden muuttujien perustalta. Voisi melkeinpä sanoa, että sikäli kuin analyysiapparaatille asetuvat vaatimukset täyttyvät, tekee LCA mahdolliseksi nähdä metsää puilta.

Kaikki edellä sanottu on käsittääksemme totta, kaunista ja oikein. Olemme silti päätymässä hiukan hämmentävään loppupäätelmään eri ryhmittelymetodien suhteista. Nykyään on ehkä tapana korostaa määrällisten ja laadullisten menetelmien rauhanomaista rinnakkaineloa, keskinäistä avunantoa ja omia sovellutusalueita. Vaikuttaisi kuitenkin siltä, että LCA:n kaltaisen tilastomatemaattisen ryhmittelymenetelmän ja hiukan toisenlaisten ryhmittelymenetelmien (joihin kuuluisivat sekä laadullinen tyypittely että edellä mainitut etäisyysindekseihin perustuvat menetelmät) välillä olisi varsin huomattava periaatteellinen ero.

Ajatelkaamme aluksi lehtijuttujen sisältöpiirteiden määrittelyn ja kuvailun avulla etenevää tyypittelyä. Tällainen tyypittely saattaisi jakaa aineiston kauniisti osajoukoiksi, jotka koostuisivat kukin sisällöltään suhteellisen samankaltaisista jutuista. Olisi myös mahdollista osoittaa, mitkä havaintoyksiköt kuuluvat mihinkin osajoukkoon. Analyysissä olisi siis koko ajan kyse havaintojen samankaltaisuuksista ja niihin perustuvista luokista (näin meneteltiin myös edellä kuvatussa etäisyysindeksiä käyttävässä ryhmittelyssä). Tämän jälkeen voidaan sitten jatkaa millaisin metodein tahansa.

LCA:n kaltaisen tilastomatemaattisen menetelmän laita on toisin. Se perustuu aksiomaattiseen oletukseen (paikallisen riippumattomuuden periaate) siitä, miten välittömän havainnon tavoittamattomissa oleva käyttäytyy kun se (esimerkiksi testein) pakotetaan välittymään ilmiömaailman piiriin. Luokittelu ei tällöin perustu samankaltaisten havaintojen ryhmittelyyn vaan oletukseen samankaltaisten havaintoyksiköiden käyttäytymisestä ja havaintojen ryhmittelyyn tämän oletuksen perusteella. Tällöin erilaiset havainnot saattavat hyvinkin tulla sijoitetuiksi samaan luokkaan, joskaan – toisin kuin laita on havaintoihin perustuvassa luokittelussa - ei voida aina varmuudella määrittää, mikä havaintoyksikkö kuuluu mihinkin luokkaan. Selviä tyyppitapauksia voidaan tietysti esittää niistä luokista, jotka koostuvat suhteellisen samankaltaisista havainnoista. Tällaisten tyyppitapauksien analyysia voidaan myös hyvin jatkaa vaikkapa laadullisella erittelyllä. Kyseenalaisempaa taas on kohdistaa laadullista jatkoerittelyä sellaiseen luokkaan, jossa erilaisia havaintoja on kutakuinkin yhtä paljon.

Havaintoihin perustuvan luokittelun kannalta tilastomatemaattisesti perustellut luokat voivat vaikuttaa älyttömiltä ja puhtaan teoreettisilta (kuten on väliluokan laita taulukossa 2). Tilastomatemaattiselta kannalta taas havaintoihin perustuva luokittelu saattaa vaikuttaa harhaiselta tai ainakin pinnalliselta. Riippuvuuksien taustalla olevat latentit rakenteet jäävät löytämättä. Kun alkuun on päästy jatkaa voidaan millaisin metodein vain, mutta lähtökohdan valinta vaikuttaisi ole-

van jokseenkin filosofinen kysymys. Jos taas filosofiat jättää sikseen, voi ehkä todeta, että mikäli tekijä on kohtuullisen varma siitä, millaisin muuttujin aineistoa on mielekästä haravoida voi LCA olla hyväkin työkalu. Muussa tapauksessa jokin aineistolähtöisempi työtapa tuskin on ainakaan huonompi vaihtoehto.

Viitteet

1 Sanoipa Berelson asiasta kannakseen mitä hyvänsä, hän on itse ollut intohimoinen luokittelija ja taulukoiden laatija. Kuvaillessaan määrällisen ja "laadullisen" sisällönanalyysin välisiä eroja ja yhtäläisyyksiä esitys etenee tietysti huolellisesti numeroituja eroavaisuuksia toteamalla ja kommentoimalla. Esitellessään sisällönanalyysin historiaa Berelson laatii viisivuotiskausiin perustuvan taulukon, joka kirjaa sisällönanalyyttisten tutkimusten määrällisen kehityksen. Kuvatessaan sisällönanalyysin käyttöalaa Berelson toteaa, että kentän laajuuden vuoksi ei ole helppoa esittää yksinkertaista luokittelua sisällönanalyysin eri muodoista. "Yksityiskohtaista kuvausta ja pohdintaa varten on kuitenkin erotettu seitsemäntoista sisällönanalyysin käyttötapaa (tai sovellusta tai funktiota)" (emt., 26). Muutenkaan teksti ei sanottavasti kärsi taulukoiden muodossa esiintyvän esimerkkimateriaalin puutteesta.

2 Antaessaan kirjansa alussa esimerkkejä siitä, millaisiin kohteisiin metodia on sovellettu, Berelson kuitenkin mainitsee seuraavankaltaisia esimerkkejä: Mitkä ovat Shakespearen näytelmien keskeiset kielikuvat? Miten kirjoittajan persoonallisuus heijastuu siinä, mitä hän kirjoittaa? Miten etnisiä vähemmistöjä kuvataan populaarilehdistön novelleissa? Listaa voisi jatkaa...

3 Lacord tekee ryhmittelyn koodauskuvioiden laskennallisten todennäköisyyksien perusteella, ei ryhmittelemällä todellisia havaintoyksikköjä. Tämän vuoksi kaikissa ratkaisuissa todellisia havaintoyksikköjä ei voi ryhmitellä täsmälleen sellaisiin luokkiin kuin Lacord esittää. Muuttuja-arvojen jakaumat sen sijaan täsmäävät aineiston kanssa kaikissa Lacordin esittämissä ratkaisuissa.

4 Moniluokkaisten ratkaisuissa täytyy koodauskuvioiden todennäköisyysarvoja laskettaessa ottaa huomioon luokkien määrä ja niiden koko. Esimerkiksi kaksiluokkaisessa ratkaisussa koodauskuvion 100 todennäköisyysarvo saadaan laskemalla ensin todennäköisyydet erikseen molemmissa luokissa:

luokassa 1: 0,075 * 0,875 * 1,000 = 0,065625
ja luokassa 2: 0,031 * 0,594 * 0,002 = 0,000037

Tämän jälkeen todennäköisyydet kerrotaan luokkakoolla ja lasketaan yhteen:

0,555 * 0,065625 + 0,445 * 0,000037 = 0,03643834

Koodauskuvion 100 todennäköisyys kaksiluokkaisessa ratkaisussa on siis 0,03643834. Koko kaksiluokkaisen ratkaisun tunnusluku (LOG-Like) saadaan laskemalla jokaisesta havainnosta vastaava todennäköisyysarvo, ottamalla siitä logaritmi (ln) ja laskemalla nämä yhteen. Nämä laskutoimitukset on esitetty taulukossa 5.

5 LOG-Like -arvojen erot taulukoissa 4 ja 5 esitettyihin johtunevat siitä, että laskennassa lienee käytetty eri määrää desimaaleja.

6 Toisinaan pääkomponenttianalyysiä pidetään omana, faktorianalyysiä muistuttavana mutta siitä kuitenkin selvästi poikkeavana analyysitapana (esim. Ranta & Rita & Kouki 1991). Toisinaan pääkomponenttianalyysiä pidetään pikemminkin yhtenä faktorianalyysin muotona (esim. Everitt & Hay 1992, 112-114). Erona menetelmillä on se, että varsinainen faktorianalyysi kiinnittää huomiota muuttujien kovarianssirakenteeseen, kun pääkomponenttianalyysi pyrkii selittämään muuttujien vaihtelua ylipäätään (Ranta & Rita & Kouki 1991, 459).

7 LCA:lia ei voinut tuottaa kolmiluokkaista ratkaisua, koska silloin ratkaisun vapausarvot olisivat menneet negatiivisiksi, eikä tulos olisi ollut pätevä. LCA kyllä tulosti 'laittoman' kolmiluokkaisen ratkaisunkin, mutta se ei ollut järkevästi tulkittavissa eikä se muistuttanut faktoripistemäärien avulla tuotettua ratkaisua

8 Faktorianalyysikin on laatueromuuttujista mahdollinen, mutta tällöin muuttujat pitää ensin dikotomisoida. Toisin sanoen jokaisesta 'vaihtoehdosta' tehtäisiin oma muuttujansa.

9 Kaikkiaan nämä teemat käsittelivät Iranin ja Kuwaitin konfliktin syntyä, varsinaisen Persianlahden sodan syntyä, sodan uhrien ja sen tuottamien kärsimysten kuvaamista, tapahtuneita sotarikoksia sekä konfliktin ratkaisuvaihtoehtojen käsittelyä.

Lähteet

Aitkin, Murray & Anderson, Dorothy & Hinde, John (1981)
Statistical Modelling of Data on Teaching Styles. Journal of the Royal Statistical Society 144(1981):4.

Alkula, Tapani & Pöntinen, Seppo & Ylöstalo, Pekka (1994)
Sosiaalitutkimuksen kvantitatiiviset menetelmät. Helsinki ja Juva: WSOY.

Andersen, Erling B (1982)
Latent Structure Analysis: A Survey. Scandinavian Journal of Statistics. 9(1982).

Berelson, Bernard (1952)
Content Analysis in Communication Research.

Everitt, Brian S. (1995)
Cluster Analysis. London, Sidney, Auckland: Arnold.

Everitt, Brian & Hay, Dale (1992)
Talking about Statistics. A Psychologist's Guide to Data Analysis. London, Melbourne, Auckland: Edward Arnold.

Heinonen, Ari (1997)
Sanomalehdistö ja Internet - toiveita, huolia, epätiotoisuutta. Journalismin tutkimuksen ja kehitystyön yksikön raportti. Tampereen yliopiston tiedotusopin laitoksen julkaisuja C 21.

Kempf, Wilhelm (1994)
Towards an Integration of Quantitative and Qualitative Content Analysis in Propaganda Research. Diskussionsbeiträge Nr 27/1994 der Projektgruppe Friedensforschung Projekt 13/85, Universität Konstanz.

Kempf, Wilhelm & Luostarinen, Heikki (1997)
New World Order Rhetorics. A Comparative Study of American and European Media During the Gulf War. Diskussionsbeiträge Nr 35/1997 der Projektgruppe Friedensforschung Projekt 13/85 & 590/95, Universität Konstanz.

Kempf, Wilhelm:
Escalating and deescalating aspects in the coverage of the Bosnian conflict - a comparative study. Teoksessa Kempf Wilhelm & Heikki Luostarinen (eds): Journalism in the New World Order - Studying War and the Media. Ilmestyy

Kracauer, S. (1990)
Fur eine qualitative Inhaltsanalyse. Teoksessa. Kracauer, Siegfried. Schriften. (Band 5). Frankfurt am Main: Suhrkamp.

Lazarsfeld, Paul F. (1950)
Logical and Mathematical Foundations of Latent Structure Analysis. In Stouffer, Samuel A. & al (eds.) Measurement and Prediction. Princeton & New Jersey: Princeton University Press.

Lazarsfeld, Paul F. (1959)
Latent Structure Analysis. In Koch, Sigmund (ed.) Psychology: A Study of a Science. Vol 3. Formulations of the Person and the Social Context. New York & Toronto & London: McGraw-Hill Book Company.

Luostarinen, Heikki (1986)
Perivihollinen. Tampere: Vastapaino

Nohrstedt, S & Ottosen, R. (eds) (1998)
Journalism in the New World Order. Gulf War, National News Discourses and Globalization. Volume 1 (ilmestyy).

Pietilä, Veikko (1997)
Joukkoviestinnän valtateillä. Tampere: Vastapaino.

Pietilä, Veikko (1970)
Johdaltusta sisällönerittelyyn II. Tiedotusopin laitoksen opetusmoniste. Tampereen yliopisto

Ranta, Esa & Rita, Hannu & Kouki, Jari (1991)
Biometria. Tilastotiedettä ekologeille. Helsinki: Yliopistopaino.

Sclove, Stanely L. (1987)
Application of Model-Selection Criteria to Some Problems in Multivariate Analysis. Psychometrica 52(1987):3.

Suhonen, Pertti (1994)
Mediat, me ja ympäristö. Helsinki: Hanki ja jää.