

II Digitaaliset tutkimusmenetelmät historiantutkimuksessa

SA-kuva-arkistoa louhimassa

Digitaaliset tutkimusmenetelmät valokuvatutkimuksen tukena

KIMMO ELO ja OLLI KLEEMOLA

Digitalisaation myötä myös historian tutkimuksen kannalta kiinnostavien aineistojen tarjonta sähköisessä muodossa lisääntyy jatkuvasti, mikä yhtäältä tuo tukijoiden ulottuville aiemmin vaikeasti saatavissa olleita aineistoja, toisaalta lisää tarvetta tietokoneavusteisten tutkimusmenetelmien soveltamiselle myös historian tutkimuksen puitteissa. Valokuvat ja muut niin kutsutut visuaaliset aineistot ovat yksi kiinnostava aineistoryhmä, joiden tietokoneavusteinen tutkiminen on kuitenkin edelleen varsin haasteellista. Tässä artikkelissa tarkastellaan mahdollisuutta hyödyntää digitalisoituihin valokuva-aineistoihin liitettyjä metatietoja kuvien visuaalisen sisällön tietokoneavusteisen tutkimuksen tukena. Artikkelin aineistona käytetään äskettäin digitalisoituna julkaistua puolustusvoimien 2. maailmansodan aikaista SA-kuvakokoelmaa. Saatujen tulosten valossa tietokoneavusteiset aineistonlouhinta ja -analyysimenetelmät tarjoavat varteenotettavan vaihtoehdon visuaalisia aineistoja hyödyntävälle tutkimukselle.

Johdanto

Viimeisen parinkymmenen vuoden aikana ”digitaaliset ihmistieteet” ovat paradigmana saavuttaneet yhä laajempaa huomiota myös historian tutkimuksen ja sen lähitieteiden piirissä. Yhtenä merkittävänä tätä kehitystä tukevana tekijänä voidaan nähdä sekä meillä että maailmalla käynnissä olevat digitalisointihankkeet, joiden kautta yhä kaukaisempiin aikakausiin liittyvät aineistot ovat yhä laajemmin saatavissa myös digitaalisessa muodossa. Digitaaliset arkistot eivät ainoastaan lyhennä (tai kokonaan poista) etäisyyttä historioitsijan ja hänen primaarilähteidensä väliltä, vaan myös luovat edellytykset uudentyyppisten digitaalisten tutkimusmenetelmien soveltamiselle myös sellaisten tieteenalojen puitteissa, jotka perinteisesti ovat tutkimustyössään tukeutuneet aineistojen manuaaliseen läpikäyntiin esimerkiksi lähiluennan avulla.

Useimmiten, etenkin juhlapuheissa, digitalisaatio näyttäytyy lähes yksinomaan positiivisena ilmiönä. Digitalisaatio, yhdessä avoimen datan

lisääntymisen kanssa, ei vain loisi tutkimukselle uusia lähdeaineistoja, vaan osaltaan vahvistaisi demokraattista toimintakulttuuria ja läpinäkyvyyttä, kun tutkimusaineistot ovat internetin kautta kaikkien ulottuvilla. On kiistämättä selvää, että digitalisaatio tulee poistamaan monia esteitä, jotka aiemmin liittyivät etenkin aineistojen keräämiseen. Mutta samalla digitalisaatio myös luo uusia hierarkioita ja epätasa-arvon lähteitä. Ensinnäkin, digitalisaation hyödyntäminen tutkimuksessa ei voi tapahtua ilman riittävää resursointia. Jo yksinomaan aineistojen digitalisointi on kallista ja aikaa vievää, mutta tämän lisäksi suurten aineistojen käsittely edellyttää panostusta asianmukaiseen ja riittävään laskentatehoon tietokoneiden osalta. Toiseksi, aineiston siirtäminen verkkoon tekee niiden käyttämisen riippuvaiseksi sekä tietokoneista että (ja etenkin) laadukkaista tietoliikenneyhteyksistä. Vaikka useimmat nk. länsimaiset yliopistot ovat tähän asti kyenneet pysyttelemään myös teknisen kehityksen mukana ja tarjoamaan tutkijoille asianmukaiset tietotekniset valmiudet, ei tilanne suinkaan ole yleistettävissä koko globaalia tiedeyhteisöä koskevaksi. Ei tarvitse siirtyä kovinkaan kauas itään tai etelään Euroopasta huomatakseni, että toimiva IT-infrastruktuuri ei suinkaan ole itsestäänselvyys.

Kolmas, ja samalla aivan oma tekijänsä on aineistojen hyödyntämisessä vaadittava menetelmäosaaminen – tai pikemminkin sen puute. Kuten esimerkiksi Jon Olsen äskettäin julkaistussa kirjoituksessaan totesi, digitaaliset ihmistieteet vaativat tutkijalta sellaisia kompetensseja, joita ei tällä hetkellä tarjota kuin poikkeustapauksissa perinteisen yliopistollisen tutkintorakenteen puitteissa.¹ Tähän lienee monia syitä, joista vähäisin ei liene eri tieteenalojen pyrkimys suojella omaa tieteenalaidentiteettiään myös tieteenalaspesifien tutkimusmenetelmien avulla. Mutta kuten Olsen huomauttaa, digitaaliset ihmistieteet ovat ensisijaisesti joukkuepeliä, joka pyrkii tuomaan yhteen eri tieteenalojen osaamista yhteisten ongelmien ratkomiseksi. Rohkenemme väittää, että niin kauan kuin eri oppiaineiden menetelmäopetus hyljeksii digitaalisia tutkimusmenetelmiä, digitalisaation hyödyntäminen tutkimuksessa jää puolinaiseksi.

1 Ks. tark. Olsen 2015.

Vielä huolestuttavampi mahdollisuus on, että eri oppiaineet alkavat kehittää omia määritelmiään digitaalisille ihmistieteille – tällaisesta kehityksestä on valitettavasti joitakin merkkejä suomalaisessa tiede- ja yliopistokentässä. Näin ei pelkästään käperrytä omaan erinomaisuuteen, vaan myös irrottaudutaan niistä kansainvälisistä keskusteluista ja kehityskuluista, joita digitaalisten ihmistieteiden ympärillä on viimeisen 10–15 vuoden aikana käyty.

Tätä kirjoitettaessa suomalainen yliopistomaailma etsii tulevaisuuttaan niukkenevien resurssien ja kasvavien (tulos)odotusten ristipaineessa. Samaan aikaan julkisen vallan taholta lisätään painetta tuottaa sekä digitalisaatioon liittyen että digitalisaatiota hyödyntäen loistavia uusia innovaatioita – ja nämä mieluummin heti tänään kuin huomenna. Valitettavasti digitalisaatio näyttäytyy hallituspolitiikassa lähinnä yhtenä säästökeinona, mikä ei ole missään suhteessa niihin resurssitarpeisiin, joita digitaalisen tutkimuskulttuurin ja tähän liittyvän *tieteellisen* osaamisperustan rakentaminen sekä ajallisesti, rahallisesti että muilla resursseilla mitattuna vaatisi. Nykyisen säästöpolitiikan avulla luodaan suomalaisille yliopistoille ja korkeakouluille lähinnä erinomaiset edellytykset pudota digitaalisten ihmistieteiden kehityksen kerkasta.

Edellä sanottu pätee luonnollisesti myös historiatieteisiin. Vaikka historian tutkimuksen piirissä digitalisaatiokeskustelun yhtenä pääjuonteena onkin ollut juuri kysymys uusien, aiemmin hankalasti hyödynnettävien aineistojen saatavuuden parantumisesta, digitalisaation redusointi pelkäksi aineistokysymykseksi olisi mielestämme virheellistä. Digitalisaatiossa on kyse koko tieteen kenttään vaikuttavasta ilmiöstä, joka väistämättä tulee myös muuttamaan suhtautumistamme tieteen perusteisiin. Hieman kärjistäen ilmaistuna digitaalisten aineistojen tuottama lisäarvo realisoituu vasta silloin, kun tutkijoilla on käytössään menetelmät, joiden avulla digitaalisten aineistojen mahdollisuuksia voidaan täysimääräisesti hyödyntää. Siinä missä tutkimusaineiston odotetaan sisältävän tutkimusongelman kannalta relevanttia tietoa, on valitulla tutkimusmenetelmällä avainrooli tämän tiedon esiin kaivamisessa.

Toki digitaalisia aineistoja voidaan tutkia myös perinteisten – siis ei-digitaalisten menetelmien – avulla ja sitä kautta saavuttaa uutta tietoa. Näiden rinnalle tarvitaan kuitenkin myös digitaalisia, käytännössä siis

tietokoneavusteisia tutkimusmenetelmiä, joita soveltamalla aineistoista voidaan saada esiin sellaista tietoa, jota perinteisin menetelmin olisi vaikea, jopa mahdoton saavuttaa. Mutta tämän ”louhinnan” lisäksi digitalisaation voidaan odottaa muuttavan myös käsitystämme ja tapaamme esittää uutta tietoa ja tutkimustuloksia. Digitaalisuuden rinnalla olisikin hyvä kuljettaa myös *visuaalisuuden* käsitettä: digitaalisten menetelmien yksi keskeinen ero perinteiseen tekstimuotoiseen esitykseen liittyy juuri menetelmien tarjoamiin mahdollisuuksiin esittää monimutkaisia kokonaisuuksia graafisesti, jolloin uusi tieto on luonteeltaan visuaalista.

Nyt käsillä oleva artikkeli osoittaa, millä tavoin suurten, strukturoimattomien aineistojen louhintaa voidaan käyttää valokuvien tutkimuksen tukena. Artikkelin perustana on ymmärrys, jonka mukaan digitaalisten tutkimusmenetelmien ja perinteisen humanistisorientoituneen tutkimuksen välillä ei vallitse nollasummapeli, vaan molempia hyödyttävä *win-win* -tilanne. Siinä missä digitaalinen historian tutkimus epäilemättä hyötyy tietojenkäsittelytieteiden piirissä kehitetyistä analyysimenetelmistä, edellyttää näiden menetelmien soveltaminen kuitenkin niiden adaptoimista, sovittamista soveltumaan historian tutkimukselle ominaisten kysymyksenasettelujen piirissä. Metodologista sillanrakennusta tarvitaan kuitenkin mielestämme juuri tietokoneavusteisen aineiston louhinnan jatkohyödyntämiseen: koska aineiston louhinnan tulokset eivät ole tutkijan itsensä tulkintaa, vaan tietokoneen kylmän laskennan tulosta, tutkijan on välttämätöntä ymmärtää, millä tavoin tietokone tulokset tuottaa. Tämän ymmärryksen tuottaminen on myös yksi tämän artikkelin tavoitteista, kun pyrimme analyysin yhteydessä myös avaamaan sitä, millä tavoin tietokoneavusteinen analyysi on toteutettu.

Laajemmin tarkasteltuna artikkelimme sijoittuu eksploratiivisen tutkimuksen kenttään. Eksploratiivisella tutkimuksella tarkoitamme tutkimusta, jonka tavoitteena on kehittää uusia tapoja lähestyä olemassa olevia aineistoja uusien kysymysten, näkökulmien ja tutkimusalueiden löytämiseksi.² Kirjoittajien suhde digitaaliseen tutkimukseen on siten

2 Ks. esim. Budrick ym. 2012; Clavert ja Noiret 2013; Ukkonen 2013.

yhtäältä tekemistä ja kehittämistä painottava, toisaalta välineellinen.³ Ensimmäinen siksi, että artikkelimme keskiössä ovat historiallisista lähteistä muodostettavien määrällisten digitaalisten aineistojen analyysitavat ja näiden soveltuvuuden arviointi. Jälkimmäinen puolestaan siksi, että sovellamme digitaalisia analyysivälineitä sekä löytääksemme uusia näkökulmia aineistoon (eksploraatiivisuus) että rakentaaksemme menetelmiä hyödyntämällä tietynlaista sotahistoriallista tulkintaa aineistomme pohjalta. Näin palaamme edellä esitettyyn argumenttiin menetelmäkysymyksen keskeisyydestä digitaalisen historian tutkimuksen kehittymiselle. Menetelmät – siis se, *miten* katsomme – vaikuttavat olennaisesti siihen, *mitä* näemme.

Artikkelin aluksi esittelemme käyttämämme tutkimusaineiston sekä sovellettavat tutkimusmenetelmät. Artikkelin ytimen muodostavassa analyysiluvussa havainnollistamme konkreettisesti aineiston metatietojen tietokoneavusteisen eksploraatiivisen louhinnan ja analysoinnin toteuttamisvaiheet ja sen tuottamat tulokset. Tulosten esittelyssä ja niiden jatkotulokinnan perustana käytämme graafisia visualisointeja, erityisesti verkostanalyysiä. Visuaalisen analyysin keskiössä ovat kuvateksti-analyysin kautta havainnoidut klusterit, jotka jakavat kuva-aineiston toisistaan poikkeaviin ryhmiin. Tulkitsemalla eri klustereista valikoituja kuvia niiden historiallisessa kontekstissa osoitamme, että kuvatekstien eksploraatiivisen louhinnan avulla on mahdollista jakaa kuvakorpus sel-laisiin temaattisiin kokonaisuuksiin, jotka luotettavasti vastaavat kuvien visuaalista sisältöä. Artikkelin lopuksi summaamme keskeiset havainnotomme ja suhteutamme ne osaksi digitaalisten ihmistieteiden laajaa kenttää.

Haluamme tässä kohdin vielä erikseen korostaa, että artikkelimme ensisijainen tavoite on menetelmällinen. Tavoitteena ei ole löytää uusia sotahistoriallisia faktoja tai tulkintoja, mikä näin perusteellisesti tutkitun ja keskustellun aihepiirin osalta tuskin edes olisi mahdollista. Sotahistoriaa on kuitenkin tutkittu pääosin perinteisten historian tutki-

3 Itsepositioidinnin perustana käytämme Haverisen ja Suomisen (2015) artikkelissaan esittämää digitaalisen tutkimuksen digitaalisten ihmistieteiden tutkimuksen nelikenttämallia.

muksen menetelmien avulla, mikä on artikkelimme tavoitteiden kannalta erittäin tärkeä seikka. Osoittamalla digitaalisten tutkimusmenetelmien avulla esiin louhittujen tulosten nivoutuvan osaksi historian tutkimuksen kokonaiskuva voimme todistaa, että esittelemiemme digitaalisten tutkimusmenetelmien avulla on *yleisesti* mahdollista löytää myös historian tutkimuksen näkökulmasta relevantteja tuloksia, jotka toimivat perustana historiallisesti tiedostavalle analyysille ja tulkinnalle.

Aineisto ja menetelmät

Artikkelimme primaariaineistona toimii 27.4.2013 digitalisoituina kuva-arkistona avoimeen, julkiseen käyttöön lanseerattu Suomen armeijan 2. maailmansodan aikainen sotavalokuvien kokoelma, ”SA-kuvaportaali”.⁴ Ennen digitaalisen SA-kuva-arkiston avautumista armeijan sodanaikaiset kuvakokoelmat olivat tutkijoiden käytössä vain paperikuvina, joiden arkistoinnista ja hallinnoinnista vastasi puolustusvoimien kuvakeskus.

SA-kuva-aineisto muodostui 2. maailmansodan aikana suomalaisten sotakuvaajien toimesta. Vielä talvisodan aikana riviisotilaat hoitivat sotavalokuvauksen eikä erikoistuneita sotavalokuvaajia ollut. Systemaattinen TK-kuvaustoiminta (TK tarkoittaa tiedotuskomppaniaa, tiedottamiseen ja propagandaan erikoistunutta Suomen armeijan sodanaikaista yksikköä)⁵ sai alkunsa, kun suomalaiset havahtuivat talvisodan kokemusten perusteella kuvallisen tiedotustoiminnan välttämättömyyteen. Jatkosotaan lähtivätkin Saksan mallin mukaisesti organisoidut tiedotuskomppaniat, joiden tehtäväkenttä oli kuitenkin merkittävästi saksalaisia vastineitaan, propagandakomppanioita, monipuolisempi. Suomessa ei sotavalokuvilla pyritty ainoastaan luomaan ja levittämään propagandaa vaan keskeisenä tavoitteena oli kansatieteen ja sotahistorian näkökulmasta dokumentoida sodankäyntiä sekä suomalaisten miehittämän alueen väestön tapoja ja oloja. Näiden tavoitteiden puitteissa TK-kuvaajien toivottiin keskittävän dokumentoimaan ja kuvaamaan erityisesti:

4 ”SA-kuvaportaali”.

5 Sodanaikainen kirjoitusasu yksikköjen nimelle oli tiedotuskomppania.

- sotatoimien kulkua;
- suomalaista sotilasta siten että hänelle ominaiset piirteet ja taistelutavat tulevat esille;
- aseita toiminnassa erilaisissa tehtävissä ja vaihtelevassa maastossa;
- omia ja vihollisen linnoituslaitteita ja kenttävarustustöitä, myös siten että aseiden aiheuttamat vauriot ilmenivät;
- eri aselajien toimintaa mahdollisimman monipuolisesti;
- takaisin vallattuja alueita siinä kunnossa kuin ne ovat joutuneet suomalaisjoukkojen käsiin;
- vallattuja alueita myös puhtaasti maasto- ja maisemakuvina.

Erityisohjeistus koski Itä-Karjalaa, jonka asujaimistosta toivottiin kuvamateriaalia, joka todistaisi alueen asukkaiden ja suomalaisten sukulaisuuden. Yleisellä tasolla kuvaajia myös ohjeistettiin dokumentoimaan aiheita, jotka todistaisivat vihollisen julmuutta ja viholliselle aiheutettujen tappioiden suuruutta.⁶

Vaikka SA-kuvaportaaliin sisältyvä aineisto on laaja ja digitoituihin kuviin on myös liitetty kuvaan liittyviä lisätietoja kuten kuvauspäivä ja kuvaajan nimi, aineiston hyödyntämistä hankaloittaa se, että SA-portaalin kautta ei ole mahdollista ladata kuvatietueita jatkoanalyysiä varten. Kuvatietueella tarkoitetaan tässä sekä itse digitalisoitua valokuvaa että siihen liitettyjä metatietoja. Metatiedolla tarkoitetaan tietoa tiedosta, jonka avulla metatiedon kohteena olevalle hyvin erilaisissa muodoissa (esim. dokumentti, kuva, audiovisuaalinen muoto, piirustus) esiintyvälle tiedolle luodaan yhteinen ja keskenään vertailukelpoinen luokittelu- ja kuvailuterminologia. Metatieto voi siten pelkästään kuvailla kohteensa keskeisen sisällön (esim. asiasanat) tai rikastaa tietoa sellaisella informaatiolla, joka ei itsessään sisälly metatiedon kohteena olevaan objektiin (esim. paikkatiedot valokuvissa).

Koska tietokoneavusteinen analyysi edellyttää aineiston muuntamista sovellettavan digitaalisen menetelmän ymmärtämään muotoon, SA-kuvaportaalin tiedot oli tavalla tai toisella saatava siirrettyä jatko-

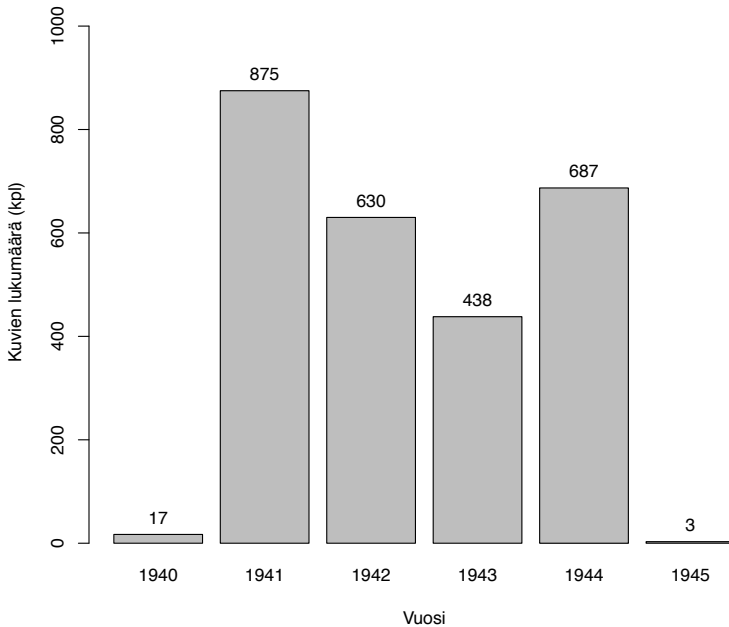
6 Tiedotuskomppanioista ja niiden toiminnasta ks. Kleemola 2013.

analyysiä varten. SA-kuva-arkistossa on reilusti yli 100 000 digitoitua kuvaa, mistä syystä korpuksen kerääminen käsin poimimalla ei ollut todellinen vaihtoehto. SA-kuva-arkistoa voi kuitenkin käyttää myös Suomen museoiden, arkistojen ja kirjastojen aineistojen yhteisen FINNA-hakuportaalien (www.finna.fi) kautta, mikä mahdollisti kuvatietueiden keräämisen tätä tarkoitusta varten koodatun pienen tietokoneohjelman avulla. Ohjelma latsi kuvanumeroiden perusteella kaksi tiedostoa FINNA-palvelusta. Ensimmäinen oli varsinaisen digitalisoitu valokuva, toinen kuvan metatiedot sisältävä HTML-muotoinen sivu. Metatietosivut prosessoitiin Python-kielellä⁷ laaditulla ohjelmalla, joka erotti HTML-koodista tarvitsemamme metatiedot ja tallensi ne taulukkona siten, että taulukon jokainen rivi vastasi yhtä kuvaa. FINNA-portaalissa SA-kuva-arkiston valokuvien tärkeimpiin metatietoihin kuuluvat kuvan ottamisajankohta, kuvaaja sekä kuvateksti. Aineiston erityispiirteistä johtuen metatiedoissa on jonkin verran puutteellisuutta, esimerkiksi kaikkien kuvien osalta tiedossa ei ole kuvan syntyhetkeä. Näin saatu taulukko, jossa oli tiedot yhteensä noin 100 000 kuvasta, siirrettiin R-tilasto-ohjelmistoon jatkoanalyysiä varten.⁸

Tätä artikkelia varten suodatimme aluksi keräämästämme aineistokorpuksesta kaikki ne kuvatietueet, joiden kuvatekstissä esiintyi saksatermi tai sen johdannainen. Suomessa toimineiden saksalaisten joukkojen kuvaaminen suomalaisten TK-kuvaajien toimesta on mielenkiintoinen ja tähän mennessä vähän tutkittu teema suomalaisessa jatkosodan ajan propagandatoiminnassa. Suomen tulenarka poliittinen asema heijastui nimenomaan saksalaisten kuvaamiseen. Esimerkiksi 2.7.1941 Päämaja kehotti suomalaisia TK-miehiä ryhtymään ”terve-

7 Python on moderni, nykyaikainen ohjelmointikieli, jonka suhteellisen helppo omaksuttavuus sekä tarjolla olevat laajennukset mm. tekstiaineistojen tai sosiaalisen median käsittelyyn ovat tehneet siitä suosittu myös digitaalisista ihmistieteistä kiinnostuneiden tutkijoiden piirissä. Ks. tark. ”Python-verkkosivu”.

8 R Core Team 2013. R on avoimeen lähdekoodiin perustuva, laajassa tieteellisessä käytössä oleva tilastoanalyysiohjelmisto. Ohjelmisto tarjoaa erinomaiset edellytykset laajojen tilastoaineistojen käsittelylle ja analysoinnille. R on laajasti käytössä myös esimerkiksi kielentutkijoiden sekä muiden, tekstiaineistojen tutkimisesta kiinnostuneiden piirissä, koska se tarjoaa erilaisia työkaluja mm. tekstiaineistojen louhintaan ja analysointiin.

Saksa-teemaan liittyvien kuvien määrät 1940–1945

Kuva 1. Saksa-teemaan liittyvien kuvien ajallinen jakauma 1940–1945.

henkiseen ja toverilliseen kilpailuun” saksalaisten PK-miesten⁹ kanssa, tavoitteenaan että ”saksalaiset liittolaisemme” joutuisivat mahdollisimman paljon lainaamaan suomalaista aineistoa.¹⁰ Näin toimien pyrittiin välttämään kiusallisia tilanteita, ettei Suomessa oleva saksalaisten joukkojen tiedotustoiminta tulisi rinnastetuksi suomalaiseen tiedotustoimintaan. Ohjekäskyssä annettu kehoitus heijastaa hyvin ulkopoliittisen tilanteen hankaluutta: suomalaisten piti kaikin keinoin, siis myös

9 PK-lyhenne viittaa saksalaisen propagandayksikön nimeen ”Propagandakompanie” (suom. propagandakomppania).

10 Päämajan ohjekäskyt tiedotuskomppanioille, käsky 6. Päämajan tiedotusosaston arkisto, kansio T20680/13, Kansallisarkisto (Sörnäisten toimipiste).

kuvapropagandassa, pyrkiä korostamaan oman sotansa ”erillisyyttä” ulkopoliitikassa keskeisen erillissotateesin uskottavuuden lujittamiseksi. Erillissotateesin uskottavuuden ylläpitämiseksi saksalaisten runsas käsittely TK-kirjoituksissa tai -kuvisa ei ollut missään vaiheessa mitenkään toivottua, joskaan sitä ei ennen kesää 1944 suoraan kiellettykään.¹¹

Aineistoomme sisältyy yhteensä 2650 kuvaa, joiden kuvatekstissä viitataan saksalaisiin eli niissä esiintyy joko sana ”Saksa” tai jokin saksatermin johdannainen (saksalainen, saksalais- jne.). Aihepiiriä näytetään myös kuvatun melko tasaisesti läpi sotavuosien eli vuosina 1941–1944 (Kuva 1). Jos tätä tasaisuutta ajatellaan vasten sotatapahtumien aikajanaa, herää luonnollisesti kysymys, millaisissa yhteyksissä saksalaisia kuvattiin ja miten kuvien tematiikat linkittyvät osaksi sotahistoriallista kokonaiskuvaa. Louhimalla ja analysoimalla saksalaisiin liittyviä sotakuvia osoitamme, että tietokoneavusteiset menetelmät tarjoavat yhden tavan löytää lähtökohtia tähän kysymykseen vastaamiseksi.

Koska artikkelimme on johdannossa sijoitettu myös *visual history* -kenttään, tarkempaan analyysiin valittujen kuvien poiminta puhtaasti kuvatekstien perusteella saattaa tuntua jonkin verran erikoiselta, eihän valinnassa siis näytettäisi huomioitavan lainkaan kuvien visuaalista tematiikkaa. Miksi emme siis ole käyttäneet valinnassa hyväksemme kuvien *visuaalista* sisältöä ja valinneet kuvia, joissa on saksalaisia elementtejä? Kuvatekstien käyttämiseen perustana on kaksi yksinkertaista syytä. Ensinnäkin, mikään tällä hetkellä käytössä oleva kuvantunnistus-tekniologia ei kykene automaattisesti tunnistamaan ”saksalaisia” elementtejä kuvista. Käytännössä kaikki kuvantunnistusmenetelmät perustuvat ohjelmiston opettamiseen tunnistamaan eri elementeille tyypillisiä kaavamaisuuksia (*pattern*) tuhansien, jopa miljoonien esimerkkikuvien avulla,¹² mikä ei nyt käsillä olevan artikkelin osalta olisi ollut mahdollista toteuttaa sen paremmin taloudellisesti kuin ajallisestikaan. Toisaalta, artikkelin yhtenä tavoitteena on nimenomaan osoittaa, että niin kutsuttujen tekstimuotoisten metatietojen avulla voidaan muodostaa kohtuullisen luotettava käsitys visuaalisen aineiston sisällöistä.

11 Pilke 2011, 121.

12 Kuvien visuaalisen sisällön muuttamisesta tekstimuotoon ks. esim. Vinyals ym. 2015.

Samalla on toki perusteltua lähteä siitä, että käyttämämme melko yksinkertainen louhintamenetelmä ei varmastikaan ole löytänyt jokaista kuvaa, jossa esiintyy saksalaisia ”elementtejä”. Löydettyjen kuvien joukossa on myös kuvia, joissa Saksa-yhteys syntyy esimerkiksi ”saksalainen IT-tykki”-tyyppisten merkintöjen kautta. Koska sen paremmin FINNA-kuin SA-kuva-portaaliakaan ei tarjoa esimerkiksi asiasanoitettua metatietoa, olisi molempien vinoumien eliminointi käytännössä vaatinut aineiston manuaalista läpikäyntiä. Löydettyjen kuvien lukumäärä on kuitenkin niin suuri, että virheellisesti mukaan päätyneiden kuvien muodostama vinouma ei ole kovinkaan merkittävä. Samalla otoksen koko on myös tilastollisesti katsoen riittävä, jotta sen perusteella voidaan tehdä koko aineistoon yleistettävissä olevia johtopäätöksiä. Ja lisäksi jäljempänä esiteltävien analyysitulosten yksi tärkeä tehtävä on antaa pohjaa myös siihen kysymykseen vastaamiseksi, voidaanko sovellettua poimintamenetelmää pitää luotettavana kuvien visuaalisen sisällön analysoimiseksi.

Aineiston analysointi rakentuu kaksivaiheisena prosessina. Ensimmäinen, tietokoneavusteinen analyysi perustuu SA-kuva-aineistoon sisältyvien metatietojen, erityisesti kuvatekstien tietokoneavusteiseen louhintaan ja analysointiin. Kuvatekstien louhinnassa sovellamme eksploratiivista, puoliautomaattista ryhmittelyä eli klusterointia, jonka avulla pyrimme löytämään aineistoomme sisältyvien kuvatekstien perusteella toisilleen samankaltaiset – siis samaan ryhmään eli klusteriin kuuluvat – kuvat. Tekstinlouhinta (*text mining*) itsessään on yläkäsite tutkimusmenetelmille, joiden avulla tekstiaineistosta pyritään tietokoneavusteisesti löytämään ja erottelemaan aiemmin tuntemattomia tai muulla tavoin kiinnostavia rakenteita ja systemaattisia piirteitä. Tekstinlouhintaa voidaan tehdä myös ennalta määrättyjen kriteerien perusteella, jolloin aineistoa pyritään myös luokittelemaan tai ryhmittelemään. Useimmiten tekstinlouhintaa sovelletaan sellaisten suurten aineistojen kohdalla, joiden työstäminen manuaalisesti esimerkiksi lähiluennan kautta olisi joko erittäin työlästä tai kokonaan mahdotonta. Tekstinlouhinnassa voidaan soveltaa hyvinkin erilaisia menetelmiä kuten klusterointia, semanttisia verkostoja tai itseorganisoituvia karttoja.¹³

13 Ks. esim. Berry ja Kogan 2010; Weiss ym. 2010; Leetaru 2012.

Tekstinlouhinnan tulosten esittämisessä ja analysoinnissa tukeudumme tilastollisten menetelmien ohella myös verkostoanalyysiin. Verkostoanalyysin formaalina perustana on graafiteoria, jonka mukaan verkostolla tarkoitetaan pisteistä (*nodes*, *vertices*) ja niitä yhdistävistä viivoista (*edges*) muodostuvaa kuviota eli graafia, jolla on tietty rakenne eli topologia.¹⁴ Sekä pisteisiin että viivoihin voidaan liittää niiden ominaisuuksia tarkentavia määreitä (attribuutit), joita voidaan edelleen hyödyntää varsinaisessa analyysissä. Attribuuttien avulla verkoston rakennetta on siis mahdollista tarkentaa vastaamaan mahdollisimman hyvin kohteena olevaa ilmiötä. Vaikka verkostoajattelun juuret ulottuvat aina 1800-luvun loppupuolelle,¹⁵ vasta tietokoneiden yleistyminen ja erityisesti niiden laskentatehon kasvu sekä verkostojen graafiseen esittämiseen eli visualisointiin tarkoitettujen helppokäyttöisten ohjelmistojen tarjonnan laajeneminen on lisännyt humanististen ja yhteiskuntatieteiden kiinnostusta verkostoanalyysiä kohtaan. Kehitystä ovat tukeneet ne tietoteknisesti orientoituneet digi-ihmistieteilijät, jotka ovat pyrkineet rätätälöimään perinteisen verkostoanalyysin menetelmiä soveltumaan paremmin ihmistieteille relevanttien kysymysten tutkimiseen.¹⁶ Tieteenalan erityispiirteet huomioiden ei liene kovinkaan yllättävää, että suurin kiinnostus on kohdistunut tekstiaineistojen kannalta relevanttien tutkimusmenetelmien – esimerkiksi tekstinlouhinta tai semanttinen analyysi – edelleen kehittämiseen.

Vaikka verkostoanalyysi menetelmänä analysoi, mallintaa ja visualisoi tutkimusaineistoon sisältyviä vuorovaikutussuhteita, tutkimuskohteen itsessään ei tarvitse (mutta se toki voi) olla verkosto. Monissa tutkimuksissa verkostoanalyysiä sovelletaan samalla eksploratiivisella perusajatuksella kuin tässä artikkelissa eli sovelletaan verkostoanalyysimenetelmiä aineistossa piilevien yhteyksien löytämiseksi, analysoimiseksi ja graafiseksi esittämiseksi. Vaikka tämä toki edellyttää aineiston

14 Hyvänä johdatuksen graafiteoriaan toimii esim. ks. Ruohonen 2013. Verkostoajattelusta ja verkostoanalyysistä laajemmin ks. esim. Prell 2012; McCulloh 2013; Eriksson 2015.

15 Ks. Eriksson 2015.

16 Hyvänä katsauksena tähän kehitykseen ks. Schultz-Jones 2009.

valmistelua muuttamalla aineisto verkostanalyysin edellyttämään muotoon sekä mahdollisesti aineiston rikastamista muista lähteistä saaduilla aineistoilla, varsinainen aineistoon sisältyvien suhteiden analysointi tapahtuu verkostanalyysimenetelmiä hyödyntämällä.

Tietokoneavusteisen sisällönlouhinnan ja verkostanalyysin ohella sovellamme tässä artikkelissa *visual history* -tutkimusparadigmaan liittyviä menetelmiä. Kyseinen paradigma syntyi 1990-luvun loppupuolella reaktionä siihen tosiseikkaan, että yhä suurempi osa yhteiskunnallisesta vuorovaikutuksesta tapahtui joko suoraan tai välillisesti visuaalisuuteen perustuvana. Tutkimusalueen uranuurtajana on ollut erityisesti Flensburgin yliopiston professori Gerhard Paul, joka on tutkimuksissaan hyödyntänyt kuva-aineistoja monipuolisesti nimenomaan tutkimuksen kohteina ja primäärilähteinä. Etenkin saksalaisen tiedeyhteisön piirissä tutkimussuuntaus on saavuttanut etabloituneen aseman.¹⁷ Tieteellisille paradigmoille tyypillisesti *visual history* -kaan ei muodosta metodi-, teoria- tai aineistopohjaltaan yhtenäistä, monoliittista koulukuntaa, vaan toimii pikemminkin eräänlaisena kehyksenä tai ”sateenvarjona”, jonka alle mahtuu monentyyppistä tutkimusta. Yhteistä näille on kuitenkin se, että niiden kohteena ovat visuaaliset aineistot julisteista, valokuvista ja elokuvista aina postimerkkeihin asti. Paradigman piirissä tehtävää tutkimusta yhdistää ajatus, että kuvia ei tule tarkastella pelkkinä taideteoksina, joita arvioidaan taiteellisen arvon mukaan, eikä myöskään pelkkinä tekstuaalisiin lähteisiin perustuvien tutkimusten kuvittamiseen soveltuvina illustraatioina. Tämän sijaan *visual history* tarkastelee kuvia itsenäisinä tiedon ja merkitysten tuottajina, joita voidaan tarkastella tekstuaalisten lähteiden kanssa samanarvoisina historiallisina dokumentteina.¹⁸

17 *Visual history*n suosiosta kertoo kesäkuussa 2013 avattu, Leibnitz-yhteisön rahoittama *Visual history* -internetportaali (visual-history.de) sekä vuonna 1981 perustetun *Fotogeschichte*-lehden vakiintunut asema saksalaisten tiedejulkaisujen kentällä. Osaltaan *visual history* -suuntauksen vakiintuneesta asemasta kertoo myös se, että sille on julkaistu omia metodioppaita (esim. Jäger 2009 ja Brocks 2012) ja sitä käsitellään myös yleisissä historiantutkimuksen metodioppaissa (esim. Bösch & Danyel 2012).

18 ”Visual history”-verkkosivut.

Kuvatekstien louhinnan tulosten jatkoanalysoinnissa olemme hyödyntäneet R-tilasto-ohjelmistoa aineiston tilastollisten analyysien osalta sekä Gephi-verkostoanalyysiohjelmistoa¹⁹ verkostoanalyysien ja -visualisointien toteuttamisessa. Molemmat ohjelmistot ovat käyttäjärjestelmäriippumattomia, avoimeen lähdekoodiin perustuvia ilmaisohjelmistoja, joita käytetään laajasti monilla tieteenaloilla.

Analyysi ja tulokset

Kuvatekstien tietokoneavusteisen analysoinnin aloitimme kuvateksteissä esiintyvien termien tarkastelulla. Ensimmäisessä vaiheessa metatietoihin tallennetut kuvatekstit prosessoitiin hyödyntämällä Turun yliopiston tietojenkäsittelytieteiden laitoksen BioNLP-tutkimusryhmän kehittämää automaattista, vapaasti levitettävää *Finnish Dependency Parser* -ohjelmaa, jonka avulla suomenkielisissä teksteissä esiintyvät sanat voidaan palauttaa perusmuotoon sekä rikastaa termejä kielitieteellisillä metatiedolla kuten sanaluokalla (adjektiivi, substantiivi jne.).²⁰ Ohjelmaa kehitetään osana luonnollisen kielen tutkimushanketta, ja sen tulokset perustuvat syöteaineiston tilastolliseen analysointiin. Parserin tulosten perusteella laadimme aluksi saksalaisiin liittyvää kuvamateriaalia kuvaavan sanapaketin (*bag-of-words*). Sanapaketin laadinta aloitettiin luomalla lista kaikista yksittäisistä termeistä, jotka esiintyivät kuvateksteissä. Tästä listasta poistettiin sellaiset termit, joilla ei ole selkeää semanttista merkitystä (*stopwords*) ja jotka siten voivat vinouttaa analyysiä.²¹ Näin saadussa listassa oli 3928 yksittäistä termiä. Tämä lista käytiin vielä manuaalisesti läpi ja siitä poistettiin muun muassa lyhenteet, numerot sekä muut, ei-yksiselitteiset termit, mikä tuloksena lopulliseen listaan jäi 3484 erillistä termiä. On hyvä tiedostaa, että yksikään automaattisesti toimiva parseri

19 Ks. tark. Bastian ym. 2009.

20 Lisätietoja ohjelmasta löytyy ”BioNLP”-verkkosivulta sekä ”Finnish-dep-parser”-verkkosivulta.

21 Internetissä on tarjolla useita valmiita stopword-listoja hyvin laajalle kielivalikoimalle. Tässä artikkelissa käytimme Snowball-projektin tarjoamaa, vapaasti ladattavissa olevaa sanalistaa (ks. ”Snowball”).

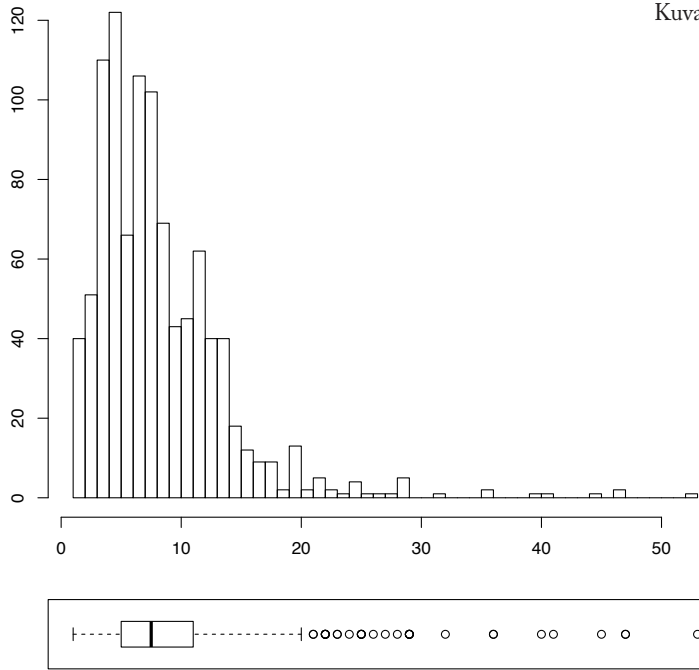
ei tuota 100-prosenttisesti virheetöntä tulosta johtuen mm. kirjoitusvirheistä, polysemiasta eli monimerkityksellisyydestä tai yksinkertaisesti siitä, ettei parserin sanasto tunne termiä. Käyttämämme parserin luotettavuus on kuitenkin osoitettu niin hyväksi,²² että sen avulla valmisteltua aineistoa voidaan pitää luotettavana pohjana jatkoanalyyysille.

Kuvatekstien louhinnalle oman haasteensa asettaa kuvatekstien suppea esitysmuoto. Kun kuvateksteistä oli aineiston valmistelun aikana poistettu analyysin kannalta merkityksettömät täytesanat, yhteen kuvatekstin jäi keskimäärin 8,7 sanaa. Samalla kuitenkin kuvatekstien pituuden jakauma on melko vahvasti vinoutunut lyhyempien kuvatekstien suuntaan: puolessa analysoituja kuvatekstejä on 5–11 termiä (Kuva 2, kuvio (a)). Neljännes kuvateksteistä sisältää 1–4 termiä, neljännes puolestaan yli 11 termiä. Aineistomme pisin kuvateksti sisältää 53 termiä. Myös termien sisällä jakauma on kaikkea muuta kuin tasainen, kun 75 prosenttia termeistä esiintyy kuvatekstikorpuksessa enintään neljä (4) kertaa (Kuva 2, kuvio (b)). Aineistomme ylivoimaisesti yleisin termi on ”saksalainen” (878 esiintymää). Yli viisikymmentä kertaa esiintyvät sellaiset termit kuin ”eversti” (77), ”suomalainen” (72), ”lento|kenttä” (70), ”kenraali” (69), ”tykki” (60), ”vierailu”/”vieras” (112), ”asema” (55), ”sankarihauta” (53) ja ”ilmatorjuntatykki” (51). Näiden lukujen valossa myös aineistomme yleisimpien termien sisällä voidaan nähdä melko voimakasta hajontaa.

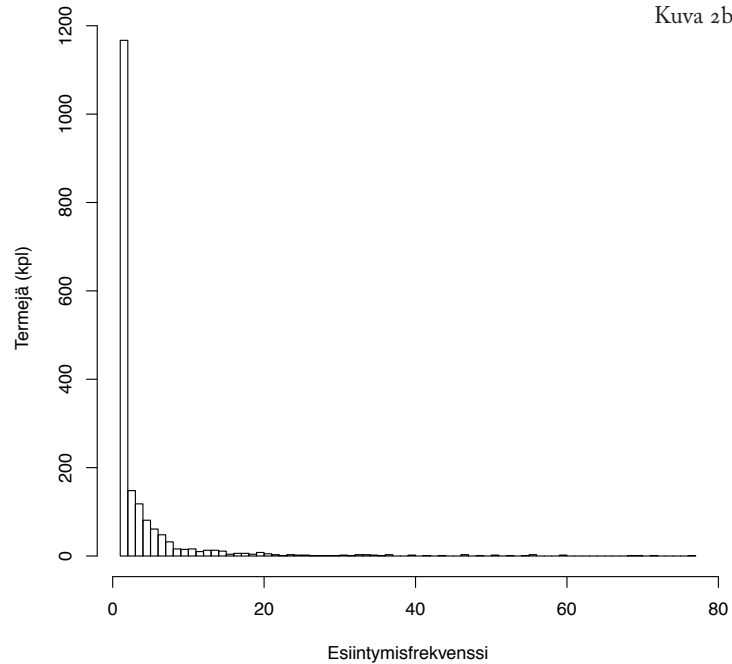
Alkuperäisten kuvatekstien yksipuolisuutta selittää osaltaan niiden syntykonteksti. Koska kuvaajan piti laatia kuvatekstit ottamiinsa kuviin, heijastavat SA-kuva-arkistoon kootut kuvien alkuperäiset kuvatekstit nimenomaan kuvaajan käsitystä siitä, mitä kuva esittää. Niitä ei siis ole laadittu myöhemmin esimerkiksi arkistoinnin yhteydessä, vaan TK-miehet tekivät tekstit itse ottamiinsa kuviin, ja monesti niiden laadussa oli suurtakin vaihtelua. Myös Päämaja kiinnitti tähän huomiota useaan otteeseen. Ohjekäskyssä 1 (23.6.1941) ohjeistetaan seuraavalla tavalla:

22 Ks. esim. Bohnet ym. 2013; Haverinen ym. 2013.

Kuva 2a



Kuva 2b



Kuva 2c



Kuva 2. (a) Termien määrä/kuvateksti-jakauma, (b) termistön esiintymistiheys-jakauma ja (c) 100 yleisimmän termin perusteella muodostettu sanapilvi.

Kuvaselostusten on oltava täydellisesti annettujen ohjeiden mukaisia, ettei enää tule vaillinaisia ja melkein olemattomia selostuksia, kuten nyt on muutamilta ryhmiltä saapunut. Selostuskaavakkeet täyttöohjeineen lähetetään heti niiden valmistuttua.

Tilanne ei ilmeisesti korjaantunut, koska ohjekäskyssä 83 (20.5.1943) Päämaja ohjeistaa uudelleen:

Lehdissä julkaistuissa TK-kuvissa on joskus ollut virheellisiä, jopa naurettavia tekstejä. Nämä virheet panee yleisö helposti kuvaajan syyksi, vaikka kuvaajan selostukseensa kirjoittama teksti on saattanut olla aivan oikein. Epäkohta johtuu siitä, että levitykseen menevien kuvien taakse laaditaan teksti aivan selostuksen mukaisesti. Lehdet kuitenkin korjaavat mielellään tällaiset laimeat tekstit kaunokirjallisempaan tai journalistisempaan muotoon, ja muuttavat ne silloin asiantuntemattomasti usein vääristellyiksi. Epäkohdan korjaamiseksi on kuvaselostukset jo tiedotuskompanian portaassa laadittava kirjallisempaan asuun, jolloin niitä ei enää toimituksissa tarvitse kirjoittaa uudelleen.

Ohjekäskyä 83 täydennettiin vielä ohjekäskyssä 85 (21.6.1943):

Viitaten ohjekäskyyn 84/20.5.43 tähdennetään vielä seuraavaa: vaikka kuvaselostukseen kirjoitettavat tekstit nyt pyritäänkin laatimaan muokattuun muotoon, on niihin silti edelleen saatava mukaan myös mahdollisimman paljon oikeita asiatietoja kuvien alkuperän selvittämiseksi arkistoitessa. Kysymykseen tulevat selostuksia laadittaessa tiedot 1) jotka ovat puhtaasti salaisia, 2) joita ei voida vielä julkaista, 3) joita ei kannata vielä julkaista ja 4) jotka voidaan julkaista. Esim. seuraavat tiedot ovat suotavia: joukko-osasto, henkilöiden nimet ja kotiseutu, aseitten tekn. määritelmät jne. Ne asiat, joita ei ole tarkoitettu julkisuuteen, on pantava sulkumerkkien väliin. Tekstit on laadittava niin, että sensuurin suorittamien poistojen jälkeenkin vielä jää jäljelle riittävä aineisto kuvan julkaisutekstin laatimista varten. Siitä huolimatta ei teksti saa olla ylettömän pitkä.

Sotavalokuvien kuvateksteillä oli monesti sotapropagandistinen funktio, joskin – kuten myöhemmin osoitamme – tämä funktio ei ollut mitenkään korostuneesti esillä suomalaisessa sotavalokuvauksessa. Propagandistisesta näkökulmasta kuvatekstit pyrkivät kiinnittämään kuvaa katselevan henkilön huomion tiettyyn, propagandan kannalta tärkeään asiaan, ja siten saamaan katsojan näkemään esitetty asia propagandistisesti oikeassa valossa.²³ Saksalaisten PK-miesten muisteluissa kerrotaan,

23 Kuvatekstin merkityksestä lehdistökuvan tulkinnessa ks. Heikkilä 2006; Sontag 1977; Glasenapp 2012.

että kuvatekstejä saatettiin jopa tietoisesti väärentää propagandatavoitteiden saavuttamiseksi: jos kuvan esittämä todellisuus ei sopinut yhteen saksalaisen propagandan kanssa, kuvan tulkintaa ohjattiin toisille urille uuden kuvatekstin avulla.²⁴ Tyypillistä propagandakuvien kuvateksteille oli myös tietty jargon, joka osaltaan selittää kuvatekstien yksipuolisuutta.

Yksittäisten termien merkitystä voidaan tarkastella myös sanapilven (*wordcloud*) avulla (Kuva 2, kuvio (c)). Sanapilven etu on siinä, että sen avulla voidaan havainnollistaa eri termien suhteellista merkitystä aineistokorpuksessa.²⁵ Yleisimpien termien perusteella voidaan jo tässä vaiheessa vetää tiettyjä johtopäätöksiä. Sanat ”tykki” ja ”asema” viittaavat siihen tosiasiaan, että Suomi sai ja osti Saksalta runsaasti aseistusta ennen jatkosotaa ja sen aikana.²⁶ Termi ”lentokenttä” puolestaan selittynee sillä, että Saksan ilmavoimat, *Luftwaffe*, käytti lukuisia Suomen lentokenttiä tukikohtinaan.²⁷ Korkeampaan upseeristoon liittyvät termit sekä termit ”sankarihauta” ja ”vierailu” puolestaan viittaavat sellaisiin ”seremoniakuviin”, joita kaksi ystävällismielisissä väleissä keskenään olevaa valtiota tuottaa. Kun Saksa ja Suomi talvisodanaikaisen ”jääkauden” jälkeen alkoivat taas lämmitellä suhteitaan, muodollisemmat tapaamiset sisälsivät monesti esimerkiksi saksalaisten upseerien tai suomalais-saksalaisista sotilashenkilöistä muodostuneen lähetystön vierailun vuoden 1918 sankarihaudoilla. Vierailuja jatkettiin myös jatkosodan ollessa jo käynnissä.

Vaikka yksittäisten termien esiintymistiheyden voidaankin katsoa kuvailevan aineiston sisältöä yleisellä tasolla, ei niiden perusteella voida juurikaan analysoida aineiston sisäistä rakennetta. Jos hyväksytään se taustaoletus, että kuvateksti – suppeudestaan huolimatta – pyrkii kertomaan kuvan olennaisen sisällön, yksittäisen kuvatekstin semanttisen rakenteen voidaan tulkita kuvaavan tiettyä tematiikkaa. Tämän ajatuksen mukaisesti tematiikaltaan samankaltaisten kuvien kuvatekstien voidaan olettaa olevan myös semanttiselta rakenteeltaan samankaltaisia eli niiden

24 Ks. esim. Schmidt-Scheeder 1997, 273–278.

25 Esim. Jockers 2014, 155.

26 Jokisipilä ja Könönen 2013, 226–232.

27 Valtonen 1997.

kuvatekstit ovat termistöltään samankaltaisia. Menetelmällisesti vakiintuneimpana tapana tekstien semanttisten rakenteiden analysoimiseksi voidaan pitää termien yhteisesiintymien (*co-occurrence*, suomeksi myös kollokaatiot) analysointia. Tämän menetelmän on useissa tutkimuksissa havaittu soveltuvan etenkin aineistojen temaattisten rakenteiden ja niiden taustalla vaikuttavien tai niihin yhteydessä olevien tekijöiden löytämiseksi.²⁸ Taustalla on siis oletus, että yhteisesiintymät muodostavat sosiaalisten verkostojen kaltaisia semanttisia vuorovaikutussuhteita, jotka auttavat havaitsemaan kuvien välillä vallitsevia temaattisia polkuja ja kuvakorpuksen temaattista rakennetta määrittäviä terminologisia solmukohtia eli ”hubeja”.

Edellä kuvatun ajatuksen mukaisesti oletamme tässä artikkelissa analysoidun kuvakorpuksen koostuvan erilaisista temaattisista lohkoista tai klustereista, joiden pääsisältö on kiteytetyssä muodossa löydettävissä korpukseen kuuluvista kuvateksteistä. Tässä kohdin teemme oletuksen, että kuvateksteissä on käytetty samankaltaisia termejä kuvaamaan samoja – tai ainakin samantyyppisiä – sisältöjä vastaavalla tavalla kuin esimerkiksi internet-aineistoissa käytetään tageja.²⁹ Looginen seuraus tästä on, että mikäli muodostamme yksittäiselle kuvatekstille kollokaatioverkoston, tematiikaltaan samantyyppisten kuvien kollokaatioverkostojen voidaan olettaa olevan samankaltaisia – edustavathan ne samantyyppistä sisältöä. Koska kuvateksteissä esiintyvien termien joukko on melko suppea, oletamme verkostoanalyysin löytävän materiaalista temaattisia polkuja eli keskeisiä avainsanoja, jotka toimivat linkeinä korpukseen kuuluvien kuvien välillä ja siten paljastavat tematiikalta samankaltaisia kuvajoukkoja.

Koska ensisijainen kiinnostuksemme kohdistuu saksalaistematiikkaan, päädyimme muodostamaan kollokaatiot saksa-johdannaisten termien välittömässä läheisyydessä sijaitsevien termien pohjalta. Näin pyrimme varmistamaan sen, että kollokaatioverkostot kuvaisivat mahdollisimman hyvin saksa-johdannaisten termien semanttista kontekstia.

28 Stuart ja Botella 2009; Brier ja Hopp 2011; Lee ym. 2010; Diesner ym. 2012; Hsu ja Kao 2013; Feicheng ja Yating 2014; Elo 2014; Elo 2015.

29 Tagien käytöstä verkostoanalyysissä ks. Feicheng ja Yating 2014.

Menetelmällisesti kollokaatiot muodostettiin kunkin kuvatekstin osalta erikseen seuraavalla tavalla. Aluksi kuvatekstistä etsittiin saksa-johdannainen termi. Tämän jälkeen luotiin ”termi-ikkuna”, johon kuuluivat tätä saksa-johdannaista termiä välittömästi edeltävät kaksi termiä sekä sitä välittömästi seuraavat kaksi termiä. Mikäli saksa-johdannainen termi esiintyi kuvatekstin ensimmäisen, toisena, toiseksi viimeisenä tai viimeisenä terminä, termi-ikkunaa pienennettiin vastaavasti. Tämän jälkeen ikkunan muut kuin saksa-johdannaiset käsitteet linkitettiin toisiinsa pareittain, minkä tuloksena luotiin saksa-johdannaisen termin välittömässä läheisyydessä olevan kollokaatioverkosto. Kukin kollokaatio varustettiin vielä kahdella attribuutilla – SA-kuvatunnuksella sekä vuosiluvulla – jatkoanalyysiä varten. Soveltamamme menetelmä on laajasti käytössä esimerkiksi tietokoneavusteisessa tekstianalyysissä, koska sen on havaittu tuottavan varsin luotettavia tuloksia tekstidokumentteihin sisältyvien teemojen luokittelun osalta.³⁰ Menetelmän tuottamat tulokset ovat kuitenkin kohtuullisen herkkiä ”ikkunan” koon suhteen, ja päädyimme pienen kokeilun jälkeen käyttämään viiden termin ikkunaa, koska sen pohjalta laaditut klusterit osoittautuivat aineistoa parhaiten erotteleviksi.

Esimerkkinä voidaan tarkastella SA-kuva-arkiston kuva numero 100886, jonka alkuperäinen kuvateksti kuuluu seuraavasti: ”Itsenäisyyspäivän vietto. Seppelen lasku saksalaisten haudalle Ristimäen hautausmaalla.” Parseroinnin ja puhdistamisen jälkeen analyysivalmis kuvateksti on muotoa: ”itsenäisyys|päivä vietto seppel lasku saksa|lainen hauta risti|mäki hautaus|maa”.³¹ Kollokaatioverkoston muodostamisen perustana toimivat termi-ikkuna puolestaan on: ”seppel lasku saksalainen hauta risti|mäki”, jonka perusteella muodostetaan seuraavat kollokaatiot: ”seppel”–”lasku”, ”seppel”–”hauta”, ”seppel”–”risti|mäki”, ”lasku”–”hauta”, ”lasku”–”risti|mäki” sekä ”hauta”–”risti|mäki”. Tärkeätä on huomata, että saksa-johdannainen termi jätetään kollokaatioiden ulkopuolelle, millä pyritään varmistamaan se, että verkosto antaisi kuvaa siitä laajemmasta kontekstista, jossa saksa-johdannaiset termit esiintyvät.

30 Esim. Schütze 1998.

31 Käytetty parseri merkitsee yhdyssanat lisäämällä yhdyssanan muodostavien sanojen väliin ”|”-merkin.

Kuvakorpuksen perusteella laaditut kollokaatioverkostot on visualisoitu oheisessa kuvasarjassa (Kuva 3). Kaikkien kuvan 3 graafien kohdalla olemme soveltaneet samaa voimaohjattua (*force directed*) visualisointimallia, joka pyrkii latomaan verkoston siten, että keskenään tiiviimmässä vuorovaikutuksessa olevat verkoston osat sijoittuvat lähelle toisiaan ja mahdollisesti muodostuvat aliverkostot sijoittuvat solmukohtansa (*hub*) ympärille. Lisäksi olemme hyödyntäneet visualisointitehosteita korostamaan tiettyjä verkoston rakenteeseen liittyviä ominaisuuksia, joita mitataan verkostometriikan tunnusluvuilla. Termin otsikkotekstin koko on suhteutettu kyseisen termin kytkösten määrään (*degree*) eli mitä suuremmalla tekstillä termin otsikko näkyy, sitä useamman muun termin kanssa se esiintyy yhdessä kuvateksteissä. Tekstin koon perusteella on siis mahdollista identifioida ne asiasanat, jotka esiintyvät monien muiden asiasanojen kanssa. Termejä yhdistävien viivojen paksuus puolestaan kuvastaa yhteisesiintymän yleisyyttä kuvakorpuksessa eli on riippuvainen yhteyden painoarvosta (*weight*). Mitä useammin kaksi asiasanaa esiintyy yhdessä, sitä painavampi on niiden suhde ja sitä paksumpi niitä yhdistävä viiva. Näin useimmin yhdessä esiintyvät asiasanat, niin kutsutut ydin-termit, on helppo tunnistaa graafista. Yksittäisiä termejä symbolisoivien solmujen koko puolestaan suhteutuu kyseisen termin asemaa ja sijaintia koko verkostossa – verkostotermein ilmaistuna keskeisyyttä (*centrality*) – ilmaisevaan *betweenness centrality* -arvoon³². Näin voidaan löytää ne asiasanat, jotka toimivat kuvakorpuksen sisältyviä semanttisia klustereita toisiinsa yhdistävinä linkkeinä ja siten indikoivat korpuksen sisältyviä temaattisia polkuja.³³

Aloitamme kuvatekstien kollokaatioverkostojen tarkastelun koko korpuksen tasolta (Kuva 3, graafi (a)), joka siis kattaa vuodet 1940–1945. Jo tämän graafin kohdalla voidaan havaita selkeätä klusteroitumista ja

32 *Betweenness centrality* -arvo lasketaan sen perusteella, kuinka usein solmu sijoittuu kahta muuta solmua yhdistävälle polulle siten, että tämä polku on samalla lyhin näitä kahta muuta solmua yhdistävä polku. Ks. myös Prell 2012, 103–104; Hsu ja Kao 2013; Huhtamäki ja Parviainen 2013.

33 Paranyushkin 2012. Ks. myös Feicheng ja Yating 2014.



suomalainen

divisiona

laatokka



käyttö

rakentaa

yttää

tuhotā

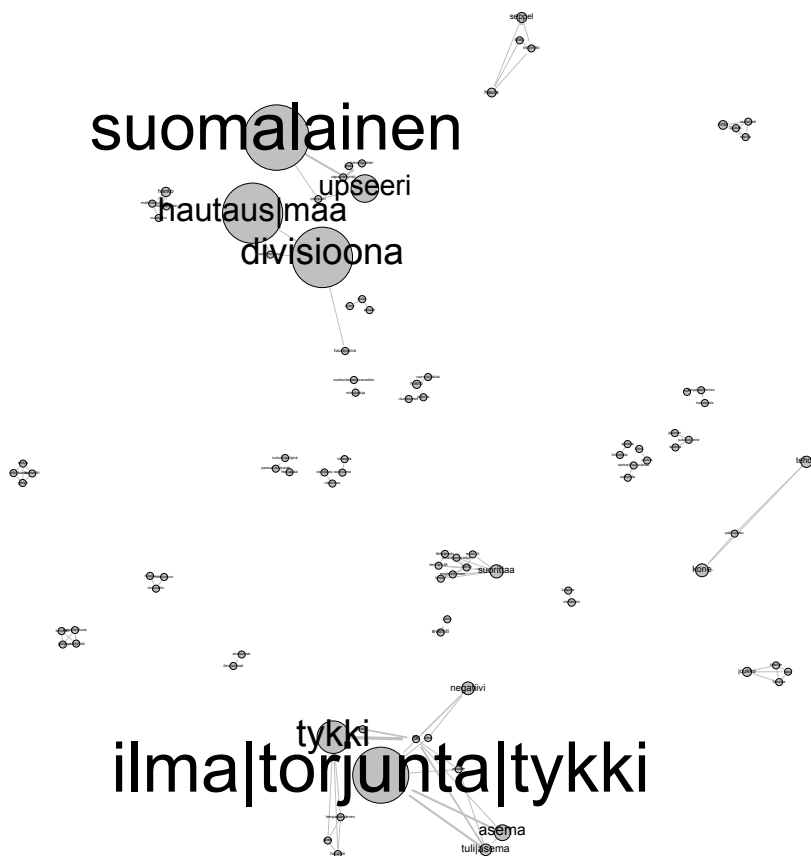
[illegible]

Kuva 3a. Saksaan liittyvien termien kollokaatioverkostot 1940–1945.

tiettyjen temaattisten polkujen olemassaoloa. Verkoston alareunassa voidaan havaita selkeä osaverkosto, jonka keskeiset termit liittyvät tykistöön ja ilmatorjuntaan. Verkoston vasemmasta yläreunasta löytyy toinen mielenkiintoinen ryhmittymä, jonka ydinkäsitteisiin kuuluu ”suomalainen”, joka terminä esiintyy yhdessä sellaisten termien kuin ”sotilas”, ”aseveli”/”aseveljeys” ja ”upseeri” kanssa.

Seuraavat kaksi verkostovisualisointia (Kuva 3, graafit (b) ja (c)) havainnollistavat verkoston dynamiikkaa. Graafi (b) esittää kollokaatioverkoston rakenteen vuosina 1940–1942 otettujen kuvien kuvatekstien perusteella, graafi (c) puolestaan vastaavan verkoston vuosien 1943–1945 osalta. Ensimmäisen huomio kiinnittyy vuosien 1940–1942 kollokaatioverkoston suppeaan kokoon. Tämä on merkillepantava asia, sillä vuosina 1940–1942 otettiin lähes 400 saksalaisiin liittyvää kuvaa *enemmän* kuin vuosina 1943–1945 (ks. myös kuvio 1). Kun vuosien 1940–1945 aikana saksa-johdannaisten termien läheisyydessä esiintyy yhteensä 722 termiä, vuosina 1940–1942 termejä esiintyy ainoastaan 94 kappaletta. Siinä missä tykistöön ja ilmatorjuntaan liittyvä tematiikka on havaittavissa jo vuoden 1940–1942 kuvakorpuksessa, puuttuu aseveljeys-tematiikkaa näiden vuosien kuvamateriaalista. Vuosien 1943–1945 kohdalla kuvatekstien terminologia laajenee huomattavasti: kuvissa esiintyy yhteensä 656 eri termiä eli suurin osa kuvatekstien temaattisten erojen kannalta relevantista terminologiasta muodostui vasta sodan loppuvaiheessa.

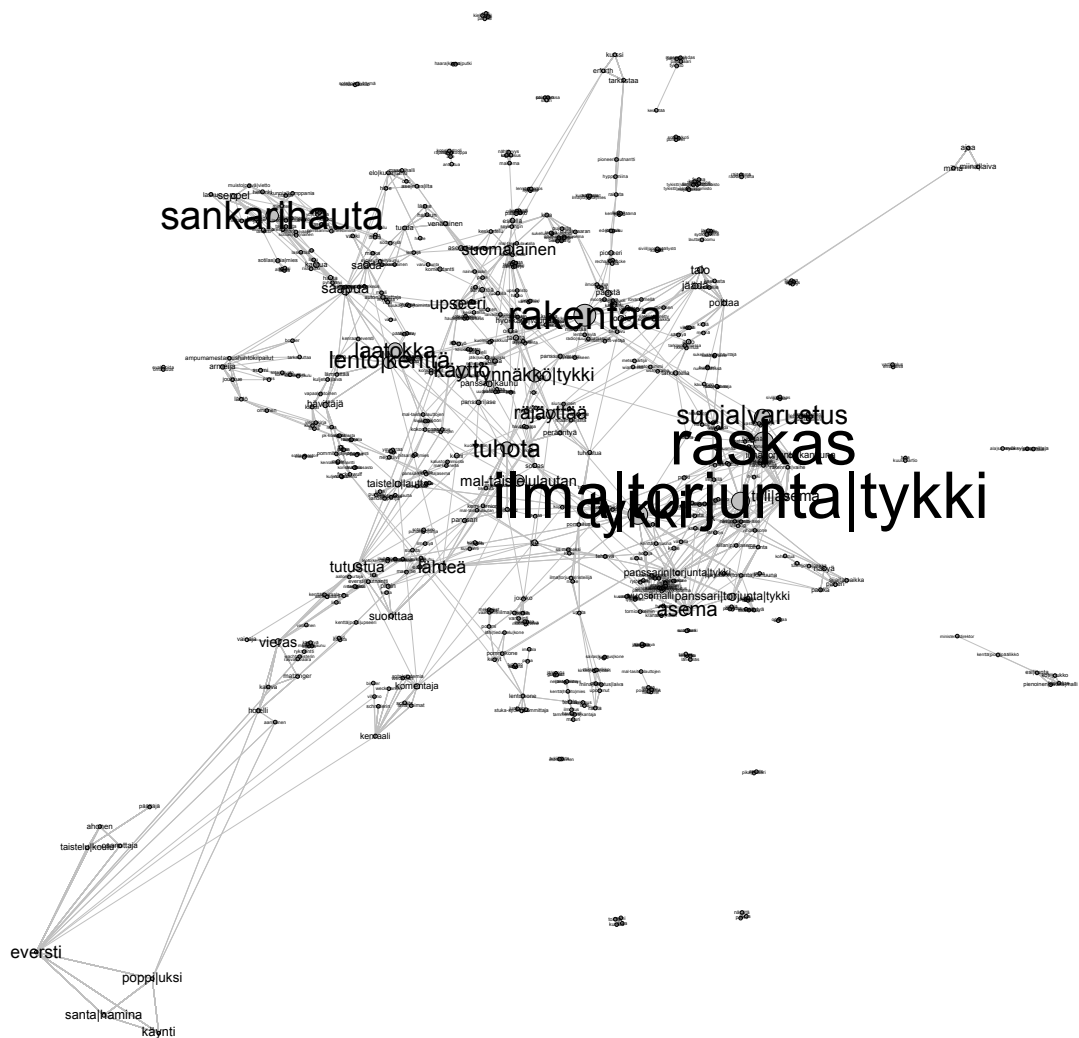
Kahden aikaikkunan välillä havaittu ero käytetyn terminologian laajuudessa ei kuitenkaan ole mitenkään poikkeavaa, jos sitä tarkastellaan suhteessa sotatapahtumiin. Kun sodan alkupuolella Saksan joukot vastasivat omasta rintamaosastaan ja suomalaiset omastaan ja kun lisäksi aiheen ulkopoliittinen tulenarkuus rajoitti suomalaisten mahdollisuuksia saksalaisten kuvaamiseen, aseveljeyden kuvaaminen redusoitui monesti ainoastaan formaalien tapahtumien, kuten yhteisten juhlien tai seremonioiden taltioimiseen. Tähän seikkaan palaamme tarkemmin alla klusterianalyysin tulosten kohdalla. Kun sitten sodan jälkimmäisellä puoliskolla – siis juuri vuosina 1943–1945 – saksalaiset joukot, esimerkiksi 122. jalkaväkidivisioona sekä lento-osasto Kuhlmeijer taistelivat kesällä 1944 yhdessä suomalaisten kanssa Karjalan kannaksella Neuvostoliiton suurhyökkäyksen pysäyttämiseksi, heijastui taistelutoiminta osaltaan myös



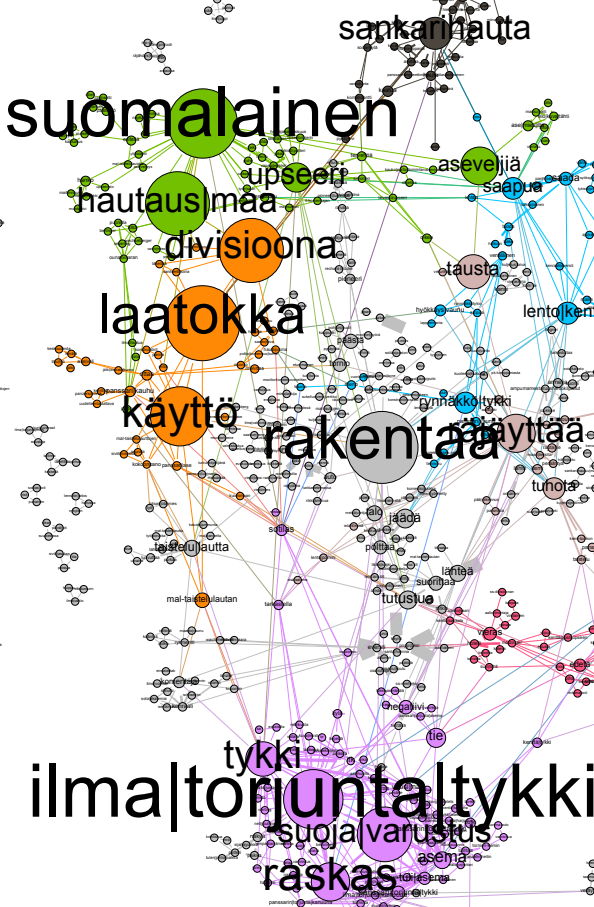
Kuva 3b. Saksaan liittyvien termien kollokaatioverkostot 1940–1942.

kuvateksteihin. Tästä todisteena voidaan pitää muun muassa viidenkymmenen yleisimmän termin joukossa olevia termejä ”perääntyä” ja ”tuhota”. Syytä on kuitenkin myös huomata, että etenkin kesän 1944 suurhyökkäyksen aikana otettiin lopulta verrattain vähän valokuvia muun muassa TK-kuvaajien filmipulasta johtuen.³⁴ Saksalaisten kuvaamista rajoitti myös tehokkaasti 2.7.1944 annettu käsky, jossa TK-kuvaajia kiellettiin

34 Päämajan ohjekäskyt tiedotuskompanioille, käsky 90. Päämajan tiedotusosaston arkisto, kansio T20680/13, Kansallisarkisto (Sörnäisten toimipiste).



Kuva 3c. Saksaan liittyvien termien kollokaatioverkostot 1943–1945.



Kuva 3d. Saksaan liittyvien termien kollokaatioverkostot 1940–1945 (modularisoitu).

ottamasta kuvia saksalaisista, sillä ”niistä huolehtivat saksalaisten omat PK-miehet”.³⁵

Toki on syytä huomata, että taistelutoimintaan liittyvät termit saattoivat liittyä myös Saksan ja Suomen välillä syksyllä 1944 alkaneisiin Lapin sodan taisteluihin. Lapin sodassa saksalaiset noudattivat poltetun maan taktiikkaa ja tuhosivat suurimman osan Lapin rakennuskannasta,³⁶ mikä osaltaan selittää esimerkiksi sellaisten termien kuten ”polttaa” ja ”Muonio” sijoittumista varsin korkealle termien esiintymistiheydellä mitattuna. Kokonaisuutena Lapin sotaan liittyvien valokuvien osuus koko kuvakorpuksesta on kuitenkin melko suppea, mihin pääsy lienee kuitenkin operatiivinen: Lapin sotaan lähti vain yksi tiedotuskomppania, ja kuvantuotanto oli sen mukaisesti verrattain vähäistä.³⁷

Edellä tarkastellut ajalliset kollokaatioverkostot antavat viitteitä siitä, että kuvien tematiikka muuttui sotavuosien aikana. Analyttiseltä kannalta tämä tarkoittaa sen oletuksen vahvistumista, jonka mukaan analysoimamme sotavalokuvakorpus koostuisi temaattisesti erilaisista ryhmistä eli klustereista. Verkostoanalyysin termein ilmaistuna kyse on siitä, voidaanko verkostosta löytää toisistaan erottuvia yhteisöjä, klustereita tai klikkejä. Tämän selvittämiseksi annoimme käyttämämme verkostoanalyysiohjelmiston suorittaa aineiston modularisointianalyysin, joka pyrkii löytämään aineistosta sellaisia ryhmiä, joiden keskinäiset suhteet ovat tiiviimmät kuin muiden verkoston osien kanssa.³⁸

35 Päämajan ohjekäsky tiedotuskomppanioille, käsky 100. Päämajan tiedotusosaston arkisto, kansio T20680/13, Kansallisarkisto (Sörnäisten toimipiste).

36 Lapin sodasta ks. Kulju 2013.

37 Päämajan ohjekäsky tiedotuskomppanioille, käsky 103. Päämajan tiedotusosaston arkisto, kansio T20680/13, Kansallisarkisto (Sörnäisten toimipiste).

38 Tässä yhteydessä on syytä huomauttaa, että klusterointi on edelleen yksi verkostoanalyysin haasteellisimpia sovelluksia eikä yksiselitteistä tapaa klusteroinnin toteuttamiseksi ole olemassa. Eri klusterointimenetelmät myös tuottavat hyvinkin erilaisia tuloksia. Lisää klusterointiin liittyvistä problematiikasta ks. esim. Fortunato 2010; Wu ym. 2013.

Klusteroinnin tulokset on esitetty kuviossa 3 (graafi (d)), jossa kukin klusteri on esitetty omalla värillään siten, että samaan klusteriin kuuluvat termit on helppo tunnistaa.

Kaikkiaan käyttämämme menetelmä löysi aineistosta 84 klusteria, mikä osaltaan kertoo aineiston olevan temati-

kaltaan melko hajanainen. Suurimpaan klusteriin kuuluu 12 prosenttia, toiseksi suurimpaan vajaan 9 ja kolmanneksi suurimpaan noin 6 prosenttia termeistä – eli yhteensä kolme suurinta klusteria sulkevat sisäänsä reilun neljänneksen (27 prosenttia) kaikista termeistä. Kolmen suurimman klusterin vahvimmat termit on esitetty oheisessa taulukossa (Taulukko 1). Kolmen suurimman klusterin eniten kytköksiä omaavien termien muodostamat semanttiset kokonaisuudet näyttäisivät myös poikkeavan toisistaan melko selkeästi, mikä osaltaan vahvistaa klusteroinnin luotettavuutta. Luotettavuuden arvioinnin kannalta tärkeä mittari on kuitenkin nk. modulaarisuusarvo (Q), joka kuvaa sitä, miten vahvasti aineistosta on löydettävissä jonkinasteinen yhteisö rakenne. Aiempien tutkimusten perusteella luotettavina voidaan pitää klusterointoja, joiden $Q > 0,4$.³⁹ Tämän artikkelin osalta kuvatekstien termien kollokaatioverkon klusteroinnin $Q = 0,87$, eli identifioituja klustereita voidaan pitää varsin luotettavina. Samalla korkea Q -arvo on osoitus siitä, että aineiston tietokoneavusteinen luokittelu voi tarjota perinteiselle lähiluentaiselle sisällönanalyysille tietokoneavusteisen vaihtoehdon, jonka avulla isoista aineistosta on mahdollista nopeastikin löytää yhdistäviä tekijöitä ja siten päästä helpommin kiinni aineistoon sisältyviin teemoihin.

Analyysin viimeisessä vaiheessa siirrymme tarkastelemaan kuvakorpuksestamme kuvatekstianalyysin tulosten perusteella valikoituja

Klusteri 1	Klusteri 2	Klusteri 3
ilmatorjuntatykki	upseeri	sankarihauta
raskas	suomalainen	seppel
tykki	kisa	hauta
asema	ounasvaara	laskea
panssarintorjuntatykki	tervehtiä	helsinki

Taulukko 1. Kolmen suurimman klusterin vahvimmat termit (eniten kytkentöjä omaavat termit).

39 Esim. Newman 2006; Paranyushkin 2011, 18.



Kuva 4a. Esimerkkikuvat, klusteri 1: SA-kuva 153704: ”68.Rask.It.Ptri. Tykkien betoniset suojavarustukset. Kuvassa on saksalainen 88 mm:n raskas ilmatorjuntatykki vuodelta 1937 (88 ItK 37 RMB).” (Lähde: FINNA).

kuvia *visual history* -näkökulmasta. Esimerkissä on kyse eksploraatiiviselle tutkimukselle luonteenomaisesta jatkoanalyysistä, jossa aineiston louhinnan kautta esille nousseet havainnot toimivat pohjana tiettyjen aineiston osien valitsemisesta tarkemmin luupin alle otettaviksi. Tässä artikkelissa rajaamme tarkastelumme kolmeen suurimpaan klusteriin. Suodatimme kuvakorpuksesta ne valokuvat, joiden kollokaatiomatriisissa esiintyi vähintään yksi klusterin viidestä yleisimmästä termistä. Näin valikoituja kuvia voidaan siis pitää oman klusterinsa edustajina, joiden kautta visuaalinen aineisto kiinnittyy sekä louhintamme tuloksiin että varsinaiseen kuvakorpukseen.

Ensimmäisen klusterin kuvatekstit kiinnittyvät terminologisesti tykistöön, ennen kaikkea ilmatorjuntaan. Kuvien visuaalinen sisältö tukee vahvasti tätä havaintoa, sisältyhän klusteriin runsaasti tykistöön liittyvää kuvamateriaalia. Klusteriin liittyvät kuvat voidaan visuaalisesti jakaa kahteen pääryhmään, joista kummastakin oheisessa kuvassa esimerkki (Kuvat 4a ja 4b). Ensimmäiseen pääryhmään kuuluvat asetta ja sen tuliasemaa esittävät kuvat, joiden keskiössä on juuri varsinainen sota-



Kuva 4b. Esimerkkikuvat, klusteri 1: SA-kuva 151114: "Ilmatorjuntatykki torjumas-
sa vihollisen maataistelukoneitten hyökkäystä Rajakonnun kylän liepeillä. Kuvassa
20 mm:n saksalainen ilmatorjuntatykki 20 ItK 30 BSW vuodelta 1930" (Lähde:
FINNA).

kalusto, siis tässä tapauksessa tykki (Kuva 4a). Tämän tyyppiset kuvat on tarkoitettu ns. teknisiksi kuviksi, joilla ei ole varsinaista propaganda-tarkoitusta, vaan joita käytetään esimerkiksi tykkimallien manuaaleissa havainnollistavana kuvituksena. Tämä on loogista, koska suomalaisten TK-kuvaajien velvollisuuksiin kuului myös tekninen kuvaaminen.⁴⁰ Tyypillistä tähän ryhmään kuuluville kuville onkin nimenomaan itse tykin kuvaaminen ilman miehistöä, jolloin esimerkiksi suuntauslaitteisto sekä tykin sijoitus asemaansa erottuvat selvästi.

Toisen pääryhmän muodostavat tilannekuvat, joissa varsinaiseen aseeseen lisäksi esitetään myös sen käytöstä vastaava miehistö, useimmiten ainakin näennäisesti ampumavalmiudessa tai taistelutehtävissä (Kuva 4b). Näiden kuvien visuaalinen viesti on jo selvästi propagandistinen,

40 Majuri G. Waseniuksen komennuskunnan ja kuvaosaston Saksan matkojen matkakertomukset. Päämajan tiedotusosaston arkisto, kansio T20681/26, Kansallisarkisto (Sörnäisten toimipiste).



Kuva 5. Esimerkkikuva, klusteri 2: SA-kuva 120226: "Ounasvaaran kisat. Saksalainen Berlinger pujottelee" (Lähde: FINNA).

kuvien avulla pyrittiin manifestoimaan Suomen ilmapuolustuksen korkeaa aineellista ja moraalista tasoa aikana, jolloin talvisodanaikaisen ilmatorjunnan puutteellisuus oli hyvin todennäköisesti vielä useimpien mielessä.

Toisen pääklusterin kuvat ovat, verrattuna edellä käsiteltyyn klusteriin 1, hyvin toisenlaisia ja heijastelevat lähinnä saksalaisten ja suomalaisten sotilaiden yhteisten hiihtokisojen kuvastoa (Kuva 5). Hiihtokisat, jotka tässä siis olivat suomalaisten ja saksalaisten sotilaiden yhteiset, kuuluivat kiinteästi jatkosodan asemasotavaiheen talvikuvastoon. Suomen rintamilla alkoi joululta 1941 noin kaksi ja puoli vuotta kestänyt asema-

sotavaihe, jolloin ei juuri taisteltu, ja miehille oli kehitettävä ajanvietettä. Kaikenlaiset urheilukilpailut olivat jopa niin suosittuja, että päämaja joutui jo vuonna 1942 rajoittamaan TK-kuvauksia urheilukisoista, koska kotirintamalle oli levinnyt käsitys, että samaan aikaan kun siellä kärsitään työvoimapulasta, rintamalla vain urheillaan.⁴¹ TK-kuvaajat joutuivatkin asemasotavaiheessa monessa mielessä puun ja kuoren väliin: päämaja saattoi vaatia näyttäviä taistelukuvia, vaikka rintamalla oli hiljaista, ja toisaalta rintamasotilaiden arjesta, kuten urheilukilpailuista, kertovat kuvat eivät läpäisseet sotasensuuria.⁴²

Myös kolmannen klusterin osalta visuaalinen analyysi vahvistaa kuvatekstien louhinnassa saatuja tuloksia. Klusterin kuvatekstien kollokaatiorakenne (Taulukko 1) tuottaa ennakko-oletuksen, että klusteriin päätyneet kuvat liittyisivät lähinnä korkean upseeriston ”seremoniallisesta” aseveljeydestä kertovaan kuvanarratiiviin, mihin olennaisesti liittyvät saksalaisten vierailut Suomessa. Ennakko-olettamusten mukaisesti klusteriin kuuluu runsaasti seppeltenlaskua sankari- ja muilla haudoilla esittävää kuvamateriaalia (Kuva 6a). Klusterin viiden vahvimman termin joukkoon kuuluva ”Helsinki” saa sekin visuaalisen analyysin kautta vahvistuksen, kun kuva-aineistoon sisältyy kuvamateriaalia etenkin Saksan merivoimien alusten mutta myös korkean upseeriston vierailuista Helsingissä (Kuva 6b).

Kun analyysituloksia tarkastellaan kokonaisuutena, voidaan todeta, että tekstinlouhintamenetelmien ja verkostanalyysin avulla toteutettu aineiston eksploratiivinen analyysi tuotti – kuvateksteihin liittyvät haasteet huomioiden – jopa hämmästyttävän luotettavan kuvan aineiston temaattisesta rakenteesta. Kolmea pääklusteria edustaneiden esimerkkivalokuvien visuaalinen analysointi osoitti, että myös kuvien visuaalinen sisältö vastasi klusterien terminologisesti määriteltä tematiikkaa. Toki on syytä muistaa, että tarkasteluun valikoidut kuvat edustivat klusterin vahvinta terminologiaa. Jos tietyn klusterin kuvien suodattamisessa käytettäisiin esimerkiksi klusterin kymmentä tai kahtakymmentä yleisintä termiä, samaan klusteriin kuuluvien kuvien visuaalisten sisältöjen hetero-

41 Pilke 2011, 128.

42 Kleemola 2013, 14–16.



Esimerkkikuvat, klusteri 3. (6a) SA-kuva 146124: "Sankarivainajien Muistopäivän vietto. Seppelten laskua saksalaisella sankarihaudalla" ja (6b) SA-kuva 124942: "Saksalainen miinanraivaaja Helsingin satamassa, taustalla saksalainen torpeedovene". (Lähde: FINNA)

geenisyyssä tulisi hyvin todennäköisesti kasvamaan. Kuitenkin myös tässä pätee sama periaate kuin esimerkiksi tilastollisten mallinnusten kohdalla. Selittävien tekijöiden määrän kasvattaminen lisää mukaan mahtuvien havaintojen määrää, mutta samalla vähentää mallin selitysvoimaa. Tässä artikkelissa esitetyn analyysimenetelmän näkökulmasta tämä tarkoittaa, että laajentamalla kuvien valintakriteerejä voidaan aineistosta löytää enemmän kuvia, mutta samalla kuvien temaattinen koherenssi pienenee. Kyse on aina tutkijan omaan harkintaan liittyvästä valinnasta, kumpaa hän omassa tutkimuksessaan painottaa. Edellä olemme osoittaneet, että kuvatekstien klustereiden kärkiterminologia muodostaa vähintäänkin kohtuullisen luotettavan pohjan löytää laajasta kuva-aineistosta sellaisia temaattisia eroja, jotka tarjoavat hyvät lähtökohdat valikoida materiaalia esimerkiksi kuvatutkimuksen menetelmiin ja teorioihin tukeutuvalla jatkoanalysoinnille.

Keskustelu

Tässä artikkelissa olemme havainnollistaneet, miten valokuvien tutkimuksessa voidaan hyödyntää verkostanalyysiin perustuvaa sisällönanalyysiä, kun tutkimusaineiston on joko rakenteeltaan tai kooltaan – tai molemmiltaan – sellainen, ettei sitä ole mahdollista hallita manuaalisen luokittelun tai perinteisen lähiluennan keinoin. Kun huomioidaan se tosiseikka, että jo aivan näköpiirissä olevassa tulevaisuudessa historian tutkimuksen kannalta kiinnostavien digitaalisten aineistojen määrä ja aineistokorpusten koot tulevat kasvamaan, tulee jatkossa olemaan entistä suurempi tarve ei vain aineistojen tehokkaassa hallinnassa avustaville menetelmille, vaan myös sellaisille tutkimusmenetelmille, joiden avulla aineistoista voidaan louhia esiin oman tutkimusongelman kannalta tarkoituksenmukaisia otoksia, yhdistellä erilaisia aineistoja ja tehdä aineistoon sellaisia metodologisia tutkimusmatkoja, jotka luovat pohjaa aivan uusien tutkimuskysymysten esittämiselle.

Mitä tulee edellä esitettyihin konkreettisiin analyysituloksiin, voidaan jo niiden katsoa osoittavan käyttämiemme digitaalisten tutkimusmenetelmien soveltuvan myös historian tutkimukselle ominaisten tutkimuskysymysten tutkimiseen. Erilaiset tekstiaineistot lukeutuvat edelleen

historiantutkimuksen keskeisiin lähdeaineistoihin, ja näiden osalta analyysimme osoitti tekstinlouhintamenetelmien kykenevän vähintäänkin kohtuullisen luotettavasti tunnistamaan tekstiin sisältyviä piileviä rakenteita, joiden kautta voidaan päästä kiinni aineiston temaattiseen rakenteeseen.

Valokuvatutkimuksen kannalta merkittävänä voidaan pitää sitä, että analyysitulokset osoittivat kuvatekstien louhinnan tuloksena saatujen klusterien vastanneen pääosin hyvin myös klustereihin sisältyvien valokuvien visuaalista sisältöä. Tätä taustaa vasten digitaaliset tutkimusmenetelmät avaavat uusia mahdollisuuksia entistä suurempien aineistojen käsittelyyn ja hallintaan, mutta myös rakenteeltaan erilaisten aineistojen yhdistelyyn. Samalla saatujen tulosten yleistettävyyys paranee merkittävästi, kun digitaalisin menetelmin esiin louhittujen tematiikkojen piiristä valittujen esimerkkikuvien tulkinnat voidaan – normaalit luotettavuuskriteerit huomioiden – yleistää koskemaan koko klusteria.

Vaikka esittelemämme tietokoneavusteinen menetelmä tuottikin mielenkiintoisia ja jatkoanalyysille hedelmällisiä lähtökohtia tarjonneita tuloksia, digitaalisia menetelmiä ei kuitenkaan tule pitää minään metodisena taikasauvana, jonka avulla oman julkaisuluettelon kasvattaminen onnistuu sekä konkreettisesti että kuvainnollisesti ”nappia painamalla”. Vaikka tietokone on nopea ja tehokas lajittelija, on se samalla sokea aineiston todelliselle sisällölle. Tulokset ovat täysin riippuvaisia siitä, kuinka hyvin tai huonosti *tutkija* määrittelee tietokoneavusteisen tutkimusvaiheen parametrit ja kriteerit. Huonosti valmistelu aineisto tai analyysi tuottaa huonoja ja käyttökelvottomia tuloksia – jotka tosin saattavat näyttää järkeviltä, etenkin jos tutkija ei ole edes viitsinyt kovin hyvin perehtyä aineistonsa kontekstiin. Esimerkiksi tekstin eri nyanssien, kuten vaikkapa propagandististen elementtien tai ironian tunnistaminen ei ole yksikertainen tehtävä.

Toinen tärkeä, huomioitava seikka liittyy historiallisten aineistojen luonteeseen. Siinä missä monet historiallisista aineistoista ovat puutteellisia, fragmentoituneita tai muuten vinoutuneita, useimmat digitaaliset menetelmät olettaen asian olevan aineiston osalta juuri päinvastoin. Meidänkin aineistostamme puuttuvat ne kuvat, jotka eivät syystä tai toisesta koskaan päätyneet osaksi kokoelmia. Vaikka esiin louhitun

kuvanarratiivin tematiikka näyttäisikin sopivan hyvin osaksi sotahistoriallisissa tutkimuksissa saatuja tuloksia, jäämme väkisinkin pohtimaan, olisivatko puuttuvat kuvat voineet tuoda kuvanarratiiviin jotakin selailaista, jota aiempi tutkimus ei ole havainnut. Puuttuvan datan ongelma on digitaalisen historian tutkimuksen kohdalla olennaisen tärkeä ja vaatii tutkimusasetelman huolellista suunnittelua. Esimerkiksi verkostanalyysissä jo muutama puuttuva solmu voi muuttaa koko verkoston rakennetta niin paljon, että verkostolle laskettuihin tunnuslukuihin tulee suhtautua vähintäänkin kriittisesti. Tässäkin korostuu tutkijan vastuu analyysin tuloksista. Piiloutuminen sen argumentin taakse, että ”kone nämä tulokset laski”, on lähinnä osoitus puutteellisesta ammattitaidosta.

Kaiken edellä sanotun perusteella kiteytämme artikkelimme pääviestin seuraavasti: digitaaliset menetelmät ovat jo nyt hyvä, jopa erinomainen apuväline, jonka avulla tutkijalle avautuu uusia mahdollisuuksia suurten aineistojen hallitsemiseksi, eri aineistojen yhdistelemiseksi ja aineistoon suuntautuvien löytöretkien toteuttamiseksi. Samalla digitaaliset menetelmät alleviivaavat ja korostavat tutkijan ammattitaidon merkitystä: mitä keskeisemmän roolin tietokone aineiston analyysissä ottaa, sitä tärkeämmäksi muodostuvat analyysiä edeltävä valmistelu sekä sitä seuraava tulkintavaihe. Molemmissa vaiheissa tutkijan kompetenssi eli suvereeni oman tutkimuskentän hallinta tulevat korostumaan. Sillä loppujen lopuksi tutkija sekä määrittää, mitä kone etsii, että tulkitsee sen, mitä kone löysi. Tässä prosessissa kiteytyy digitaalisen historian tutkimuksen ydin.

Lähteet

- Bastian, Mathieu; Heymann, Sebastien & Jacomy, Mathieu: Gephi: An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Weblogs and Social Media* 2009. <<http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>>.
- Berry, Michael W. & Kogan, Jacob (toim.): *Text Mining: Applications and Theory*, Wiley, Chichester 2010.
- ”BioNLP”. BioNLP-tutkimushankkeen verkkosivut. <<http://bionlp.utu.fi/finnish-parser.html>> 5.9.2016.

- Bohnet, Bernd, Nivre, Joakim, Boguslavsky, Igor, Farkas, Richárd, Ginter, Filip, Hajič, Jan: Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *Transactions of the Association for Computational Linguistics* 1(2013), 415–428.
- Brier, Alan; Hopp, Bruno: Computer assisted text analysis in the social sciences. *Quality & Quantity* 45(1) 2011, 103–128.
- Burdick, Anne; Drucker, Johanna; Lunenfeld, Peter; Presner, Todd; Schnapp, Jeffrey: *Digital Humanities*. MIT Press, Cambridge MA 2011.
- Clavert, Frédéric; Noiret, Serge (toim.): *L'histoire contemporaine à l'ère numérique / Contemporary History in the Digital Age*. Peter Lang, Bruxelles 2013.
- Diesner, Jana; Carley, Kathleen M.; Tambayong, Laurent: Extracting socio-cultural networks of the sudan from open-source, large-scale text data. *Computational and Mathematical Organization Theory* 12(3) 2012, 328–339.
- Elo, Kimmo: A Spider Spinning its Web. East German HUMINT Networks on Nordic Affairs. Teoksessa *Need to Know. Eastern and Western Perspectives*. Toim. Thomas Wegener Friis ja Wladyslaw Bulhak. University Press of Southern Denmark, Odense 2014, 61–79.
- Elo, Kimmo: The Content Structure of Intelligence Reports. *Connections* 35(1) 2015, 20–38.
- Eriksson, Kai (toim.): *Verkostot yhteiskuntatutkimuksessa*. Gaudeamus, Helsinki 2015.
- Feicheng, Ma; Yating, Li: Utilising Social Network Analysis to Study the Characteristics and Functions of the Co-occurrence Network of Online Tags. *Online Information Review* 38(2) 2014, 232–247.
- ”Finnish-dep-parser”. Finnish-dep-parser-projektin verkkosivut. <<http://turkunknlp.github.io/Finnish-dep-parser/>>. 5.9.2016.
- Fortunato, Santo: Community detection in graphs. *Physics Reports* 486/2010, 75–174.
- Glasenapp, Jörn: Vergangenheit und Zukunft. Wie fotografischer Sinn entsteht. *Fotogeschichte* 124/2012, 5–12.
- Haverinen, Anna; Suominen, Jaakko: Kirjoittamisen vuoropuhelu? – Mitä on digitaalinen humanistinen tutkimus, *Ennen ja Nyt Online* 2015. <<http://www.ennenjanyt.net/2015/02/koodaamisen-ja-kirjoittamisen-vuoropuhelu-mita-on-digitaalinen-humanistinen-tutkimus/>>. 01.06.2015.
- Haverinen, Katri; Nyblom, Jenna; Viljanen, Timo; Laippala, Veronika; Kohonen, Samuel; Missilä, Anna; Ojala, Stina; Salakoski, Tapio; Ginter, Filip: Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation* 48(3) 2013, 493–531.

- Heikkilä, Elina: *Kuvan ja tekstin välissä. Kuvateksti uutiskuvan ja lehtijutun elementtinä*. SKS, Helsinki 2006.
- Hsu, Yi-Yu; Kao, Hung-Yu: Coin. A network analysis for document triage. *Database* 2013, 1–11.
- Huhtamäki, Jukka, Parviainen, Olli: Verkostoanalyysi sosiaalisen median tutkimuksessa. Teoksessa *Otteita verkosta. Verkon ja sosiaalisen median tutkimusmenetelmät*. Toim. Salla-Maaria Laaksonen, Janne Matikainen ja Minttu Tikka. Vastapaino, Tampere 2013, 245–273.
- Jockers, Matthew L.: *Text Analysis with R for Students of Literature*. Springer, London 2014.
- Jokisipilä, Markku; Könönen, Janne: *Kolmannen valtakunnan vieraat. Suomi Hitlerin Saksan vaikutuspiirissä 1933–1944*. Otava, Helsinki 2013.
- Jäger, Jens: *Fotografie und Geschichte*. Campus, Frankfurt 2009.
- Kleemola, Olli: Kamera aseena. Teoksessa Helena Pilke & Olli Kleemola (toim.) *Suomi taisteli. Kuvat kertovat*. Readme.fi, Helsinki 2013, 10–21.
- Kulju, Mika: *Lapin sota 1944–45*. Gummerus, Helsinki 2013.
- Lee, Sangno; Song, Jaeki; Kim, Yongjin: An Empirical Comparison of Four Text Mining Methods. *Journal of Computer Information Systems* 51(1) 2010, 1–10.
- Leetaru, Kalev: *Data Minig Methods for the Content Analysis*, Routledge, New York & London 2012.
- McCulloh, Ian: *Social Network Analysis with Applications*. John Wiley & Sons, Somerset 2013.
- Newman, M. E. J.: Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103(23) 2006, 8577–8582.
- Olsen, Jon: *Digital History als Mannschaftssport*, H-Soz-Kult Online Forum 23.11.2015. <<http://www.hsozkult.de/debate/id/diskussionen-2894>>. 25.4.2016.
- Paranyushkin, Dmitry: *Identifying the Pathways for Meaning Circulation using Text Network Analysis*. Nodus Labs Research Report, 2012. <<http://nodus-labs.com/publications/Pathways-Meaning-Text-Network-Analysis.pdf>>. 3.6.2015.
- Pilke, Helena: *Julkaiseminen kielletty. Rintamakirjeenvaihtajat ja Päämajan sensuuri*. SKS, Helsinki 2011.
- Prell, Christina: *Social Network Analysis. History, theory and methodology*. SAGE, London 2012.
- ”Python-verkkosivu”, Python-ohjelmointikielen www-sivut. <<https://www.python.org/about/>>.

- R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna 2013. <<http://www.R-project.org/>>. 25.4.2016.
- Ruohonen, Keijo: *Graafiteoria*. Tampere: Tampereen teknillinen yliopisto 2013 <<http://math.tut.fi/~ruohonen/GT.pdf>>.
- "SA-kuvaportaali", Suomen armeijan valokuva-arkiston [www.sivut. www.sa-kuva.fi](http://www.sivut.www.sa-kuva.fi). 5.9.2016.
- Schmidt-Scheeder, Georg: *Reporter der Hölle. Die Propaganda-Kompanien im 2. Weltkrieg*. Motorbuch Verlag, Stuttgart 1997.
- Schultz-Jones, Barbara: Examining information behavior through social networks. An interdisciplinary review. *Journal of Documentation* 66(4) 2009, 592–631.
- Schütze, Hinrich: Automatic word sense discrimination. *Computational Linguistics* 24(1) 1998, 97–124.
- "Snowball". Snowball-tutkimushankkeen verkkosivu. <<http://svn.tartarus.org/snowball/trunk/website/algorithms/finnish/>>. 5.9.2016.
- Sontag, Susan: *Über Fotografie*. Fischer, Frankfurt 1977.
- Stuart, Keith; Botella, Ana: Corpus Linguistics, Network Analysis and Co-occurrence Matrices. *International Journal of English Studies* 51/2009, 1–20.
- Ukkonen, Antti: Big Data ja laskennalliset menetelmät. Teoksessa *Otteita verkosta. Verkon ja sosiaalisen median tutkimusmenetelmät*. Toim. Salla-Maaria Laaksonen, Janne Matikainen ja Minttu Tikka. Vastapaino, Tampere 2013, 274–304.
- Valtonen, Hannu: *Luftwaffen pohjoinen sivusta: Saksan ilmavoimat Suomessa ja Pohjois-Norjassa 1941–1944*. Keski-Suomen ilmailumuseo, Tikkakoski 1997.
- Vinyals, Oriol; Toshev, Alexander; Bengio, Samy & Erhan, Dumitru: Show and Tell: A Neural Image Caption Generator. *Computer Vision and Pattern Recognition* 2015. <<http://arxiv.org/abs/1411.4555v2>>. 25.4.2016.
- "Visual history". Visual History -tutkimusalueen verkkosivu. <http://documentation.de/zg/Visual_History_Version_3.0_Gerhard_Paul>. 5.9.2016.
- Weiss, Solomon M.; Indurkha, Nitin; Zhang, Tong: *Fundamentals of Predictive Text Mining*. Springer, London 2010.
- Wu, Zhiang; Cao, Jie; Wu, Junjie; Wang, Youquan; Liu, Chunyang: Detecting Genuine Communities from Large-Scale Social Networks. A Pattern-Based Method. *The Computer Journal* 2013, 1–15.