Sana-assosiaatioverkostojen Analyysi

Emmi Rytkölä

Pro Gradu -tutkielma
Huhtikuu 2024

Tarkastajat:
prof. Kari Auranen
apulaisprof. Janne Kujala

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO

Matematiikan ja tilastotieteen laitos
Emmi Rytkölä: Sana-assosiaatioverkostojen analyysi
Pro gradu -tutkielma, 28 s., 14 liites.
Tilastotiede
Huhtikuu 2024

Tässä pro gradu -tutkielmassa käytetään Turun Yliopston psykologian laitoksen keräämää sana-assosiaatioaineistoa ja pyritään tekemään graafiteoriaan perustuvaa analyysia tästä aineistosta.

Ihmiset oppivat uusia sanoja sen kontekstin kautta, jossa sana esiintyy. Lapsi oppii ensimmäisiä sanoja fyysisen kontekstin kautta, mutta vanhempana hän voi päätellä uusien sanojen merkityksen siitä, minkä tunnettujen sanojen kanssa ne usein esiintyvät. Näin ihminen muodostaa assosiaatioita sanojen välille, ja nämä assosiaatiot muuttuvat iän myötä. Sanojen välisiä assosiaatioita voidaan tutkia graafiteorian avulla.

Graafiteoria on matematiikan osa-alue, jonka avulla tutkitaan asioiden välisiä yhteyksiä piirtämällä yhteyksistä graafeja (tai verkostoja). Sana-assosiaatioaineistosta piirretty graafi on kaksimuotoinen graafi, jonka on projisoitava yksimuotoiseksi graafiksi jotta sen analysointi olisi helpompaa. Tästä graafista voidaan laskea graafitason suureita: asteen keskeisyysmitta, klusterointikerroin ja polun pituus. Klusterointikertoimen ja polun pituuden avulla voidaan myös tehdä arvio siitä, onko graafi pieni maailma -graafi. Lisäksi voidaan laskea sanatason suure, entropia.

Kun graafiteoriaa sovellettiin sana-assosiaatioaineistoon, saatiin näyttöä siitä, että ikääntyneillä aikuisilla on keskivertaisesti vähemmän yhteyksiä sa-

nojen välillä kuin nuorilla aikuisilla ja että molempien ikäryhmien sana-assosiaatioverkostot ovat pieni maailma -graafeja. Pienen koehenkilömäärän takia tarvitaan kuitenkin lisää tutkimusta, jotta voitaisiin varmuudella puhua näiden ikäryhmien välisestä erosta kielen organisoinnissa.

Asiasanat: graafiteoria, verkkoteoria, sana-assosiaatio, kielen organisointi.

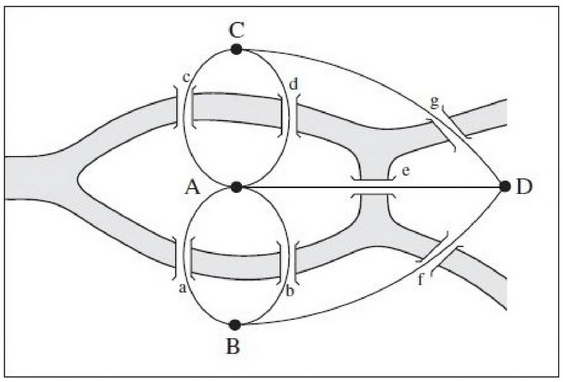
Sisällys

1 Johdanto

Tässä tutkielmassa tarkastellaan sitä, miten ihmiset organisoivat kieltä ja miten tämän organisoinnin voi esittää verkoston eli graafin avulla. Tutkielma alkaa lyhyellä johdannolla sana-assosiaatiotutkimuksiin luvussa 2, jonka jälkeen siirytään graafiteoriaan sekä sen peruskäsitteiden ja tutkielmassa käytettyjen suureiden esittelyyn luvussa 3. Luvussa 4 käytetään näitä peruskäsitteitä ja suureita sana-assosiaatioaineiston analyysiin. Luvussa 5 on pohdintaa aineistosta, sen analyysista ja tuloksista. Tämän tutkielman tavoitteena on esittää graafiteoriaa ja sen sovellusta sana-assosiaatiotutkimuksessa.

Graafiteoria on matematiikan osa-alue, jonka avulla tutkitaan asioiden välisiä yhteyksiä piirtämällä yhteyksistä graafeja [1]. Graafiteorian alku voidaan jäljittää Leonhard Eulerin 1741 kirjoittamaan paperiin, joka pyrki vastaamaan Königsbergin siltaongelmaan. Königsbergin (nykyisin Kaliningrad) jakaa kahteen Pregel-joki, jossa on kaksi saarta. Joen yli kulkee yhteensä seitsemän siltaa. Tuon ajan matemaatikoilta kysyttiin, onko mahdollista kävellä Königsbergin läpi kulkemalla jokaisen sillan yli vain kerran. Matemaatikot väittivät, että tämä kysymys ei kuulu matematiikan alaan, mutta kun Euler tutki asiaa, hän piirsi kuvan ongelmasta käyttäen graafia (kuva 1), ja täten syntyi graafiteorian alku. Hän pystyi graafin avulla todistamaan, että ei ole mahdollista kulkea Königsbergin läpi käyttämällä jokaista siltaa vain kerran. Tämän jälkeen graafiteoriaa on sovellettu moneen eri alaan topologian ulkopuolella, kuten tietotekniikkaan (esimerkiksi internetti), sosiologiaan, ja kielitieteeseen. [2][3]

Tutkielman päälähteenä on käytetty Dubossarskyn ym. artikkelia *Quantifying the Structure of Free Association Networks Across the Life Span* [4], jossa käsitellään sana-assosiaatioita eri ikäisillä hollantilaisilla koehenkilöillä. Vastaavanlaisia tutkimuksia on myös tehty mm. englannin kielellä [5] ja Brasilian portugalin kielellä [6], mutta ei suomen kielellä. Tässä tutkielmassa käytetty aineisto on meneillään olevasta Minna Lehtosen ja Kati Renvallin suomen kieleen liittyvästä tutkimusprojektista.



Kuva 1: Königsbergin sillat; kuva on kirjasta *Linkit : verkostojen uusi teoria* [3].

2 Sana-assosiaatiotutkimuksista

Ihmiset oppivat kieltä kielelle altistumisen kautta eli kuuntelemalla toisten puhetta ja päättelemällä sanojen tarkoituksen kontekstista. Varsinkin varhaislapsuudessa konteksti on suurimmaksi osaksi fyysinen, mutta kun ihminen on jo oppinut joitakin sanoja, hän voi päätellä uusien sanojen merkityksen sen perusteella, minkä tunnettujen sanojen kanssa ne usein esiintyvät samanaikaisesti [7]. Kun lapsi altistuu yhä enemmän kielelle keskusteluun osallistumisen ja myöhemmin lukemisen välityksellä, hänen sanavarastonsa kasvaa nopeasti. Vaikka kasvu on vähäisempää lapsuuden jälkeen, ihmisen sanavarasto kehittyy koko eliniän. Koska ihmiset oppivat uusia sanoja koko elämän läpi, myös sana-assosiaatiot muuttuvat iän myötä [4]. Tällaisen muutoksen tutkimiseen voi käyttää sana-assosiaatiotutkimuksia [4][5][6].

Yleisesti sana-assosiaatiotutkimuksissa käytetään kahta eri tyyppistä sana-assosiaatiotehtävää. Vapaassa assosiaatiotehtävässä (*free association task*) koehenkilölle annetaan jokin ärsykesana ja hänen on nimettävä, mikä sana (tai mitä sanoja) hänelle tulee ensimmäiseksi mieleen. Näitä koehenkilön nimeämiä sanoja kutsutaan kohdesanoiksi. Semanttinen assosiaatiotehtävä (*semantic association task*) on muuten samanlainen, mutta ärsykesanan ja kohdesanan yhteyden on perustuttava ärsykesanan merkitykseen. Tässä tutkielmassa käsitelty aineisto on kerätty käyttäen vapaata assosiaatiotehtä-

vää. Tämä tarkoittaa, että assosiaatio voi olla mikä tahansa suhde sanojen välillä; esimerkiksi niillä voi olla syy-seuraussuhde (*paperi – haava*), ne voivat olla synonyymejä (*puhua – jutella*) tai vastakohtia (*oikea – vasen*), ne voivat usein ilmestyä yhdessä (*taito – luistelu*), tai ne voivat vain kuulostaa/näyttää samalta (*pää – jää*). [6]

Ennen oletettiin, että sanojen välisiä yhteyksiä voidaan kuvata puun avulla. Nyky-ymmärrys kuitenkin on, että sanojen yhteyksiä voidaan kuvata parhaiten graafiteoriaan perustuvien verkkojen eli graafien avulla. Graafin tarkoitus on edustaa, miten sanat kognitiivisesti järjestyvät aivoissamme. Kielen rakenteen tutkimuksen mukaan tämä järjestyminen voi vaikuttaa mm. kognitiivisen toiminnan nopeuteen, esimerkiksi muistiprosesseihin. Kun tutkitaan eri ikäisten assosiaatioverkostoja, voidaan siis paremmin ymmärtää mm. kielen oppimista ja ikääntymiseen liittyvää muistin heikkenemistä. [4][6]

3 Graafiteoriaa

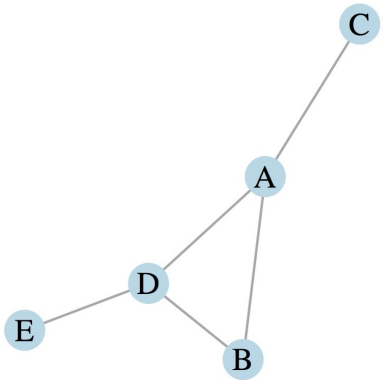
Graafiteoria (tai verkkoteoria) on matematiikan osa-alue, jonka avulla tutkitaan asioiden välisiä yhteyksiä piirtämällä yhteyksistä graafeja eli verkostoja [1]. Tässä tutkielmassa käsitellään graafeja, jotka kuvaavat sanojen välisiä assosiaatioita. Tällaista graafia kuvaavat mitat voidaan jakaa sanatason (*word-level*) ja graafitason (*network-level*) suureisiin. Sanatason suureet lasketaan jokaiselle ärsykesanalle erikseen, ja niiden tarkoitus on tutkia, kuinka annettujen kohdesanojen monipuolisuus muuttuu iän myötä. Graafitason suureet kuvaavat graafin rakennetta ja sitä, miten tämä muuttuu iän myötä. Tässä tutkielmassa ainoa käytetty sanatason suure on entropia (luku 3.8); muut käytetyt suureet (luvut 3.3–3.7) ovat graafitason suureita.

Tässä tutkielmassa käytetty suomenkielinen graafiteorian sanasto perustuu suurimmaksi osaksi Koiviston ja Niemistön graafiteoriaa koskevaan opetusmonisteeseen [8]. Se, miten tässä tutkielmassa sovelletaan graafiteoriaa sana-assosiaatioaineiston analyysiin, perustuu Dubossarskyn ym. artikkeliin [4].

3.1 Graafiteorian peruskäsitteitä

Aloitetaan erilaisten graafien määritelmistä.

Määritelmä 3.1. Yksinkertainen graafi (*simple graph*) G on pari (V, E), jossa $V \neq \emptyset$ on äärellinen joukko ja E on äärellinen joukko järjestämättömiä pareja $\{u, v\}$, $u, v \in V$, $u \neq v$. Joukon V alkioita kutsutaan solmuiksi (*node; vertex*) ja joukon E alkioita särmiksi (*edge*).



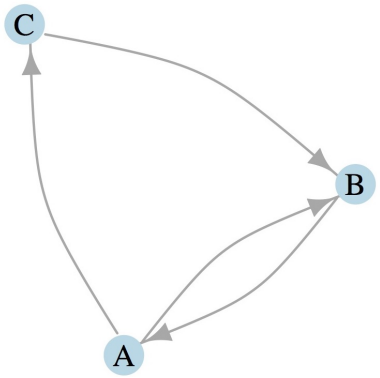
Kuva 2: Yksinkertainen graafi, jossa solmujen joukko $V = \{A, B, C, D, E\}$ ja särmien joukko $E = \{\{A, B\}, \{A, C\}, \{A, D\}, \{B, D\}, \{D, E\}\}$.

Määritelmä 3.2. Suunnattu graafi (*directed network*) G on pari (V, E), jossa $V \neq \emptyset$ on äärellinen joukko ja E on äärellinen joukko järjestettyjä pareja (u, v), $u, v \in V$. Suunnatun graafin särmiä kutsutaan myös kaariksi.

Kuva 2 esittää yksinkertaista graafia, jolla on viisi solmua ja viisi särmää. Kuva 3 esitää suunnattua graafia, jolla on kolme solmua ja neljä kaarta.

Määritelmä 3.3. Painotettu graafi (*weighted network*) on graafin yleistys, jossa jokaiseen särmään on liitetty jokin luku (paino). Merkitään särmän $\{i, j\}$ painoa w_{ij}.

Kuva 4 esittää painotettua graafia, jossa on kolme solmua ja kolme särmää. Graafin voi myös esittää matriisimuodossa; tätä kutsutaan vierusmatriisiksi.



Kuva 3: Suunnattu graafi, jossa solmujen joukko $V = \{A, B, C\}$ ja kaarien joukko $E = \{(A, B), (B, A), (A, C), (C, B)\}$.

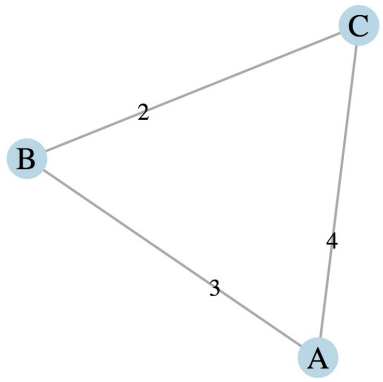
Määritelmä 3.4. Olkoon $G = (V, E)$ n-solmuinen graafi, jossa $V = \{v_1, v_2, \ldots, v_n\}$. Graafin G vierusmatriisi (*adjacency matrix*) A on $n \times n$-matriisi, jonka alkio (i, j) on

$$a_{ij} = \text{särmien lukumäärä solmusta } v_i \text{ solmuun } v_j.$$

Määritelmä 3.4 on yleinen eli se pätee erityyppisille graafeille. Jos G on painottamaton graafi, a_{ij} voi olla vain 0 tai 1. Jos G on painotettu graafi, alkiot a_{ij} ovat särmien painoja w_{ij}. Jos G on suuntaamaton graafi, A on symmetrinen matriisi. Jos G on suunnattu graafi, A ei välttämättä ole symmetrinen, ja tässä tapauksessa A on käytännössä melkein aina epäsymmetrinen. On myös olemassa graafeja, joissa solmusta voi olla särmä tai särmiä itseensä (näitä särmiä kutsutaan luupeiksi). Emme käsittele tällaisia graafeja tässä tutkielmassa, joten tässä tutkielmassa kaikkien vierusmatriisien diagonaalialkiot ovat aina nollia.

Esimerkki 3.5. Kuvan 3 suunnatun graafin $G = \{\{A, B, C\}, \{\{A, B\}, \{B, A\}, \{A, C\}, \{B, C\}\}\}$ vierusmatriisi on

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$



Kuva 4: Painotettu graafi, jossa solmujen joukko $V = \{A, B, C\}$ ja särmien joukko $E = \{\{A, B\}, \{A, C\}, \{B, C\}\}$. Särmien painot ovat $w_{AB} = 3$, $w_{AC} = 4$, ja $w_{BC} = 2$.

Esimerkki 3.6. Kuvan 4 painotetun, suuntaamattoman graafin $G = \{\{A, B, C\}, \{\{A, B\}, \{A, C\}, \{B, C\}\}\}$ vierusmatriisi on

$$A = \begin{pmatrix} 0 & 3 & 4 \\ 3 & 0 & 2 \\ 4 & 2 & 0 \end{pmatrix}.$$

3.2 Kaksimuotoisen graafin projisointi yksimuotoiseksi

Tämän tutkielman tavoitteena on tehdä graafiteoriaan perustuvaa analyysia sana-assosiaatioaineistosta. Tutkielman aineistossa on kahden tyyppisiä sanoja, ärsykesanoja ja kohdesanoja, joten vastaavasti myös aineistosta luodussa graafissa on lähtökohtaisesti kahden tyyppisiä solmuja. Lisäksi aineistosta luodussa graafissa kaikki kaaret kulkevat ärsykesanoista kohdesanoihin (eikä esimerkiksi ärsykesanasta toiseen ärsykesanaan). Tällainen graafi on kaksimuotoinen graafi (*two-mode network; bipartite network*). Suurin osa graafiteoriasta perustuu oletukseen, että graafi on yksimuotoinen graafi (*one-mode network*), jolloin graafissa on vain yhdenlaisia solmuja. Analyysin helpottamiseksi kaksimuotoinen graafi on siis projisoitava yksimuotoiseksi painotetuksi (suunnatuksi) graafiksi. Yksimuotoisessa projektiossa solmut ovat ärsykesanoja. Kaarien painot w_{ij} perustuvat siihen, kuinka monta yhteistä kohdesanaa ärsykesanalla i on ärsykesanan j kanssa, normeerattuna

kohdesanojen yleisyydellä:

$$w_{ij} = \sum_{p=1}^{P} \frac{w_{i,p}}{N_p - 1} \tag{1}$$

jossa summaus on ärsykesanojen i ja j yhteisten kohdesanojen yli ja

$w_{ij} =$ kaaren paino yksimuotoisessa graafissa solmusta v_i solmuun v_j,
$P =$ kuinka monta yhteistä kohdesanaa ärsykesanoilla i ja j on,
$w_{i,p} =$ kuinka monta kertaa ärsykesana i sai ärsykesanan j kanssa yhteisen kohdesanan p,
$N_p =$ kuinka monta ärsykesanaa sai kohdesanan p vastauksena.

On luonteenomaista, että luodussa yksimuotoisessa suunnatussa graafissa kaaren paino w_{ij} ei useinkaan ole sama kuin paino w_{ji}. [4]

Esimerkki 3.7. Käytetään esimerkkinä ärsykesanoja *nainen*, *lapsi*, ja *ahdas*. Oletetaan, että joku koehenkilö antaa ärsykesanalle *nainen* kohdesanat {*mies*, *tyttö*, *poika*}, ärsykesanalle *lapsi* kohdesanat {*pieni*, *tyttö*, *poika*}, ja ärsykesanalle *ahdas* kohdesanat {*ahdistus*, *pieni*, *kapea*}. Tällöin yksimuotoisessa graafissa voidaan piirtää kaari sanasta *nainen* sanaan *lapsi*, sanasta *lapsi* sanaan *nainen*, sanasta *lapsi* sanaan *ahdas*, sekä sanasta *ahdas* sanaan *lapsi*.

Oletetaan, että yhteensä koko aineistossa 3 koehenkilöä antoi ärsykesanalle *nainen* kohdesanat {*tyttö*, *poika*}, 5 koehenkilöä antoi ärsykesanalle *lapsi* kohdesanan *pieni*, 3 antoi kohdesanan *tyttö* ja 4 kohdesanan *poika*, sekä 2 koehenkilöä antoi ärsykesanalle *ahdas* kohdesanan *pieni*. Oletetaan myös, että joku koehenkilö antoi myös ärsykesanoille *kevyt* ja *iso* kohdesanan *pieni*.

Tällöin yksimuotoisen graafin kaarille saadaan painot:

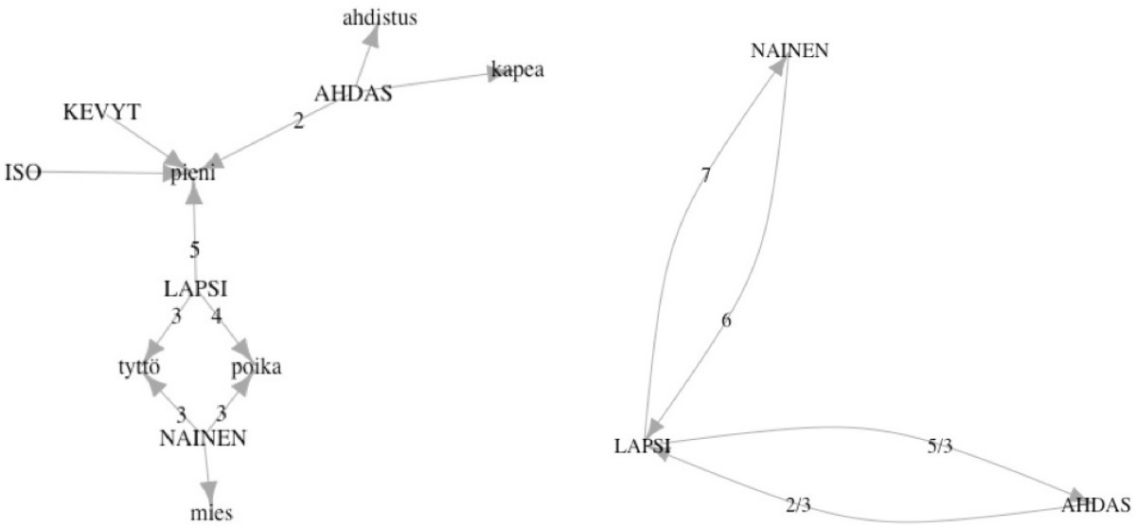
$$w_{12} = \sum_{p=1}^{2} \frac{w_{1,p}}{N_p - 1} = \frac{3}{2-1} + \frac{3}{2-1} = 6 \text{ sanasta } \textit{nainen} \text{ sanaan } \textit{lapsi},$$

$$w_{21} = \sum_{p=1}^{2} \frac{w_{2,p}}{N_p - 1} = \frac{3}{2-1} + \frac{4}{2-1} = 7 \text{ sanasta } \textit{lapsi} \text{ sanaan } \textit{nainen},$$

$$w_{23} = \sum_{p=1}^{1} \frac{w_{2,p}}{N_p - 1} = \frac{5}{4-1} = \frac{5}{3} \text{ sanasta } \textit{lapsi} \text{ sanaan } \textit{ahdas}, \text{ ja}$$

$$w_{32} = \sum_{p=1}^{1} \frac{w_{3,p}}{N_p - 1} = \frac{2}{4-1} = \frac{2}{3} \text{ sanasta } \textit{ahdas} \text{ sanaan } \textit{lapsi}.$$

Esimerkin visualisoi kuva 5.



Kuva 5: Vasemmalla on tekstissä esitetyn esimerkin mukainen kaksimuotoinen graafi ja oikealla tästä luotu yksimuotoinen, painotettu suuntaamaton graafi.

On huomioitava, että toisin kuin kaksimuotoinen graafi, luotu yksimuotoinen graafi ei edusta sanojen kognitiivista järjestymistä. Sen sijaan yksimuotoinen graafi kuvaa sana-assosiaatioaineiston rakenteellisia ominaisuuksia. [4]

Seuraavaksi määritellään solmun aste, solmun voima, asteen keskeisyysmitta, klusterointikerroin, ja polun pituus. Näiden laskemisessa käytetään projisoitua yksimuotoista graafia.

3.3 Aste, voimakkuus ja asteen keskeisyysmitta

Määritelmä 3.8. Solmun lähtöaste (*out-degree*) k^{out} on niiden kaarien lukumäärä, jotka lähtevät solmusta. Solmun tuloaste (*in-degree*) k^{in} on niiden kaarien lukumäärä, jotka saapuvat solmuun. Lähtö- ja tuloasteet ovat siis

$$k_i^{out} = \sum_{j=1}^{n} x_{ij}, \qquad k_j^{in} = \sum_{i=1}^{n} x_{ij},$$

jossa $x_{ij} = 1$, jos on olemassa kaari joka kulkee solmusta v_i solmuun v_j, ja $x_{ij} = 0$ muuten; n on solmujen lukumäärä [9]. Solmun aste k on niiden kaarien lukumäärä, jotka ovat kytkeytyneet solmuun [10].

Solmun aste siis kuvaa vain sitä, kuinka monta kaarta on kytkeytynyt solmuun. Kun kyseessä on painotettu graafi, solmun kytkeytyvyyttä voidaan kuvata myös kaarien painojen summana.

Määritelmä 3.9. Solmun voimakkuus (*node strength*) on solmuun kytkeytyneiden kaarien painojen summa. Kun on kyseessä suunnattu graafi, voidaan laskea lähtö- ja tulovoimakkuudet s^{out} ja s^{in}:

$$s_i^{out} = \sum_{j=1}^{n} w_{ij}, \qquad s_j^{in} = \sum_{i=1}^{n} w_{ij},$$

jossa w_{ij} ovat painotetun suunnatun graafin vierusmatriisin alkioita; toisin sanoen, w_{ij} on kaaren paino solmusta v_i solmuun v_j.

Koska solmun aste sekä voimakkuus ovat tärkeitä indikaattoreita siitä, kuinka kytkeytynyt solmu on, on kehitetty kaava, joka ottaa huomioon molemmat. Tätä menetelmää kutsutaan Opsahlin menetelmäksi [10].

Määritelmä 3.10. Asteen keskeisyysmitta $C_D^{w\alpha}(i)$ (*degree centrality measure*) on solmun asteen ja voimakkuuden tulo, jota säädetään viritysparametrilla:

$$C_D^{nw\alpha}(i) = k_i^{1-\alpha} \times s_i^{\alpha},$$

jossa

$$k_i = \text{solmun } i \text{ aste},$$
$$s_i = \text{solmun } i \text{ voimakkuus},$$
$$\alpha = \text{positiivinen viritysparametri}.$$

Voidaan myös laskea lähtö- ja tuloasteiden keskeisyysmitat samalla tavalla eli

$$C^{w\alpha}_{D-out}(i) = (k_i^{out})^{1-\alpha} \times (s_i^{out})^{\alpha} \qquad \text{ja}$$
$$C^{w\alpha}_{D-in}(i) = (k_i^{in})^{1-\alpha} \times (s_i^{in})^{\alpha}.$$

Viritysparametrin α valinta riippuu siitä, halutaanko painottaa kaarien määrää (solmun astetta) vai kaarien painoja (solmun voimakkuutta), vai halutaanko ottaa molemmat huomioon tasapuolisesti. Jos valitaan $\alpha = 0$, menetelmä jättää huomioimatta solmun voimakkuuden. Jos valitaan $\alpha = 1$, menetelmä jättää huomioimatta solmun asteen. Kun $\alpha = 0.5$, aste ja voimakkuus otetaan tasavertaisesti huomioon. Jos valitaan $\alpha > 1$, solmun aste pienentää asteen keskeisyysmittaa. [10][11]

Dubossarsky ym. [4] kutsuvat mittaa $C_D^{w\alpha}(i)$ asteeksi, mutta tässä tutkielmassa käytetään kuitenkin Opsahlin käyttämää termiä *asteen keskeisyysmitta*.

Asteen keskeisyysmitta lasketaan jokaiselle solmulle (eli ärsykesanalle) erikseen. Tämän jälkeen voidaan laskea koko graafin keskimääräinen asteen keskeisyysmitta summaamalla kaikkien solmujen asteen keskeisyysmitat ja jakamalla solmujen lukumäärällä.

3.4 Klusterointikerroin

Klusterointikerroin c_i mittaa solmun yhteiskytkeytyvyyttä solmun naapureiden kanssa. Se mittaa sitä, kuinka paljon painoista, jotka lähtevät solmusta i sen naapureihin, jää paikalliseen naapurustoon. [4]

Määritelmä 3.11. Solmun i painotettu klusterointikerroin (*weighted clus-*

tering coefficient) on:

$$c_i = \frac{1}{s_i^{out}(k_i - 1)} \sum_{\substack{1 \leq j,\, l \leq n \\ j,\, l \neq i}} \frac{w_{ij} + w_{il}}{2} a_{ij} a_{il} a_{jl},$$

jossa

k_i = solmun i aste,
s_i^{out} = solmun i lähtövoimakkuus,
w_{ij} = kaaren (i, j) paino,
$A = (a_{ij})$ = graafin painoton vierusmatriisi, jonka alkiot $a_{ij} = 1$ jos $w_{ij} \geq 0$ ja $a_{ij} = 0$ muuten.

On huomioitava, että $a_{ij}a_{il}a_{jl} = 1$ jos ja vain jos solmut i, j ja l muodostavat kolmion graafissa. [12][13]

Kun graafi on luotu oikeasta aineistosta, klusterointikerroin käyttäytyy epätriviaalisesti seuraten potenssilakia asteen k funktiona [13]. Keskimääräinen klusterointikerroin C saadaan laskemalla kaikkien solmujen klusterointikertoimien keskiarvo [14].

3.5 Polun pituus

Määritelmä 3.12. Polun pituus (*shortest path*) l kahden solmun välillä on lyhyimmän mahdollisen polun pituus, kun kuljetaan kaaria pitkin. [4]

Kun kyseessä on graafi, joka ei ole painotettu, polun pituus solmusta i solmuun j voidaan laskea seuraavasti:

$$d(i, j) = \min(x_{ih_1} + \cdots + x_{h_{l-1}j}),$$

jossa h_1, ..., h_{l-1} ovat välisolmuja solmun i ja j välisillä poluilla ja x_{ih_1}, ..., $x_{h_{l-1}j}$ ovat kaikki 1. Tätä voidaan kutsua binääriseksi polun pituudeksi (*binary shortest distance*). [10]

Kun kyseessä on painotettu graafi, myös painot pitää ottaa huomioon siten, että mitä suurempi paino, sitä lyhyempi polku on. Polun pituus painotetussa graafissa voidaan määritellään seuraavasti:

$$d^w(i,j) = \min(\frac{1}{w_{ih_1}} + \cdots + \frac{1}{w_{h_{l-1}j}}),$$

jossa h_1, ..., h_{l-1} ovat välisolmuja solmun i ja j välisillä poluilla. [10]

Tässä tutkielmassa käytetään kuitenkin normalisoituja kaarten painoja, jakamalla kaarien painot graafin painojen keskiarvolla [4]. Polun pituus voidaan täten määritellä

$$d^w(i,j) = \min(\frac{\bar{w}}{w_{ih_1}} + \cdots + \frac{\bar{w}}{w_{h_{l-1}j}}),$$

jossa $\bar{w}$ on graafin painojen keskiarvo.

Keskimääräinen polun pituus L saadaan laskemalla kaikkien solmuparien polun pituuksien keskiarvo [14]. Tämä luku ottaa huomioon koko graafin.

Voidaan myös laskea lokaali polkujen pituus (*local shortest path*) l_i jokaiselle solmulle i erikseen, laskemalla keskiarvo niiden polkujen pituuksista, jotka lähtevät solmusta i [15]. Tämä luku on indikaattori siitä, kuinka hyvin solmu on yhteydessä mihin tahansa muuhun solmuun graafissa.

3.6 Satunnainen graafi

Satunnaisen graafin voi generoida käyttämällä Erdös-Rényin satunnaisen graafin mallia. Erdös-Rényi mallista on kaksi eri versiota. Tässä tutkielmassa käytetään mallin versiota $\gamma_{n,M}$, jossa graafi valitaan satunnaisesti kaikista mahdollisista graafeista, joilla on n solmua ja M särmää. Toinen vaihtoehto olisi ottaa huomioon särmän olemassa olemisen todennäköisyys särmien määrän sijaan. [16]

Määritelmä 3.13. Olkoon $E_{n,M}$ kaikkien graafien joukko, joilla on n solmua ja M särmää. Satunnainen graafi $\gamma_{n,M}$ on joukosta $E_{n,M}$ satunnaisesti valittu alkio. Jokaisella joukon $E_{n,M}$ alkiolla on sama todennäköisyys tulla valituksi $(1/\binom{\binom{n}{2}}{M})$. [16]

Erdös-Renyin satunnaisesta graafista $\gamma_{n,M}$ laskettu aste k_{rand} seuraa binomijakaumaa Bin(n, p), jossa $p = 1/\binom{n}{2}_M$. Sen keskimääräinen aste voidaan laskea $K_{rand} = p(n-1)$. Graafin keskimääräistä astetta käyttämällä voidaan estimoida sen keskimääräinen klusterikerroin ja polun pituus:

$$C_{rand} = p \approx \frac{K_{rand}}{n}, \quad \text{ja}$$
$$L_{rand} \approx \log_{K_{rand}}(n) = \frac{\log(n)}{\log(p(n-1))}.$$

Kun solmujen määrä n kasvaa, Erdös-Renyin satunnaisen graafin keskimääräinen klusterointikerroin lähestyy nollaa, ja sen keskimääräinen polun pituus lähestyy jotain vakiota. [19]

Erdös-Rényin satunnainen graafi voi olla suunnattu, mutta se on aina painottamaton. On myös esitetty tapa generoida painotettu satunnainen graafi vastaavanlaisesti [18], mutta tässä tutkielmassa käytetään vain Erdös-Rényin graafia.

3.7 Pieni maailma

Pieni maailma -graafi (*small-world newtwork*) on verkosto, jolla on tiiviisti toisiinsa liittyvät solmuklusterit ja jonka polkujen pituudet ovat keskimäärin lyhyitä. Tällaisen graafin kriteerinä on pidetty, että sen keskimääräinen polun pituus L on lähes sama kuin vastaavanlaisella satunnaisella gaafilla ja sen keskimääräinen klusterointikerroin C on suurempi kuin vastaavanlaisella satunnaisella graafilla. Voidaan siis käyttää kriteerejä $L \geq L_{rand}$ ja $C > C_{rand}$. Se, onko kyseessä pieni maailma -graafi, voidaan myös määrittää käyttämällä pieni maailma -indeksiä. Verkostoa voidaan kutsua pieni maailma -graafiksi jos $SWI > 1$. [14]

Määritelmä 3.14. Pieni maailma -indeksi SWI (*small-world index*) saadaan normalisoimalla keskimääräinen klusterikerroin C ja keskimääräinen polun pituus L jakamalla nämä vastaavan satunnaisen graafin samoilla mitoilla ja jakamalla normalisoitu klusterikerroin normalisoidulla keskimääräisellä polun pituudella, eli

$$SWI = \frac{C}{C_{rand}} / \frac{L}{L_{rand}},$$

jossa C ja L ovat keskimääräinen klusterikerroin ja keskimääräinen polun pituus, ja C_{rand} ja L_{rand} ovat samat mitat sellaisessa Erdös-Rényin satunnaisessa graafissa, jolla on sama koko ja tiheys. [4]

3.8 Entropia

Entropialla lasketaan, mikä osuus vastauksista tietylle ärsykesanalle on sama kohdesana. Entropia kuvaa siis vastauksien monipuolisuutta; entropia on pieni sellaisille ärsykesanoille, joille moni koehenkilö antoi samoja vastauksia, ja suuri ärsykesanoille, jotka saivat paljon erilaisia vastauksia. Toisin kuin asteen keskeisyysmitan, klusterointikertoimen ja polun pituuden laskelmissa, joissa käytetään projisoitua yksimuotoista graafia, entropian laskennassa käytetään alkuperäistä kaksimuotoista graafia.

Jokaiselle ärsykesanalle lasketaan erikseen vastausvektori $(p(x_1), \ldots, p(x_n))$, jossa n on vastaustyyppien (erilaisten kohdesanojen) lukumäärä ja $p(x_k)$ on kohdesanan x_k lukumäärä jaettuna ärsykesanan kaikkien kohdesanojen lukumäärällä.

Määritelmä 3.15. Ärsykesanan i normalisoitu entropia (*normalised entropy; metric entropy*) h_i on sen vastausvektorin keskiarvo, eli

$$h_i = -\sum_{k=1}^{n_i} \frac{p(x_k)\log(p(x_k))}{\log(n_i)},$$

jossa

$n_i =$ ärsykesanalle i annettujen ainutlaatuisten kohdesanojen lukumäärä,

$p(x_k) =$ kohdesanan x_k osuus kaikista kohdesanoista, jotka annettiin vastauksena ärsykesanalle k.

Keskimääräinen entropia H saadaan laskemalla kaikkien ärsykesanojen entropioiden keskiarvo. [4]

3.9 Tilastolliset menetelmät

Tilastollisena menetelmänä tässä tutkielmassa käytetään monen muuttujan varianssianalyysiä (MANOVA; *multivariate analysis of variance*), jossa se-

littävä muuttuja on ikäryhmä ja selitettävät muuttujat ovat tuloasteen keskeisyysmitta, lähtöasteen keskeisyysmitta, klusterointikerroin ja lokaali polkujen pituus. Lisäksi käytetään yksisuuntaista varianssianalyysia (ANOVA; *analysis of variance*) jokaiselle suurelle erikseen, kun selittävä muuttuja on ikäryhmä ja selitettävä muuttuja on entropia, tuloasteen keskeisyysmitta, lähtöasteen keskeisyysmitta, klusterointikerroin tai lokaali polkujen pituus. Jos suureet eivät seuraa normaalijakaumaa, käytetään myös epäparametrista Mann–Whitneyn U-testiä eli Wilcoxonin järjestyssummatestiä. Normaalisuutta testattiin Shapiro–Wilk -testillä.

3.10 Analyysiohjelmistot

Aineiston analyysissä käytin erityisesti R Studion paketteja *tnet* ja *igraph*. Klusterointikerroimet laskettiin funktiolla *clustering_w* (kun kyseessä on painotettu graafi) paketissa *tnet* ja funktiolla *ClustBCG* paketissa *igraph*. Aste ja voima laskettiin funktiolla *degree_w* ja polun pituudet funktiolla *distance_w* (kun kyseessä on painotettu graafi) paketissa *tnet*. Paketissa *igraph* on funktio *erdos.renyi.game*, joka generoi Erdös-Rényin satunnaisen graafin. Yksisuuntainen varianssianalyysi tehtiin käyttämällä funktiota *lm* ja monen muuttujan varianssianalyysi käyttämällä funktiota *manova* paketissa *stats*. Käytetty R-koodi on liitteessä A.

4 Aineiston analyysi

4.1 Aineisto

Turun yliopiston psykologian laitoksen sana-assosiaatiotutkimukseen rekrytoitiin suomea äidinkielenä puhuvia 18–40-vuotiaita nuoria aikuisia ja yli 60-vuotiaita ikääntyneitä aikuisia, joilla ei ollut diagnosoitua kognitiivista heikentymää tai neuropsykologista häiriötä. Molemmilla ikäryhmillä käytettiin 270 sanan sanalistaa (liite B). Tämä sanalista jaettiin 24 suppeampaan sanalistaan, joissa on 45 sanaa per lista. Jokaiselle koehenkilölle arvottiin satunnaisesti yksi suppeammasta sanalistasta; tämän satunnaisuuden takia

ikäryhmä	koehenkilöitä	R1	R2	R3	yhteensä	ainutlaatuisia	suhde
18–40	62	2786	2734	2689	8209	3373	0.411
60–100	52	2335	2292	2191	6818	2574	0.378

Taulukko 1: Vastausten lukumäärät sana-assosiaatiotutkimuksessa. Sarakkeissa ”R1”, ”R2”, ja ”R3” ovat ensimmäisten, toisten, ja kolmansien vastauksien lukumäärät; ”yhteensä” on vastauksien määrä yhteensä; ”ainutlaatuisia” on sellaisten vastausten lukumäärä, joissa annettu kohdesana esiintyi koko aineistossa vain kerran; ”suhde” on ainutlaatuisten vastausten lukumäärän suhde kaikken vastausten lukumäärään.

joillekin sanoille saatiin vähemmän vastauksia kuin toisille. Koehenkilön pyydettiin antamaan kolme assosiaatiota jokaiselle ärsykesanalle. Yhdelle ärsykesanalle annettujen vastausten määrän vaihteluväli oli 9–51.

Aineisto kerättiin käyttämällä vapaata assosiaatiotehtävää. Sana-assosiaatiotehtävä suoritettiin tietokoneella, internetpohjaisella Soile-alustalla. Ärsykesanat esitettiin kirjallisesti ja tehtävään vastattiin kirjoitamalla.

Koehenkilöiden antamat assosiaatiot eli kohdesanat on muokattu korjaamalla mahdollisimman paljon kirjoitusvirheitä (esim. *nennyttä* = *mennyttä*) ja poistamalla vastauksia, jotka eivät ole sanoja (esim. jos on vahingossa kirjoittanut vain yhden kirjaimen sanan sijaan). Lisäksi kaikki sanat, jotka oli annettu monikossa, muunnettiin niiden yksikkömuotoon (esim. *marjat* = *marja*), ja kaikki taivutetut sanat on pyritty muuntamaan niiden perusmuotoon (esim. *erilaista* = *erilainen*). Vastauksien kirjainkoot muunnettiin pieniksi kirjaimiksi ja varmistettu, että jokainen vastaus on vain yksi vastaus (ei esimerkiksi kaikki kolme vastausta kirjoitettu yhdeksi vastaukseksi). Taulukko 1 esittää vastausten lukumäärät kahdessa ikäryhmässä.

Tutkimukseen osallistui 62 nuorta aikuista (18–40 vuotiaita) ja 52 ikääntynyttä aikuista (60–100 vuotiaita). Yhteensä vastauksia oli 8209 nuorelta aikuiselta ja 6818 ikääntyneeltä.

4.2 Tutkimuksen voimasta

Tässä luvussa lasketaan, kuinka suureen tilastolliseen voimaan tämän työn tutkimuksen koko riittää, kun lähtökohtana pidetään Dubossarskyn artikkelin [4] tietoja eri ikäryhmien välisistä eroista. Tarkastellaan esimerkin vuoksi klusterointikerrointa.

Dubossarskyn ym. artikkelin [4] kuvasta 4 voi arvioida klusterointikertoimelle ikäluokittaiset keskiarvot $\bar{x}_{60v} = 0.1025$, $\bar{x}_{70v} = 0.1045$, $\bar{x}_{18v} = 0.1160$, $\bar{x}_{30v} = 0.1125$, ja $\bar{x}_{40v} = 0.1095$, sekä keskiarvon keskivirheen $\sigma/\sqrt{n_D} = 0.0028$, jossa σ on klusterikertoimen hajonta ja n_D on yhdessä ikäryhmässä käytettyjen ärsykesanojen lukumäärä.

Dubosarskyn ym. artikkelin tietojen perusteella ikääntyneiden (A) ja nuorten (B) aikuisten sanaverkostojen klusterointikertoimet ja niiden hajonta arvioitiin seuraavasti:

$$\bar{x}_A = (\bar{x}_{60v} + \bar{x}_{70v})/2 = (0.1025 + 0.1045)/2 = 0.1035,$$
$$\bar{x}_B = (\bar{x}_{18v} + \bar{x}_{30v} + \bar{x}_{40v})/3 = (0.1160 + 0.1125 + 0.1095)/3 \approx 0.1127,$$
$$\sigma = 0.0028 * \sqrt{n_D} = 0.0028 * \sqrt{420} \approx 0.0574,$$

jossa $\bar{x}_A$ on keskimääräinen klusterointikerroin ikääntyneiden aikuisten ryhmässä, $\bar{x}_B$ on keskimääräinen klusterointikerroin nuorten aikuisten ryhmässä, ja σ on klusterikertoimen hajonta. Oletetaan, että klusterointikertoimet noudattavat normaalijakaumaa yllä esitetyin parametrein.

Lasketaan tutkimuksemme voima käyttäen kaavaa [17]

$$\delta_D = (q_{1-\alpha/2} + q_{1-\beta})\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}},$$

jossa

$\delta_D = |x_A - x_B| \approx 0.01 =$ ero eri ikäryhmien klusterointikertoimissa Dubossarskyn ym. artikkelissa,

$n_A = 270 =$ käytettyjen ärsykesanojen lukumäärä ikääntyneiden aikuisten ryhmässä Turun yliopiston tutkimusprojektissa,

$n_B = 270 =$ käytettyjen ärsykesanojen lukumäärä nuorien aikuisten ryhmässä Turun yliopiston tutkimusprojektissa,

$\alpha = 5\% = 0.05 =$ tyypin 1 virhe,

$q_{1-\alpha/2} = 1.96 =$ standardinormaalijakauman $1 - \alpha/2$ kvantiili,

$1 - \beta =$ voimakkuus,

$q_{1-\beta} =$ standardinormaalijakauman $1 - \beta$ kvantiili.

Tästä saadaan

$$0.01 = (1.96 + q_{1-\beta})\sqrt{\frac{0.0574^2}{270} + \frac{0.0574^2}{270}}$$
$$\implies q_{1-\beta} = \frac{0.01}{\sqrt{\frac{0.0574^2}{270} + \frac{0.0574^2}{270}}} - 1.96 \approx 0.06.$$

Koska oletetaan, että aineistomme noudattaa normaalijakaumaa ja tiedetään, että standardinormaalijakauman kvantiili $q_{0.5239} = 0.06$, saadaan tutkimuksen voimaksi klusterointikertoimen osalta

$$1 - \beta = 0.5239 \approx 52\%.$$

4.3 Entropia

Taulukkoon 2 on kirjattu entropian keskiarvot, keskiarvojen keskivirheet, sekä 95 %:n luottamusvälit molemmissa ikäryhmissä. Ikäryhmien välillä ei ollut merkittävää eroa. Entropia on aina välillä $[0, 1]$, eli saamamme entropia on suuri molemmissa ikärymissä. Yksisuuntainen varianssianalyysi sekä Mann–Whitney U-testi osoittivat, että ikäryhmät eivät eronneet entropian suhteen ($p = 0.53$ ja $p = 0.88$).

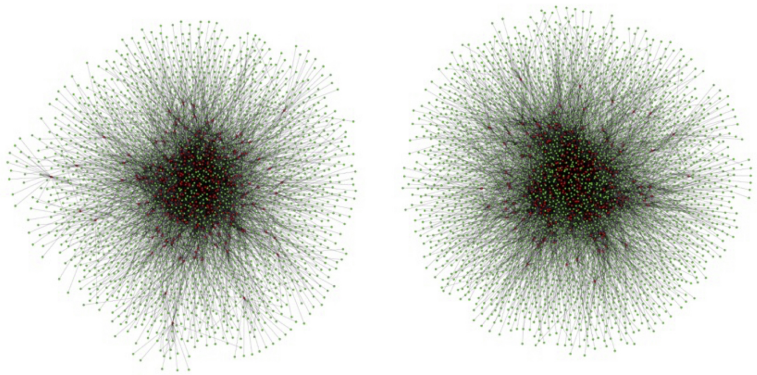
	ikääntyneet aikuiset	nuoret aikuiset
keskiarvo	0.954	0.956
keskiarvon keskivirhe	0.002	0.002
95 %:n luottamusväli	[0.950, 0.959]	[0.953, 0.960]

Taulukko 2: Entropian (H) keskiarvo, keskiarvon keskivirhe ja 95 %:n luottamusväli.

4.4 Graafianalyysi

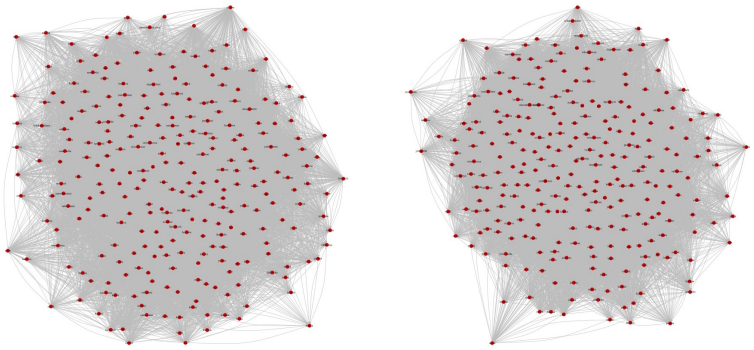
Kuva 6 esittää aineistosta tehtyjä graafeja ilman projisointia. Näiden kaksimuotoisten graafien rakentamisen jälkeen graafit projisoitiin yksimuotoisiksi luvussa 3.2 esitetyn menettelyn mukaisesti. Nämä yksimuotoiset graafit on visualisoitu kuvassa 7. Loput analyysistä käsittelee tätä projisoitua graafia.

Verrattuna nuorten aikuisten graafiin ikääntyneiden graafi (kuva 7) on hieman laajempi ja sen ulkoreunoilla näyttäisi olevan enemmän solmuja, jotka eivät ole vahvasti yhteyksissä toisiin solmuihin.



Kuva 6: Visualisaatio kaksimuotoisista graafeista; vasemmalla ikääntyneet ja oikealla nuoret aikuiset.

Taulukkoon 3 on kirjattu lähtöasteen keskeisyysmitan, tuloasteen keskeisyysmitan, klusterointikertoimen ja polun pituuden keskiarvot ja keskiarvojen keskivirheet molemmissa ikäryhmissä, sekä näiden suureiden 95%:n luottamusvälit. Sekä lähtö- että tuloasteen keskeisyysmitat olivat merkittävästi



Kuva 7: Visualisaatio projisoiduista yksimuotoisista graafeista; vasemmalla ikääntyneet ja oikealla nuoret aikuiset.

suuremmat nuorten aikuisten ryhmässä. Tämä viittaa siihen, että ikääntyneiden graafin solmuilla oli keskivertaisesti vähemmän yhteyksiä kuin nuorten aikuisten graafin solmuilla. Myös polun pituudet ovat merkittävästi pidempiä ikääntyneiden aikuisten graafissa kuin nuorten aikuisten, mikä myös viittaa siihen, että ikääntyneiden graafin solmut eivät ole yhtä hyvin yhteyksissä toisiinsa. Klusterointikertoimella ei ollut merkittävää eroa eri ikäryhmien välillä.

Yksisuuntaisen varianssianalyysiin perustuvat päätelmät olivat luonnollisesti samanlaiset kuin luottamusvälien estimointiin perustuvat. Ikäryhmät poikkesivat toisistaan tilastollisesti merkitsevästi tuloasteen ja lähtöasteen keskeisyysmitan sekä polun pituuden suhteen (kaikissa $p \leq 0.0001$). Klusterointikertoimilla ei ollut eroa ($p = 0.58$). Monen muuttujan varianssianalyysi johti samalaisiin päätelmiin ($p < 0.0001$ kun malliin otettiin mukaan klusterointikerroin, sekä silloin kun se jätettiin pois mallista). Varianssianalyysi kuitenkin olettaa vastemuuttujan normaalisuuden. Koska sekä Shapiro–Wilkin normaalisuustesti että residuaalien tarkastelu (liite C) viittasivat siihen, että normaalisuusoletus ei toteudu, tehtiin myös epäparametrinen Mann–Whitney U-testi näille parametreille. Myös tämä testi viittasi siihen, että eri ikäryhmillä erosivat merkittävästi toisistaan tulo-asteen ($W = 44969$, $p < 0.0001$), lähtö-asteen ($W = 43888$, $p = 0.0001$), ja polun pituuden ($W = 25002$, $p < 0.0001$) suhteen. Ikäryhmät eivät eronneet merkittävästi

klusterointikertoimen suhteen ($W = 38436$, $p = 0.27$).

	ikääntyneet aikuiset			nuoret aikuiset		
	keskiarvo	keskiarvon keskivirhe	95 %:n luottamusväli	keskiarvo	keskiarvon keskivirhe	95 %:n luottamusväli
lähtöasteen keskeisyysmitta ($C^{w\alpha}_{D-out}$)	30.03	0.61	[28.84, 31.22]	33.37	0.68	[32.04, 34.69]
tuloasteen keskeisyysmitta ($C^{w\alpha}_{D-in}$)	30.22	0.62	[29.01, 31.43]	33.46	0.65	[32.18, 34.73]
klusterointikerroin (C)	0.3597	0.0046	[0.3507, 0.3587]	0.3631	0.0041	[0.3550, 0.3712]
polun pituus (L)	0.7852	0.0065	[0.7725, 0.7979]	0.7302	0.0057	[0.7191, 0.7413]

Taulukko 3: Keskeisyysmittojen, klusterointikertoimen ja polunpituuden keskiarvot, keskiarvojen keskivirheet ja 95 %:n luottamusvälit ikääntyneiden aikuisten ($n = 52$) ja nuorten aikuisten ($n = 62$) ikäryhmissä.

4.4.1 Pieni maailma

Nuorten aikuisten graafin pieni maailma -indeksiksi saatiin 3.81 ja ikääntyneiden aikuisten graafille 3.87. Molempien ikäryhmien graafit ovat siis pieni maailma -graafeja koska niiden pieni maailma -indeksit ovat suurempia kuin yksi, eivätkä ne eroa paljon toisistaan tältä osin. Se, että graafien pieni maailma -indeksien välillä ei ollut eroa oli oletettavissa, koska ikäryhmien klusterointikertoimissa ei ollut merkittävää eroa.

5 Pohdinta

Tässä tutkielmassa tutkittiin kahden eri ikäryhmän (18–40-vuotiaiden ja yli 60-vuotiaiden) sana-assosiaatioverkostoja ja niiden välisiä eroja, jotta pystyisimme ymmärtämään, miten ikä vaikuttaa kielen organisointiin. Tulokset viittaavat siihen, että ikääntyneillä aikuisilla on keskivertaisesti vähemmän yhteyksiä sanojen välillä kuin nuorilla aikuisilla. Saimme myös näyttöä sii-

tä, että molempien ikäryhmien sana-assosiaatioverkostot ovat pieni maailma -graafeja eli että ne eroavat merkittävästi satunnaisesta graafista.

Dubossarskyn ym. artikkelissa [4] käytettiin samoja menetelmiä kuin tässä tutkielmassa. Suurimmat erot tutkimusten välillä olivat, että hollantilaisessa tutkimuksessa oli enemmän ikäryhmiä sekä koehenkilöitä ja ärsykesanoja per ikäryhmä ja heidän tutkimuksensa tehtiin hollannin kielellä, kuin tässä tutkielmassa käsitelty tutkimus tehtiin suomen kielellä. Suurimmaksi osaksi kuitenkin saatiin samanlaisia tuloksia kuin Dubossarskyn ym. artikkelissa.

Tuloasteen sekä lähtöasteen keskeisyysmitta olivat suurempia nuorten aikuisten ryhmässä kuin ikääntyneiden. Dubossarsky ym. saivat samantyyppisen tuloksen; keskeisyysmitat ovat pieniä lapsuudessa, suurimpia nuorena aikuisena, ja pienenevät vanhemmiten. Tässä tutkielmassa lasketut keskeisyysmitat olivat kuitenkin ylipäätänsä suurempia kuin Dubossarskyn ym. keskeisyysmitat, eli tässä tutkielmassa käsiteltyjen graafien solmuilla oli keskimääräisesti enemmän yhteyksiä. Toisin sanoen, ärsykesanoille annettiin (suhteellisesti) enemmän samoja kohdesanoja tässä tutkimuksessa kuin Dubossarskyn ym. tutkimuksessa. On mahdollista, että tämä tulos johtuu siitä, että Dubossarskyn tutkimuksessa käytettiin yli 400 ärsykesanaa ja tässä tutkielmassa käsitelty tutkimus käytti vain 270 ärsykesanaa. On myös mahdollista, että ero johtuu käytetystä kielestä tai valituista ärsykesanoista.

Polun pituudet olivat lyhyempiä nuorten aikuisten ryhmässä kuin ikääntyneiden, mikä myös viittaa siihen, että nuorten aikuisten graafin solmut ovat kytkeytyneempiä toisiinsa kuin ikääntyneiden aikuisten graafin solmut. Dubossarsky ym. saivat samantyyppisen tuloksen; polun pituudet olivat pitkiä lapsena, lyhyimpiä nuorena aikuisena, ja pidempiä ikääntyessä. Dubossarskyn ym. polun pituudet olivat pidempiä kuin meidän tutkimuksessamme. Tämä mahdollisesti johtuu vain siitä, että tässä tutkielmassa käsitellyt graafit olivat pienempiä (vähemmän ärsykesanoja) verrattuna Dubossarskyn ym. graafeihin.

Tässä tutkielmassa ei löydetty eroja klusterikertoimessa kahden ikäluokan välillä. Dubossarskyn ym. artikkeli sen sijaan kuvasi kuvattiin suuri vähenemä klusterikertoimessa iän myötä. Tätä olisi hyvä tutkia uudelleen suuremmalla koehenkilömäärällä ja kenties käyttämällä suurempaa ärsykesanalistaa.

Tässä tutkielmassa ei myöskään löydetty eroa keskimääräisessä entropiassa eri ikäryhmissä. Dubossarskyn ym. raportoivat, että entropia kasvaa iän myötä.

5.1 Pieni maailma -indeksistä

Jotkut asiantuntijat ovat kyseenalaistaneet sen, onko pieni maailma -indeksi, kuten se on tässä tutkielmassa laskettu, hyvä tapa kvantifioida pientä maailmallisuutta. Kuten luvussa 3.7 on määritelty, pieni maailma -graafilla on suuri keskimääräinen klusterointikerroin ja pieni polun pituus, ja pieni maailma -indeksi vertaa näitä satunnaisen graafin klusterointikertoimeen ja polun pituuten. Toinen tapa mitata pientä maailmallisuutta olisi verrata niitä myös hilagraafin (*lattice network*) klusterointikertoimeen ja polun pituuten. Hilagraafeilla on suuri klusterointikerroin, ja satunnaisella graafilla on pieni polun pituus, joten yksi ehdotettu tapa kvantifioida pientä maailmallisuutta on käyttää hilagraafista laskettua keskimääräistä klusterointikerrointa ja satunnaisen graafin keskimääräistä polun pituutta ja laskea erotus

$$\omega = \frac{L_{rand}}{L} - \frac{C}{C_{latt}},$$

jossa L ja C ovat kesimääräinen polun pituus ja keskimääräinen klusterointikerroin, L_{rand} on generoidun Erdös-Rényin satunnaisen graafin keskimääräinen polun pituus ja C_{latt} on generoidun hilagraafin keskimääräinen klusterointikerroin. Käyttäen tällaista määritelmää voidaan sanoa, että graafi on pieni maailma jos $\omega \approx 0$, satunnainen jos $\omega > 0$, ja hilagraafi jos $\omega < 0$. Mitta ω voi olla parempi kuin käyttämämme indeksi SWI sen vuoksi, että graafin koko vaikuttaa pieni maailma -indeksiin SWI voimakkaasti, mikä vaikeuttaa graafien vertailua; graafin koko ei vaikuta paljon pieni maailma -mittaan ω. Lisäksi graafit, joissa on erittäin pieni klusterointikerroin, voidaan virheellisesti määritellä pieniksi maailmaksi, kun luotetaan pienen maailma -indeksiin SWI, ja tämäkin ongelma voidaan ratkaista käyttämällä mittaa ω. [20]

Toinen ehdotettu tapa kvantifioida pientä maailmallisuutta on käyttää

kahden graafin normalisoitua indeksiä (*double-graph normalised index*)

$$SWI_2 = \frac{L - L_{latt}}{L_{rand} - L_{latt}} \times \frac{C - C_{rand}}{C_{latt} - C_{rand}},$$

jossa L ja C ovat keskimääräinen polun pituus ja keskimääräinen klusterointikerroin, L_{rand} ja C_{rand} ovat vastaavat mitat generoidussa Erdös-Rényin satunnaisessa graafissa ja L_{latt} ja C_{latt} ovat vastaavat mitat generoidussa hilagraafissa. Kahden graafin normalisoitu indeksi $SWI_2 = 1$ vain jos $L = L_{rand}$ ja $C = C_{latt}$, mikä on ominaista pieni maailma -graafille, ja lähellä nollaa, jos vain yksi tai ei kumpikaan kriteeri täyty. Voidaan siis sanoa, että graafi on pieni maailma jos $SWI_2 \approx 1$. [21]

Syy, miksi tässä tutkielmassa käytettiin indeksiä SWI yllä esiteltyjen vaihtoehtojen sijasta, on se, että tätä indeksiä yhä käytetään mm. neurobiologian ja psykologian aloilla [20] ja muut sana-assosiaatiotutkimukset käyttivät tätä indeksiä [4]. Lisäksi sana-assosiaatioaineistosta luodut graafit ovat suurinpiirtein saman kokoisia, koska niillä on yhtä paljon solmuja (ärsykesanoja), joten graafien vertailu ei ole ongelma. Tällaiset vaihtoehdot kannattaisi kuitenkin ottaa huomioon myös tällä tieteenalalla.

5.2 Rajoituksia

Tässä tutkielmassa käytetyn aineiston keruussa oli vaikeuksia varsinkin ikääntyneiden aikuisten ikäryhmässä, mikä johti siihen, että tutkimukseen ei saatu haluttua koehenkilömäärää. Tästä syystä voima näyttää ikäryhmien välinen ero klusterointikertoimessa laskettiin luvussa 4.2. Voima oli pieni, mikä viittaa siihen, että mahdollisuus osoittaa ero klusterointikertoimessa oli huono. Lisäksi on mahdollista, että juuri pienen koehenkilömäärän vuoksi lasketut suureet eivät seuraa normaalijakaumaa, mikä tuottaa lisää ongelmia tulosten tulkinnassa. Aineistossa oli vielä vähemmän vastauksia ikääntyneiden aikuisten ikäryhmältä kuin nuorten aikuisten, mikä saattaa myös olla syynä joihinkin ikäryhmien välisiin eroihin. Ikääntyneiden aikuisten graafin solmuilla ei ollut yhtä paljon yhteyksiä kuin nuorten aikuisten graafin solmuilla. On mahdollista, että tämä seurasi siitä, että tutkimukseen saatiin mukaan vähemmän ikääntyneitä aikuisia ($n = 52$) kuin nuoria aikuisia ($n = 62$);

enemmän koehenkilöitä tietenkin johtaa siihen, että saadaan enemmän kohdesanoja. Tässä tutkielmassa ei myöskään otettuu huomioon yksilöiden välisiä eroja ryhmän sisällä, joten on myös mahdollista, että pienen koehenkilömäärästä johtuen yksi poikkeava havainto johti merkittäviin eroihin ryhmien keskiarvojen välillä.

Käytetty rekrytointiprosessi myös johti siihen, että koehenkilöt eivät edusta populaatiota. Paljon koehenkilöitä rekrytoitiin samasta ryhmästä (mm. yliopiston tai harrastuksien kautta), mikä on saattanut aiheuttaa paljon homogeenisuutta tutkimusryhmän sisällä. Lisäksi koehenkilöt olivat suurimmaksi osaksi myös sellaisia henkilöitä, joita kiinnosti aihealue tai tutkimukseen osallistumisesta saatu palkinto (mahdollisuus voittaa lahjakortti).

5.3 Lopuksi

Tässä tutkielmassa pyrittiin vastaamaan kysymykseen, miten ihmiset organisoivat kieltä, sekä visualisoimaan tämän organisoinnin verkoston eli graafin avulla. Vaikka onnistuttiin soveltamaan graafiteoriaa valittuun tutkimusalueeseen ja visualisoimaan sana-assosiaatiotutkimuksen tuloksia graafin avulla, tämä tutkielma ei antanut selviä vastauksia siitä, miten kieli organisoituu aivoissa (ja miten tämä eroaa eri ikäryhmissä). Konkreettisemman vastauksen saamiseen tarvittaisiin suurempi aineisto eli aineisto jossa on enemmän koehenkilöitä, mutta mahdollisesti myös suurempi ärsykesanalista.

Viitteet

[1] J.D. Nystuen, M.F. Dacey. A graph theory interpretation of nodal regions. *Papers in Regional Science* **7** (1961), No. 1, 29-42.

[2] K. Kohlstedt, *The Seven Bridge Problem: How an Urban Puzzle Inspired a New Field of Mathematics.* 99 % Invisible. <https://99percentinvisible.org/article/the-seven-bridge-problem-how-an-urban-puzzle-inspired-a-new-field-of-mathematics/>.

[3] A.L. Barabási. *Linkit – Verkostojen uusi teoria.* Helsinki, Terra Cognita, 2002. (Suomentanut Kimmo Pietiläinen. Alkuperäinen: *Linked: The New Science of Networks*, Brockman Inc.)

[4] H. Dubossarsky, S. De Deyne, T.T. Hills. Quantifying the Structure of Free Association Networks Across the Life Span. *Developmental Psychology* **53** (2017), No. 8, 1560-1570.

[5] D.L. Nelson, C.L. McEvoy, T.A. Schreiber. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* **36** (2004), 402–407. <https://doi.org/10.3758/BF03195588>.

[6] M. Zortea, B. Menegola, A. Villavicencio, J. Fumagalli de Salles. Graph Analysis of Semantic Word Association among Children, Adults, and the Elderly. *Psicologia: Reflexão e Crítica* **27** (2014), No. 1, 90-99.

[7] T.T. Hills, M. Maouene, J. Maouene, A. Sheya, L. Smith. Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science* **20** (2009b), No. 6, 729–739.

[8] P. Koivisto, R. Niemistö. *Graafiteoriaa.* Tampere, Tampereen yliopisto. 2018, 2. painos. (Tampereen yliopisto, Opintomoniste).

[9] M.E.J Newman, *Networks: An Introduction.* Oxford University Press. 2010, 1. painos.

[10] T. Opsahl, F. Agneessens, J. Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks* **32** (2010), 245-251.

[11] J. Holloway, H. Sposito, J. Tan, B. Bieri, Package 'migraph'. R-package version 1.3.2, (2024). <https://cran.r-project.org/web/packages/migraph/migraph.pdf>.

[12] N. Masuda, M. Sakaki, T. Ezaki, T. Watanabe. *Clustering Coefficients for Correlation Networks.* Frontiers in Neuroinformatics. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863042/>.

[13] A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* **101** (2004), No. 11, 3747–3752. <https://www.pnas.org/doi/10.1073/pnas.0400087101>.

[14] M.D. Humphries, K. Gurney. Network 'Small-World-Ness': A Quantitative Method for Determining Canonical Network Equivalence. *PLoS ONE* **3** (2008), No. 4: e0002051. <https://doi.org/10.1371/journal.pone.0002051>.

[15] *Clustering Coefficient in Graph Theory.* Geeks for Geeks. <https://www.geeksforgeeks.org/clustering-coefficient-graph-theory/>.

[16] M. Karoński, A. Ruciński. The Origins of the Theory of Random Graphs. In: R.L. Graham, J. Nešetřil (eds), *The Mathematics of Paul Erdös I. Algorithms and Combinatorics*. Springer. 1997, 13. painos.

[17] C. DiMagio. *Power Tools for Epidemiologists.* ICEPaC. <http://www.columbia.edu/ cjd11/charles_dimaggio/DIRE/styled-4/code-12/>.

[18] D. Garlaschelli. The weighted random graph model. *New Journal of Physics* **11** (2009). <https://iopscience.iop.org/article/10.1088/1367-2630/11/7/073005>.

[19] C.L. Hsu, *Erdos-Renyi Random Graph.* An Explorer of Things. <https://chih-ling-hsu.github.io/2020/05/15/Gnp>.

[20] Q.K. Telesford, K.E. Joyce, S. Hayasaka, J.H. Burdette, P.J. Laurienti. The Ubiquity of Small-World Networks. *Brain Connectivity* **5** (2011), No. 1, 367-375. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604768/>.

[21] Z.P. Neal. How small is it? Comparing indices of small worldliness. *Network science.* **5** (2017), No. 1, 30-44.

Liite A R-koodi

```
library(operators)
library(tidyverse)
library(statnet)
library(plyr)
library(migraph)
library(igraph)
library(tnet)
library(DirectedClustering)
library(effectsize)
library(ggplot2)
library(bootstrap)
library(boot)
library(vcd)
library(ggpubr)

## 1. datan organisointi ######################
datadf <- as.data.frame(data)
data000 <- datadf %>% filter(V2 != "0")
data_nozero <- data000[!(data000$V3==""), ]
data_notext <- data_nozero[!(data_nozero$V1=="stim"), ]
f1 <- function(data){
  df <- data.frame()
  for (i in 1:nrow(data)) {
    df[i,1] <- as.character(data[i,1])
    df[i,2] <- as.character(data[i,5])
  }
  for (j in 1:nrow(data)) {
    df[nrow(data)+j,1] <- as.character(data[j,1])
    df[nrow(data)+j,2] <- as.character(data[j,8])
  }
  for (k in 1:nrow(data)) {
    df[nrow(data)*2+k,1] <- as.character(data[k,1])
    df[nrow(data)*2+k,2] <- as.character(data[k,11])
  }
  colnames(df) <- c("arsykesana", "kohdesana")
  df_ordered <- df[order(df$arsykesana),]
  return(df_ordered)
}
f2 <- function(data){
  df <- data.frame(matrix(NA, nrow = 270, ncol = 52))
  for (i in 2:nrow(data)) {
    stringword <-  paste("^", data[i,1], "$", sep="")
    if(any(str_detect(df[,1], stringword), na.rm = TRUE)){
      J <- which(df == data[i,1], arr.ind=TRUE)
      j <- J[1,1] #df[j, ?]
      count=1
      while(is.na(df[j,count])==FALSE) {count <- count + 1}
      df[j,count] <- data[i,2]
    } else {
      camt=1
      while(is.na(df[camt,1])==FALSE) {camt <- camt + 1}
      df[camt,1] <- data[i,1]
      df[camt,2] <- data[i,2]
    }
  }
  colnames(df) <- c("arsykesana", "X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9",
       "X10", "X11", "X12", "X13", "X14", "X15", "X16", "X17", "X18", "X19", "X20", "X21
       ", "X22", "X23", "X24", "X25", "X26", "X27", "X28", "X29", "X30", "X31", "X32", "
       X33", "X34", "X35", "X36", "X37", "X38", "X39", "X40", "X41", "X42", "X43", "X44",
        "X45", "X46", "X47", "X48", "X49", "X50", "X51")
  return(df)
}

#### A ik  ntyneet ####
```

```
data_A <- data_notext %>% filter(V3 == "a")
TA <- f1(data_A)
TA_1 <- TA[!(TA$kohdesana=="-"), ]
A_ikaantyneet <- TA_1[!(TA_1$kohdesana==""), ]
A_ikaantyneet <- mutate_all(A_ikaantyneet, .funs=tolower)
A_ikaantyneet2 <- f2(A_ikaantyneet)

#### B nuoret ####
data_B <- data_notext %>% filter(V3 == "b")
TB <- f1(data_B)
TB_1 <- TB[!(TB$kohdesana=="-"), ]
B_nuoret <- TB_1[!(TB_1$kohdesana==""), ]
B_nuoret <- mutate_all(B_nuoret, .funs=tolower)
B_nuoret2 <- f2(B_nuoret)

## 2. kaksimuotoinen graafi ######################
colors <- c("green", "red")
shapes <- c("circle", "square")

#### A ikaantyneet ####
A_matrix <- as.matrix(A_ikaantyneet)
gA <- graph.data.frame(A_matrix, directed = FALSE)
V(gA)$type <- V(gA)$name %in% A_matrix[,1]
V(gA)$size <- 1
plot(gA, layout = layout_with_graphopt, vertex.label = NA, vertex.color = colors[V(gA)
    $type+1], vertex.shape = shapes[V(gA)$type+1], edge.color = "black", edge.width =
    0.1)

#### B nuoret  ####
B_matrix <- as.matrix(B_nuoret)
gB <- graph.data.frame(B_matrix, directed = FALSE)
V(gB)$type <- V(gB)$name %in% B_matrix[,1]
V(gB)$size <- 1
plot(gB, layout = layout_with_graphopt, vertex.label = NA, vertex.color = colors[V(gB)
    $type+1], vertex.shape = shapes[V(gB)$type+1], edge.color = "black", edge.width =
    0.1)

## 3. yksimuotoinen painotettu graafi ######################
#### painot ####
f3 <- function(data, data2) {
  net_weighted <- data.frame(matrix(NA, nrow=270, ncol=270))
  colnames(net_weighted) <- data2[,1]
  rownames(net_weighted) <- data2[,1]
  df = subset(data2, select = -arsykesana )
  for (i in 1:270) {
    for (j in 1:270) {
      if(i==j) {
        net_weighted[i,j] = 0
        print(c(i,j))
    } else {
        VEC <- c()
        for (v in 1:ncol(df)) {
          if (df[i,v] %in% df[j,] == TRUE & is.na(df[i,v]) == FALSE) {
            VEC <- append(VEC, df[i,v])
          }
        }
        unique_VEC <- unique(VEC)
        P <-  length(unique_VEC)
        if(P==0) {
          net_weighted[i,j] = 0
          print(c(i,j))
      } else {
          w_ij = 0
          for(p in 1:P){
            w_ip = 0
            for (bb in 1:length(VEC)) {
              if(VEC[bb] == unique_VEC[P]) {
```

```
            w_ip=w_ip+1
          }
        }
        N_p = 0
        stringword <-  paste("^", unique_VEC[P], "$", sep="")
        for (k in 1:nrow(df)) {
          if(any(str_detect(df[k,], stringword), na.rm = TRUE)){
            N_p = N_p +1
          }
        }
        w_ij = w_ij + w_ip / (N_p - 1)
      }
      net_weighted[i,j] = w_ij
      print(c(i,j))
    }
  }
 }
 return(net_weighted)
}
A_weights_onemode_ikaantyneet <- f3(data=A_ikaantyneet, data2=A_ikaantyneet2)
B_weights_onemode_nuoret <- f3(data=B_nuoret, data2=B_nuoret2)

#### graafin piirtminen ####
####### A ikaantyneet ####
matAA <- as.matrix(A_weights_onemode_ikaantyneet)
gAA <- graph_from_adjacency_matrix(matAA, mode="directed", weighted=TRUE, diag=FALSE)
plot(gAA, layout = layout_with_fr(gAA), edge.width = 0.1*log(E(gAA)$weight), edge.arrow.
    size=0.5, vertex.color="red", vertex.size=2, vertex.frame.color="red", vertex.label.
    color="black", edge.color = "grey", vertex.label.cex=0.8, vertex.label.dist=0, edge.
    curved=0.2)

####### B nuoret ####
matBB <- as.matrix(B_weights_onemode_nuoret)
gBB <- graph_from_adjacency_matrix(matBB, mode="directed", weighted=TRUE, diag=FALSE)
plot(gBB, layout = layout_with_fr(gBB), edge.width = 0.1*log(E(gBB)$weight), edge.arrow.
    size=0.5, vertex.color="red", vertex.size=2, vertex.frame.color="red", vertex.label.
    color="black", edge.color = "grey", vertex.label.cex=0.8, vertex.label.dist=0, edge.
    curved=0.2)

## 4. laskennat ######################
tnetA_ikaantyneet <- as.tnet(A_weights_onemode_ikaantyneet)
tnetB_nuoret <- as.tnet(B_weights_onemode_nuoret)

#### aste ja voimakkuus ####
# tuloaste ja -voimakkuus
kands_A_in <- degree_w(tnetA_ikaantyneet, measure=c("degree","output"), type="in", alpha
    =1)
kands_B_in <- degree_w(tnetB_nuoret, measure=c("degree","output"), type="in", alpha=1)

# l ht aste ja -voimakkuus
kands_A_out <- degree_w(tnetA_ikaantyneet, measure=c("degree","output"), type="out",
    alpha=1)
kands_B_out <- degree_w(tnetB_nuoret, measure=c("degree","output"), type="out", alpha=1)

#### asteen keskeisyysmitta ####
alpha = 0.5
centrality_A_in <- vector(mode="double",length = 270)
for (i in 1:270) {centrality_A_in[i] = kands_A_in[i,2]^(1-alpha) * kands_A_in[i,3]^alpha
    }
centrality_A_out <- vector(mode="double",length = 270)
for (i in 1:270) {centrality_A_out[i] = kands_A_out[i,2]^(1-alpha) * kands_A_out[i,3]^
    alpha}
centrality_B_in <- vector(mode="double",length = 270)
for (i in 1:270) {centrality_B_in[i] = kands_B_in[i,2]^(1-alpha) * kands_B_in[i,3]^alpha
    }
centrality_B_out  <- vector(mode="double",length = 270)
```

```
for (i in 1:270) {centrality_B_out[i] = kands_B_out[i,2]^(1-alpha) * kands_B_out[i,3]^
    alpha}

degreestrengthcentrality_A_in <- cbind(kands_A_in, centrality_A_in)
degreestrengthcentrality_A_in <- cbind(degreestrengthcentrality_A_in[,1], A_ikaantyneet2
    [,1], degreestrengthcentrality_A_in[,2], degreestrengthcentrality_A_in[,3],
    degreestrengthcentrality_A_in[,4])
colnames(degreestrengthcentrality_A_in) = c("solmu id", "  rsykesana  ", "tuloaste", "
    tulovoimakkuus", "(tuloasteen) keskeisyysmitta")

degreestrengthcentrality_B_in <- cbind(kands_B_in, centrality_B_in)
degreestrengthcentrality_B_in <- cbind(degreestrengthcentrality_B_in[,1], B_nuoret2[,1],
     degreestrengthcentrality_B_in[,2], degreestrengthcentrality_B_in[,3],
    degreestrengthcentrality_B_in[,4])
colnames(degreestrengthcentrality_B_in) = c("solmu id", "  rsykesana  ", "tuloaste", "
    tulovoimakkuus", "(tuloasteen) keskeisyysmitta")

degreestrengthcentrality_A_out <- cbind(kands_A_out, centrality_A_out)
degreestrengthcentrality_A_out <- cbind(degreestrengthcentrality_A_out[,1],
    A_ikaantyneet2[,1], degreestrengthcentrality_A_out[,2],
    degreestrengthcentrality_A_out[,3], degreestrengthcentrality_A_out[,4])
colnames(degreestrengthcentrality_A_out) = c("solmu id", "  rsykesana  ", "l ht aste", "
    l ht voimakkuus", "(l ht asteen) keskeisyysmitta")

degreestrengthcentrality_B_out <- cbind(kands_B_out, centrality_B_out)
degreestrengthcentrality_B_out <- cbind(degreestrengthcentrality_B_out[,1], B_nuoret2
    [,1], degreestrengthcentrality_B_out[,2], degreestrengthcentrality_B_out[,3],
    degreestrengthcentrality_B_out[,4])
colnames(degreestrengthcentrality_B_out) = c("solmu id", "  rsykesana  ", "l ht aste", "
    l ht voimakkuus", "(l ht asteen) keskeisyysmitta")

#### klusterointikerroin ####
CA_all <- ClustBCG(mat = matAA, type = "directed")
CA_global <- CA_all$GlobaltotalCC
CA_local <- CA_all$totalCC
CB_all <- ClustBCG(mat = matBB, type = "directed")
CB_global <- CB_all$GlobaltotalCC
CB_local <- CB_all$totalCC
# keskim  r inen klusterointikerroin:
CA <- clustering_w(tnetA_ikaantyneet, measure = "am") #am = arithmetic mean
CB <- clustering_w(tnetB_nuoret, measure = "am")

###### keskiarvo ja keskihajonta ####
CA_global
sd(CA_local)/sqrt(270)
CA_global - 1.96*(sd(CA_local)/sqrt(270))
CA_global + 1.96*(sd(CA_local)/sqrt(270))
CB_global
sd(CB_local)/sqrt(270)
CB_global - 1.96*(sd(CB_local)/sqrt(270))
CB_global + 1.96*(sd(CB_local)/sqrt(270))

#### polun pituus ####
d_A <- distance_w(tnetA_ikaantyneet, directed=TRUE, gconly=FALSE, subsample=1, seed=NULL
    )
d_B <- distance_w(tnetB_nuoret, directed=TRUE, gconly=FALSE, subsample=1, seed=NULL)
LA_global <- mean(d_A, na.rm=TRUE)
LB_global <- mean(d_B, na.rm=TRUE)
LA_local <- c()
for (i in 1:270){LA_local[i] <-  mean(d_A[i,], na.rm=TRUE)}
LB_local <- c()
for (i in 1:270){LB_local[i] <-  mean(d_B[i,], na.rm=TRUE)}

#### erd s renyi satunnainen graafi ####
f4 <- function(tnetA_or_B){
  randomX <- erdos.renyi.game(270, nrow(tnetA_or_B), type = "gnm", directed = TRUE,
      loops = TRUE)
```

```
  df <- matrix(NA, nrow = 270, ncol = 270)
  for (i in 1:270){ df[i,] <- randomX[i] }
  tnetdata <- as.tnet(df)
  return(tnetdata)
}
f5 <- function(tnetA_or_B){
  randomX <- erdos.renyi.game(270, nrow(tnetA_or_B), type = "gnm", directed = TRUE,
      loops = TRUE)
  df <- matrix(NA, nrow = 270, ncol = 270)
  for (i in 1:270){ df[i,] <- randomX[i] }
  return(df)
}

random_A <- f4(tnetA_ikaantyneet)
random_B <- f4(tnetB_nuoret)
CA_random <- clustering_w(random_A, measure = "am")
CB_random <- clustering_w(random_B, measure="am")
d_A_random <- distance_w(random_A, directed=TRUE, gconly=FALSE, subsample=1, seed=NULL)
d_B_random <- distance_w(random_B, directed=TRUE, gconly=FALSE, subsample=1, seed=NULL)
LA_random <- mean(d_A_random, na.rm=TRUE)
LB_random <- mean(d_B_random, na.rm=TRUE)
random_A_mat <- f5(tnetA_ikaantyneet)
random_B_mat <- f5(tnetB_nuoret)

#### pieni maailma indeksi ####
SMI_A <- (CA / CA_random) / (LA_global / LA_random)
SMI_B <- (CB / CB_random) / (LB_global / LB_random)

#### entropia ####
A_ikaantyneet1 <- sort(A_ikaantyneet)
B_nuoret1 <- sort(B_nuoret)

f6 <- function(data_sorted){ ## counts the totals
  totals <- data.frame()
  m=1
  n=0
  for (j in 1:270){
    m=m+n
    n=0
    for (i in 1:nrow(data_sorted)) {
      {if (data_sorted[i,1]==data_sorted[m,1]){n=n+1}}
    }
    totals[j,1] = data_sorted[m,1]
    totals[j,2] = n
  }
  totals <- sort(totals)
  return(totals)
}

f7 <- function(data_sorted, totals) {
  data3 <- matrix(NA, nrow = nrow(data_sorted), ncol = 5)
  colnames(data3) <- c("arsykesana", "kohdesana", "count", "total_for_arsykesana", "
      probs")
  data3 <- as.data.frame(data3)
  m = 1
  t = m + totals[1,2] - 1
  k = 1
  for (j in 1:269){
    fr <- count(data_sorted[m:t,2])
    data3[k:(k+nrow(fr)-1), 1] <- c(rep(totals[j,1], nrow(fr)))
    data3[k:(k+nrow(fr)-1), 2] <- fr$x
    data3[k:(k+nrow(fr)-1), 3] <- fr$freq
    data3[k:(k+nrow(fr)-1), 4] <-  c(rep(totals[j,2], nrow(fr)))
    m = m + totals[j,2]
    t = m + totals[j+1,2] - 1
    k = k + nrow(fr)
  }
```

```
  fr <- count(data_sorted[m:t,2])
  data3[k:(k+nrow(fr)-1), 1] <- c(rep(totals[270,1], nrow(fr)))
  data3[k:(k+nrow(fr)-1), 2] <- fr$x
  data3[k:(k+nrow(fr)-1), 3] <- fr$freq
  data3[k:(k+nrow(fr)-1), 4] <-  c(rep(totals[270,2], nrow(fr)))
  data3$count <- as.numeric(data3$count)
  data3$total_for_arsykesana <- as.numeric(data3$total_for_arsykesana)
  data3$probs <- data3$count / data3$total_for_arsykesana
  data3$probs <- as.numeric(data3$probs)
  return(na.omit(data3))
}

A_totals <- f6(A_ikaantyneet1)
B_totals <- f6(B_nuoret1)
A_3 <- f7(A_ikaantyneet1, A_totals)
B_3 <- f7(B_nuoret1, B_totals)

f_entropy <- function(data3, totals) {
  H <- c()
  for (j in 1:nrow(totals)) {
    C <- c()
    for (i in 1:nrow(data3)) {
      if (data3[i,1] == totals[j,1]) {
        C[i] <- data3[i,5]
      } }
    C <- C[!is.na(C)]
    n = length(C)
    h = 0
    for (k in 1:n) {
      h = h + ( C[k] * log(C[k]) ) / log(n)
    }
    h = -h
    H[j] <- h
    }
  return(H)
}
A_entropy <- f_entropy(A_3, A_totals)
B_entropy <- f_entropy(B_3, B_totals)

## 5. MANOVA/ANOVA ######################
# 1 = A (ikaantyneet)
# 0 = B (nuoret)

#### ANOVA  ####
####### tuloasteen keskeisyysmitta ####
AABB <- rep(c(1), each=270)
degrecent_A_in <- cbind(AABB, degreestrengthcentrality_A_in)
degrecent_A_in <- as.data.frame(degrecent_A_in)
colnames(degrecent_A_in) <- c("AABB", "solmu id", "  rsykesana  ", "tuloaste", "
    tulovoimakkuus", "(tuloasteen) keskeisyysmitta")
AABB <- rep(c(0), each=270)
degrecent_B_in <- cbind(AABB, degreestrengthcentrality_B_in)
degrecent_B_in <- as.data.frame(degrecent_B_in)
colnames(degrecent_B_in) <- c("AABB", "solmu id", "  rsykesana  ", "tuloaste", "
    tulovoimakkuus", "(tuloasteen) keskeisyysmitta")
degreestrengthcentrality_in <- rbind(degrecent_A_in, degrecent_B_in)

anova_cent_in <- lm(degreestrengthcentrality_in[,5] ~ degreestrengthcentrality_in$AABB)
summary(anova_cent_in)

####### l ht asteen keskeisyysmitta ####
AABB <- rep(c(1), each=270)
degrecent_A_out <- cbind(AABB, degreestrengthcentrality_A_out)
degrecent_A_out <- as.data.frame(degrecent_A_out)
colnames(degrecent_A_out) <- c("AABB", "solmu id", "  rsykesana  ", "tuloaste", "
    tulovoimakkuus", "(tuloasteen) keskeisyysmitta")
AABB <- rep(c(0), each=270)
```

```
degrecent_B_out <- cbind(AABB, degreestrengthcentrality_B_out)
degrecent_B_out <- as.data.frame(degrecent_B_out)
colnames(degrecent_B_out) <- c("AABB", "solmu id", "  rsykesana  ", "tuloaste", "
    tulovoimakkuus", "(tuloasteen) keskeisyysmitta")
degreestrengthcentrality_out <- rbind(degrecent_A_out, degrecent_B_out)

anova_cent_out <- lm(degreestrengthcentrality_out[,5] ~
    degreestrengthcentrality_out$AABB)
summary(anova_cent_out)

###### klusterointikerroin ja polun pituus ####

df_LA <- as.data.frame(LA_local)
df_CA <- as.data.frame(CA_local)
df_LC_A <- cbind(df_LA, df_CA)
AABB <- rep(c(1), each=270)
df_LC_A <- cbind(AABB, df_LC_A)

df_LB <- as.data.frame(LB_local)
df_CB <- as.data.frame(CB_local)
df_LC_B <- cbind(df_LB, df_CB)
AABB <- rep(c(0), each=270)
df_LC_B <- cbind(AABB, df_LC_B)
colnames(df_LC_A) <- c("AABB", "local_shortest", "klustering")
colnames(df_LC_B) <- c("AABB", "local_shortest", "klustering")
df_LC <- rbind(df_LC_A, df_LC_B)

######## klusterointikerroin
anova_klust <- lm(klustering ~ AABB, data = df_LC)
summary(anova_klust)

######## polun pituus
anova_length <- lm(local_shortest ~ AABB, data = df_LC)
summary(anova_length)

###### entropia ####

AABB <- rep(c(1), each=270)
A_ent <- cbind(AABB, A_entropy)
A_ent <- as.data.frame(A_ent)
colnames(A_ent) <- c("AABB", "entropy")
AABB <- rep(c(0), each=270)
B_ent <- cbind(AABB, B_entropy)
B_ent <- as.data.frame(B_ent)
colnames(B_ent) <- c("AABB", "entropy")
AB_entropy <- rbind(A_ent, B_ent)

anova_entropy <- lm(entropy ~ AABB, data = AB_entropy)
summary(anova_entropy)

#### MANOVA ####
all_measures <- cbind(degreestrengthcentrality_in$`solmu id`, df_LC$AABB,
    df_LC$local_shortest, df_LC$klustering, degreestrengthcentrality_in$`(tuloasteen)
    keskeisyysmitta`, degreestrengthcentrality_out$`(l ht asteen) keskeisyysmitta`)
colnames(all_measures) <- c("solmu_ID", "AABB", "shortestpath", "klustering", "
    in_centrality", "out_centrality")
all_measures <- as.data.frame(all_measures)
all_measures[,1] <- as.numeric(all_measures[,1])
all_measures[,2] <- as.numeric(all_measures[,2])
all_measures[,3] <- as.numeric(all_measures[,3])
all_measures[,4] <- as.numeric(all_measures[,4])
all_measures[,5] <- as.numeric(all_measures[,5])
all_measures[,6] <- as.numeric(all_measures[,6])

all_manova <- manova(cbind(shortestpath, klustering, in_centrality, out_centrality) ~
    AABB, data = all_measures)
all_manova
```

```
summary(all_manova)

effectsize::eta_squared(all_manova)
degrees_manova <- manova(cbind(in_centrality, out_centrality) ~ AABB, data =
    all_measures)
degrees_manova
summary(degrees_manova)
klustlength_manova <- manova(cbind(shortestpath, klustering) ~ AABB, data = all_measures
    )
klustlength_manova
summary(klustlength_manova)
summary.aov(all_manova)

## 8. normaalisuuden testaus ####################
degreestrengthcentrality_in[,6] <- as.numeric(degreestrengthcentrality_in[,6])
degreestrengthcentrality_out[,6] <- as.numeric(degreestrengthcentrality_out[,6])
shapiro.test(centrality_A_in)
shapiro.test(centrality_B_in)
shapiro.test(degreestrengthcentrality_in[,6])
shapiro.test(centrality_A_out)
shapiro.test(centrality_B_out)
shapiro.test(degreestrengthcentrality_out[,6])
shapiro.test(LA_local)
shapiro.test(LB_local)
shapiro.test(df_LC$local_shortest)
shapiro.test(CA_local)
shapiro.test(CB_local)
shapiro.test(df_LC$klustering)

#### residuaalit ####
degreestrengthcentrality_in[,6] <- as.numeric(degreestrengthcentrality_in[,6])
degreestrengthcentrality_out[,6] <- as.numeric(degreestrengthcentrality_out[,6])

qqnorm((degreestrengthcentrality_in[,6] - mean(degreestrengthcentrality_in[,6])) / sd(
    degreestrengthcentrality_in[,6]), main = "tuloasteen keskeisyysmitta")
abline(0,1)
ggqqplot(degreestrengthcentrality_in[,6], main = "tuloasteen keskeisyysmitta")

qqnorm((degreestrengthcentrality_out[,6] - mean(degreestrengthcentrality_out[,6])) / sd(
    degreestrengthcentrality_out[,6]), main= "l ht asteen keskeisyysmitta")
abline(0,1)
ggqqplot(degreestrengthcentrality_out[,6], main= "l ht asteen keskeisyysmitta")

qqnorm((df_LC$local_shortest - mean(df_LC$local_shortest)) / sd(df_LC$local_shortest),
    main="polun pituus")
abline(0,1)
ggqqplot(df_LC$local_shortest, main="polun pituus")

qqnorm((df_LC$klustering - mean(df_LC$klustering)) / sd(df_LC$klustering), main="
    klusterointikerroin")
abline(0,1)
ggqqplot(df_LC$klustering, main="klusterointikerroin")

qqnorm((AB_entropy$entropy -
          mean(AB_entropy$entropy)) /
         sd(AB_entropy$entropy),
       main = "entropia")
abline(0,1)
ggqqplot(AB_entropy$entropy,
         main = "entropia")

## 9. ep parametriset testit ####################
# tuloasteen keskeisyysmitta
degreestrengthcentrality_in[,5] <- as.numeric(degreestrengthcentrality_in[,5])
wilcox.test(degreestrengthcentrality_in[,5] ~ degreestrengthcentrality_in$AABB)
# l ht asteen keskeisyysmitta
degreestrengthcentrality_out[,5] <- as.numeric(degreestrengthcentrality_out[,5])
```

```
wilcox.test(degreestrengthcentrality_out[,5] ~ degreestrengthcentrality_out$AABB)
# klusterointikerroin
wilcox.test(klustering ~ AABB, data = df_LC) # p = 0.2734
# polun pituus
wilcox.test(local_shortest ~ AABB, data = df_LC)
# entropia
wilcox.test(entropy ~ AABB, data = AB_entropy)
```

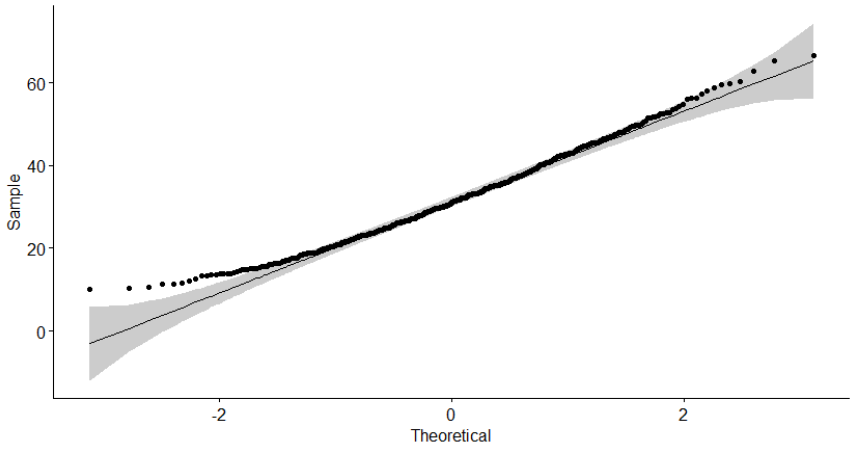
Liite B Ärsykesanat

ahdas	ainainen	ajatella	ajattelu	ajatus
alkuperäinen	anoa	anomus	antaa	apu
arvella	arvokas	asia	aurinko	auto
avara	ehta	ehto	elämä	ennakoida
epäillä	epävarma	erilainen	erinomainen	etana
etu	fakta	haastava	halu	haluta
hammas	harmaa	hauska	helppo	hymy
hyvä	ilmapallo	iloita	iso	jääkiekko
järkevä	järvi	juosta	juttu	kaatua
kakku	kalastus	kallio	kasvaa	katsoa
kaunis	kehittää	keltainen	kerätä	kevyt
kieriä	kiivetä	kirjava	kirjoittaa	kiva
kivi	koira	korjata	korkea	kova
kuolema	kuoppa	kuvitella	kylmä	kynä
kynnys	kätellä	käyttäytyä	kääntää	kömpelö
laahustaa	lahja	laiha	laki	lapio
lapsi	lasi	lehti	lihava	liikkua
litteä	loikata	luja	lukea	lukita
lumi	luonne	luotettava	lyhyt	lämmin
läpinäkyvä	maalata	mahdollinen	mahtava	maine
menestyä	meri	mieli	mielipide	mieluinen
miettiä	musta	mäki	nainen	nauraa
nauttia	negatiivinen	nopea	nostaa	nousta
nöyrä	ohjata	ohje	ohut	oikeudenmukainen

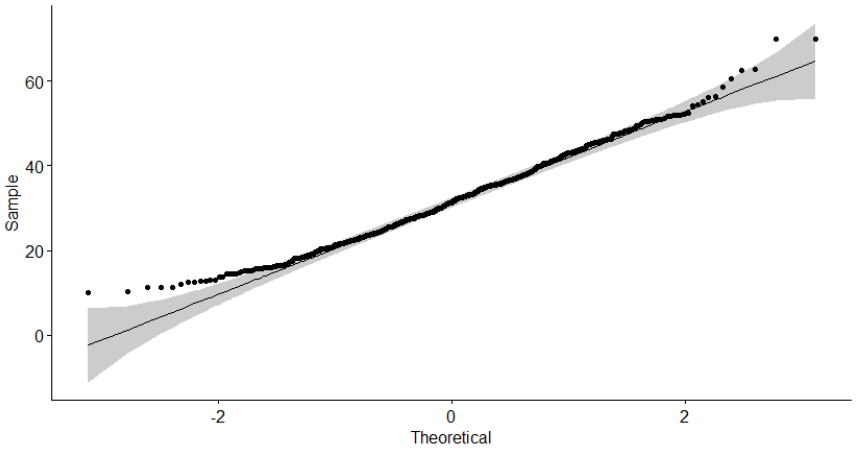
oiva	oivaltaa	olla	omena	ongelma
onnellinen	onnistua	ontto	opetella	osapuoli
osata	ostaa	outo	paha	pakko
pallo	paperi	parantaa	pelata	pieni
pimeä	pohtia	politiikka	pomppia	posti
pudota	puhdas	puhua	puolustaa	puu
pyöreä	pyörä	pystyvä	pystyä	pää
päätös	rakentaa	raskas	ratkaista	ratkaisu
rehellinen	rehti	remontoida	riittää	rikkoutua
rohjeta	rohkeus	ruoka	ruskea	rutistaa
saavuttamaton	salaisuus	sateenkaari	sattua	satuttaa
seikkaperäinen	selvittää	selvitä	selviytyä	sileä
sininen	soittaa	sopimus	surullinen	suunnitella
suuri	syventyvä	syödä	sääntö	särkyä
tahaton	taitaa	taitava	taito	taivas
tajuta	talo	talvi	tanssia	tapahtua
tarpeellinen	tarve	tasapuolinen	tavata	tehdä
tehtävä	teko	tie	tieto	tietää
tilanne	tippua	tiputtaa	tiukka	todellinen
toimia	toive	tulevaisuus	tunne	tuntematon
tutkia	tutustua	tuuli	tuulla	tyyny
työllistää	tärkeä	täysi	upea	urheilla
usko	uskoa	uskomaton	uusia	vaara
vaate	vaatia	vaatimaton	vaatimus	vaativa
vahva	vaikuttava	valhe	valkoinen	valmis
valmistautua	valoisa	vanha	varma	vasen
veitsi	vene	vesi	vihreä	vilpitön
vinkki	virhe	voida	voimakas	vuori
väärä	välttämätön	värittää	yletä	ylevä
yllätys	ymmärtää	yrittää	ystävällinen	äly

Taulukko 4: Sana-assosiaatiotutkimuksessa käytetyt ärsykesanat

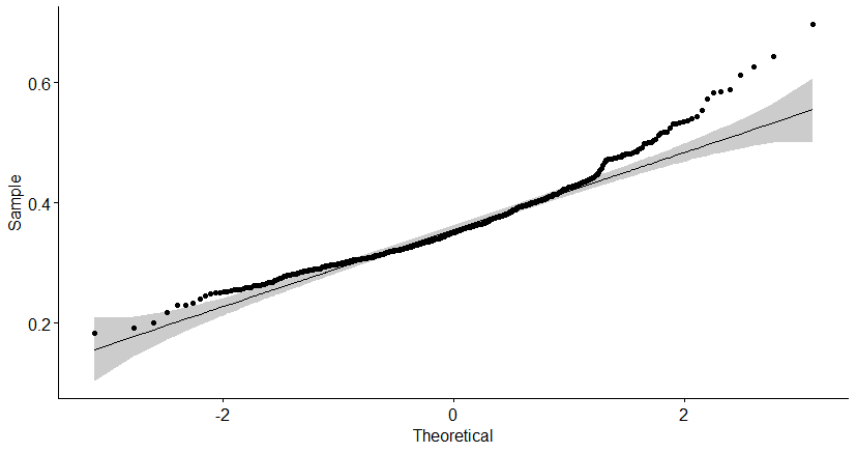
Liite C Residuaalit



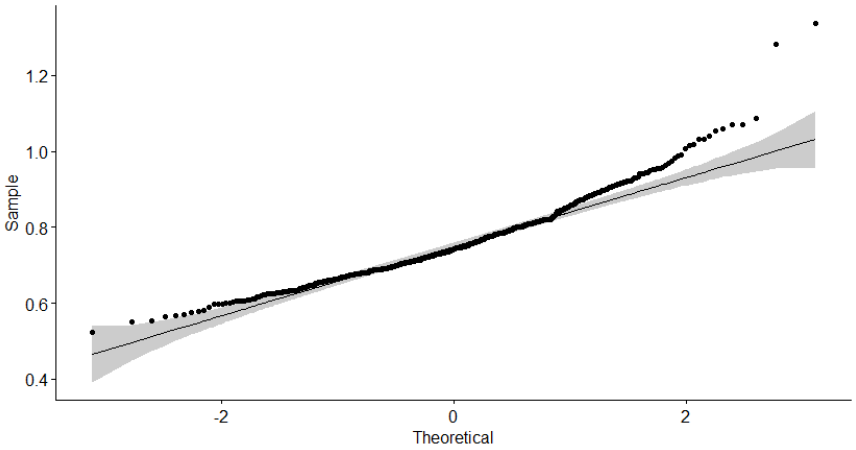
Lähtöasteen keskeisyysmitan residuaalit.



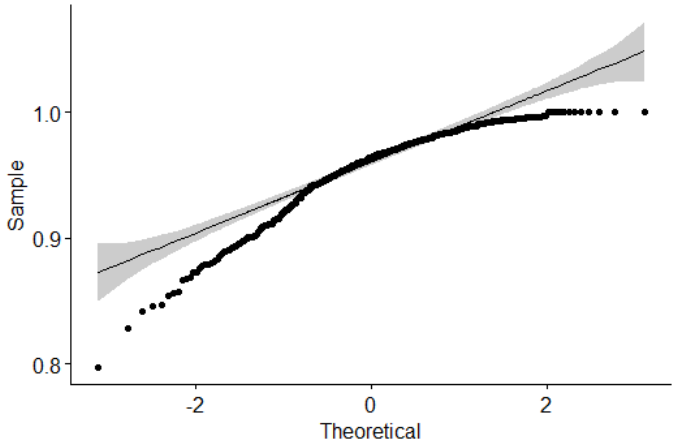
Tuloasteen keskeisyysmitan residuaalit.



Klusterointikertoimen residuaalit.



Polun pituuden residuaalit.



Entropian residuaalit.