



Korpusavusteinen diskurssianalyysi kuluttajatutkimuksen menetelmänä: Pikavippikeskustelun synty, nousu ja arkipäiväistyminen Suomi24- keskustelufoorumilla

Ella Lillqvist

ABSTRAKTI

Tämä tutkimus osallistuu viimeaikaiseen keskusteluun suurten tekstimäärien analyysimenetelmistä esittelemällä kielitieteen määrälliseen tutkimusperinteeseen pohjautuvaa korpusavusteista diskurssianalyysia (corpus-assisted discourse studies, CADS). Artikkelissa havainnollistetaan CADSin olennaisimpia menetelmiä kuluttajatutkimuksellisessa kontekstissa analysoimalla, miten Suomi24-keskustelufoorumien käyttäjät ovat keskustelleet pikalainoista ilmiön ensimmäisellä vuosikymmenellä Suomessa eli vuosina 2005–2015. Lisäksi artikkelissa kartoitetaan ja kehitetään CADSin yhteyteen soveltuvia visualisointitapoja. CADS tarjoaa monipuolisia välineitä kuluttajatutkimuksen työkalupakkiin: Sen avulla voidaan saada yleiskuva suuressa tekstimäärässä käsitellyistä teemoista esimerkiksi tarkastelemalla aineiston erityispiirteitä suhteessa vertailuaineistoon tai pidemmän aikavälin kattavassa aineistossa yleistyviä ja harvinaistuvia sanoja. Lisäksi voidaan tutkia ja havainnollistaa visuaalisesti tiettyjen tutkijaa erityisesti kiinnostavien sanojen esiintymiskonteksteja – nousivat nuo sanat sitten aineistolähtöisesti korpuksen yleistarkastelusta tai teorialähtöisesti aiemmasta tutkimuksesta. Myös siirtyminen määrällisen ja laadullisen analyysin välillä on joustavaa. Korpusanalyysiohjelmit mahdollistavat sekä valmiiden että itse kerättyjen korpusten analyysin. CADS sopii erityisesti kuluttajatutkijoille, joita kiinnostavat yleiset puhettavat isossa tekstiaineistossa koskien jotakin selkeästi sanallistettavaa teemaa kuten tiettyä tuotetta, tuotekategoriaa tai kulttuurista kulutusilmiötä.

Johdanto

Viime aikoina kuluttajatutkimuksen kentällä on kiinnostuttu menetelmistä, jotka soveltuvat suurten tekstimäärien, esimerkiksi sosiaalisen median aineistojen analysointiin (esim. Arsel & Bean, 2012; Humphreys & Wang, 2017; Isotalo, Laaksonen, Pöyry & Jallinoja, 2019; Repo & Matschoss, 2018; Ruckenstein, 2019). Tämä mielenkiinto kytkeytyy laajempaan keskusteluun ”digitaalisista

ihmistieteistä”, ”laskennallisista yhteiskuntatieteistä” ja ”big datasta” (ks. esim. Purhonen & Toikka, 2016). Tässä lehdessä Ylisiurua (2017) on pohtinut LDA-aihemallinnuksen tarjoamia mahdollisuuksia ja Lagus, Pantzar ja Ruckenstein (2018) puolestaan ovat esitelleet sentimenttianalyysejä muistuttavan tutkimustavan. Kahden viimeksi mainitun artikkelin tavoin myös nyt käsillä olevan artikkelin ensisijaisena tavoitteena on toimia menetelmällisenä avauksena.

Osallistun tähän suurten tekstimäärien analyysejä koskevaan metodologiseen keskusteluun esittelemällä kielitieteen määrälliseen tutkimusperinteeseen, *korpuslingvistiikkaan*, pohjautuvan lähestymistavan. Korpuslingvistiikka on merkityksellistä kuluttajatutkimukselle ja muille yhteiskuntatieteille, kun sitä käytetään *diskurssianalyysin* (eli *diskurssintutkimuksen*) menetelmänä – siis silloin, kun ollaan kiinnostuneita kielenkäytöstä ja merkityksistä suhteessa sosiaaliseen toimintaan ja yhteiskuntaan (diskurssianalyysistä ks. Pälli & Lillqvist, tulossa). Viittaan tähän tutkimustapaan käsitteellä *korpusavusteinen diskurssianalyysi* (*corpus-assisted discourse studies; CADs*), tosin joskus käytän myös yleisempää nimitystä *korpusanalyysi* (esim. Lillqvist & Harju, 2018). Itselleni erityisen läheinen diskurssintutkimuksen suuntaus on kriittinen diskurssintutkimus (Fairclough, 1995; van Dijk, 2001; Wodak & Meyer, 2016), jonka piirissä myös CADsia nykyään melko usein käytetään (ks. Mautner, 2016). CADsin hyödyntäminen ei kuitenkaan edellytä kriittistä näkökulmaa tai välttämättä edes eksplisiittistä *diskurssi*-termin käyttöä – olennaista on kiinnostus tietyissä konteksteissa vallalla oleviin puhetapoihin.

Toistaiseksi CADsia ei ole Suomessa tietääkseni käytetty kuluttajatutkimuksen alalla, eikä muutenkaan kovin laajasti: vertaisarvioituja artikkeleita ovat suomeksi julkaisseet ainakin Jantunen (2018) homouteen ja heterouteen liittyvistä diskursseista ja Johansson, Jantunen, Heimo, Ahonen ja Laippala (2018) *kansan* representaatioista. Lisäksi akateemisia lopputöitä ovat kirjoittaneet esimerkiksi Lehto (2018) japaninsuomalaisten kielipuheesta, Paldán (2018) seksuaalisen haluttomuuden diskursseista ja minä (Lillqvist, 2010) työttömiä koskevista puhetavoista. Näille tutkijoille on yhteistä tausta kielitieteiden piirissä. Tämän artikkelin tavoitteena on avata CADsin mahdollisuuksia kuluttajatutkijoille ja edistää vuoropuhelua muiden suurten tekstiaineistojen analyysitapojen kanssa. Esittelen ja havainnollistan CADsin keskeisimpiä menetelmiä analysoimalla pika- tai pienlainoihin – eli arkisesti ”pikavippeihin” – liittyvää keskustelua Suomi24-keskustelufoorumilla, johon perustuva korpus on saatavilla Kielipankin kautta (Aller Media, 2014; Lagus, Pantzar, Ruckenstein & Ylisiurua, 2016). Samalla pyrin kartoittamaan ja laajentamaan erilaisten visualisointitapojen käyttöä CADs-tutkimuksen yhteydessä.

Pikalainat ovat ongelmallisten piirteidensä, voimakkaan yleistymisensä, ja niihin viime vuosina kohdistuneiden lainsäädäntömuutosten vuoksi ajankohtainen kuluttajatutkimuksen aihepiiri, josta ei kuitenkaan vielä ole kovin paljon tutkimusta Suomen kontekstissa (ks. kuitenkin Kaartinen & Lähteenmaa 2006; Valkama & Muttilainen, 2008; Kaartinen, Autio & Lähteenmaa 2009; Autio, Wilska, Kaartinen & Lähteenmaa, 2009). Pikavippiyritysten toiminta alkoi Suomessa vuonna 2005, joten

vuodesta 2001 alkava Suomi24-korpus mahdollistaa niitä koskevan keskustelun seuraamisen alusta lähtien. Keskustelufoorumien osallistujat eivät yleensä ole alan asiantuntijoita, vaan käsittelevät aihepiiriä pääosin asiakkaan tai kuluttajan näkökulmasta. Aiemmasta tutkimuksesta tiedetään, että kuluttajien velka- ja talousasiat ymmärretään usein moraalisisina ja stigmatisoituina (esim. DuFault & Schouten, 2018), ja että jotkut kokevat erityisesti pikalainojen ottamisen hävettävänä (Burton, 2010). Tästäkin syystä Suomi24-aineisto on kiinnostava: Suomi24:ään kirjoitetaan usein anonyymina (Lagus ym., 2016), mikä saattaa edistää kokemusten ja mielipiteiden avointa jakamista.

Seuraavaksi taustoitan ensin CADSiä ja sitten hieman myös pikalainoja ilmiönä ennen kuin siirryn käsittelemään aineistoa. Sen jälkeen yhdistän menetelmän esittelyn ja havainnollistavat tulokset neljässä luvussa, joissa käsittelem sanojen frekvenssejä (yleisyydet), avainsananoja (vertailuaineistoon nähden erityisen yleiset sanat), kollokaatteja (hakusanan läheisyydessä esiintyvät sanat) ja ajallisia trendejä (tarkastelujakson aikana yleistyvät ja vähentyvät sanat).

Korpuksset ja niiden analyysi

Korpuksella tarkoitetaan yleensä laajaa ja järjestelmällisesti kerättyä sekä koneluettavassa muodossa olevaa kokoelmaa näytteitä luonnollisesta kielestä – korpuslingvistinen analyysi (esim. CADS) puolestaan on tällaisen tekstikokoelman tietokoneavusteista määrällistä ja laadullista analyysia (Baker & McEnery, 2015; Hoey, Mahlberg, Stubbs & Teubert, 2007; McEnery, Xiao & Tono, 2006). Onkin tärkeää huomata, että CADSissa tarvitaan myös laadullista analyysia (Mautner, 2016). Vastaavasti tarvetta laadullisuudelle suurten tekstimäärien analyysin yhteydessä ovat korostaneet myös esimerkiksi Purhonen ja Toikka (2016), Ylisiurua (2017) sekä Lagus ja kumppanit (2018). Korpuslingvistinen analyysi ei ole perinteisesti hyödyntänyt esimerkiksi koneoppimista (ks. kuitenkin Brookes & McEnery, 2019; Murakami, Thompson, Hunston & Vajn, 2017). Voikin pitää nyrkkisääntönä, että korpusanalyysi on pääosin tilastollista analyysia, jossa tutkijan tulkinta on ensisijaisen tärkeää.

Korpuksia on monenlaisia: jotkin ovat tasapainotettuja, mahdollisimman edustavia otoksia tietyistä kielestä tai kielimuodosta, toiset taas ovat pienempiä, johonkin erityistarkoitukseen (kuten tiettyyn tekstilajiin, aihepiiriin tai tapaukseen) keskittyviä aineistoja (Baker, 2006, s. 26-27; McEnery ym., 2006, s. 5). CADSissa on usein käytetty jälkimmäisen tyyppin korpuksia. Nykyisin korpuksella viitataan yhä useammin internetistä kerättyihin erittäin suuriin, jopa monen miljardin sanan tekstikokoelmiin. Näiden hyvinä puolina ovat ajankohtaisuus ja laajuus, sen sijaan edustavuudesta ei varsinaisesti voi niiden kohdalla puhua, koska aineiston valinta on perustunut helppoon saatavuuteen eikä tarkkaan harkittuun tasapainotukseen. Tulosten yleistämisessä on silloin oltava varovainen.

Visualisointi ei ole perinteisesti ollut kovin merkittävässä asemassa korpuslingvistisessä tutkimuksessa (Anthony, 2018). Viime aikoina monet tutkijat

ovat kuitenkin korostaneet, että visuaalinen havainnollistus on tärkeä askel suurten tekstiaineistojen ymmärtämisessä (Anthony, 2018; Brezina, McEnery & Wattam, 2015; Hilpert, 2011).

Korpusavusteisuuden edut

Diskurssintutkimusta on joskus syytetty analyysin vinoutumisesta tutkijan näkemysten mukaan, ja ratkaisuksi on esitetty muun muassa korpuslingvistiikan määrällisiä menetelmiä (esim. Stubbs, 1997; Widdowson, 1998), sillä ne mahdollistavat laajempien ja mahdollisesti edustavampien aineistojen käytön ja tietyllä tavalla systemaattisemman analyysin. Samat syytökset ovat toki kohdanneet yleisemminkin perinteisiä laadullisia analyysitapoja (ks. esim. Purhonen & Toikka, 2016, s. 12). En varsinaisesti usko, että laadullinen diskurssintutkimus olisi muita menetelmiä alttiimpi analyysin tai johtopäätösten vääristymiselle tutkijan ennakkokäsityksiin perustuen. Oikeastaan on päinvastoin, sillä diskurssintutkimuksen piirissä korostetaan usein tutkijan oman position reflektointia ja avointa esiin tuomista, jolloin lukija pystyy arvioimaan kriittisesti analyysin näkökulmaa. Siitä huolimatta pidän järkevänä ajatusta, että määrällinen analyysi oikein käytettynä tuo diskurssintutkimukseen hyödyllisen lisän.

Mahdollisuutta entistä paljon laajempien aineistojen hyödyntämiseen on pidetty tärkeänä, koska diskursiiviset merkitykset ovat luonteeltaan kumulatiivisia: yksittäinen sana, lause tai ilmaus ei käy näytöksi diskurssista, vaan nimenomaan samojen ja samankaltaisten ilmaisujen ja merkitysten johdonmukainen *toistuminen* antaa diskurssille sen voiman muokata ajattelua ja sen mukana yhteiskuntaa (Baker, 2006). Kun tekstejä on käytettävissä paljon, on tällaisten toistuvuuksien löytäminen helpompaa. Lisäksi se mahdollistaa luotettavamman diakronisten eli pitkittäisten vertailujen tekemisen (Partington, 2010). CADSin avulla voi myös nopeuttaa analyysia, ainakin verrattuna yksinomaan laadullisten menetelmien käyttöön.

Peruskäsitteitä, menetelmiä ja työkaluja

Korpusanalyysia varten on kehitetty monia ohjelmistoja. Osa näistä on omalle työasemalle asennettavia, kuten #Lancsbox¹, Antconc², ja WordSmith Tools³, ja osa toimii verkossa, kuten Sketch Engine⁴, Korp⁵, ja CQPweb⁶. Osin näiden ominaisuudet vaihtelevat, mutta niillä on kuitenkin yhteinen menetelmällinen pohja.

¹ <http://corpora.lancs.ac.uk/lancsbox/>

² <http://www.laurenceanthony.net/software/antconc/>

³ <https://www.lexically.net/wordsmith/>

⁴ <https://www.sketchengine.eu/>

⁵ Suomen versio: <https://korp.csc.fi/>; Ruotsin versio: <https://spraakbanken.gu.se/korp/>

⁶ <https://cqpweb.lancs.ac.uk/>

Avainsana-analyysia käytetään CADSissa usein antamaan yleiskuva aineiston aiheista tai muista erityispiirteistä, joko yhdessä muokkaamattomien sanamäärien eli *frekvenssien* kanssa tai niiden sijasta (Gabrielatos, 2018). Myös aineistossa esiintyviä kahden tai useamman sanan *klustereita* voidaan käyttää tällaiseen eksploratiiviseen tarkasteluun. Avainsanat (*keywords*) lasketaan vertaamalla sanojen yleisyyttä yhdessä korpuksessa niiden yleisyyteen toisessa korpuksessa; tämä tuo esiin, mitkä sanat ovat tarkasteltavassa korpuksessa erityisen yleisiä tai harvinaisia (Kilgariff, 2012). Avainsanat riippuvat siis olennaisesti vertailuun käytettävästä korpuksesta ja onkin syytä valita vertailukorpus, jonka kielimuoto on mahdollisimman samankaltainen kuin tarkasteltavan korpuksen kielimuoto esimerkiksi genren ja aikakauden osalta. Jos näin ei ole, jotkin tuloksista saattavat kertoa enemmän näistä kontekstuaalisista eroista kuin kiinnostuksen kohteena olevien diskurssien eroista. Kuitenkaan aina ei voi löytää täydellistä vertailukorpusta valmiina ja uuden korpuksen kerääminen pelkkää vertailua varten on harvoin realistista. Tämä rajoite on siis pidettävä mielessä tulkintoja tehtäessä.

”Avaimuuden” käsite (*keyness*) voidaan laajentaa myös aineiston muihin piirteisiin, kuten kieliopillisiin tai semanttisiin luokkiin – tällöin on tietenkin välttämätöntä, että nämä metatiedot on ensin merkitty (eli annotoitu) kumpaankin korpuksen jokaisen sanan yhteyteen (Rayson, 2008). Sanaluokkien avaimuus voisi kiinnostaa diskurssintutkijaa tyylin kuvaajana. Semanttisten luokkien avaimuus puolestaan kertoisi kahden korpuksen aihepiirien välisistä eroista. Wmatrix⁷ on internetissä toimiva työkalu, jolla on mahdollista tehdä englanninkielisille teksteille automaattinen semanttinen annotointi eli aihehuokittelu. Suomen kielelle ei tällaista tietääkseni ole vielä käytössä (enkä siksi käsittele sitä tässä enempää), vaikka Löfberg (2017) on kehittänyt sen mahdollistavaa suomenkielistä sanastoa väitöskirjassaan.

Aineiston yleisen tarkastelun jälkeen edetään usein yksityiskohtaisempaan tarkasteluun esimerkiksi *kollokaattien* avulla. Kollokaatit ovat sanoja, jotka esiintyvät oletusarvoa useammin tarkastelun kohteena olevan sanan läheisyydessä, ja niitä analysoimalla voidaan saada ainakin alustava käsitys kyseiseen sanaan liittyvistä merkityssuhteista. Kollokaatit voivat esiintyä missä vain tutkijan määrittelemän hakuikkunan sisällä (esim. hakusanasta kolme sanaa vasemmalle ja kolme oikealle, eli 3:3). Siinä ne eroavat klustereista, jotka ovat toistuvasti vierekkäin ja nimenomaan tietyssä järjestyksessä esiintyviä sanaryhmiä. Klustereita voidaan myös etsiä automatisoidusti ilman tietyn hakusanan käyttöä, kun taas kollokaatit lasketaan tietylle hakusanalle. Eräs kollokaattianalyysin muoto ryhmittelee kollokaatit sanojen kieliopillisen funktion ja/tai sijainnin mukaan (esimerkiksi mitkä adjektiivimääritteet ja mitkä verbit esiintyvät hakusanan oikealla ja vasemmalla puolella). Edellä mainituista työkaluista tällainen *sanakuva*-toiminto on toteutettu Sketch Engineissä ja Korpissa.

⁷ <http://ucrel.lancs.ac.uk/wmatrix/>

Hakutuloksia, joissa hakusana näytetään välittömässä kontekstissaan, sanotaan *konkordansseiksi*⁸. Usein konkordanssit esitetään yhden rivin pituisina ja hakusana on niissä keskitettynä, jolloin sanan lähikontekstista voi saada yleiskuvan nopeasti. Toinen esittämisvaihtoehto on hakusanan esittäminen virke- tai kappalekontekstissaan, ja yleensä kontekstin tarkempaa tutkimista varten on mahdollista avata tarkasteltavaksi se teksti, jossa yksittäinen hakutulos esiintyy. Korpusohjelmistoissa ”laadullinen loikka” (Ylisiurua, 2017) kohti tekstikontekstin ymmärtämistä on siis tehty mahdollisimman helpoksi.

Jotkin korpusohjelmistot, ainakin #Lancsbox ja Sketch Engine, auttavat myös uuden korpuksen valmistelussa käyttöön annotoimalla kieliopilliset metatiedot kunkin sanan yhteyteen eli tekemällä aineistoon automaattisen lemmauksen (tieto kunkin sanan perusmuodosta) ja lauseenjäsennyksen.

Tutkimuskonteksti: ongelmallinen pikalaina-ala Suomessa

Suomessa vakuudettomien, lyhytaikaisten ja korkeakorkoisten pika- tai pienlainojen – arkikielessä ”pikavippien” – myöntäminen alkoi vuonna 2005 (Autio, Wilska, Kaartinen & Lähteenmaa, 2009; Makkonen, 2014; Valkama & Mutttilainen, 2008). Pikavippejä haetaan joko tekstiviestillä tai internetissä. Yhdysvalloissa ja Iso-Britanniassa, missä tämän tyyppisillä lainoilla on pidemmät periteet, niistä käytetään usein nimitystä *payday loan* (esim. Rowlingson, Appleyard & Gardner, 2016; Stegman & Faris, 2003) – ainakin alun perin periaatteena oli, että laina maksetaan takaisin seuraavana palkkapäivänä. Nykyisin ainakaan Suomessa näin ei ole, ja ainoana myöntökriteerinä näyttää yleensä olevan maksuhäiriömerkinnän puuttuminen. Näiden lainojen saaminen on siis huomattavasti helpompaa kuin perinteisten kulutusluottojen, mikä johtaa myös useammin maksuvaikeuksiin.

Isossa-Britanniassa on todettu, että 67 prosentilla pikalainojen ottajista vuositulot ovat alle 25 000 £ (noin 29 000 €) ja Yhdysvalloissa vastaava osuus on lähempänä 75 prosenttia (Burton, 2010). Suomesta taas tiedetään muun muassa, että pikalainojen ottajat ovat tyypillisesti 18-29-vuotiaita nuoria (Rantala & Tarkkala, 2010). Pikalainoja on havaittu käytettävän usein perustarpeisiin, esimerkiksi ruokaan ja vuokranmaksuun, kun taas muita kulutusluottoja käytetään useammin suurempien hankintojen rahoittamiseen (Kaartinen & Lähteenmaa, 2006).

Pikavipit johtavat monien kuluttajien kohdalla velkakierteeseen (Rantala, 2012; vrt. Stegman & Faris, 2003). Muita ikäviä seurauksia voivat olla maksukyvyttömyys ja maksuhäiriömerkintä, joka vaikeuttaa jatkossa paitsi lainan myös esimerkiksi vuokra-asunnon saamista. Tällaiset lainat on maailmalla yhdistetty jopa terveysongelmiin (Sweet, Kuzawa & McDade, 2018). Pikavippien monet ongelmalliset piirteet ovatkin vuosien varrella olleet esillä julkisessa keskustelussa

⁸ Konkordanssiriveihin viitataan myös lyhenteellä KWIC – ”key word in context”: tätä ei pidä sekoittaa avainsanoihin, sillä ”key word” merkitsee tässä yhteydessä vain hakusanaa.

ja niihin on puututtu myös lainsäädännöllä useaan otteeseen: vuonna 2010 (helmikuussa ja joulukuussa), vuonna 2013 (Majamaa, Sarasoja & Rantala, 2017; Makkonen, 2014) ja viimeksi vuonna 2019 (mm. Laki kuluttajansuojalain muuttamisesta). Näin monet sääntely-yritykset ovat olleet tarpeen, sillä ala tuntuu aina löytävän keinot sopeutua ja tarjota hieman uudenlaisia mutta yhtä lailla kuluttajien kannalta ongelmallisia tuotteita.

Aineisto

Tässä artikkelissa analysoin pikavippejä koskevaa kuluttajakeskustelua Suomi24:ssä, johon perustuva annotoitu korpus on saatavissa Kielipankin kautta (Aller Media, 2014; Lagus ym., 2016). Korpusta päivitetään ja siitä on siis olemassa useita eri versioita, joiden kattavuus vaihtelee (eikä mikään niistä sisällä ”koko” Suomi24:ää edes tietyltä aikaväliltä – puuttuvasta aineistosta ks. Lagus ym., 2016, s. 19–20; Pink, Ruckenstein, Willim & Duque, 2018, s. 6–8). Tässä tutkimuksessa käytettiin Suomi24-korpuksen versiota 2015H1, tarkemmin sanottuna siitä tehtyä tietokantaversiota, josta on mahdollista ladata hakutuloksia koko viestikontekstissaan (kun Kielipankin versiosta voi toistaiseksi ladata osumia vain yhden tekstikappaleen kontekstissa).

Suomi24 on Suomen suurin keskustelufoorumi, joka käsittää yli 3 000 alafoorumia (Lagus ym., 2016). Sen käyttö on ilmaista, ja sinne on mahdollista kirjoittaa viestejä joko omalla nimellä, kirjautuneena nimimerkillä tai täysin anonyymisti tilapäisellä nimimerkillä (mt.). Pikavippejä ei ole tähän mennessä tarkasteltu tällaisen aineiston avulla ja se tarjoaa siten ainutlaatuisen ikkunan kuluttajien pikavippipuheeseen. Toki tulosten tulkinnassa on huomioitava, että keskustelijat eivät ole edustava otos suomalaisista – keskustelijoilla on todennäköisesti keskimääräistä useammin kokemusta pikalainoista joko suoraan tai välillisesti, mikä selittäisi heidän kiinnostuksensa osallistua aihetta koskevaan keskusteluun.

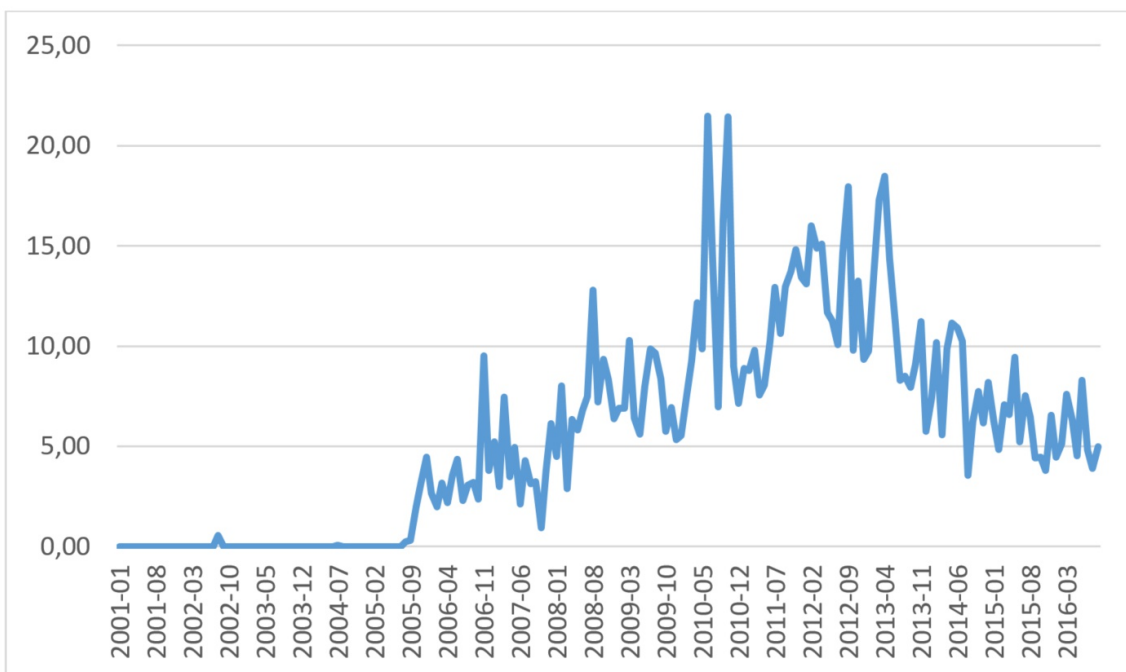
Aineiston ajallisen laajuuden ansiosta on mahdollista mennä ajassa taaksepäin aina pikavippiyritysten rantautumiseen Suomeen: Suomi24-korpuksen versio 2015H1 sisältää tekstejä vuoden 2001 tammikuusta vuoden 2015 kesäkuuhun. Aineistomäärät ovat seuraavat:

- 48 423 611 viestiä (aloitusviestit ja vastaukset)
- 231 373 748 lausetta
- 2 385 073 226 sanetta (juoksevaa sanaa ja välimerkkiä).

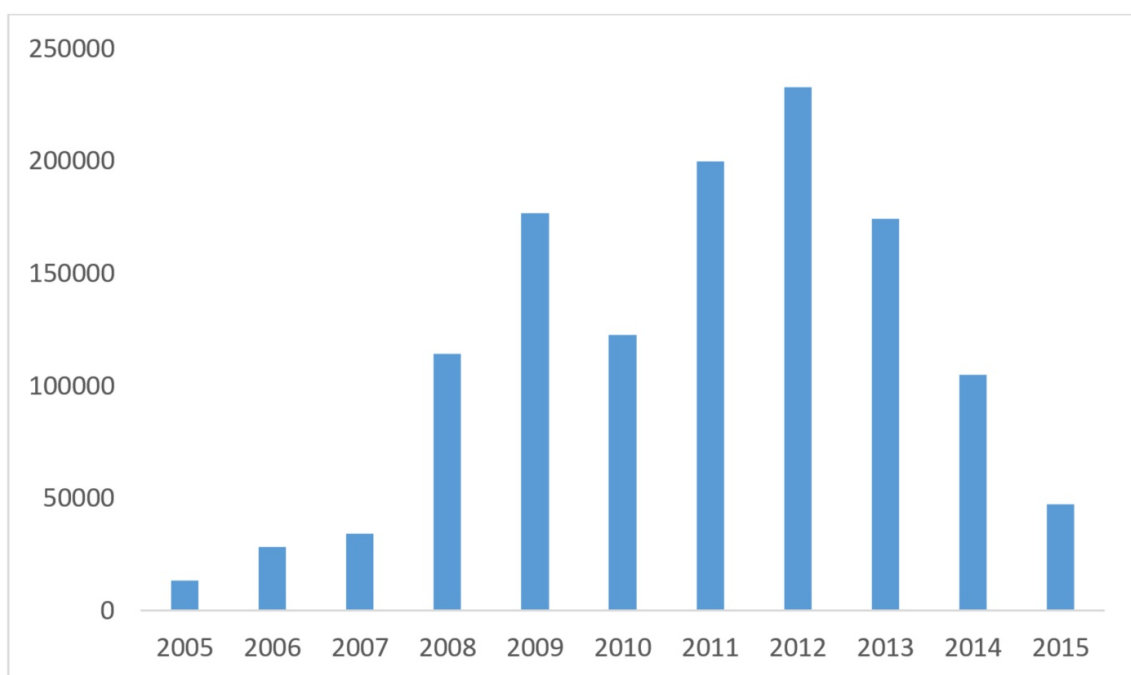
Tähän tutkimukseen on Suomi24-korpusta rajattu hakemalla aineiston tietokantaversiosta merkkijonoa *pikavip*⁹ (kuviossa 1 nähdään kyseisen merkkijonon suhteellinen yleisyys vuosittain koko korpuksessa). Tällä tavalla löydettiin kaikki sanan *pikavippi*

⁹ Kursivoin analysoitavan kieliaineksen.

taivutusmuodot ja yhdyssanat, joissa *pikavippi* on osana. Kyseinen merkkijono ei käytännössä esiinny muissa sanoissa, joten se rajasi aineiston tehokkaasti vain mielenkiinnon kohteena oleviin viesteihin. Ennen vuotta 2005 oli päivätty neljä viestiä, joissa hakusana esiintyi; on mahdollista, että näiden aikaleimat olivat virheellisiä (pikalainoja ei siis myönnetty Suomessa ennen tuota vuotta). Aikaleimavirheitä on huomattu Suomi24:ssä aiemminkin (Lagus ym., 2016).



Kuvio 1. Merkkijonon *pikavip* suhteelliset frekvenssit miljoonaa sanetta kohden koko Suomi24-korpuksessa.



Kuvio 2. Tarkasteltavan vippikorpuksen sanemäärät vuotta kohden.

Nämä viestit ladattiin ja vuotta 2005 edeltävät neljä viestiä poistettiin. Näin muodostui analyysin kohteena oleva uusi, rajatumpi korpus – sekaannusten välttämiseksi kutsun tätä osa-aineistoa tästä eteenpäin vippikorpuksiksi. Vippikorpuksessa on 16 223 viestiä ja 1 248 771 sanetta (vain juoksevat sanat, ei sisällä välimerkkejä). Kuvioista 2 nähdään saneiden jakautuminen eri vuosille. On huomattava, että ensimmäinen viesti on päivätty 16.8.2005 ja viimeinen 5.7.2015, eli ensimmäiseltä ja viimeiseltä vuodelta on selvästi muita vuosia vähemmän aineistoa.

Menetelmät ja Suomi24:n pikavippikeskusteluiden analyysia

Käyn seuraavaksi tarkemmin läpi joitakin CADSin menetelmiä ja havainnollistan niitä esimerkeillä vippikorpuksen analyysista. Aluksi tarkastelen frekvenssi- ja avainsanalistoja, jotka kertovat aineiston yleispiirteistä. Avainsanojen jälkeen tarkastelen kollokaatteja, eli tutkittavan sanan lähiympäristössä esiintyviä sanoja. Lopuksi tuon esiin erään tavan tarkastella ajallista muutosta korpuksessa.

Frekvenssit

Frekvenssilistoissa esitetään korpuksen sanojen absoluuttinen ja/tai suhteellinen yleisyys (esim. miljoonaa sanaa kohden) ja sijaluku. Yleiset sanat kertovat jotain aihepiireistä, jotka korostuvat aineistossa. Kielen yleisimmät sanat (esim. *olla*, *ja*, *ei*) nousevat kuitenkin aina näiden listojen kärkeen, joten eri korpusten frekvenssilistat saattavat ensi vilkaisulla näyttää hyvinkin samanlaisilta. Tietyille korpukselle tyypilliset sanat voidaan seuloa näistä paremmin esiin poistamalla listasta yleiset kieliopilliset sanat (ns. *stop word list*) tai keskittymällä esimerkiksi vain substantiiveihin, joiden voidaan olettaa kertovan aineiston aiheista enemmän kuin monien muiden sanaluokkien sanojen. Tässä tutkimuksessa käytin jälkimmäistä tapaa saadakseni yleiskuvan vippikorpuksen sisällöstä. Sanalistan voi tuottaa millä tahansa yleisistä korpustyökaluista, mutta tässä käytin *#Lancsbox*-ohjelmistoa.

Tarkastelin sataa yleisintä substantiivin taivutusmuotoa perustuen automaattisen kielenjäsenyyksen tuloksiin. Automaattinen jäsenyys oli mennyt vikaan joidenkin sananmuotojen kohdalla, jotka eivät olleet substantiiveja – jätin nämä pois tarkastelusta. Luokittelin sen jälkeen sananmuodot ”laadullisesti” temaattisiin ryhmiin, mikä helpottaa aineiston tulkitsemista lukijalle. Havainnollistan näiden ryhmien osuuksia visuaalisesti *treemap*-kuvioilla (kuvio 3). Kunkin laatikon koko kuvaa sen sisältämien saneiden (sanaesiintymien) määrää suhteessa koko kuvion edustamiin saneisiin.

Kuvio 3 auttaa muodostamaan yleiskuvan siitä, millaisista asioista aineistossa keskustellaan eniten. Toisaalta näin löydettyt sanat ovat melko yllätyksettömiä, kun otetaan huomioon tarkasteltavan korpuksen luonne. On itsestään selvää, että sanat *pikavippi* ja *vippi* sekä niiden kanssa samaan semanttiseen kenttään kuuluvat *laina*



Kuvio 3. Yleisimmät substantiivit temaattisiin ryhmiin luokiteltuna.

ja *velka* kuuluvat yleisimpiin substantiiveihin, eikä myöskään ole yllättävää, että niitä seuraa muita *rahaan* liittyviä sanoja sekä pankkitoimintaan liittyvät *korko*, *kulu*, *pankki* ja *tili*. Diskurssintutkijan näkökulmasta on sen sijaan kiinnostavaa, että joukosta erottuu sanoja, jotka liittyvät ongelmiin ja apuun: tällaisiksi tulkitseen sanat *pakko*, *ongelma*, *kunto*, *apu*, *onni* ja *kiitos*, ja erilliseksi ryhmäksi kokoamani *ulosotto*, *perintä* ja *luottotieto*. Laajemmassa analyysissä olisi mahdollisesti kiinnostavaa tutkia tarkemmin sanoja *elämä*, *työ* ja *mieli* (erityisesti muodossa *mielestä* + possessiivisuffiksi).

Yleisiä sanoja tarkasteltaessa on huomattava, että yleisyys ei ole merkityksellistä tietyn aineiston kannalta, jos kyseinen sana on muuallakin suunnilleen yhtä yleinen. On siis tärkeää välttää johtopäätösten vetämistä olemattomasta näytöstä. Monet kuvion 3 sanoista ovat tosiaan likimain yhtä yleisiä myös muissa suomenkielisissä teksteissä: tällaisia ovat esimerkiksi *aika*, *asia*, *osa* ja *mies* (ks. esim. Kotimaisten kielten keskus, n.d.). Nämä sanat eivät siis kerro mitään kiinnostavaa tästä nimenomaisesta korpuksesta. Jos en olisi tietoinen tästä, saattaisin esimerkiksi olettaa sen liittyvän jotenkin pikavippikeskusteluun, että kuviossa 3 esiintyy *mies* mutta ei *nainen*. Näiden sanojen suhteen tarkasteltava aineisto ei kuitenkaan olennaisesti poikkea muista suomenkielisistä teksteistä. Toinen esimerkki tällaisesta virheellisestä päätelmästä olisi ajatella, että aineiston tekstit ovat varmaankin hyvin negatiivisia, koska *ei* on kolmanneksi yleisin sana – vaikka niin on käytännössä aina. Negatiivisuutta voi kyllä olla, mutta se näkyy muualla kuin kieltoverbin yleisyydessä.

Frekvenssien analysointi vaatii siis tutkijalta jonkinasteista kykyä suhteuttaa

korpuksen sanojen yleisyyksiä siihen, mikä olisi odotuksenmukaista. Useimmiten on kuitenkin epärealistista, että tutkija tietäisi ulkomuistista tiettyjen sanojen tarkkoja sijoituksia suomen kielen taajuussanastossa tai ehtisi vertailla yleisyyksiä yksitellen. Niinpä käsittelen seuraavaksi sitä, kuinka vertailuaineiston käyttöä voidaan automatisoida avainsana-analyysin muodossa.

Avainsanat

Avainsana-analyysissa verrataan tarkasteltavan korpuksen (eli kohdekorpuksen) sanojen yleisyyttä joko kokonaiseen vertailukorpukseen tai sellaisesta tehtyyn frekvenssilistaan. Avainsanat lasketaan jollakin tarkoitukseen kehitetyistä tilastollisista menetelmistä (ks. esim. Brezina, 2018b, s. 83–85). Tuloksena on lista sanoista, jotka ovat merkitsevästi yleisempiä (tai harvinaisempia) kohdekorpuksessa kuin vertailukorpuksessa. Samalla tämä karsii listasta yleensä myös hyvin yleiset sanat kuten ja tai olla, koska tällaisten sanojen yleisyys ei juuri vaihtelee korpuksesta toiseen.

Pyrin löytämään vertailukorpuksen, joka muiden kuin kiinnostuksen kohteena olevan ominaisuuden osalta muistuttaa kohdekorpusta mahdollisimman paljon. Sketch Engine korpustyökalu tarjosi käyttöön melko sopivan vertailukorpuksen, *Finnish Web 2014 corpus (FiTenTen2014)*, joka on internetistä kerätty erittäin laaja (noin 1,4 miljardia sanetta) kokoelma suomenkielisiä tekstejä. Vertailukorpus edustaa siis kohdekorpuksen tavoin internetissä käytettävää suomea viime vuosilta; kielimuotojen erot eivät näin nouse tuloksissa ainakaan kovin vahvasti esille. Laskin avainsanat lemminoita (eli sisältäen kaikki taivutusmuodot) ja isoja kirjaimia huomioimatta. Sketch Enginen tekemä lemmaus kuitenkin valitettavasti petti muiden muassa juuri sanan *pikavippi* kohdalla. *Pikavipin* taivutusmuodot jätinkin listasta kokonaan pois – tiesinhän jo, että kyseistä sanaa esiintyy aineistossa suhteettoman paljon. Näin käsitellyn listan ensimmäiset sata sanaa ovat nähtävissä liitteessä.

Sketch Enginessä avainsanat lasketaan *Simple Maths* -nimellä tunnetulla laskutavalla (Kilgarriff, 2009, 2012). Kilgarriffin (2005) mukaan perinteisesti käytetyt monimutkaisemmat menetelmät (esim. *log-likelihood*) eivät oikeasti ole hyödyksi avainsanojen laskemisessa, koska niillä pyritään kumoamaan nollahypoteesi, joka ilmiselvästi on väärä – nimittäin että sanojen esiintyminen kielessä olisi sattumanvaraista. Kilgarriffin yksinkertaisessa laskutavassa sanan normalisoitu frekvenssi kohdekorpuksessa jaetaan sen normalisoidulla frekvenssillä vertailukorpuksessa, kuitenkin niin, että molempiin lisätään parametri N (esim. 1, 100 tai 1000). $N:n$ voi määritellä tutkimuksen tarpeiden mukaan; pienet $N:n$ arvot tuovat listan kärkeen vertailukorpuksen perusteella harvinaisempia sanoja, isommat taas yleisempiä sanoja. Asetin $N:n$ arvoksi 1000, koska *pikavipin* muodot ja siihen liittyvät muut harvinaiset sanat kuten pikalainayritysten nimet ja internetosoitteet olisivat muuten vallanneet listan alkupään.

Avainsanalistan kärjessä on jälleen odotetusti paljon sanoja, jotka viittaavat

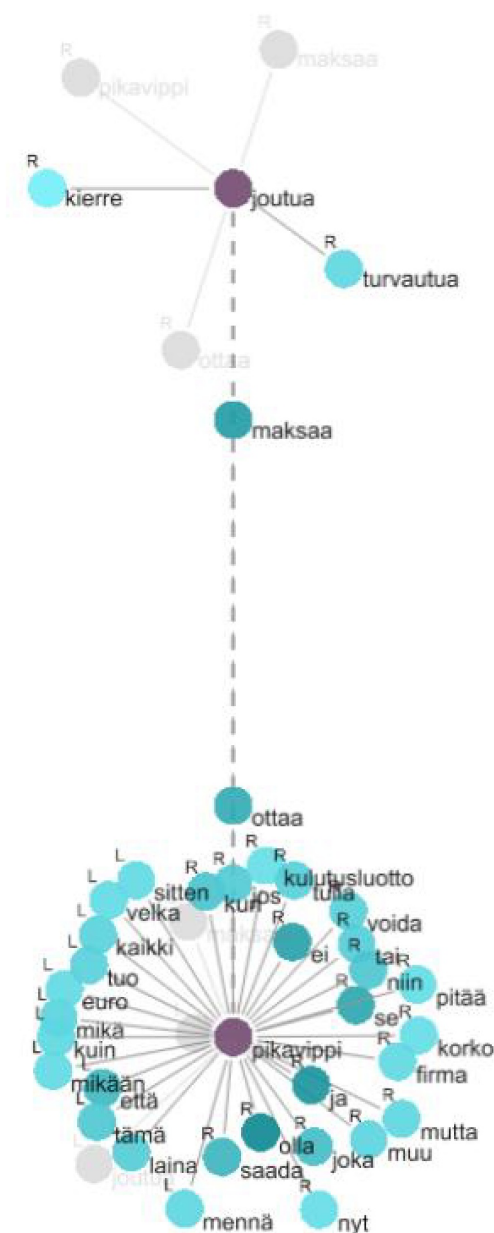
velkaan tai rahaan, kuten *laina*, *maksaa*, *raha*, *velka*, *pankki* ja *euro* (ks. liite). On myös sanoja, jotka viittaavat velkaan liittyviin ongelmiin (suluissa liitteen rivinumerot): *ulosotto* (11), *luottotieto* (13), *perintä* (29) ja *kierre* (92). *Joutua* (36) näyttää myös ainakin osaksi liittyvän keskusteluun ongelmista – tämä arvio perustuu nopeaan tutustumiseen konkordanssiriveihin eli lausekonteksteihin. Analyysin mahdollinen seuraava askel olisi tutustua valittujen avainsanojen esiintymiskonteksteihin tarkemmin sen selvittämiseksi, miten ja millaisista velkaongelmista korpuksessa puhutaan. Muita avainsanoja, joita voisi olla kiinnostavaa tutkia tarkemmin ovat esimerkiksi *kallis* (74) ja *halpa* (91), sillä ne voisivat valottaa sitä, kuinka kuluttajat käsittävät pikavippien hinnan.

Muita avainsanoja, jotka herättävät diskurssintutkijan huomion, ovat *köyhä* (51), *työtön* (52) ja *tyhmä* (93), sillä ne voivat kertoa jotain siitä, mitä pikavippien ottajista aineistossa sanotaan. *Kannattaa* (43) puolestaan näyttää liittyvän tapauksiin, joissa kuluttajat neuvovat toisiaan (vrt. Ylisiurua, 2017). *Pystyä*-verbin (53) tarkempi analyysi puolestaan saattaisi johtaa kiinnostavan keskustelun jäljille esimerkiksi yksilöiden toimintamahdollisuuksiin liittyen. *Kieltää* (83) ja *laki* (84) taas viittaavat kriittiseen, yhteiskunnallisemman tason keskusteluun ongelmien mahdollisista ratkaisuista.

Kollokaatit

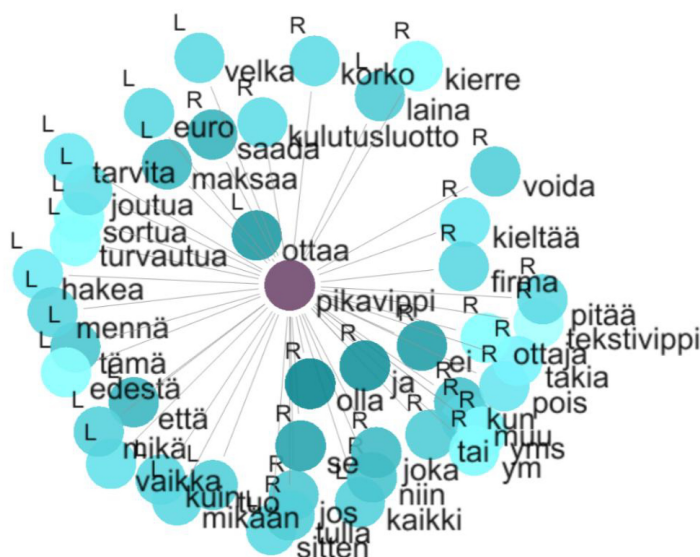
Tässä tutkimuksessa kollokaatit on (ellei muuta mainita) laskettu käyttäen *logDice*-menetelmää (Rychlý, 2008), joka tasapainottelee yhdessäesiintymisen yleisyyden (yleinen, mutta kollokaatti esiintyy usein muidenkin sanojen kanssa) ja eksklusiivisuuden välillä (kollokaatti esiintyy erityisen usein juuri hakusanan kanssa) (Brezina, 2018a). Hakuikkuna oli kolme sanaa vasemmalle ja kolme oikealle. Käytin visualisointiin #Lancsbox-korpusohjelmistoa, jonka innovaationa ovat kollokaattikaaviot ja verkostot (Brezina ym., 2015). Näissä kuvioissa hakusana on keskellä ja kollokaatit asettuvat sen ympärille; keskimääräinen sijainti hakusanan oikealla tai vasemmalla puolella merkitään joko suoraan sijainnilla tai tarvittaessa, tilankäytöllisistä syistä, R- ja L-kirjaimilla (right, left). Yleisemmät kollokaatit merkitään tummemmalla värillä ja tilastollinen yhteys näkyy etäisyytenä pääsanasta. Yksinkertaisesta graafista voidaan edetä verkostoksi klikkaamalla kollokaatteja ”auki”, jolloin kyseisen sanan kollokaatit lisätään kuvioon.

Koska *pikavippi* on vippikorpuksessa erittäin yleinen sana, kollokattianalyysin tulokset muistuttavat edellä esitettyjä avainsanoja. Kollokaatit ovat kuitenkin pikavipin välittömässä läheisyydessä, joten niillä on todennäköisemmin myös suora merkitysyhteys tekstissä. Kuvio 4 on esimerkki kollokaattiverkostosta. Koska laskutapa korostaa jonkin verran yleisyyttä, kuviossa on mukana sellaisia sanoja kuin *olla* ja *mutta*, ja toisaalta esimerkiksi *kulutusluotto*, *velka* ja *laina*. Kiinnostavampi on jo avainsanoissa mukana ollut *joutua*, jota pääsin nyt tutkimaan tarkemmin avaamalla myös sen kollokaatit näkyviin kuvioon 4. Yhteisiä kollokaatteja ovat katkoviivalla yhdistetyt *ottaa* ja *maksaa*. *Joutua* sanan kollokaatteja ovat myös *kierre* ja *turvautua*. Vertailun vuoksi tarkastelin *pikavipin*



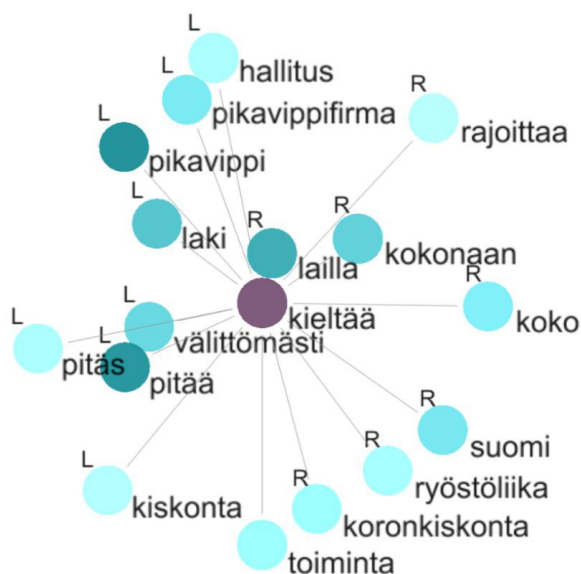
Kuvio 4. Kollokaattiverkosto: *pikavippi* ja *joutua*. Hakuikkuna: 3:3. Laskentatapa: logDice, minimiarvo: 9. Tehty lemmeille. L=vasemmalla, R=oikealla.

kollokaatteja myös hieman erilaisella laskentatavalla (MI2, eräs ns. keskinäisen informaation variantti) (kuviosta 5 näkyy, että *turvautua* on nyt myös suoraan *pikavipin* kollokaatti, samoin *tarvita* ja negatiivissävytteinen *sortua*, joka sisältää ajatuksen huonosta päätöksestä, ehkä arvostelukyvyn tai itsehillinnän petttämisestä. Sekä *joutua* että *turvautua* viittaavat vastentahtoisuuteen tai pakkoon, (hyvien) vaihtoehtojen puuttumiseen. Kaikki nämä sanat yhdessä – *tarvita*, *joutua*, *kierre*, *turvautua*, *sortua* – tuntuisivat viittaavan ajatukseen, että ihmisillä on rahantarve, mutta heille on tarjolla tai he pystyvät tarttumaan vain huonoihin vaihtoehtoihin, joista on myöhemmin vaikea päästä irti.



Kuvio 5. Pikavipin kollokaatteja. Hakuikkuna 3:3. Laskentatapa: MI2, minimiarvo: 10,5. Tehty lemmeille.

Kuten avainsanoissakin, myös kuviossa 5 on mukana sana *kieltää*. Tutkin sitä nyt tarkemmin tekemällä sille oman kollokaattikuvion (kuvio 6): *kieltää*-verbin yleisimmät kollokaatit vippikorpuksessa ovat *pikavippi* ja *pitää*, sen jälkeen *laki* ja *lailla* (tässä on automaattinen lemmatisointi taas pettänyt). Muutkin kollokaatit ovat kiinnostavia: *hallitus* on mahdollisesti ehdotettu toimija, *välittömästi* ja *kokonaan* viittaavat mielipiteen ehdottomuuteen, ja *kiskonta* ja *koronkiskonta* esiintyvät vertailukohtina. Tarkastelin *kieltää*-verbin kollokaatteja myös konkordanssirivien avulla, jolloin varmistui, että suurimmassa osassa tapauksista tosiaan toivotaan pikavippien kieltämistä tai ennakoidaan, että niin todennäköisesti tehdään jossain vaiheessa. Vain pienessä vähemmistössä esitetään, että pikavippejä *ei* pitäisi kieltää.



Kuvio 6. Kieltää. Hakuikkuna: 3:3. Laskentatapa: logDice, minimiarvo: 8. Tehty lemmeille.

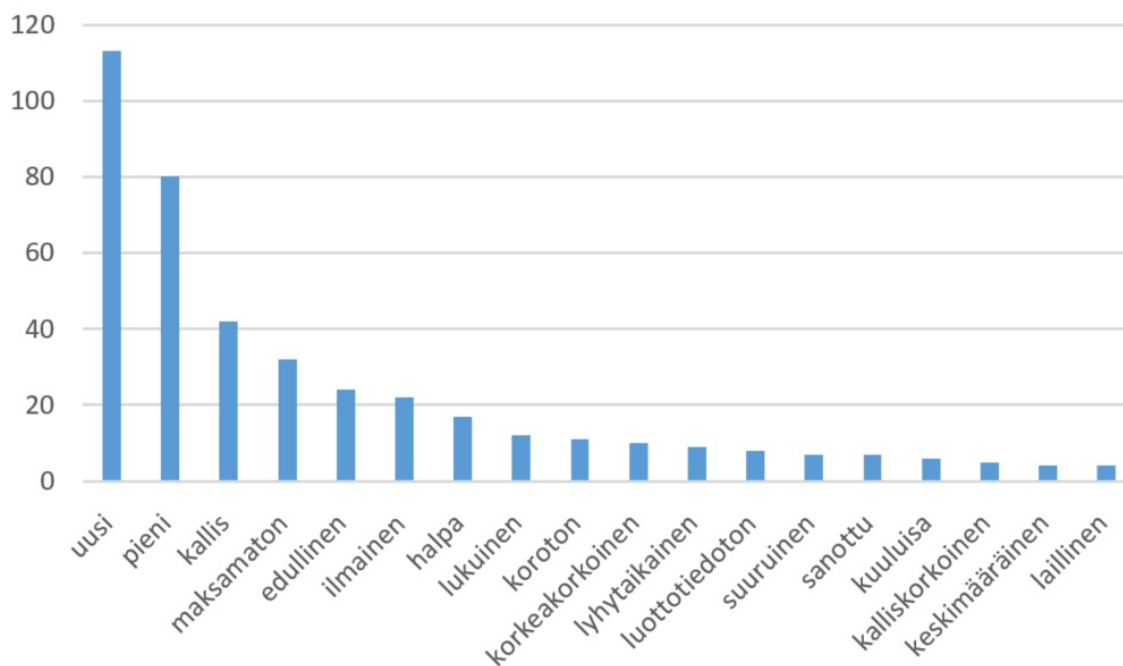
Tällaisten kuvioiden rajoitteena on, että sanan ympärille mahtuu vain tietty määrä kollokaatteja. Listalla hieman alempana voi kuitenkin olla vähintään yhtä kiinnostavia sanoja. Yhtenä ratkaisuvaihtoehtona voisi olla suodattaa tai jälkikäteen poistaa kieliopilliset sanat kuten konjunktiot, mutta #Lancsbox ei toistaiseksi tarjoa sitä mahdollisuutta. *Kieltää*-sanalle saatiin selkeämpi kuvio, varmaankin koska se ei ole yhtä yleinen eli sillä on kokonaisuudessaan vähemmän kollokaatteja. Ehkä sen esiintymiskontekstit ovat myös luonteeltaan homogeenisempia.

Etumäärärite	pikavippi	Jälkimäärärite	pikavippi	verbi	Verbi	pikavippi	
1. euro (substantiivi)	144	1. firma (substantiivi)	57	1. kieltää (verbi)	89	1. ottaa (verbi)	2616
2. ottaa (verbi)	70	2. kierre (substantiivi)	30	2. erääntyä (verbi)	14	2. maksaa (verbi)	366
3. maksamaton (adjektiivi)	32	3. ulosotto (substantiivi)	27	3. ottaa (verbi)	101	3. myöntää (verbi)	92
4. satanen (substantiivi)	37	4. velka (substantiivi)	27	4. tuhota (verbi)	15	4. kieltää (verbi)	112
5. tonni (substantiivi)	27	5. maksaa (verbi)	52	5. loppua (verbi)	33	5. hakea (verbi)	155
6. kallis (adjektiivi)	42	6. edestä (adverbi)	12	6. jättää (verbi)	23	6. maksella (verbi)	16
7. http://www.ostosraha.fi/ (muu)	7	7. yhtiö (substantiivi)	15	7. pelastaa (verbi)	17	7. markkinoida (verbi)	13
8. koroton (adjektiivi)	11	8. korko (substantiivi)	13	8. velkaannuttaa (verbi)	3	8. otella (verbi)	5
9. kysely (substantiivi)	10	9. koronkiskonta (substantiivi)	6	9. monetti (verbi)	2	9. pääösän (verbi)	2
10.edullinen (adjektiivi)	24	10.helposti (adverbi)	10	10.kokutella (verbi)	3	10.mainostaa (verbi)	20
11.onnilaina (substantiivi)	6	11.paikko (substantiivi)	14	11.maksaa (verbi)	61	11.vertailla (verbi)	15
12.korkeakorkoinen (adjektiivi)	10	12.vippi (substantiivi)	7	12.rahoittaa (verbi)	6	12.vipata (verbi)	3
13.luottotiedoton (adjektiivi)	8	13.paikka (substantiivi)	26	13.mahdollistaa (verbi)	7	13.listata (verbi)	8
14.fi (substantiivi)	10	14.perintä (substantiivi)	7	14.kertyä (verbi)	6	14.tehtailla (verbi)	7
15.kalliskorkoinen (adjektiivi)	5	15.lisää (adverbi)	37	15.pilata (verbi)	7	15.rahottaa (verbi)	2

Kuvio 7. Pikavipin sanakuva Kielipankin Korp-korpustyökälusta. Tässä vain 15:n kärki.

Toinen kollokaattianalyysin tapa on niin sanottu sanakuva, jossa kollokaatit järjestetään hakusanan ympärille kieliopillisen funktion ja/tai sijainnin mukaisiin ryhmiin. Sanakuvia voi tehdä suoraan esimerkiksi Korp-työkälulla, joten tein sillä sanakuvan *pikavipille* Suomi24-korpuksessa (ks. kuvio 7). Korpissa on laskutapana käytössä *Lexicographers Mutual Information* -niminen keskinäisen informaation muunnos (Bordag, 2008). En kommentoi sanakuvaa kokonaisuudessaan sillä se sisältää paljon sellaista, mitä olen jo käsitellyt. Korpissa sanakuvan kategoriat ovat ainoastaan etumäärärite, jälkimäärärite, verbi sanan edessä ja verbi sanan jäljessä, vaikka hienojakoisempi ryhmittely olisi selkeämpi. Tarkastelen lähemmin vain Korpin antamia, *pikavipin* eteen sijoittuvia adjektiivimääritteitä, joiden määrät on esitetty pylväsdiagrammina kuviossa 8. Tästä on poistettu väärin jäsenneetyt sanat (jotka eivät olleet adjektiiveja). *Luottotiedoton* ei aineistossa varsinaisesti määritä *pikavippiä*, mutta jätin sen kuvioon, koska luottotietojen menetyksen tai luottotiedottomana olon voi ajatella liittyvän pikaluottojen ottamiseen, vaikka tässä ei kieliopillista dependenssiä olekaan.

Kuviosta 8 nähdään, että yleisin adjektiivimäärärite pikavipin kollokaattina on *uusi*, mikä viitanee toistuvaan lainojen ottamiseen – kuten myös *lukuinen*. Ajatus kierteestä on siis tässäkin läsnä. *Pieni* ja *lyhytaikainen* ovat muita pikavipin tyypillisiä ominaisuuksia. Kollokaatit keskittyvät myös vahvasti hintaan tai korkotasoon: *kallis*, *korkeakorkoinen*, *kalliskorkoinen* tai toisaalta *edullinen*, *ilmainen*, *halpa* ja *koroton*. Ymmärrettävästi hinta on ominaisuus, joka kuluttajia erityisesti kiinnostaa.



Kuvio 8. Yleisimmät adjektiivimääritteet *pikavipin* kollokaatteina. Y-akselilla kollokaation absoluuttinen frekvenssi. Alle 4 kertaa esiintyvät adjektiivit on jätetty pois kuvioista.

Trendit

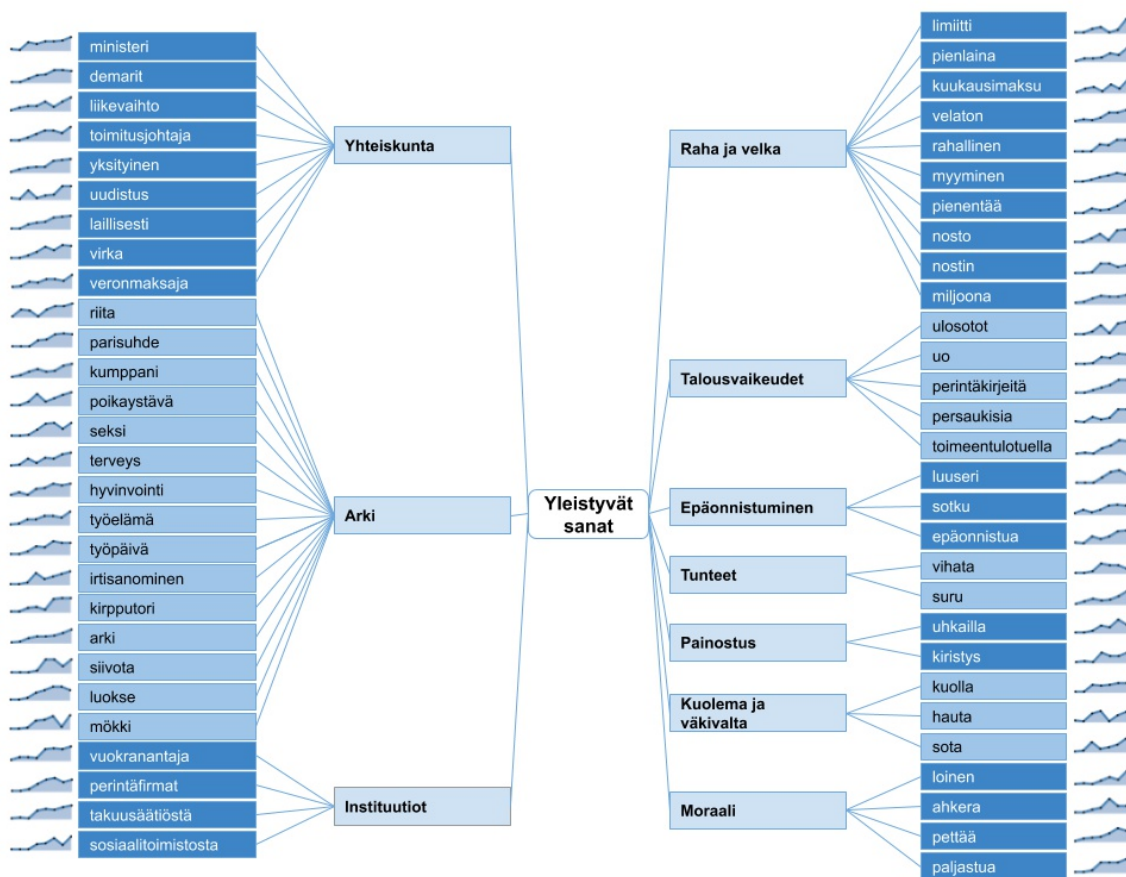
Ajallisten systemaattisten muutosten tarkasteluun käytin Sketch Enginen trendityökalua (viittaan tässä siis *trendeillä* nimenomaan tähän analyysimenetelmään ja sen tuloksiin). Yksittäisten sanojen osalta on helppo ilman erityistä työkaluakin laskea muutokset sanojen normalisoiduissa frekvensseissä ajanjakson aikana, esimerkiksi eri vuosina tai kuukausina (muutoksia yksittäisten sanojen käytössä ovat analysoineet esim. Lagus ym., 2018, ja Soila & Kaarakainen, 2018). Sketch Enginen trendityökalun etuna on, että se laskee automaattisesti trendit koko aineistolle ja seuloo esiin ne yleistyneet ja harvinaistuneet sanat, joiden yleisyydessä on ajanjakson aikana tapahtunut merkittävin muutos. Valitsin laskutavaksi Mann-Kendall ja Theil-Sen menetelmien yhdistelmän (Kilgarriff, Busta & Rychlý, 2015) ja asetin minimifrekvenssin melko korkeaksi (20) ja p-arvon matalaksi ($p < 0.01$). Tuloksille eli yleistyneiden ja harvinaistuneiden sanojen listoille tein "laadullisen" ryhmittelyn ja piirsin yleistyneistä sanoista mind map tyyllisen kuvion (kuvio 9).

Erityisesti arjen ja yhteiskunnan semanttiset kentät ovat voimistaneet asemiaan (ks. kuvio 9). Sen sijaan monet sanat, jotka liittyvät läheisesti pikavippi-ilmioon, ovat suhteessa vähentyneet: muiden muassa monien lainojen myöntäjien nimet sekä esimerkiksi *maksupäivä* (), *kännykkä* () ja *vuosikorkoja* (). Uskon tämän viittaavan siihen, että pikavipit ovat ilmiönä tulleet tutuksi ja niihin viitataan entistä kontekstualisoidummin. Keskustelijoilla on näkemyksiä ja kokemuksia niiden roolista ihmisten arjessa ja

yhteiskunnassa – ja osin he keskustelevat niistä myös abstraktimmalla tasolla, mihin viittaisi aiemmin esiin tullut puhe niiden kieltämisestä. Ilmiön alkuvaiheessa sen sijaan ihmiset usein kyselivät toisiltaan tietoa ja neuvoja pikavippeihin ja niiden tarjoajiin liittyen, mikä ei ehkä enää ole samassa mittakaavassa tarpeen.

Olen visualisoinut trendit pienillä sparkline-kaavioilla, jotka otin suoraan Sketch Enginestä. Ne kuvaavat kunkin sanan kehitystä pienessä tilassa. Kuviota 9 varten kävin läpi yleistyneiden sanojen listan 85 ensimmäistä sanaa. Jätin näistä kuitenkin pois erisnimet ja useita sanoja, jotka liittyvät käyttäytymiseen keskustelufoorumeilla, esim. *nimetön* ja *aloittaja*, sekä joitakin sanoja, jotka nopean konkordanssien tarkastelun perusteella ovat aineistossa hyvin monimerkityksisiä ja siten mahdottomia luokitella (esim. *sormi*, *kato*).

Kuvion 9 yleistyneissä sanoissa huomion kiinnittävät monet kielteiset sanat, liittyen niin työelämään ja talousasioihin kuin henkilökohtaiseen elämään ja tunteisiin – esimerkiksi *irtisanominen*, *uo* (eli ulosotto), *riita*, *epäonnistuminen*, *suru* ja jopa *kuolema*. On huomionarvoista, että mukana on myös voimakkaan moraalisen latauksen sisältäviä sanoja kuten *loinen* (vrt. Saari, Behm & Lagus, 2017), joka on tämän analyysin mukaan *limiitin* jälkeen toiseksi voimakkaimmin yleistynyt sana. Kaikkia näitä sanoja voisi lähteä tutkimaan tarkemmin joko kollokaattien tai konkordanssien laadullisen analyysin avulla, mutta tässä olen tyytynyt katsomaan



Kuvio 9. Yleistyviä sanoja ryhmittelevä mind map.

konkordansseja nopeasti saadakseni käsityksen siitä, ovatko tulkintani oikeansuuntaisia. Esimerkiksi *loinen*-sanana kohdalla tämä tarkastelu vahvisti, että sanaa on käytetty vain metaforisesti ja puhuttaessa ihmisistä tai muista toimijoista kuten pikavippejä tarjoavista yrityksistä.

Lopuksi

Tässä artikkelissa olen pyrkinyt osoittamaan, että kielitieteen määrälliset menetelmät ovat hyödynnettävissä suurten tekstiaineistojen analyysiin myös kuluttajatutkimuksen kontekstissa. Korpusavusteinen diskurssianalyysi eli CADS sopii erityisesti kuluttajatutkijoille, joita kiinnostavat yleiset puhettavat isossa tekstiaineistossa koskien jotakin selkeästi sanallistettavaa teemaa, kuten tiettyä tuotetta, tuotekategoriaa tai kulttuurista kulutusilmiötä – sillä määrällinen analyysi perustuu sanoihin ja niiden yhdessäesiintymiseen.

Tässä artikkelissa olen soveltanut CADSin keskeisiä menetelmiä pikavippejä koskevan keskustelun analysointiin Suomi24-korpuksessa. Ajallinen tarkastelu toi esiin sen, että siinä missä pikavipit olivat aluksi uusi ilmiö, josta etsittiin tietoa, ajanjakson aikana ne ovat tulleet tutuksi ja limittyneet vahvasti monen keskustelijan arkielämään. Samalla pikalainojen kielteiset puolet ovat nousseet tarkastellun ajanjakson (2005–2015) aikana kuluttajien tietoisuuteen, ja taloudellisia ja psyykkisiä ongelmia liitetään niihin keskustelufoorumilla yhä enemmän. Yksi keskeinen tulos on, että vaatimus pikalainojen kieltämisestä lainsäädännöllä nousee keskusteluissa voimakkaasti esiin. Monet keskustelijat näyttävät myös ymmärtävän pikalainojen ottajia: *pikavippi*-sanana yhteydessä käytetyt ilmaisut kuten *tarvita* ja *turvautua* viittaavat välttämättömyyteen ja vaihtoehtottomuuteen (ks. myös esim. Kaartinen ym., 2009; Rowlingson, Appleyard & Gardner, 2016). Vastuu ongelman ratkaisemisesta sijoittuu puheissa siis enemmän yhteiskunnalle kuin yksilöille, vaikka pikalainojen otto esitetään myös esimerkiksi *tyhmänä*.

Tulokset ovat samansuuntaisia kuin monet aiempien tutkimusten havainnot, kuten että pikalainoilla on yhteys velkakierteen syntymiseen (esim. Rantala, 2012; Stegman & Faris, 2003) ja maksuvaikeuksiin (esim. Valkama & Muttilainen, 2008; Rantala & Tarkkala, 2009; Majamaa, Lehtinen & Rantala, 2019) – aineistossa puhutaan paljon muun muassa *ulosotosta* ja *luottotiedoista*. Myös lainan hinta (esim. sanat *kallis* ja *halpa*) on keskusteluissa usein esillä. Pikavippien ottajissa on havaittu olevan paljon työttömiä ja huonotulaisia (Kaartinen & Lähteenmaa, 2006, Rantala & Tarkkala, 2009; Autio ym., 2009), ja ne kohdentuvat erityisesti Suomen köyhimmille alueille (Valkama ja Muttilainen, 2008); myös Suomi24:n keskusteluissa muiden muassa sanat *työtön* ja *köyhä* nousivat avainsanojen joukkoon. Tulokset siis toisaalta tukevat aiempia tutkimuksia ja viittaavat niissä havaittujen suuntausten jatkumiseen, toisaalta osoittavat, että CADSin avulla voidaan saada relevantteja tuloksia.

Oikeastaan CADS ei ole yksi ainoa menetelmä vaan jatkuvasti kehittyvä joukko

määrällisiä ja laskennallisia lähestymistapoja kielen ja diskurssin analyysiin. CADSin vahvuus verrattuna muihin viimeaikaisessa keskustelussa esillä olleisiin suurten tekstiaineistojen analyysimenetelmiin onkin sen monipuolisuus. Esimerkiksi suosittu aihemallinnus tarjoaa lähinnä yleisnäkömään aineistoon (esim. Nelimarkka, 2019; Purhonen & Toikka, 2016; Repo & Matschoss, 2018), vaikka siinä voidaan toki myös vertailla aiheiden eroja eri osa-aineistoissa (Laaksonen & Nelimarkka, 2018) tai sitä voidaan yhdistää laadulliseen tarkasteluun (Ylisiurua, 2017). Aihemallinnuksella itsessään on kuitenkin vaikea porautua syvemmälle esiin nouseviin kiinnostaviin teemoihin – tässä voisi auttaa esimerkiksi kollokaattien tarkastelu ja visualisointi. Siinä missä aihemallinnuksen tuloksiin eivät vaikuta ”sanajärjestys, sanamuodot tai yksittäisen sanan lähikonteksti” (Ylisiurua, 2017, s. 45), CADSissa nuo kaikki voivat vaikuttaa, riippuen valitusta analyysitavasta. CADSissa voidaan myös yhdistää aineistolähtöisiä tutkimustapoja, kuten frekvenssien ja avainsanojen tarkastelu (vrt. aihemallinnus), ja teorialähtöisiä tutkimustapoja, kuten tiettyjen sanojen valinta analyysin kohteeksi aiemman kirjallisuuden pohjalta (vrt. Lagus ym., 2018; Soila & Kaarakainen, 2018).

CADS mahdollistaa myös joustavan siirtymisen määrällisen ja laadullisen tarkastelun välillä. Koska kielen kontekstuaalisuus on huomioitu korpustyökaluissa, esimerkiksi tietystä avainsanasta on yleensä helppoa siirtyä tarkastelemaan lause- ja kappale- ja tekstitaso konteksteja, joissa kyseinen sana esiintyy. Kuten muidenkin menetelmien osalta on todettu, laadullista analyysia myös tarvitaan (Lagus ym., 2018; Ylisiurua, 2017). Vähintäänkin nopea tekstikontekstien tarkastelu on tarpeen, jotta nähdään, onko tietyn sanan pääasiallinen merkitys tulkittu oikein. Tässä artikkelissa olen tilan rajallisuuden vuoksi tyytynyt tämän tason laadulliseen analyysiin, mutta jos artikkelin päätavoitteena ei niinkään olisi ollut CADSin esittely vaan pikavippikeskustelun syvälinen ymmärtäminen, olisin todennäköisesti valinnut vain yhden tai kaksi määrällistä menetelmää ja yhdistänyt niihin perusteellisempaa laadullista diskurssianalyysia.

CADSin soveltaminen edellyttää tiettyjä perustietoja kielestä, mikä voi asettaa haasteita joillekin kuluttajatutkijoille. Toisaalta se on melko helposti lähestyttävä menetelmä, etenkin verrattuna eräisiin muihin digitaalisten ihmis- ja yhteiskuntatieteiden lähestymistapoihin: esimerkiksi koodaustaidot eivät ole CADSin käyttäjälle välttämättömiä, sillä valmiita ohjelmistoja analyysin avuksi on olemassa moneen tarpeeseen. Toinen haaste suomalaisille tutkijoille on suomenkielinen aineisto. Automaattinen lemmaus ja kielenjäsenitys on saatavilla suomen kielelle, mutta erilaisten sananmuotojen runsaus tekee automaattisen analyysin virheistä – joita joka tapauksessa tulee – häiritsevempiä kuin esimerkiksi englannissa. Suomeksi on myös tarjolla rajallisesti valmiita korpuksia. Nykyiset korpustyökalut mahdollistavat kuitenkin myös itse kerättyjen korpusten analyysin, joten ei ole välttämätöntä tyytyä muiden keräämiin aineistoihin.

Lähteet

- Aller Media Oy (2014). Suomi 24 -korpus (2015H1) [tekstikorpus]. Kielipankki. Saatavilla: <http://urn.fi/urn:nbn:fi:lb-201412171>
- Anthony, L. (2018). Visualisation in corpus-based discourse studies. Teoksessa C. Taylor & A. Marchi (toim.), *Corpus Approaches to Discourse: A Critical Review*. Milton, UK: Routledge.
- Arsel, Z. & Bean, J. (2012). Taste regimes and market-mediated practice. *Journal of Consumer Research*, 39(5), 899–917. doi:10.1086/666595
- Autio, M., Wilska, T.-A., Kaartinen, R. & Lähteenmaa, J. (2009). The use of small instant loans among young adults – a gateway to a consumer insolvency? *International Journal of Consumer Studies*, 33(4), 407–415. doi:10.1111/j.1470-6431.2009.00789.x
- Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.
- Baker, P. & McEnery, T. (toim.). (2015). *Corpora and discourse studies: integrating discourse and corpora*. Basingstoke: Palgrave Macmillan.
- Bordag, S. (2008). A comparison of co-occurrence and similarity measures as simulations of context. *International Conference on Intelligent Text Processing and Computational Linguistics*, (s. 52–63). Berlin, Heidelberg: Springer.
- Brezina, V. (2018a). Statistical choices in corpus-based discourse analysis. Teoksessa C. Taylor & A. Marchi (toim.), *Corpus Approaches to Discourse: A Critical Review* (s. 259–280). Milton, UK: Routledge.
- Brezina, V. (2018b). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.
- Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173.
- Brookes, G. & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. 21(1), 3–21. doi:10.1177/1461445618814032
- Burton, M. (2010). *Keeping the plates spinning*. London: Consumer Focus.
- DuFault, B. L. & Schouten, J. W. (2018). Self-quantification and the datapreneurial consumer identity. *Consumption Markets & Culture*. Ennakkojulkaisu verkossa. doi:10.1080/10253866.2018.1519489
- Fairclough, N. (1995). *Critical Discourse Analysis: The Critical Study of Language*. London: Longman.
- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. Teoksessa C. Taylor & A. Marchi (toim.), *Corpus Approaches to Discourse: A Critical Review* (s. 225–258). Milton, UK: Routledge.
- Hilpert, M. (2011). Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 16(4), 435–461. doi:10.1075/ijcl.16.4.01hil
- Hoey, M., Mahlberg, M., Stubbs, M. & Teubert, W. (toim.) (2007). *Text, discourse and corpora: theory and analysis*. London: Continuum.
- Humphreys, A. & Wang, R. J.-H. (2017). Automated text analysis for consumer research. *Journal of Consumer Research*, 44(6), 1274–1306. doi:10.1093/jcr/ucx104
- Isotalo, V., Laaksonen, S.-M., Pöyry, E. & Jallinoja, P. (2019). Sosiaalisen median ennustekyky kaupan myynnissä - esimerkkinä veganismi ja vegaanisat ruuat. *Kansantaloudellinen aikakauskirja*, 115(1), 91–112.

- Jantunen, J. H. (2018). Homot ja heterot Suomi24:ssä: analyysi digitaalisista diskursseista. *Puhe ja kieli*, 38(1), 3–22.
- Johansson, M., Jantunen, J. H., Heimo, A., Ahonen, M. & Laippala, V. (2018). Verkkokeskustelujen kansa: korpusavusteinen diskurssianalyysi Suomi24-keskustelupalstasta. *Sananjalka*, 60, 96–117.
- Kaartinen, R., Autio, M., & Lähteenmaa, J. (2009). Pikavipeillä vaikeuksiin – nuorten luotonkäyttäjien pakkotilanteet, vippikierteet ja huono-osaisuus. Teoksessa M. Laine, V. Keskinen, M. Tuominen & T. Hakkarainen (toim.), *Köyhyyden monet kasvot* (s. 119–129). Helsinki: Helsingin kaupungin tietokeskus.
- Kaartinen, R. & Lähteenmaa, J. (2006). Miten ja mihin nuoret käyttävät pikavippejä ja muita kulutusluottoja? Helsinki: Kauppa- ja teollisuusministeriö.
- Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), 263. doi:10.1515/cllt.2005.1.2.263
- Kilgarriff, A. (2009). Simple maths for keywords. *Proceedings of the Corpus Linguistics Conference*. Liverpool, UK. Saatavilla: <https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf>
- Kilgarriff, A. (2012). Getting to Know Your Corpus. Teoksessa P. Sojka, A. Horák, I. Kopeček & K. Pala (toim.), *Text, Speech and Dialogue. TSD 2012. Lecture Notes in Computer Science*, vol 7499 (s. 3-15). Berlin, Heidelberg: Springer.
- Kilgarriff, A., Busta, J. & Rychlý, P. (2015). DIACRAN: a framework for diachronic analysis. *Corpus Linguistics (CL2015)*. Lancaster, UK.
- Kotimaisten kielten keskus (n.d.). Kirjoitetun suomen kielen sanojen taajuuksia. Saatavilla: <http://kaino.kotus.fi/sanat/taajuuslista/parole.php>
- Laaksonen, S.-M. & Nelimarkka, M. (2018). Omat ja muiden aiheet: laskennallinen analyysi vaalijulkisuuden teemoista ja aiheomistajuudesta. *Politiikka: Valtiotieteellisen yhdistyksen julkaisu*, 60(2), 132–147.
- Lagus, K., Pantzar, M. & Ruckenstein, M. (2018). Kansallisen tunnemaiseman rakentuminen: Pelon ja ilon rytmit verkkokeskusteluissa. *Kulutustutkimus.Nyt*, 12(1–2), 62–83.
- Lagus, K., Pantzar, M., Ruckenstein, M. & Ylisiurua, M. (2016). SUOMI24: Muodonantaa aineistolle. Helsinki: Valtiotieteellinen tiedekunta, Helsingin yliopisto.
- Laki kuluttajansuojalain muuttamisesta, 596/2019 (2019).
- Lehto, L.-M. (2018). Korpusavusteinen diskurssianalyysi japaninsuomalaisten kielipuheesta. (Väitöskirja), Oulun yliopisto, Oulu. Saatavilla: <http://urn.fi/urn:isbn:9789526219097>
- Lillqvist, E. (2010). Työtön suomalaisessa diskurssissa: korpusavusteinen tutkimus työttömiä koskevista puhetavoista. (Pro gradu -tutkielma), Aalto-yliopisto, Helsinki. Saatavilla: <http://hsepubl.lib.hse.fi/FI/ethesis/id/12418>
- Lillqvist, E. & Harju, A. A. (2018). Discourse of enticement: How Facebook solicits users. *Critical Approaches to Discourse Analysis Across Disciplines (CADAAD journal)*, 10(1), 63–80.
- Löfberg, L. (2017). Creating large semantic lexical resources for the Finnish language. (Väitöskirja), Lancaster University, Lancaster. Saatavilla: <https://eprints.lancs.ac.uk/id/eprint/85960/1/2017lofbergphd.pdf>
- Majamaa, K., Lehtinen, A.-R. & Rantala, K. (2019). Debt judgments as a reflection of consumption-related debt problems. *Journal of Consumer Policy*. doi:10.1007/s10603-018-9402-3
- Majamaa, K., Sarasoja, L. & Rantala, K. (2017). Viime vuosien muutokset

vakavissa velkaongelmissa. Analyysi velkomustuomioista. *Yhteiskuntapolitiikka*, 82(6), 676–686.

- Makkonen, A. (2014). Instant loans: Problems and regulations in Finland. *Juridica International*, 22, 96–119.
- Mautner, G. (2016). Checks and balances: How corpus linguistics can contribute to CDA. Teoksessa R. Wodak & M. Meyer (toim.), *Methods of critical discourse studies* (3. painos, s. 122–143). London: SAGE Publications.
- McEnergy, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.
- Murakami, A., Thompson, P., Hunston, S. & Vajn, D. (2017). 'What is this corpus about?' Using topic modelling to explore a specialised corpus. *Corpora*, 12(2), 243–277. doi:10.3366/cor.2017.0118
- Nelimarkka, M. (2019). Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa. *Politiikka: Valtiotieteellisen yhdistyksen julkaisu*, 61(1), 6–33.
- Paldán, M. (2018). Seksuaalisen haluttomuuden diskurssit Suomi24-korpusaineistossa. (Pro gradu -tutkielma), Jyväskylän yliopisto, Jyväskylä. Saatavilla: <http://urn.fi/URN:NBN:fi:jyu-201810044348>
- Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: an overview of the project. *Corpora*, 5(2), 83–108. doi:10.3366/cor.2010.0101
- Pink, S., Ruckenstein, M., Willim, R. & Duque, M. (2018). Broken data: Conceptualising data in an emerging world. *Big Data & Society*, 5(1), 1–13. doi:10.1177/2053951717753228
- Purhonen, S. & Toikka, A. (2016). "Big datan" haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät: esimerkkitapauksena aihemallianalyysi tasavallan presidenttien uudenvuodenpuheista 1935-2015. *Sociologia*, 53 (1), 6–27.
- Pälli, P. & Lillqvist, E. (tulossa). Diskurssianalyysi. Teoksessa M. Luodonpää-Manni, M. Hamunen, R. Konstenius, M. Miestamo, U. Nikanne & K. Sinnemäki (toim.), *Kielentutkimuksen menetelmiä*.
- Rantala, K. (2012). Vippikierteen muotokuva. *OPTL:n verkkokatsauksia*, 24/2012, 1–27.
- Rantala, K. & Tarkkala, H. (2010). Luotosta luottoon: velkaongelmien dynamiikka ja uudet riskiryhmät yhteiskunnan markkinalogiikan peilinä. *Yhteiskuntapolitiikka* 75, 1.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519–549. doi:10.1075/ijcl.13.4.06ray
- Repo, P. & Matschoss, K. (2018). Citizen visions for European futures—methodological considerations and implications. *European Journal of Futures Research*, 6(1), 20. doi:10.1186/s40309-018-0149-5
- Rowlingson, K., Appleyard, L. & Gardner, J. (2016). Payday lending in the UK: the regul(arisation) of a necessary evil? *Journal of Social Policy*, 45(3), 527–543. doi:10.1017/S0047279416000015
- Ruckenstein, M. (2019). Tracing medicinal agencies: Antidepressants and life-effects. *Social Science & Medicine*, 235, 112368. doi:10.1016/j.socscimed.2019.112368
- Rychlý, P. (2008). A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*. Saatavilla: <https://nlp.fi.muni.cz/raslan/2008/raslan08.pdf#page=14>
- Saari, J., Behm, M. & Lagus, K. (2017). *Sosiaalipummi! Moraalipaniikki 2010-luvun*

- Suomessa. Teoksessa J. Saari (toim.), *Sosiaaliturvariippuvuus: sosiaalipummit oleskeluyhteiskunnassa?* (s. 207–232). Tampere: Tampere University Press.
- Soila, T. & Kaarakainen, M. (2018). Potilaasta kuluttajaksi? Sosiaali- ja terveystalouden asiakkuuksien rakenteellinen muutos mediassa 2007–2017. *Hallinnon tutkimus*, 37(4), 270–285.
- Stegman, M. A. & Faris, R. (2003). Payday lending: A business model that encourages chronic borrowing. *Economic Development Quarterly*, 17(1), 8–32. doi:10.1177/0891242402239196
- Stubbs, M. (1997). Whorf's children: Critical comments on critical discourse analysis (CDA). Teoksessa A. Ryan & A. Wray (toim.), *Evolving Models of Language* (s. 100–116). Clevedon: Multilingual Matters.
- Sweet, E., Kuzawa, C. W. & McDade, T. W. (2018). Short-term lending: Payday loans as risk factors for anxiety, inflammation and poor health. *SSM – Population Health*, 5, 114–121. doi:10.1016/j.ssmph.2018.05.009
- Valkama, E. & Muttilainen, V. (2008). *Maksuvaikeudet pikaluottomarkkinoilla*. Helsinki: Oikeuspoliittinen tutkimuslaitos.
- van Dijk, T. A. (2001). Critical discourse analysis. Teoksessa D. Tannen, D. Schiffrin & H. Hamilton (toim.), *Handbook of Discourse Analysis* (s. 352–371). Oxford: Blackwell.
- Widdowson, H. G. (1998). The theory and practice of critical discourse analysis. *Applied Linguistics*, 19(1), 136–151. doi:10.1093/applin/19.1.136
- Wodak, R. & Meyer, M. (2016). Critical discourse analysis: History, agenda, theory and methodology. Teoksessa R. Wodak & M. Meyer (toim.), *Methods of Critical Discourse Studies* (3. painos, s. 1–22). London: Sage.
- Ylisiurua, M. (2017). Aihemallinnuksen mahdollisuudet sosiaalisen median aineistojen jäsentämisessä – terveyskeskustelu Suomi24-verkkopalstalla. *Kulutustutkimus.Nyt*, 11(2), 44–67.

Kirjoittajatiedot

KTT Ella Lillqvist työskentelee tutkijatohtorina Kuluttajatutkimuskeskuksessa, Helsingin yliopiston valtiotieteellisessä tiedekunnassa. Hänen kiinnostuksenkohteisiinsa kuuluvat kuluttajien varallisuus ja velka, vuorovaikutus digitaalisessa mediassa sekä menetelmistä erityisesti diskurssianalyysi. ella.lillqvist@helsinki.fi

English abstract

Corpus-assisted discourse studies as a method in consumer research: Exploring talk about instant loans on the Suomi24 discussion forum

This study presents corpus-assisted discourse studies (CADS) that draws on the tradition of quantitative linguistics, and contributes thus to recent discussion concerning methods for analysing large quantities of text. The article illustrates CADS techniques relevant to consumer research by analysing how users of the Suomi24 discussion forum have discussed instant loans (similar to payday loans in English-speaking countries) during the first decade of the phenomenon in Finland, 2005–2015. In addition, the study applies and develops various types of data

visualization suitable for CADs. CADs offers versatile tools for consumer research: It helps get an overview of themes present in a corpus for example by examining the lexical characteristics of the target corpus in relation to a reference corpus or, in a longitudinal corpus, by analysing words that become more or less common over time. Moreover, it is possible to observe and visualize the text contexts of certain words that are of interest to the researcher; those words may arise inductively from a general examination of the corpus or deductively from previous research. CADs also enables moving flexibly between quantitative and qualitative analysis. Corpus analysis software can be used for both pre-existing corpora and corpora collected by the researcher. CADs is suitable especially for consumer researchers who are interested in common discourses in large quantities of text, relating to some theme that is manifested in specific vocabulary, such as a product, product category or cultural consumption phenomenon.

Liite: Avainsanoja

* = Kirjoittaja korjannut automaattisen kielenanalyysin virheen (tämän vuoksi sama lemma listassa kaksi kertaa).

	Avainsana	Avainsanaisuusarvo (Simple Maths)	Frekvenssi vippikorpuksessa	Frekvenssi vertailukorpuksessa (FiTenTen2014)
1	laina	4.910	6226	112296
2	maksaa	4.700	8099	653153
3	raha	4.000	6509	602978
4	velka	3.960	4659	86591
5	pankki	2.740	2857	124915
6	ottaa	2.590	7560	2339654
7	euro	2.580	3902	703246
8	lasku	2.490	2495	142199
9	e	2.360	2326	161527
10	korko	2.290	1979	43210
11	ulosotto	2.240	1828	5076
12	jos	1.920	9211	4728091
13	luottotieto	1.890	1319	3328
14	tulo	1.830	1869	413563
15	summa	1.820	1351	92548
16	lainata	1.810	1329	84626
17	vippi	1.780	1148	2833
18	korko*	1.750	1131	21605
19	luotto	1.750	1143	24094
20	maksaa*	1.750	1207	68511
21	saada	1.740	11946	7226347
22	kulu	1.720	1167	74141
23	pois	1.720	2327	859310
24	tili	1.720	1160	63192
25	kk	1.700	1198	111076
26	ei	1.690	32514	21584868
27	maksu	1.680	1188	125577
28	mennä	1.660	5571	3220198
29	perintä	1.660	976	7377
30	takaisin	1.630	1647	512716
31	kuukausi	1.610	1671	549557
32	vuokra	1.600	942	42638
33	nuo	1.590	1773	652626
34	mikään	1.580	3659	2062431
35	sinä	1.580	2576	1254624
36	joutua	1.570	1987	842782
37	kulutusluotto	1.560	834	5452
38	firma	1.530	846	55084
39	niin	1.530	9384	6526602
40	hoitaa	1.520	1386	478578
41	hakea	1.510	1493	576066
42	asunto	1.500	1027	221517
43	kannattaa	1.500	1883	882700
44	vipejä	1.490	719	563
45	myöntää	1.480	1065	284175
46	edes	1.470	1991	1031844
47	palkka	1.470	913	171821
48	enää	1.460	2011	1048134
49	luottokortti	1.460	698	14846

	Avainsana	Avainsanaisuusarvo (Simple Maths)	Frekvenssi vippikorpuksessa	Frekvenssi vertailukorpuksessa (FiTenTen2014)
50	kun	1.450	10482	7846718
51	köyhä	1.450	787	98356
52	työtön	1.430	716	71843
53	pystyä	1.420	1391	636809
54	eli	1.410	2690	1709909
55	ne	1.400	8151	6250986
56	sitten	1.400	5408	3994716
57	vaan	1.400	3638	2510569
58	ole	1.390	2027	1211807
59	vara	1.390	908	283215
60	joku	1.380	2322	1466978
61	talous	1.380	850	247202
62	elämä	1.370	2129	1332762
63	meno	1.370	940	327481
64	tietää	1.370	2425	1584528
65	vipin	1.370	540	1093
66	nyt	1.360	5552	4265640
67	työ	1.360	2455	1643859
68	asia	1.350	4238	3182869
69	elää	1.350	1092	497279
70	eräpäivä	1.350	523	7722
71	vipit	1.350	513	562
72	kukaan	1.340	1220	622405
73	itse	1.330	4934	3852727
74	kallis	1.330	713	192894
75	merkintä	1.330	538	47528
76	mikä	1.330	7837	6410850
77	perintätoimisto	1.330	484	991
78	tonni	1.330	539	42238
79	asuntolaina	1.320	484	11404
80	ostaa	1.320	1396	808843
81	tilanne	1.320	1172	621003
82	vuosikorko	1.320	467	2591
83	kieltää	1.310	703	222923
84	laki	1.310	1103	564327
85	takia	1.310	1004	482242
86	vaikka	1.310	3479	2672498
87	viesti	1.310	771	271638
88	kuu	1.300	710	244527
89	palsta	1.300	487	40679
90	selvitä	1.300	721	244093
91	halpa	1.290	626	183399
92	kierre	1.290	442	16931
93	tyhmä	1.290	514	82664
94	kaikki	1.280	6418	5433749
95	koska	1.280	2600	1992128
96	kyllä	1.280	2823	2194280
97	v	1.280	606	179293
98	apu	1.270	756	329559
99	eikä	1.270	2287	1724942
100	että	1.270	15591	13828312