HELSINGIN YLIOPISTO

# Bayesiläinen ennustava lukija ja Aleksis Kiven *Seitsemän veljestä*

Laskennallinen näkökulma kaunokirjallisen teoksen analysoimiseen

Kirjallisuudentutkimuksen maisteriohjelma
Maisterintutkielma

Laatija:
Jussi Määttä

Ohjaaja:
professori Riikka Rossi

5.5.2024
Helsinki

**Tiedekunta:** humanistinen tiedekunta

**Koulutusohjelma:** kirjallisuudentutkimuksen maisteriohjelma

**Opintosuunta:** kotimainen kirjallisuus

**Tekijä:** Jussi Määttä

**Työn nimi:** Bayesiläinen ennustava lukija ja Aleksis Kiven *Seitsemän veljestä*. Laskennallinen näkökulma kaunokirjallisen teoksen analysoimiseen

**Työn laji:** maisterintutkielma

**Kuukausi ja vuosi:** toukokuu 2024

**Sivumäärä:** 40

**Avainsanat:** laskennallinen kirjallisuudentutkimus, kognitiivinen kirjallisuudentutkimus, ennustava käsittely, bayesiläinen tilastotiede, Aleksis Kivi, Seitsemän veljestä

**Ohjaaja tai ohjaajat:** professori Riikka Rossi

**Säilytyspaikka:** Helsingin yliopiston kirjasto

**Muita tietoja:** –

**Tiivistelmä:**

Tutkielmassa tarkastellaan laskennallista näkökulmaa kaunokirjallisuuden analysoimiseen kognitiivisen kirjallisuudentutkimuksen ja bayesiläisen tilastotieteen lähtökohdista ja sovelletaan näiden synteesiä Aleksis Kiven romaanin *Seitsemän veljestä* (1870) analyysiin.

Keskeisenä teoriakehyksenä tutkielmassa on kognitiivinen kirjallisuudentutkimus, erityisesti Karin Kukkosen (2020) esittämä todennäköisyysrakenteiden (probability designs) analyysi. Kukkosen lähestymistapa liittyy läheisesti kognitiotieteen ennustavan käsittelyn hypoteesiin, joka puolestaan kytkeytyy bayesiläiseen tilastotieteeseen. Tutkielmassa tehdään tämä yhteys näkyväksi ja rakennetaan sen pohjalta kytkentä konkreettisiin laskennallisiin malleihin.

Sovelluskohteena tutkielmassa käytetään *Seitsemästä veljeksestä* johdettua sekundääristä aineistoa, puhujasekvenssiä, joka perustuu teoksen dialogeihin ja sisältää tietoa vain repliikkien esittäjistä. Tämän puhujasekvenssin analysoimiseen esitellään räätälöity laskennallinen malli, joka tunnistaa puhujasekvenssistä muutoskohtia. Mallin antamien tulosten pohjalta tutkielmassa erotetaan Kiven romaanista kahdeksan osin päällekkäistä jaksoa, jotka voidaan motivoida lähilukemisella ja aiemman tutkimuksen kautta.

Tutkielma osoittaa, että kognitiivisen kirjallisuudentutkimuksen ennustavan käsittelyn mallin pohjalta voidaan rakentaa laskennallisia apuvälineitä kaunokirjallisen teoksen analysoimiseen, ja että tällaisella lähestymistavalla voidaan tuottaa mielekästä ja hyödyllistä informaatiota *Seitsemästä veljeksestä*.

# Sisällys

# Luku 1

# Johdanto

> AAPO. Mutta jo on aika viisastua, aika on panna kaikki halut ja himot järjen ikeen alle ja etunenässä tehdä se, joka tuo hyötyä, vaan ei sitä joka makeammalle maistuu. [– –]
>
> (SV, 23–24.)

## 1.1 Tutkielman aihe ja tutkimuskysymykset

Aleksis Kiven *Seitsemän veljestä* (= SV, 2020/1870) on ensimmäinen suomenkielinen romaani ja kotimaisen kirjallisuuden merkkiteos[1]. Sitä lukevat ja tulkitsevat yhä uudet sukupolvet kouluissa, vapaa-ajalla ja kirjallisuudentutkimuksen piirissä. Tässä tutkielmassa laajennan teoksen lukutapojen kirjoa esittelemällä laskennallisen mallin *Seitsemän veljeksen* analysoimisen avuksi. Kehittämäni malli ikään kuin lukee *Seitsemän veljeksen* repliikkien puhujien nimistä koottua sekundääristä aineistoa ja tekee sen pohjalta ennusteita teoksen rakenteesta. Kyse ei kuitenkaan ole varsinaisesta ihmislukijan lukemisesta (ks. esim. Kortekallio & Ovaska 2020, 57), vaan teorian tasolla analogisesta konstruktiosta, laskennallisesta apuvälineestä.

Kirjallisuudentutkimus pitää sisällään monenlaisia lähestymistapoja lukijuuden ja lukemisen määrittelyyn ja tutkimiseen. Tutkielmassani otan lähtökohdakseni kognitiivisen kirjallisuudentutkimuksen, erityisesti Karin Kukkosen (2014a, 2014b, 2014c, 2020) *ennustavaan käsittelyyn* (*predictive processing*) perustuvan todennäköisyysrakenteiden (*probability designs*) mallin. Kukkosen mallin juuret ovat osin *bayesiläisessä tilastotieteessä*, ja hän viittaa alaa määrittävään Bayesin teoreemaankin (Kukkonen 2014a, 721; Kukkonen 2014c, 58; ks. myös luku 2.2.2). Tutkielmassani kehittelen Kukkosen mallia kvantitatiiviseen suuntaan ja luon sitä kautta yhteyden kognitiivisen kirjallisuudentutkimuksen ja laskennallisen kirjallisuudentutkimuksen välille.

Tutkimusaineistonani käytän Kiven *Seitsemää veljestä* paitsi sen keskeisen aseman ja tunnettuuden vuoksi, myös siksi, että teoksen rakenne soveltuu erityisen hyvin laskennallisen kirjallisuudentutkimuksen koetuskentäksi. Romaanin dialogivoittoisuus tekee sen rakenteesta poikkeuksellisen soveltuvan laskennallisten menetelmien sovelluskohteeksi. Lisäksi seitsemän päähenkilön tiiviistä yhteydestä seuraa, että dialogien muodostama aineisto on verrattain laaja ja rikas ja – kuten pyrin osoittamaan – heijastaa perustavanlaatuisella tavalla teoksen rakennetta.

Tutkielmani laajentaa ja kehittää kirjallisuudentutkimuksen teoriakenttää ja tuo samalla uusia näkökulmia *Seitsemään veljekseen* osoittamalla, että teoksen rakenteesta voi tehdä laskennallisen analyysin avulla mielenkiintoisia ja valaisevia havaintoja, joita voidaan myös monessa tapauksessa suhteuttaa aiempaan tutkimukseen. Tutkielmassani pyrin vastaamaan seuraaviin tutkimuskysymyksiin:

1. Kuinka kognitiivisen kirjallisuudentutkimuksen ennustavan käsittelyn mallia voidaan käyttää perustana laskennallisten teosanalyysin apuvälineiden kehittämiseen?
2. Soveltuuko *Seitsemän veljeksen* dialogien puhujien sekvenssi tällaisen laskennallisen analyysin kohteeksi, ja tuottaako se hyödyllistä informaatiota teoksesta?

## 1.2 Tutkielman ala ja aiempi tutkimus

Kognitiivisen kirjallisuudentutkimuksen tarkkarajainen määrittely on osoittautunut vaikeaksi: Lisa Zunshine (2015, 1) lainaa Alan Richardsonin toteamusta, jonka mukaan ala kattaa kaiken kognitiotieteestä kiinnostuneiden kirjallisuudentutkijoiden tutkimustyön monine eroavaisuuksineen. Tämän tutkielman kannalta keskeistä on lukijan ja lukemisen tutkimus kognitiivisen kirjallisuudentutkimuksen näkökulmasta; tätä käsittelen tarkemmin luvussa 2.1.

Laskennallisella kirjallisuudentutkimuksella (engl. *computational literary studies*) tarkoitan algoritmista ja tilastollista näkökulmaa kirjallisuudentutkimukseen. Ala kytkeytyy ja on osin synonyyminen digitaaliseen kirjallisuudentutkimukseen ja sitä kautta digitaalisiin ihmistieteisiin (Parente-Čapková 2021, Tolonen & Lahti 2015). Sen eräs näkyvä saavutus on ollut kirjallisuushistorian uudelleentarkastelu laajojen aineistojen kautta, vakiintuneen kaanonin hyläten (ks. esim. Moretti 2000, Bode 2018), mutta tässä tutkielmassa näkökulma kohdistuu yksittäisen teoksen analysoimiseen. Laskennallisessa näkökulmassa keskeistä on kytkeä kvantitatiivisen analyysin löydökset teoksen tulkintaan. Niklas Alén ym. (2018, 135) kirjoittavatkin: ”Jos veljesten välisiä eroja tarkastelee etäännytetysti [– –] heidän ominaislaaduistaan nousee esille monia sellaisia ominaisuuksia, joita voi erikseen tutkia *dialogeja lukien*” (kursivointi minun). Franco Moretti (2011, 84) on todennut vastaavasti, että laskennallisten mallien tuottamat reduktiot ja abstraktiot tekevät mahdolliseksi nähdä tekstissä piileviä rakenteita, joiden tunnistaminen voi ohjata tulkintaa. Tämä lähestymistapa väistää

myös Nan Dan (2019, 638) laskennallista kirjallisuudentutkimusta kohtaan esittämän kritiikin: reduktionismi ei ole itsetarkoitus vaan väline.

Tutkielmani kohdeteosta, Kiven *Seitsemää veljestä*, on toki tutkittu runsaasti. Viljo Tarkiaisen väitöskirja (1910) on edelleen kelpo katsaus romaaniin. Aarne Kinnunen on käsitellyt teosta peräti kahden monografian verran (1987, 2002) uuskriittisestä näkökulmasta, Pirjo Lyytikäinen (2004) taas on pohtinut *Seitsemän veljeksen* lajia. Kiven romaania ovat tarkastelleet varsinkin sen rakenteen osalta Rafael Koskimies (1958) ja Lauri Viljanen (1963). Tämän tutkielman kannalta on kiinnostavaa myös se, että *Seitsemää veljestä* on tutkittu kvantitatiivisestikin: Sirppa Kauppinen (1969) on tutkielmassaan pohtinut, miten veljesten ryhmätoiminnan tilastollinen analyysi voi vahvistaa impressionistisia havaintoja ja paljastaa uutta. Alén ym. (2018) ovat puolestaan tarkastelleet monipuolisesti muun muassa teoksen repliikkien jakautumista ja laajuutta, henkilödynamiikkaa sekä veljesten persoonapronominien käyttöä. Tämä tutkielma asettuu osaksi *Seitsemän veljeksen* tutkimuksen jatkumoa suhteuttamalla tutkimustuloksia aiempiin havaintoihin ja laajentamalla teoksen analysoimiseen soveltuvien laskennallisten apuvälineiden valikoimaa.

## 1.3 Tutkielman rakenne

Luvussa 2 esittelen tutkielman teoreettisen taustakehikon. Luku 2.1 käsittelee lukijan käsitettä erityisesti kognitiivisen kirjallisuudentutkimuksen näkökulmasta, luku 2.2 taas johdattaa kirjallisuudentutkimukseen perehtyneen lukijan matemaattisille poluille bayesiläisen tilastotieteen maailmaan. Viimeksi mainitussa pyrin melko didaktiseen esitystapaan.

Siirryn sitten käsittelemään nimenomaan *Seitsemää veljestä* luvussa 3. Kuvaan teoksesta muodostamani sekundäärisen aineiston, niin sanotun puhujasekvenssin, ja kerron, kuinka sen muutoskohtia on mahdollista mallintaa laskennallisesti. Perustelen myös, miksi lähestymistapani on mielekäs rajoituksistaan huolimatta.

Luvussa 4 tarkastelen edellisessä luvussa esittelemäni lähestymistavan tuottamia tuloksia. Osoitan *Seitsemästä veljeksestä* kahdeksan osittain päällekkäistä jaksoa, joiden merkitystä pohdin sekä lähilukemisen avulla että aiemman tutkimuksen valossa. Luvussa 5 arvioin tulosteni merkitystä tutkimuskysymyksieni valossa ja esitän suuntaviivoja jatkotutkimukseen.

# Luku 2

# Teoriakehys

> JUHANI. Mahdotonta oppia lukemaan.
> AAPO. Ihmisten tekemänä on tämä konsti ollut ennenkin.
>
> (SV, 37.)

## 2.1 Lukija ja todennäköisyys

### 2.1.1 Ennustava lukija

Kirjallisuudentutkimuksessa *lukijan* käsite ei rajoitu pelkästään todellisiin, fyysisiin lukijoihin, vaan lukija voi olla myös abstrakti konstruktio tai malli, joka tavalla tai toisella dekoodaa tai tulkitsee jotakin tekstiä tai tekstejä. Tällainen voi olla esimerkiksi tekijän ideaalinen lukija, joka ymmärtää kaikki tekijän vihjeet, viittaukset ja intentiot. (Prince 2013/2011.) Tässä tutkielmassa tarkastelen lukijaa *kognitiivisen kirjallisuudentutkimuksen* näkökulmasta (ks. esim. Zunshine 2015).

Kognitiivisen kirjallisuudentutkimuksen niin kutsutut toisen sukupolven lähestymistavat korostavat kehollisuuden keskeistä merkitystä (Kukkonen 2020, 65; ks. myös Kortekallio & Ovaska 2020, 59–60) ja asettuvat 4E-kognition sateenvarjokäsitteen alle (ks. esim. Newen ym. 2018). Tämän tutkielman kannalta erityisen tärkeä on *ennustava käsittely* (*predictive processing*), joka perustuu bayesiläiseen päättelyyn ja jolla on läheinen yhteys 4E-kognitioon (Hohwy 2018). Ennustavassa käsittelyssä oletetaan sisäinen malli, joka tuottaa ennusteita ja jota päivitetään uuden evidenssin valossa. Kirjallisuudentutkimuksen kontekstissa tätä voidaan havainnollistaa Kukkosen (2020, 5) *ennustaen käsittelevän lukijan* (*predictive processing reader*) kautta; jäljempänä käytän lyhennettyä käsitettä *ennustava lukija*.

Ennustavan lukijan malli perustuu käsitykseen, että kaikki ajattelu muovautuu todennäköisyysarvioiden ja niihin perustuvien ennusteiden kautta. Lukeminen hahmotetaan

lineaarisesti etenevänä prosessina, jossa lukija tekee ja päivittää ennusteitaan jatkuvasti edetessään tekstissä. Ennusteet voivat koskea tekstin eri tasoja, esimerkiksi yksittäisiä tavuja ja sanoja taikka juonta eri aikaskaaloineen. Eri tasojen ennusteet myös kytkeytyvät toisiinsa. (Kukkonen 2020, 1–3.) Malli muistuttaa Stanley Fishin (1970, 125–127, 135) kuvaamaa lukiessaan merkityksiä luovaa lukijaa, mutta Kukkosen lähestymistapa rakentuu kognitiotieteelliselle pohjalle ja korostaa erityisesti *ennustevirheiden* merkitystä.

Lukijan tekemien ennustevirheiden tärkeys tulee ilmeiseksi, kun huomataan, että kirjallisuus ei saa olla täysin ennustettavaa, koska muutenhan sitä ei tarvitsisi lukea (Kukkonen 2020, 1). Ennustevirheitä ei pidäkään tulkita lähtökohtaisesti negatiivisiksi tapahtumiksi (mts. 3). Lukijan tekemät ennustevirheet saavat aikaan yllätyksiä, uteliaisuutta ja jännitystä (Kukkonen 2014a, 726). Kukkonen (2020, 15–17) käyttää esimerkkinä laajalti tunnettua Charles Perrault'n versiota *Tuhkimo*-sadusta: Sadun alussa lukijalla on joukko hypoteeseja, joista lähtötilanteen perusteella todennäköisin on se, ettei Tuhkimo pääse osallistumaan tanssiaisiin. Kun haltiakummi muuttaa yllättäen taikasauvallaan kurpitsan vankkureiksi, lukija huomaa tehneensä ennustevirheen ja joutuu arvioimaan eri hypoteesien todennäköisyydet uudelleen. Tämän jälkeen hiirien muuttaminen hevosiksi ei ole enää erityisen yllättävää, ei myöskään se, että samat transformaatiot toistuvat seuraavana iltana tanssiaisiin lähdettäessä.

Lukijan tekemien ennusteiden ei välttämättä tarvitse olla sanallistettuja propositioita, vaan ne voivat olla myös ei-tietoisesti tehtyjä arvioita ja arveluja siitä, millaisia asiat ja tilanteet ovat (Kukkonen 2014b, 8; Kukkonen 2020, 1–2). Kukkonen havainnollistaa asiaa metaforalla sulkahattuisesta naisesta: hatun kantaja osaa varjella sulkaa osumasta mihinkään tekemättä erillisiä laskelmia, ja samaan tapaan ennustava lukija tekee ennusteensa näennäisen vaivattomasti ja kehollisuutta hyödyntäen (Kukkonen 2020, 7).

### 2.1.2 Todennäköisyysrakenteet

Lukijan tekemät ennustevirheet eivät synny sattumanvaraisesti, vaan kerronta on rakennettu tuottamaan ennustevirheitä tietyn ”lentoradan” mukaisesti. Nämä ennustevirheet syntyvät suunnitellussa järjestyksessä ja useilla eri tasoilla, ja niiden tarkoituksena on saada lukija korjaamaan ennusteitaan kerronnan edetessä ja kertomus päätymään tiettyyn lopetukseen. (Kukkonen 2020, 30.) Tämä rakenne voidaan samaistaa juoneen, ja juonenkäänteet voidaan yhdistää yllättäviin havaintoihin, jotka aiheuttavat lukijan ennustevirheen ja pakottavat tämän uudelleenarvioimaan eri hypoteesien todennäköisyyksiä (mts. 30, 45–48). Tällaiset *todennäköisyysrakenteet* (*probability designs*) eivät välttämättä ole kaunokirjallisen tekstin tekijän yksityiskohtaisesti suunnittelemia, mutta niiden avulla voidaan konstruoida malli lukijasta (mts. 5, 40).

Kukkonen esittää jaon kolmen eri kertaluvun todennäköisyysrakenteisiin. Edellä olen käsitellyt ennen kaikkea *ensimmäisen kertaluvun* todennäköisyysrakenteita, jotka liittyvät

juoneen ja kerronnan tapahtumiin ja saavat lukijan päivittämään todennäköisyysarvioitaan. *Toisen kertaluvun* todennäköisyysrakenteet taas liittyvät tyyliin, fokalisaatioon ja muihin kieleen kytkeytyviin keinoihin, ja niiden kautta syntyy yhteys kehollisuuteen muun muassa liikeverbien ja käsitemetaforien kautta (Kukkonen 2020, 4, 62, 65). *Kolmannen kertaluvun* todennäköisyysrakenteet puolestaan tuovat mukaan intertekstuaalisuuden (mts. 5, 105). Tässä tutkielmassa huomio kiinnittyy näistä tasoista vain ensimmäiseen, mutta aikaisempi *Seitsemän veljestä* -tutkimus liittyy myös esimerkiksi toisen kertaluvun todennäköisyysrakenteisiin Alénin ym. (2018, 123–129) sanastoanalyysin osalta ja kolmannen kertaluvun todennäköisyysrakenteisiin Lyytikäisen (2004) lajihistoriallisen näkökulman kautta.

Nämä kolmen eri kertaluvun todennäköisyysrakenteet voidaan ymmärtää myös ennustavan käsittelyn kautta. Ensimmäinen kertaluku koskee ennusteita ja ennustevirheitä, toisen kertaluvun tasolla moduloidaan ennusteiden *tarkkuutta* (*precision*) juonen ulkopuolisilla keinoilla (Kukkonen 2020, 3). Kolmas kertaluku kytkee mukaan tekstin ulkopuoliset, esimerkiksi genreen ja interteksteihin liittyvät odotukset siitä, mitkä hypoteesit ovat enemmän tai vähemmän todennäköisiä (mts. 105, 108–110). Kolmiportainen, todennäköisyysrakenteisiin perustuva malli tuo lisää syvyyttä perinteisiin hermeneuttisiin malleihin ja mahdollistaa lukuprosessin mallintamisen lukemisen eri tapojen ja tasojen kautta (Kukkonen 2020, 143–144).

## 2.2 Laskennallinen näkökulma ja bayesiläinen tilastotiede

### 2.2.1 Laskennalliset apuvälineet kirjallisuudentutkimuksessa

Kuten edellä on todettu, ihminen tekee ennusteitaan paljolti ei-tietoisesti. Analyyttisimpia lukijoita lienevät kirjallisuudentutkijat, jotka tekevät tutkimistaan teoksista monenlaisia havaintoja ja muodostavat ja sanallistavat kokonaistulkintoja. Tätä taitoa kutsutaan kirjalliseksi kompetenssiksi (ks. esim. Culler 1975, 113–114; Stockwell 2002, 20). Kuitenkin tutkijoillakin kokonaistulkintaan johtavat havainnot ja päätelmät voivat joskus olla melko impressionistisia. Tämä ei tarkoita, etteivätkö ne olisi myös osuvia ja oikeita, mutta ainakin lienee selvää, että tällaiset tulkinnat voisivat hyötyä myös täsmällisemmästä evidenssistä.

Tässä tutkielmassa pyrin osoittamaan, että todennäköisyysrakenneanalyysiä on mahdollista laajentaa edelleen laskennallisten menetelmien avulla. Todennäköisyysrakenteidenhan ei Kukkosen (2020, 178) mukaan ole tarkoitus kuvata todellisia kognitiivisia prosesseja, vaan nimenomaan kirjallisuudelle ominaisia ajattelun tapoja. Yhtä lailla laskennalliset menetelmätkään eivät simuloi todellisia kognitiivisia prosesseja, mutta ne voivat tunnistaa ja tuoda näkyville sellaisia kaunokirjalliseen tekstiin sisältyviä todennäköisyysrakenteita, joita ihmislukijan on vaikea tunnistaa. Näin analyysiin saadaan lisää tarkkuutta ja moniulotteisuutta.

Kukkosen todennäköisyysrakenteet ja merkittävä osa kognitiotieteestä ylipäänsä perustuvat *bayesiläiseen* tilastotieteeseen (Kukkonen 2014a, 721; Hohwy 2018; Clark 2013). Bayesiläinen lähestymistapa perustuu todennäköisyyden tulkintaan subjektiivisena epävarmuuden mittarina, ei toistokokeiden kautta saatavana suureena kuten frekventistinen tilastotiede (Gelman ym. 2013, 11–13). Tämä vaikuttaa luonnolliselta kirjallisuudentutkimuksen kannalta jo siksi, että yksittäisen teoksen lukemiseen liittyviä todennäköisyyksiä ei ole mielekästä hahmottaa toistokokeiden kautta, koska peräkkäiset lukukerrat eivät ole toisistaan riippumattomia.

### 2.2.2 Bayesiläinen tilastotiede

Bayesiläinen tilastotiede perustuu Thomas Bayesin (1763) esittämään Bayesin teoreemaan

$$P(H \mid D) = \frac{P(D \mid H)\, P(H)}{P(D)}.$$

Näin muotoiltuna $H$ kuvaa jotakin hypoteesia ja $D$ kuvaa jotakin havaintoa tai dataa. Teoreeman avulla voidaan laskea todennäköisyys $P(H \mid D)$, joka kertoo, millä todennäköisyydellä hypoteesin $H$ arvioidaan pitävän paikkansa, kun on tehty havainto $D$. Tämä *posterioritodennäköisyys* saadaan laskettua teoreeman oikealla puolella olevien suureiden avulla: $P(D \mid H)$ kuvaa, kuinka todennäköinen havainto $D$ on, jos hypoteesi $H$ pätee, ja $P(H)$ eli niin sanottu *priori* kuvaa ennakko-odotuksia hypoteesin pätevyydestä ennen kuin havainto on tehty. Bayesiläiset mallit pyritään konstruoimaan sellaisiksi, että $P(D \mid H)$ ja $P(H)$ voidaan määrittää laskennallisesti. Lausekkeen $P(D)$ arvo ja merkitys voidaan tässä yhteydessä sivuuttaa matemaattisista syistä (ks. esim. Gelman ym. 2013, 7).

Bayesin teoreeman käyttöä voidaan havainnollistaa soveltamalla Kukkosen (2020, 15–17) edellä jo sivuttua esimerkkiä *Tuhkimosta*. Sadun alkuvaiheilla voidaan tunnistaa kaksi vaihtoehtoista hypoteesia: Tuhkimo ei pääse tanssiaisiin ($H_a$) ja Tuhkimo pääsee tanssiaisiin ($H_b$). Nimetään vielä sadun tapahtumat peräkkäisiksi havainnoiksi $E_1, E_2, \ldots, E_n$. Prioritodennäköisyydet $P(H_a)$ ja $P(H_b)$ kuvaavat lukijan ennakko-odotuksia hypoteeseista ennen kuin hän on aloittanut sadun lukemisen. Kun lukija etenee sadun ensimmäiseen tapahtumaan $E_1$, kummankin hypoteesin arvioiduiksi todennäköisyyksiksi saadaan Bayesin teoreemalla

$$P(H_a \mid E_1) = \frac{P(E_1 \mid H_a)\, P(H_a)}{P(E_1)} \quad \text{ja}$$
$$P(H_b \mid E_1) = \frac{P(E_1 \mid H_b)\, P(H_b)}{P(E_1)}.$$

Matemaattisista syistä suureesta $P(E_1)$ ei tarvitse välittää (sen arvo ei riipu siitä, mitä hypoteesia tarkastellaan), vaan oleellista on vain arvioida, kuinka todennäköiseksi hypoteesit

$H_a$ ja $H_b$ ennustavat tapahtuman $E_1$. Toisin sanoen käytössä oleva matemaattinen malli tuottaa arviot todennäköisyyksistä $P(E_1 \mid H_a)$ ja $P(E_1 \mid H_b)$, ja näiden ja prioritodennäköisyyksien avulla saadaan laskettua tapahtuman $E_1$ valossa päivitetyt todennäköisyydet hypoteeseille $H_a$ ja $H_b$.

Kun sitten lukija etenee sadun seuraavaan tapahtumaan $E_2$, Bayesin teoreeman soveltaminen jatkuu:

$$P(H_a \mid E_1, E_2) = \frac{P(E_1, E_2 \mid H_a)\, P(H_a)}{P(E_1, E_2)},$$

minkä voidaan osoittaa olevan yhtäpitävä seuraavan muodon kanssa:

$$P(H_a \mid E_1, E_2) = \underbrace{P(E_2 \mid H_a, E_1)}_{\text{(i)}} \cdot \underbrace{\frac{P(E_1)}{P(E_1, E_2)}}_{\text{(ii)}} \cdot \underbrace{\frac{P(E_1 \mid H_a)\, P(H_a)}{P(E_1)}}_{\text{(iii)}}.$$

Tässä yhtäsuuruusmerkin oikea puoli jakautuu kolmeen lausekkeeseen. Lauseke (i) on ennuste tapahtuman $E_2$ todennäköisyydestä, jos oletetaan hypoteesi $H_a$ ja on jo havaittu tapahtuma $E_1$. Lauseke (ii) voidaan sivuuttaa, koska se ei riipu tarkasteltavasta hypoteesista. Lauseke (iii) taas on sama kuin edellä saatu arvio hypoteesin $H_a$ todennnäköisyydestä, kun on havaittu vasta tapahtuma $E_1$. Koska siis lausekkeella (ii) ei ole merkitystä ja lauseke (iii) on laskettu jo aiemmin, riittää laskea seuraavan yksittäisen tapahtuman $E_2$ arvioitu todennäköisyys, jotta saadaan päivitetty arvio hypoteesin $H_a$ todennäköisyydestä. Vastaavat yhtälöt toteutuvat yhtä lailla hypoteesille $H_b$.

Näin rakentuvassa laskennallisessa prosessissa on kyse nimenomaan ennustavasta käsittelystä, ja malli vastaa ennustavan lukijan mallia: uusia havaintoja käsitellään yksi kerrallaan ja ennusteita päivitetään jatkuvasti. Edellä esitetty laskennallinen prosessi ei rajoitu pelkästään ensimmäiseen ja toiseen havaintoon, vaan se yleistyy tapahtumasekvenssille $E_1, E_2, \ldots, E_n$:

$$\begin{aligned} P(H_a \mid E_1, E_2, \ldots, E_n) = {} & P(E_n \mid H_a, E_1, E_2, \ldots, E_{n-1}) \\ & \cdot \frac{P(E_1, E_2, \ldots, E_{n-1})}{P(E_1, E_2, \ldots, E_n)} \\ & \cdot P(H_a \mid E_1, E_2, \ldots, E_{n-1}). \end{aligned}$$

Kukkosta (2020, 15–17) seuraten voidaan siis olettaa, että $P(H_a)$ on suuri ja $P(H_b)$ on pieni, toisin sanoen ennakko-odotus eli priori on, että Tuhkimo ei pääse tanssiaisiin. Merkitään kurpitsan muuttumista vankkureiksi tapahtumalla $E_k$. Ennen tätä havaintoa hypoteesien $H_a$ ja $H_b$ todennäköisyysarviot vastaavat suurin piirtein ennakko-oletuksia: kaikilla $i < k$ saadaan arviot, että $P(H_a \mid E_1, E_2, \ldots, E_i)$ on suuri ja $P(H_b \mid E_1, E_2, \ldots, E_i)$ on pieni. Mutta kurpitsan muuttuminen vankkureiksi on todennäköisempää hypoteesin $H_b$ kuin hypoteesin $H_a$ vallitessa, toisin sanoen todennäköisyys $P(E_k \mid H_b, E_1, E_2, \ldots, E_{k-1})$ on selvästi suurempi

kuin $P(E_k \mid H_a, E_1, E_2, \ldots, E_{k-1})$. Tätä kautta todennäköisyys $P(H_b \mid E_1, E_2, \ldots, E_k)$ nouseekin suuremmaksi kuin todennäköisyys $P(H_a \mid E_1, E_2, \ldots, E_k)$. Ennen tapahtumaa $E_k$ todennäköisemmältä näyttänyt hypoteesi (Tuhkimo ei pääse tanssiaisiin) johtaa siis ennustevirheeseen, joka pakottaa radikaaliin hypoteesien todennäköisyyksien uudelleenarvioimiseen. Lukija alkaa arvella, että Tuhkimo saattaa sittenkin päästä tanssiaisiin.

### 2.2.3 Laskennallisen mallin laatiminen

Kuinka edellä kuvatun kaltainen bayesiläinen laskennallinen työkalu sitten voidaan rakentaa? Lähtökohtana voi toimia esimerkiksi, että ensin valitaan tutkittavasta tekstistä jokin piirre

- joka voidaan koodata koneluettavaan muotoon (automaattisesti tai ihmisvoimin)
- jonka vaihtelu aineistossa on rikasta ja monimuotoista
- jonka osalta vaikuttaa uskottavalta, että sen vaihtelu sisältää tulkinnallisesti mielenkiintoista informaatiota tutkittavasta teoksesta.

Nämä ehdot täyttäviä piirteitä voivat olla tilanteesta riippuen esimerkiksi virkkeen pituus, eri vokaalien esiintymistiheydet (vrt. Rossi 2022) tai *Seitsemän veljeksen* tapauksessa se, kuka veljeksistä on milloinkin äänessä. Tällä tavoin saadaan alkuperäisestä aineistosta muodostettua sekundäärinen aineisto, joka esimerkiksi viimeksi mainitun piirteen tapauksessa koostuu pitkästä luettelosta veljesten nimiä (jäljempänä kutsun tällaista luetteloa *puhujasekvenssiksi*).

Tämän jälkeen tarvitaan valittua piirrettä kuvaava bayesiläinen todennäköisyysmalli. *Seitsemän veljeksen* puhujasekvenssin osalta malli voi olla esimerkiksi sellainen, joka ennustaa, millä todennäköisyydellä kukin veljeksistä on seuraavan repliikin esittäjä. Luonnollinen priori eli ennakko-odotus on se, että ennen romaanin lukemista kunkin veljeksen todennäköisyys ajatellaan yhtä suureksi (tämä on ns. epäinformatiivinen priori, ks. esim. Gelman ym. 2013, 51). Yksinkertaisimmillaan tällainen malli vain pitää kirjaa siitä, kuinka monta kertaa kukin veljeksistä on siihen saakka käyttänyt puheenvuoron, ja tekee ennusteen tämän perusteella. Jos esimerkiksi teoksen sadan ensimmäisen veljesten esittämän repliikin jälkeen Lauri on käyttänyt vaikkapa kahdeksan puheenvuoroa, malli antaa 8 %:n todennäköisyyden sille, että Lauri esittää seuraavan repliikin. Tällainen malli ei kuitenkaan ole erityisen mielenkiintoinen eikä hyödyllinen, koska se antaa yhtä suuren painoarvon teoksen kaikille repliikeille, vaikka esimerkiksi seitsemännessä luvussa aloitusluvun tapahtumat ovat jo kaukaista menneisyyttä ja kuudennen luvun tapahtumat paljon keskeisempiä.

Seuraavassa luvussa esittelen juuri *Seitsemää veljestä* varten räätälöidyn, edellistä esimerkkiä hienostuneemman todennäköisyysmallin, joka pyrkii ennustamaan, kuka veljeksistä puhuu seuraavaksi. Malli tunnistaa aineistosta kohdat, joissa veljesten suhteelliset repliikkien osuudet muuttuvat merkittävästi, toisin sanoen kohdat, joissa yksi tai useampi veljes alkaa puhua aiempaa selvästi enemmän tai vähemmän. Tulen osoittamaan, että tällaiset kohdat voivat

ilmentää oleellista muutosta esimerkiksi tarinan käsillä olevassa tilanteessa tai veljesten keskinäisessä dynamiikassa.

# Luku 3

# *Seitsemän veljeksen* puhujasekvenssin laskennallinen analyysi

> Siinä seisoi pihalla paljon kansaa, miehiä ja naisia, ja kauas kuului eräs ääni, joka räknäili: ”ensimmäinen, toinen, kolmas kerta”, ja kyseli: ”eikös kukaan enemmin lisää?”
>
> (SV, 282.)

## 3.1 Puhujasekvenssi sekundäärisenä aineistona

*Seitsemän veljestä* on romaaniksi huomattavan dialogipainotteinen, ja dialogi on lisäksi varsin riippumatonta muusta kerronnasta. Aarne Kinnunen (1987, 72) on arvioinut dialogin osuudeksi teoksessa 62 %, ja hän toteaa, että ”[teoksen dialogi] sisältää tilanteesta kaiken tarpeellisen informaation ja että yleensä muuta informaatiota ei anneta” (mts. 67; ks. myös Kinnunen 2002, 115–116). Vuorosanoissaan veljekset antavat itsensä kokonaan, kuten Koskimies (1958, 59) on todennut. Tapahtumien kulku selviää lukijalle dialogien kautta, kuten seuraavassa katkelmassa, jossa kuvataan yksinomaan dialogimuotoisella kerronnalla, kuinka ovi teljetään ja ikkuna rikotaan:

> TUOMAS. Ovi hakaan, Eero!
> EERO. Niin juuri; linnan isoportti kiini pataljonan marssiessa västingin takaportista ulos. – Ha’assa on ovi.
> AAPO. Minä varoitan teitä!
> JUHANI. Tehty on tehty. Katsos tuossa!
> AAPO. Sinä hirmuinen, julkijumalaton!
> SIMEONI. Kas niin! Se on tehty! Siinähän akkuna sälähti!

JUHANI. Akkuna sälähti ja taivas välähti, kun kerran vaan keikahti Jussin pussi!
Se oli Laiska-Jaakon mälli.

(SV, 66.)

Teoksen dialogeista voidaan muodostaa sekundäärinen aineisto tarkastelemalla yksinomaan repliikkien puhujia. Esimerkiksi yllä oleva katkelma voidaan typistää muotoon TUOMAS–EERO–AAPO–JUHANI–AAPO–SIMEONI–JUHANI. Tässä tutkielmassa tutkin aineistoa, joka käsittää tällä tavoin koko *Seitsemän veljeksen* dialogeista muodostetun puhujasekvenssin. Aineistosta on poistettu ne 7 % repliikeistä, joiden puhujana ei ole kukaan veljeksistä (Alén ym. 2018, 118). Lopputuloksena saadaan edellisen kaltainen, 2 152 nimen mittainen puhujasekvenssi. Puhujasekvenssiaineistoni pohjautuu Project Gutenbergin sähköisesti julkaisemaan *Seitsemän veljestä* -editioon (Kivi 2004/1870).

Aineisto ei siis sisällä mitään informaatiota siitä, milloin yksi dialogi tai luku vaihtuu toiseen tai milloin dialogien välissä on kertovaa tekstiä (esimerkiksi Aapon sisäkertomuksia). Myös tieto repliikkien pituuksista ja sisällöistä menetetään. Miksi sitten on perusteltua ajatella, että tätä pelkistettyä puhujasekvenssiä olisi mielekästä tutkia?

Ensinnäkin vaikuttaa uskottavalta hypoteesilta, että sillä, ketkä veljeksistä esittävät enemmän ja ketkä vähemmän repliikkejä, on yhteys romaanin juoneen. Räikein esimerkki on *Seitsemän veljeksen* kahdeksas luku, jossa muuten vähäpuheinen Lauri ”[– –] viinapäissään [– –] vilkastuu, muuttuu melkein nerokkaaksi ja laukaisee suustansa purevat, naulanpäähän iskevät arvostelut veljistänsä” (Tarkiainen 1910, 69). Tapahtumasarjan alun voi helposti nähdä Laurin runsaista repliikeistä, sen lopun (ja Laurin sammumisen) taas repliikkien loppumisesta. Puhujasekvenssi siis pitää sisällään teoksen kannalta relevanttia informaatiota. Kukkosen terminologialla (ks. luku 2.1.2) kyse on ensimmäisen kertaluvun todennäköisyysrakenteesta.

Toiseksi on mahdollista, että puhujasekvenssi pitää sisällään informaatiota veljesten keskinäisestä dynamiikasta. Tarkastellaan esimerkiksi Kinnusen (1987, 48–49) tulkintaa, jonka mukaan romaanin kuudennessa luvussa Juhani, veljeksistä vanhin, menettää keskeisen asemansa, ja johtoon astuvat Tuomas ja Aapo. Puhujasekvenssiä analysoimalla on mahdollista tutkia, tapahtuuko teoksen samoilla paikkeilla muutoksia siinä, kuinka usein kukin tästä kolmikosta lausuu repliikin.

Edelliset kaksi yritystä motivoida puhujasekvenssin tutkimusta perustuvat ajatukseen, että sekundääristä aineistoa laskennallisesti analysoimalla olisi kenties mahdollista löytää *Seitsemästä veljeksestä* aiemmin esitetyille havainnoille tai tulkinnoille laskennallista evidenssiä tai vastaevidenssiä. Toisaalta on myös ajateltavissa, että puhujasekvenssin laskennallinen tarkastelu voi tuoda kirjallisuudentutkijan näkyville teoksen rakenteesta jotakin sellaista, mitä ei ole aiemmin huomattu, ja siten ohjata lähilukemista teoksen tiettyihin osuuksiin.

## 3.2 Muutoskohdat puhujasekvenssissä

Edellä kuvatun *Seitsemän veljeksen* puhujasekvenssin voi halutessaan jakaa osiin esimerkiksi kohtauksien tai lukurajojen mukaan. Muunkinlaiset jaot ovat kuitenkin mahdollisia. Seuraavassa tarkastelen lähestymistapaa, jossa puhujasekvenssi pyritään jakamaan osiin, joiden sisällä veljesten repliikkien suhteelliset osuudet pysyvät jotakuinkin samoina.

Lähestymistapaa voi havainnollistaa yksinkertaistetulla keinotekoisella esimerkillä. Tarkastellaan puhujasekvenssiä, joka perustuu kolmen eri henkilöhahmon (A, B ja C) repliikkeihin. Repliikkejä on yhteensä 250 kappaletta, joten puhujasekvenssin pituus on noin 12 % *Seitsemän veljeksen* puhujasekvenssistä. Puhujasekvenssi on seuraava (viidelle riville jaettuna):

```
1 BCAABBACABCACABABBABAACAABAAABBAAAAABCCCABAACBABAA
2 ACBAAABCBACABAABBBCAABBACACABBBBCACAAAAAACCABAAABC
3 CABCCBCCBBBACCBACCBABCAACABCBACBBBCACCBBBABBCCBBCC
4 BBACAABACAABACCCCAACACACACCCCCCABCAABCCBCACAACBBCB
5 ACCBBCABCAAACCACACBACAAAAABCCCACBAACCABCABACBCABAA
```

Tässä puhujasekvenssissä repliikit jakautuvat siten, että A:lla on 100 repliikkiä, B:llä on 70 repliikkiä ja C:llä on 80 repliikkiä. Ihmissilmän on vaikea havaita puhujasekvenssistä mitään kiinnostavaa. Tarkastellaan sitten puhujasekvenssiä 50 repliikin katkelmina. Katkelmakohtaiset suhteelliset osuudet on esitetty taulukossa 1. Näin ryhmiteltynä huomataan, että ensimmäisten sadan repliikin (rivit 1–2) ajan noin puolet kaikista repliikeistä on A:n ja loput repliikit jakautuvat B:lle ja C:lle noin suhteessa 3:2. Repliikeissä 101–150 (rivi 3) A:n osuus pienenee noin 20 %:iin ja loput repliikeistä jakautuvat tasan B:n ja C:n välille. Repliikeissä 151–250 (rivit 4–5) C:n osuus pysyy jotakuinkin samana, A:n osuus kohoaa noin 40 %:iin ja B:n osuus laskee vastaavasti. Muutokset näyttäytyvät nyt melko selkeinä etenkin A:n osalta, jonka repliikkien suhteellinen osuus muuttuu radikaalisti kahteen otteeseen.

Jos kyseessä olisi todellinen (dialogivoittoinen) kaunokirjallinen teos, nämä havainnot voisivat ohjata kiinnittämään huomiota siihen, miten repliikkien 1–100, 101–150 ja 151–250 määrittämät osat teoksesta eroavat toisistaan. Havainnot herättävät monenlaisia kysymyksiä:

**Taulukko 1:** Luvun 3.2 esimerkkipuhujasekvenssin henkilöhahmojen suhteelliset osuudet viidenkymmenen repliikin ryhmissä.

| Rivi | Repliikit | A | B | C |
|---|---|---|---|---|
| 1 | 1–50 | 52 % | 30 % | 18 % |
| 2 | 51–100 | 48 % | 30 % | 22 % |
| 3 | 101–150 | 20 % | 40 % | 40 % |
| 4 | 151–200 | 36 % | 20 % | 44 % |
| 5 | 201–250 | 44 % | 20 % | 36 % |

Mikä saa aikaan A:n suhteellisen osuuden vaihdokset? Onko kyseessä jokin selkeä juoneen liittyvä ilmiö? Tapahtuuko henkilöhahmojen välillä jotakin merkittävää repliikkien 101 ja 151 paikkeilla?

Tässä keinotekoisessa esimerkissä A:n, B:n ja C:n repliikkien suhteelliset osuudet on asetettu vaihtumaan juuri repliikkien 101 ja 151 kohdalla. Todellisessa aineistossa, esimerkiksi *Seitsemän veljeksen* puhujasekvenssissä, ei tietenkään ole tällaista suoraviivaista ja tunnettua rakennetta. Käytännössä syntyy kaksi käytännön ongelmaa. Ensinnäkin, muutokset henkilöhahmojen repliikkien osuuksissa voivat tapahtua vaiheittain yksittäisten äkillisten muutosten asemesta, jolloin ei ole mielekästä yrittää osoittaa tiettyä yksittäistä kohtaa, jossa muutos tapahtuu. Toiseksi, ei ole etukäteen tiedossa, kuinka usein tai missä kohtaa henkilöhahmojen repliikkien suhteelliset osuudet ylipäätään vaihtuvat. Seuraavaksi esittelen mallin, joka ratkaisee molemmat ongelmat.

## 3.3 Muutoskohtien laskennallinen tunnistaminen

Muutoskohtien tunnistaminen aikasarjamuotoisista aineistoista on monitahoinen ongelma, johon on esitetty laaja joukko erilaisia ratkaisuja ja jolla on sovelluksia lukuisilla tieteenaloilla (Aminikhanghahi & Cook 2017). Tässä tutkielmassa tarkasteltava *Seitsemän veljeksen* puhujasekvenssi on myös tulkittavissa aikasarjaksi: kyseessä on aineisto, joka kuvaa tietyn järjestelmän tai prosessin aiheuttamaa havaintosekvenssiä. Kun tutkitaan luvun 3.2 tapaan sitä, milloin veljesten repliikkien suhteellinen osuus muuttuu merkittävästi, kyse on nimenomaan muutoskohtien etsimisestä.

Tämän tutkielman kannalta erityisen mielenkiintoinen on Ryan Prescott Adamsin ja David J. C. MacKayn (2007) *Bayesian Online Changepoint Detection* -malli ja siihen liittyvä algoritmi. Adamsin ja MacKayn malli vastaa rakenteeltaan luvussa 2.2.2 kuvattua bayesiläistä ennustavaa käsittelyä ja sitä on sovellettu muun muassa pörssikurssien ja kaivosonnettomuuksien analysoimiseen. Malli on suunniteltu nimenomaan havainto kerrallaan ennustamiseen (mihin sana *online* itse asiassa viittaa) ja ennusteiden päivittämiseen sitä mukaa. Sen keskeisiä oletuksia ovat, että (*i*) aineisto jakautuu peräkkäisiin, toisistaan erillisiin jaksoihin ja (*ii*) kunkin tällaisen jakson sisällä datapisteitä voidaan mallintaa jonkin tietyn todennäköisyysjakauman avulla. Jaksojen rajakohtia tai niihin liittyviä todennäköisyysjakaumia ei tunneta etukäteen, vaan malliin liittyvä algoritmi pyrkii arvioimaan niitä aineistosta sitä käsitellessään. *Seitsemän veljeksen* puhujasekvenssin tapauksessa nämä oletukset merkitsevät sitä, että teoksen oletetaan jakautuvan peräkkäisiin jaksoihin siten, että jaksosta seuraavaan siirryttäessä veljesten repliikkien suhteellinen osuus muuttuu aina merkittävästi. Tarkastelen seuraavaksi näitä oletuksia hieman tarkemmin *Seitsemän veljeksen* ja sen puhujasekvenssin kannalta.

Edellä käsitellyssä taulukon 1 esimerkissä edellä mainitut oletukset täyttyvät eksaktisti, koska kyseinen aineisto on luotu synteettisesti näiden oletusten pohjalta. Ei välttämättä tunnu lähtökohtaisesti mielekkäältä olettaa, että *Seitsemässä veljeksessä* olisi yhtä teräväarajaisia muutoskohtia, vaan mahdolliset vaihtelut henkilöhahmojen suhteellisissa osuuksissa lienevät hienovaraisempia ja häilyväisempiä. On kuitenkin muistettava, että kyseessä on malli, abstraktio, jonka tarkoituksena on tunnistaa suuria linjoja yksinkertaistamisen kautta. Jäljempänä nähdään, että bayesiläinen lähestymistapa mahdollistaa tarkan rajanvedon välttämisen sitä kautta, että mallin tuloksiin liittyy epävarmuusarvioita. Joissakin tapauksissa muutoskohtia on sitä paitsi romaanitekstin lähiluvun avulla kyllä mahdollista erottaa tarkastikin – ajatellaan esimerkiksi Laurin nukahtamista Hiidenkivellä – mutta puhujasekvenssiaineistossa tällaista informaatiota ei ole käytettävissä, vaan Laurin hiljeneminen täytyy tunnistaa vähitellen siitä, ettei hän tietystä kohdasta alkaen käytä yhtäkään repliikkiä. Tämä esimerkki havainnollistaa sitä, miten laskennallinen näkökulma voi tukea ja ohjata lähilukua: laskennallisen mallin tunnistamat yleiset suuntaviivat ohjaavat kohdistamaan lähilukua tiettyyn osaan teosta.

*Seitsemän veljeksen* puhujasekvenssin osalta oletus havaintojen mallintamisesta jollakin todennäköisyysjakaumalla kunkin jakson sisällä voidaan kuvata täsmällisemmin seuraavasti. Kutakin oletettua jaksoa vastaa jokin todennäköisyysjakauma, joka kertoo, millä todennäköisyydellä jakson sisältä satunnaisesti valittu repliikki on kunkin veljeksen esittämä. Todennäköisyysjakauma tarkoittaa siis tässä yhteydessä samaa kuin veljesten suhteelliset osuudet jakson repliikeistä. Taulukossa 2 on esitetty esimerkki tällaisesta todennäköisyysjakaumasta. Tällainen mallintaminen jakson sisällä on periaatteessa voimakas yksinkertaistus, koska se olettaa, ettei repliikin esittäjän todennäköisyyteen vaikuta lainkaan se, kuka on esittänyt edellisen repliikin. Kuitenkin tiedämme, ettei *Seitsemässä veljeksessä* sama veljes esitä tyypillisesti kahta peräkkäistä repliikkiä ja että edellisen puhujan henkilöllisyys kyllä vaikuttaa siihen, kuka luultavammin puhuu seuraavaksi (Alén ym. 2018, 121–123). Monimutkaisempi lähestymistapa kuitenkin lisäisi mallin matemaattista mutkikkuutta merkittävästi ja saattaisi helposti johtaa epäuskottaviin tuloksiin. Tämä johtuu siitä, että jo mahdollisia kahden veljeksen pareja on $7 \times 6 = 42$ kappaletta, ja näin usean todennäköisyyden arvioiminen mahdollisesti melko lyhyistäkin puhujasekvenssin jaksoista ei välttämättä ole tilastollisesti mielekästä. Lisäksi edellä esitettyjen esimerkkien valossa on selvää, että yksinkertainenkin jaksokohtainen todennäköisyysjakauma voi sisältää potentiaalisesti kiinnostavaa tietoa vaikkapa yksittäisen veljeksen repliikkien suhteellisesta osuudesta.

Tässä tutkielmassa sovellan *Seitsemän veljeksen* puhujasekvenssiin tätä tarkoitusta varten räätälöityä versiota Adamsin ja MacKayn mallista[2]. Seuraavaksi havainnollistan askel kerrallaan, kuinka malli toimii. Algoritmiseen ja ohjelmointitekniseen toteutukseen liittyvät yksityiskohdat sivuutan tässä tutkielmassa[3].

**Taulukko 2:** Esimerkki todennäköisyysjakaumasta. Tämä jakauma vastaa seitsemän veljeksen suhteellisia osuuksia veljessarjan kaikista repliikeistä koko romaanin tasolla. Ks. myös Alén ym. (2018, 119).

| **Veljes** | **Todennäköisyys** |
|---|---|
| Juhani | 40 % |
| Tuomas | 11 % |
| Aapo | 15 % |
| Simeoni | 10 % |
| Timo | 12 % |
| Lauri | 4 % |
| Eero | 8 % |
| *Yhteensä* | 100 % |

## 3.4 Analyysi käytännössä

Ennen kuin malli on havainnut yhtään *Seitsemän veljeksen* repliikkiä, tarvitaan priorijakauma eli jonkinlainen ennakko-odotus siitä, kuinka veljesten repliikit saattaisivat jakautua. Koska tarkoituksena on antaa aineiston puhua puolestaan ja tehdä ennustavaa käsittelyä havainto kerrallaan, on luonnollista aloittaa priorista, joka olettaa kaikkien veljesten suhteelliset osuudet yhtä suuriksi. Toisin sanoen jokaisen veljeksen suhteelliseksi osuudeksi oletetaan noin 14,3 % ennen kuin yhtään repliikkiä on havaittu. Matemaattisista syistä tämä priori on suoraviivaisinta toteuttaa olettamalla, että jokainen veljeksistä olisi ikään kuin ennen romaanin alkua esittänyt yhden repliikin[4]. Tätä alkutilannetta voidaan kuvata muodossa $(1, 1, 1, 1, 1, 1, 1)$, jossa on lueteltuna ikäjärjestyksessä[5] kunkin veljeksen esittämien repliikkien lukumäärä. Ennustettu todennäköisyys sille, että tietty veljeksistä on seuraavan repliikin esittäjä, saadaan jakamalla veljeksen kohdalla oleva luku kaikkien lukujen summalla: $1/(1 + 1 + 1 + 1 + 1 + 1 + 1) = 1/7 \approx 0{,}14 = 14\ \%$.

Sitten havaitaan romaanin ensimmäinen repliikki:

> JUHANI. Elettiinpä ennenkin,
> Vaikk' ojan takan' oltiin. . . . . .
> Mutta peijakas meidän kuitenkin viimein täällä perii. Se on niinkuin kädessä, te mullisaukon pojat.
>
> (SV, 19.)

Havainto ei tietenkään ole mallin näkökulmasta yllättävä, koska sen ennuste kunkin veljeksen todennäköisyydeksi oli sama. Nyt on kuitenkin havaittu Juhanilta uusi repliikki, minkä seurauksena lukuarvoa Juhanin paikalla kasvatetaan yhdellä; saadaan $(2, 1, 1, 1, 1, 1, 1)$. Siispä mallin ennuste seuraavan repliikin esittäjästä onkin, että Juhanin todennäköisyys on $2/8 = 25\ \%$ ja muiden veljesten todennäköisyys on $1/8 = 12{,}5\ \%$.

Teoksen toisen repliikin esittää Aapo. Lukuarvoa hänen paikallaan kasvatetaan yhdellä ja saadaan $(2, 1, 2, 1, 1, 1, 1)$. Kolmannen repliikin esittäjäksi ennustetaan nyt Juhania ja Aapoa todennäköisyydellä $2/9 \approx 22\ \%$ ja muita veljeksiä todennäköisyydellä $1/9 \approx 11\ \%$.

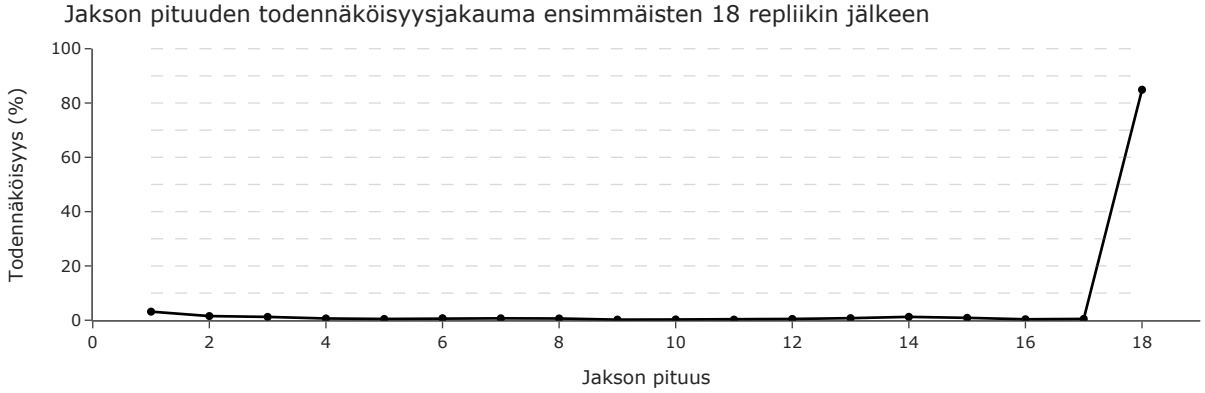
Jos mallinnuksessa ei olisi pyrkimyksenä tunnistaa, milloin veljesten suhteelliset osuudet muuttuvat, voitaisiin yksinkertaisesti jatkaa tällä tavoin teoksen loppuun saakka. Teoksen ensimmäisen dialogin päättyessä tilanne olisi $(13, 7, 6, 5, 2, 3, 1)$ ja ennuste seuraavan repliikin esittäjästä olisi seuraava: Juhani: 35 %, Tuomas: 19 %, Aapo: 16 %, Simeoni: 14 %, Timo: 5 %, Lauri: 8 % ja Eero: 3 %. Asetelma ei kuitenkaan ole aivan näin suoraviivainen, koska tavoitteena on myös löytää muutoskohtia.

Muutoskohtien tunnistaminen perustuu siihen, että kunkin repliikin kohdalla malli arvioi *senhetkisen jakson pituutta*. Tietyn repliikin kohdalla jakson pituus voi olla pienimmillään yksi, jos repliikki aloittaa uuden jakson, ja suurimmillaan koko siihenastisen puhujasekvenssin pituus, jos saman jakson katsotaan jatkuneen puhujasekvenssin alusta asti. Kunkin repliikin kohdalla arvioidaan siis rinnakkain yhtä montaa hypoteesia kuin puhujasekvenssissä on siihen mennessä ollut repliikkejä. Koska jaksojen vaihtumisesta ja pituuksista ei ole mielekästä esittää pelkän puhujasekvenssin perusteella liian varmoja väitteitä, on luonnollista, että malli tuottaa kunkin repliikin jälkeen todennäköisyysjakauman senhetkisen jakson pituudesta. Seuraavassa kutsun senhetkisen jakson pituuden todennäköisyysjakaumaa lyhyemmin *jaksonpituusjakaumaksi*. Tarkastellaan tämän merkitystä *Seitsemän veljeksen* puhujasekvenssiin perustuvien esimerkkien avulla.

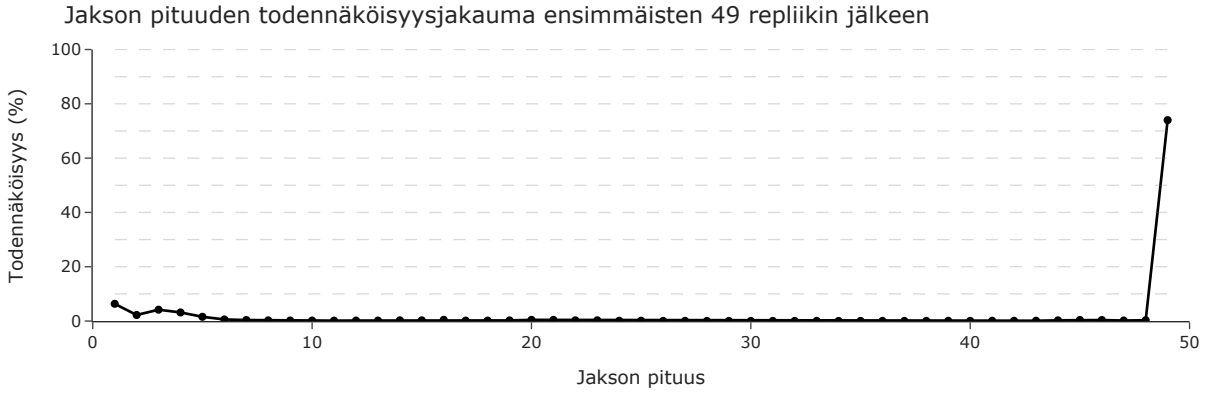
*Seitsemän veljeksen* puhujasekvenssin ensimmäiset 17 repliikkiä ovat viiden eri veljeksen esittämiä. Näistä repliikeistä jokaisen jälkeen malli arvioi yli 90 %:n todennäköisyyden sille, että senhetkinen jakso on alkanut teoksen ensimmäisestä repliikistä. Puhujasekvenssin 18. repliikki sen sijaan on Laurin ensimmäinen, ja siihen liittyvä jaksonpituusjakauma on esitetty kuvassa 1a. Laurin repliikin kohdalla malli antaa enää 85 %:n todennäköisyyden sille, että kyseessä on edelleen puhujasekvenssin ensimmäinen jakso. Koska siihen asti vaienneen Laurin ensirepliikki on hieman yllättävä tapahtuma, malli antaa 3 %:n todennäköisyyden sille, että Laurin repliikki aloittaa uuden jakson. Puhujasekvenssi on kuitenkin tässä kohtaa vielä siinä määrin lyhyt, että aloitusjakson jatkuminen saa selvästi korkeimman todennäköisyysarvion.

Seuraavan kerran aloitusjakson jatkumisen todennäköisyydessä tapahtuu huomattavaa laskua, kun Eero käyttää ensimmäisen puheenvuoronsa. Tämä tapahtuu puhujasekvenssin 49. repliikissä. Vastaava jaksonpituusjakauma on esitetty kuvassa 1b. Malli antaa 74 %:n todennäköisyyden sille, että jakson pituus on 49. Nollaa suurempia todennäköisyyksiä saavat jakson pituudet 1–6, jotka vastaavat sitä, että uusi jakso olisi alkanut joko Eeron repliikistä tai jostakin sitä edeltäneestä viidestä aiemmasta.

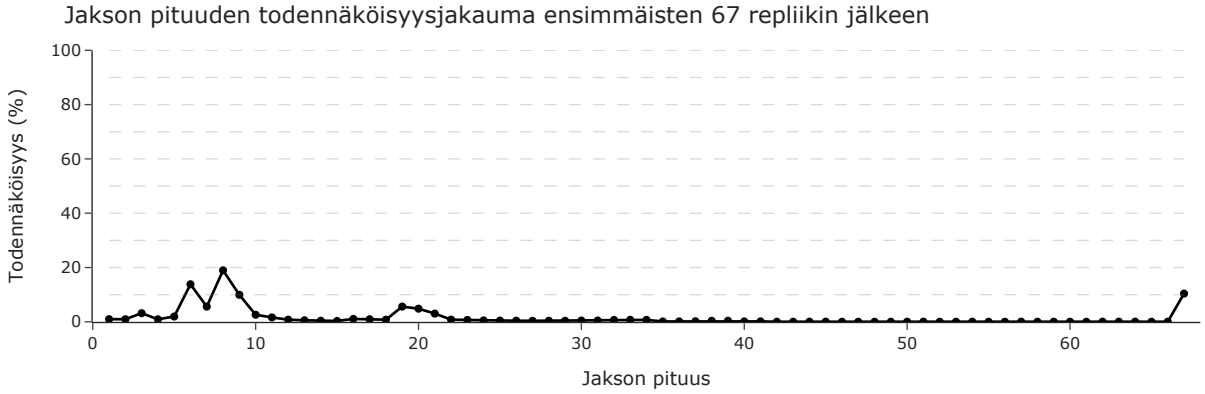
**(a)**




**(b)**




**(c)**




**Kuva 1:** Jaksonpituusjakauma *Seitsemässä veljeksessä*, kun puhujasekvenssistä on nähty ensimmäiset **(a)** 18, **(b)** 49 ja **(c)** 67 repliikkiä.

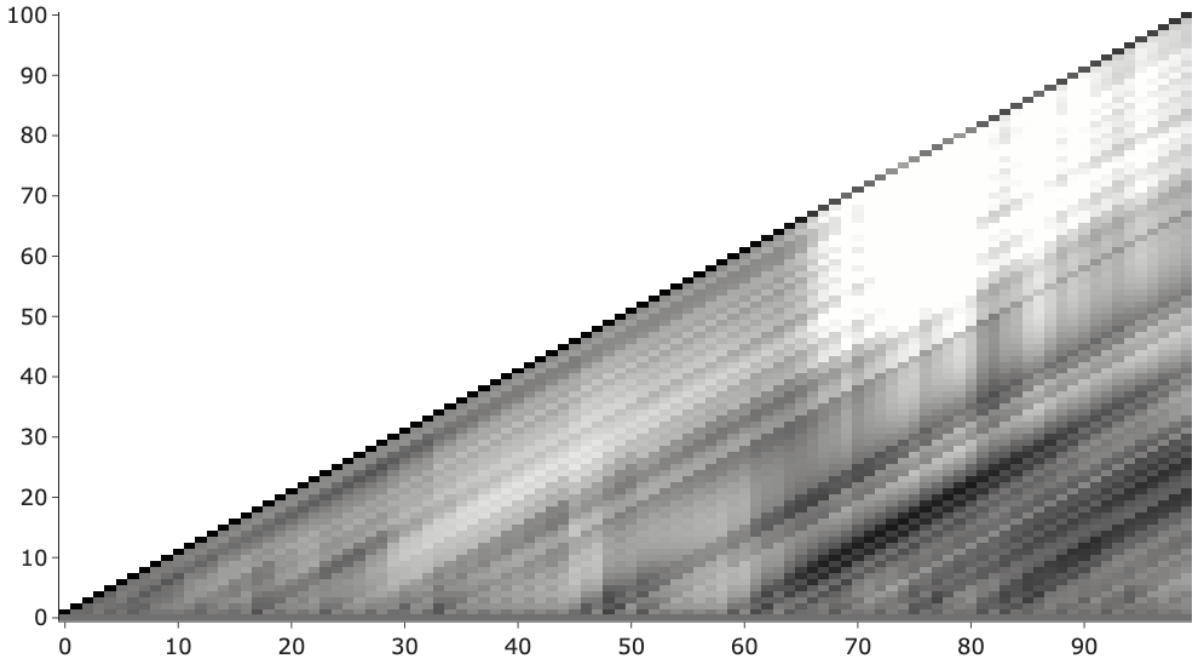
Puhujasekvenssin 67. repliikki on ensimmäinen, jonka kohdalla jaksonpituusjakauma (kuva 1c) ei enää arvioikaan, että jakso olisi todennäköisimmin jatkunut puhujasekvenssin alusta alkaen. Tämä repliikki on Eeron neljäs, joten hänen repliikkiensä osuus puhujasekvenssissä näyttää olevan kasvussa. Todennäköisyys sille, että aloitusjakso jatkuu yhä, on 10 %, kun taas todennäköisin senhetkinen jakson pituus on kahdeksan (todennäköisyydellä 19 %). Tällöin siis kyseinen jakso alkaisi puhujasekvenssin 59. repliikistä, jossa Lauri konkretisoi ehdotuksensa Impivaaraan muuttamisesta:

> LAURI. Mutta toisin tehty vielä järkevämmin tehty. Muuttakaamme metsien kohtuun ja myykäämme viheliäinen Jukola, tai pankaamme se vuokralle Rajaportin nahkapeitturille. [– –] Tehkäämme niinkuin sanon ja muuttakaamme [– –] juurelle jyrkän Impivaaran vuoren. [– –]
>
> (SV, 26.)

Puhujasekvenssin kaikkia kohtia vastaavat jaksonpituusjakaumat on mahdollista esittää visuaalisesti myös yksittäisessä kuvassa. Kuvassa 2 on tällainen *jaksokartaksi* kutsumani visualisaatio *Seitsemän veljeksen* puhujasekvenssin sadasta ensimmäisestä repliikistä. Jaksokartan vaaka-akseli kuvaa repliikin järjestysnumeroa, eli puhujasekvenssi etenee vasemmalta oikealle. Pystyakseli taas merkitsee jakson pituutta. Jaksokartan yksittäinen piste kuvaa harmaasävyn avulla, mikä on todennäköisyys sille, että vaakakoordinaattia vastaavassa puhujasekvenssin kohdassa senhetkinen jakson pituus on pystykoordinaatin osoittama. Esimerkiksi pisteissä $(67, 8)$ ja $(67, 67)$ kuva on musta tai lähes musta, mikä merkitsee verrattain korkeaa todennäköisyyttä. Pisteen $(67, 50)$ paikkeilla kuva taas on hyvin vaalea, mikä merkitsee matalaa todennäköisyyttä. Itse asiassa kuvan 2 jokainen sarake esittää saman informaation kuin kuvat 1a–c. Esimerkiksi sarake 67 vastaa kuvaa 1c.

Kuvasta 2 voidaan tehdä monenlaisia havaintoja *Seitsemän veljeksen* puhujasekvenssin alusta. Jaksokartan yläosan halki kulkee pääsääntöisesti musta diagonaali, joka osoittaa, että sadan ensimmäisen repliikin aikana säilyy kauttaaltaan melko todennäköisenä se hypoteesi, että nämä repliikit muodostavat yhtenäisen jakson. Kuitenkin varsinkin repliikin 67 kohdalta (tai hiukan sitä ennen) alkavana erottuu rinnakkainen tumma diagonaali, joka tosin alkaa haalistua repliikin 90 paikkeilla. Suunnilleen repliikkien 65–90 muodostama katkelma näyttää siis erottuvan mahdollisena omana jaksonaan, mikä voi selittyä siihen sisältyvillä, yleensä harvinaisilla Laurin repliikeillä. Tämän jakson aikana repliikkien 70–80 paikkeilla jaksokartan yläreunan diagonaali on haaleampi ja heti sen alla kokonaan valkoinen, koska osa todennäköisyysmassasta siirtyy välin 65–90 muodostamalle jaksolle.

Seuraavassa luvussa tarkastelen mallin tekemiä ennusteita ja erityisesti *Seitsemän veljeksen* jaksokarttaa useasta eri näkökulmasta.




**Kuva 2:** Jaksokartta *Seitsemän veljeksen* puhujasekvenssin ensimmäisestä sadasta repliikistä.

# Luku 4

# Tulokset

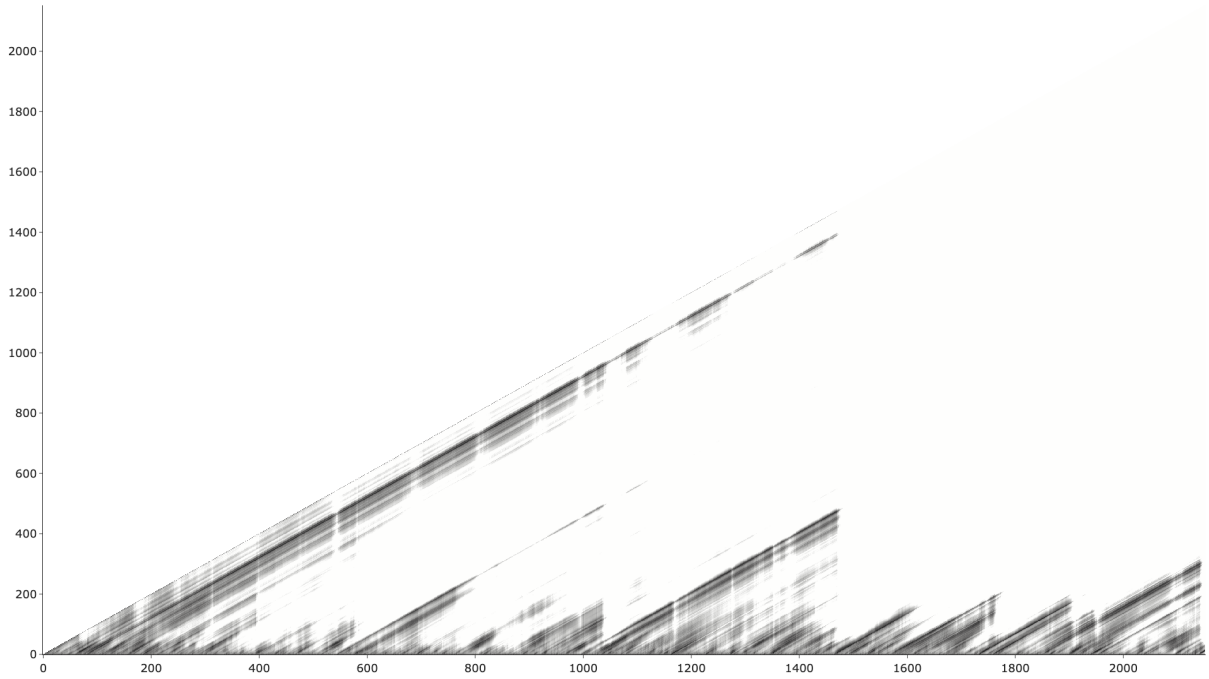
> AAPO. Ethän käsitä nyt sinäkään satua ja sen tarkoitusta.
>
> (SV, 78.)

## 4.1 Yleiskatsaus *Seitsemän veljeksen* jaksokarttaan

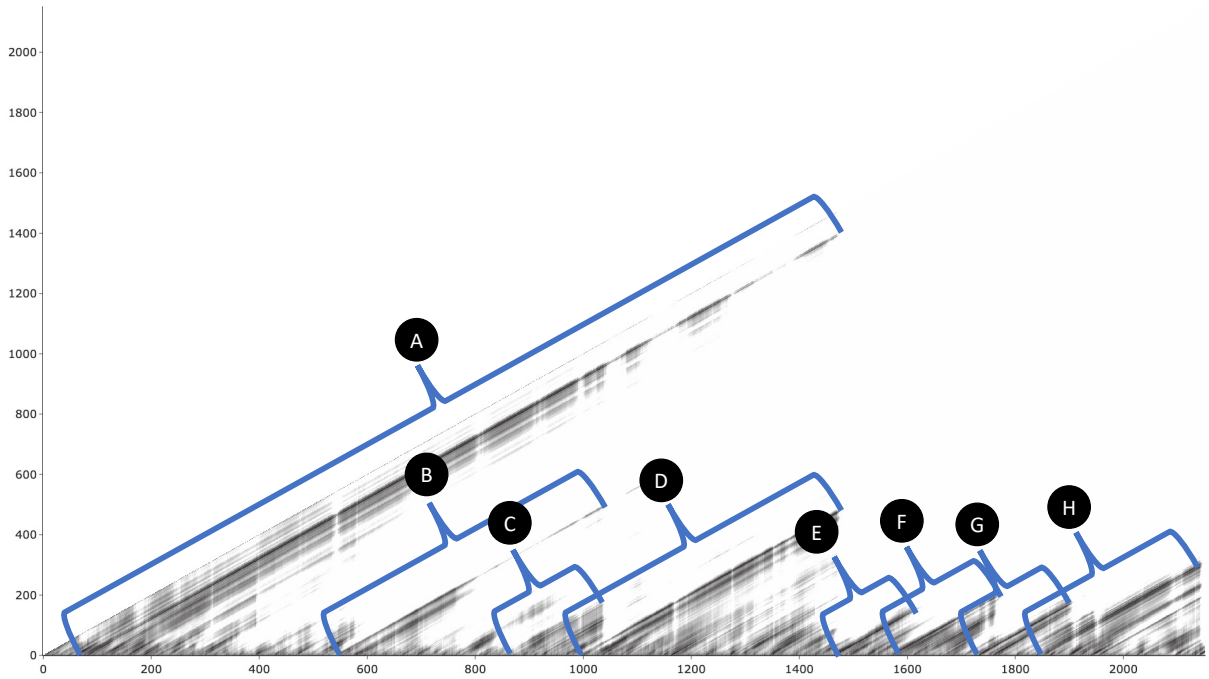
Koko *Seitsemän veljeksen* kattava jaksokartta on esitetty kuvassa 3. Jaksokartasta on mahdollista erottaa silmämääräisesti useita kiinnostavia jaksoja, jotka on osoitettu kuvassa 4 kirjaimin A–H. On syytä korostaa, että nämä jaksot on tunnistettu ja merkitty käsivaraisesti. Oleellista ei ole tässä kohden analyysiä pyrkiä suurimpaan mahdolliseen tarkkuuteen, koska jaksokartan tarkoitus on vain antaa suuntaviivoja ja toimia lähilukemisen apuvälineenä. Kuten todettua, pelkästä puhujasekvenssistä ei edes ole mahdollista osoittaa absoluuttisia muutoskohtia, ja jonkin oleellisenkin muutoskohdan jälkeen algoritmi voi tarvita kymmenienkin repliikkien verran informaatiota ennen kuin uuden jakson alkamisen todennäköisyys ylittää edellisen jakson jatkumisen todennäköisyyden.

Kuvassa 4 huomio kiinnittyy ensiksi pitkään jaksoon A, joka alkaa melkein teoksen alusta ja jatkuu pitkälti yli puoleenväliin puhujasekvenssiä. Pitkän jakson A kanssa rinnakkain erottuvat jaksokartasta lyhyemmät jaksot B, C ja D.

Jakson B alkukohta näyttää olevan siinä määrin poikkeuksellinen, että myös jakson A ja muiden mahdollisten jaksojen (haaleammat diagonaaliviivat) senhetkiset todennäköisyysarviot laskevat. Toisin sanoen jakson B alussa jaksonpituusjakauman todennäköisyydet painottuvat voimakkaasti siihen, että kohdassa alkaa uusi jakso. Pitkän jakson A sisällä rinnakkaiset jaksot B ja C näyttävät loppuvan samassa kohdassa, ja toisaalta samainen kohta on lähellä jakson D alkua.




**Kuva 3:** *Seitsemän veljeksen* jaksokartta.




**Kuva 4:** *Seitsemän veljeksen* jaksokartta, johon on merkitty kokonaisrakenteen kannalta keskeisiä jaksoja.

Osin rinnakkaiset jaksot A ja D loppuvat jotakuinkin samassa kohtaa. Niiden jälkeen jaksokartasta voidaan havaita neljä osin rinnakkaista jaksoa E–H. Jakson H jälkeen näyttää vielä alkavan hyvin lyhyt erillinen jakso.

Seuraavassa analysoin kuvaan 4 merkitty jaksoja A–H *Seitsemän veljeksen* tekstin kautta. Luvussa 4.2 pyrin tarkentamaan kunkin jakson alku- ja loppukohdat lähilukemisen avulla ja tarkastelemaan jaksojen merkitystä teoksen juonen kautta. Sen jälkeen tarkastelen luvussa 4.3 näiden havaintojen muodostamaa kokonaisuutta eri näkökulmista.

## 4.2 Jaksot

### 4.2.1 Jakso A: Jukolasta Hiidenkivelle

Kuvaan 4 merkitty jakso A rajoittuu jotakuinkin repliikkeihin 50–1 450. Kuten jo luvussa 3.4 todettiin, repliikki 49 on Eeron ensimmäinen. Luonnollisempi kohta jakson aloitukselle on kuitenkin repliikki 31. Siinä kohden romaanin alussa kerrottu kohtaus veljesten lapsuudesta (SV, 17–22) on käsitelty ja veljekset (Eero mukaan lukien) alkavat keskustella heidän äitinsä kuoleman aiheuttamasta muutoksesta ja heidän saamastaan käskystä mennä lukkarin oppiin lukutaidon hankkimista varten. Aapon repliikki pohjustaa alkavaa muutosta:

> AAPO. Sanonpa: tämä hurja elämä ei käy päisin, vaan on sen loppu viimein hävitys ja turmio. Veljet! toiset tavat ja toimet, jos toivomme onnea ja rauhaa.
>
> (SV, 23.)

Jakson loppu taas asettuu luonnollisesti linjaan luvun 7 päätöksen kanssa. Luvun 8 ensimmäinen repliikki on järjestysnumeroltaan 1463, ja repliikki 1 466 kiteyttää luvun alun tapahtumat: ”LAURI. Vimman villitty mies!” (SV, 210). Luvun alussa Lauri on ”[– –] kumonnut kurkkuunsa joltisen siemauksen viinaa” Hiidenkivellä (ibid.) ja alkaa juovuspäissään pilkata veljiään ja haastaa riitaa.

Pitkän jakson A voi siis katsoa alkavan heti *Seitsemän veljeksen* alussa olevan takauman jälkeen ja jatkuvan seitsemän ensimmäisen luvun ajan aina siihen saakka, että muuten erittäin vähäpuheinen Lauri alkaa käyttäytyä Hiidenkivellä huomattavan poikkeuksellisesti. Hiidenkivelle pakenemisen erityisyyttä *Seitsemässä veljeksessä* voidaan tarkastella myös muista näkökulmista: Kinnunen (1987, 114) on kiinnittänyt huomiota veljesten toistuviin pakoihin halki romaanin ja katsoo, että teoksen ”[– –] intensiteetti on Hiidenkivelle pyrittäessä korkeimmillaan ja laskee siitä vähitellen loppua kohden”. Myös Viljanen (1963, 11) katsoo Hiidenkiven-episodin olevan ”ratkaiseva käännekohta”, jossa ”veljekset ovat [– –] pelottavassa turvattomuudessaan kauimpana ihmisten yhteiskunnasta”.

### 4.2.2 Jakso B: Kaisan ennustusten vaikutusalue

Jakso B rajoittuu kuvassa 4 suunnilleen repliikkeihin 550–1 030. Tämän jakson aloitusta on luonteva tarkentaa repliikkiin 542, jossa veljesten ja Rajamäen rykmentin välinen kohtaus luvussa 3 on juuri päättynyt miltei väkivaltaisesti Mikon heitettyä kiven veljeksiä kohti: ”JUHANI. Tuota riivattua! Nakkasi kiven, ja liki liippasi, ettei iskenyt minua otsikkoon” (SV, 76). Kyseinen kohtaus alkaa repliikistä 514, kun Juhani tervehtii Mikkoa ja Kaisaa perheineen, ja erityisen keskeinen on kohta, jossa Kaisa ennustaa Juhanille: ”KAISA. Kuule, kuule! Tuleksi on menevä saunasi ja tuleksi tupasi myös, ja surkeassa tilassa lähdet sinä itse samoomaan metsiä, rämeitä ja soita, etsien suojaa paleltuvalle ruumiillesi. [– –]” (SV, 74).

Jakson lopetus taas tarkentuu mielekkäästi kuudenteen lukuun, jossa veljekset ovat saaneet valmiiksi Impivaaran pirtin. Jouluaterian jälkeen Juhani ehdottaa repliikiissä 1 017 painiin ryhtymistä: ”JUHANI. [– –] Yksi sana, Tuomas. Takasihan kerran veli Aapo sinun väkes ja voimas käyvän jo huikeasti yli Juhon, mutta sitä en juuri mielisi uskoa. Kuinkahan tuossa pyörähtelisimme? Koetetaan!” (SV, 148–149). Tästä alkaa tapahtumasarja, joka johtaa painiin, juomiseen, saunomiseen – ja tulipaloon, jonka seurauksena veljekset joutuvat repliikistä 1 218 alkaen pelastautumaan pakkasen halki Jukolaan.

Kaiken kaikkiaan jakso B näyttää siis muodostavan melko selkeän kokonaisuuden, joka alkaa Kaisan ennustuksesta ja päättyy Kaisan ennustuksen jälkiosan toteutumista juuri edeltäviin tapahtumiin (saunan palohan on tapahtunut jo neljännessä luvussa). Koskimies (1958, 48–49) on myös kiinnittänyt huomiota Kaisan ennustuksiin ja toteaa Kiven käyttämän ennakoinnin olevan keinona ”vanhaa eepillistä perua”. On mielenkiintoista, että näiden ennustusten vaikutusalue näkyy niin selkeästi *Seitsemän veljeksen* jaksokartassa.

### 4.2.3 Jakso C: Impivaaran ensimmäinen pirtti

Jakso C on jakson B sisältä erottuva lyhyempi jakso, joka alkaa noin repliikin 850 kohdalta ja päättyy samaan paikkaan kuin jakso B. *Seitsemän veljeksen* luku 5 alkaa repliikin 831 kohdalta, ja repliikki 850 kuuluu luvun aloitusdialogiin, joten jakson on perusteltua tulkita alkavan luvun 5 alusta, jossa veljekset muuttavat kevään tultua Impivaaraan:

> JUHANI. Ihminen on merenkulkija elämän myrskyisellä merellä. Niinhän mekin nyt purjehdimme armaista syntymänurkistamme pois, purjehdimme vankkurilaivallamme eksyttävien metsien halki Impivaaran jyrkkää saarta kohden. Ah!
>
> (SV, 115.)

Aluksi he asuvat isoisänsä vanhassa koijussa (SV, 122). Kuudennen luvun alussa pirtti on viimein valmis, ja käyttökelpoisena se säilyy edellisessä luvussa käsiteltyyn jouluiltaan saakka.

Jakso C vastaa siis varsin tarkasti Impivaaran ensimmäisen pirtin rakennustöihin ryhtymisestä alkavaa ja pirtin tuhoon päättyvää ajanjaksoa. Näin sen voi tulkita muodostavan sisällöllisesti koherentin kokonaisuuden, vaikka se onkin samalla osa kahta laajempaa jaksoa.

### 4.2.4 Jakso D: seikkailuja Impivaaralla

Myös jakso D sijoittuu jakson A sisälle. Näillä kahdella jaksolla on yhteinen päätöskohta, jonka olen edellä tarkentanut teoksen lukujen 7 ja 8 vaihteeseen. Jakson D alku asettuu noin repliikin 1 000 kohdalle. Romaanin kuudennen luvun ensimmäinen repliikki on repliikki 997, joten jakso D koostuu käytännössä teoksen luvuista 6–7.

Jakson keskeisiä tapahtumia ovat muun muassa Impivaaran ensimmäisen pirtin valmistuminen kuudennen luvun alussa (SV, 141) ja jouluyön pakomatka Jukolaan (mts. 164–173), jälleenrakennus seuraavana keväänä seitsemännessä luvussa (mts. 178), karhun kaataminen (mts. 191–192) ja lopulta pako härkälaumaa Hiidenkivelle (mts. 197).

Erityisesti jaksoa D määrittäväksi voi nostaa Tuomaan aseman. Kinnunen (1987, 48–51) on kuvannut, kuinka tulipalon syttymisen myötä Juhani menettää erityisasemansa ja Tuomas ja Aapo ottavat johdon. Nämä kaksi ”jakelevat käskyjä muille veljeksille ja saavat joukon järjestyneenä liikkeelle kohti Jukolaa” ja ”Tuomas osoittaa olevansa varsinainen johtaja” (mts. 49). Kun jakso D vaihtuu jaksoksi E (ks. alla), Tuomas joutuu kuitenkin pian kamppailuun veljiään vastaan, ja kohtauksen lopuksi ”[– –] Tuomas eristäytyy kärsittyään tappion” (mts. 51).

### 4.2.5 Jakso E: Laurin pilasaarna

Jakso E alkaa käytännössä siitä, mihin jaksot A ja D päättyvät – siis siitä, kun Lauri on juonut Hiidenkivellä itsensä humalaan. Jakso päättyy kuvan 4 perusteella pian repliikin 1 600 jälkeen. Itse asiassa repliikissä 1 600 Timo raportoi Laurin nukahtaneen (SV, 223), joten jakso E voidaan perustellusti määrittää repliikkien 1 463–1 599 muodostamaksi. Jakso erottuu muista jaksoista harvinaisen selkeästi, koska Laurin repliikkien osuus on siinä tavattoman suuri, kuten jäljempänä nähdään. Jakson keskeisyyttä korostaa sekin, että kyseessä on Lyytikäisen (2004, 160) mukaan ”[v]eljesyhteisön kannalta [– –] ainoa vakava häiriö”.

### 4.2.6 Jakso F: kiveltä kaskelle

Jakso F on osin päällekkäinen jakson E kanssa. Se alkaa jaksokartan perusteella jotakuinkin repliikkien 1 550 ja 1 600 puolivälin paikkeilla, siis Laurin juopumusjakson loppuvaiheilta. Sopiva aloitusrepliikki jaksolle on esimerkiksi repliikki 1 568 eli Juhanin määräys heittää Lauri härille: ”JUHANI. Hän viskattakoon härkien eteen, ja Jumala olkoon hänen kanssansa! Ammen. – Nyt se on sanottu. Hän menköön.” (SV, 221). Tämän repliikin myötä tilanne saa

dramaattisen käänteen, kun Tuomas ryhtyy Juhaninkin yllätykseksi toteuttamaan käskyä, ja muut saavat vain vaivoin pelastettua Laurin. Pian tämän jälkeen Timo keksii, että härät voidaan ampua.

Jakso loppuu jonkin verran ennen repliikkiä 1 800. Teoksen luvun 10 ensimmäinen repliikki on puhujasekvenssin 1 757:s, joten jakso F voidaan luontevasti täsmentää repliikkeihin 1 568–1 756. Sisällöllisesti jakso käsittää siis Hiidenkiven Laurin sammumisesta eteenpäin monine seurauksineen. Jakso loppuu sovintoon härkien omistajan, Viertolan isännän, kanssa ja siitä seuraavaan kaskiviljelyn aloituksen (SV, 256–261).

### 4.2.7 Jakso G: välienselvittelyt

Kuvasta 4 nähdään, että jakson G alku näyttää ulottuvan pitkälle osin edeltävän jakson F sisälle, lähelle repliikkiä 1 700. Tämä kohta sijoittuu lukuun 9, jonka kurranlyöntiosuus päättyy Eeron lauluun repliikissä 1 704. Kisailun jälkeen veljekset syövät ja käyvät nukkumaan. He heräävät lautamies Mäkelän ja Viertolan isännän edustajan saapumiseen, joiden kanssa keskustelu alkaa Juhanin repliikistä 1 705. Veljesten Hiidenkiven-seikkailu jälkimaininkeineen on ohi, ja on tilinteon aika.

Jakso jatkuu ohi jakso F:n lopun, jossa veljekset aloittavat kaskiviljelyn, ja ulottuu aina repliikin 1 900 tietämille saakka. Väliin mahtuvat velan sovittaminen, Simeonin ja Eeron Hämeenlinnan-matka sekä Simeonin ja Laurin näyt ja vielä verinen tappelu toukolaisten kanssa Tammiston kartanolla. Luvun 11 alussa veljekset pelkäävät rangaistusta tappelusta. Epätoivon vallassa pohditaan jopa itsemurhaa: ”JUHANI. [– –] Veitsi kurkkuun joka miehen!” (SV, 288).

Kinnunen (2002, 34–37) on myös tunnistanut näiden kahden konfliktin yhteyden. Viitaten myös jakson G alkua edeltävään välikohtaukseen Viertolan väen kanssa Kinnunen toteaa: ”Niin kuin näemme, samat kuviot toistuvat: yhteen vereen uppoutuminen sovittaisi mahdollisen miesmurhan, joka on synneistä suurimpia” (mts. 37). Jakson lopussa kuitenkin tapahtuu muutos: ”Romaanin veriset seikkailut päättyvät saman päivän iltana. Ovat veljekset päättäneet lähteä pestautumaan Heinolan pataljoonaan” (ibid.). Tiellä veljekset kohtaavatkin nimismiehen, joka veljesten yllätykseksi kertoo, ettei seuraamuksia ole tiedossa (SV, 291–295). Tämän jälkeen repliikki 1 885 aloittaa veljesten dialogin tilanteestaan uuden tiedon valossa. Repliikin 1 912 kohdalla veljekset ovat piileksineet päiväkausia, lähettäneet Aapon tiedusteluretkelle ja saaneet lopulta varmuuden armahduksestaan. Tämä kohta on teoksessa uuden elämän alku veljeksille – lukutaitokin päätetään lopulta hankkia – ja muodostaa siksi sopivan päätöksen jaksolle G.

### 4.2.8 Jakso H: yhteiskuntakelpoistuminen

Jakso H alkaa kuvan 4 perusteella noin repliikistä 1 850 ja jatkuu melkein puhujasekvenssin loppuun asti. Teoksen luku 11 alkaa repliikistä 1 834 ja luku 13 päättyy repliikkiin 2 139, joten tämä jakso asettuu jakson D tavoin luonnollisesti lukurajojen puitteisiin. Jakson ulkopuolelle jää lopetusluku 14, jossa dialogin osuus on vähäinen; veljeksistä vain Timolla on siinä repliikkejä.

Jakson alku on osin päällekkäinen jakson G kanssa, mutta laajemmin voidaan katsoa, että nämä kolme lukua muodostavat veljesten yhteiskuntakelpoistumisen kauden: nimismiehen ja muiden auktoriteettien armahdettua veljekset alkaa lukutaidon hankkiminen, peltojen raivaaminen ja soiden kuivaaminen sekä uudisrakentaminen. Tarkiainen (1910, 55) on kuvannut ilmiötä seuraavasti: ”Villin mielivallan sijaan alkaa astua järki, hetkellisten päähänpistojen ja kuvittelujen sijaan vakava tahdon suunta ja pyrkimys.”

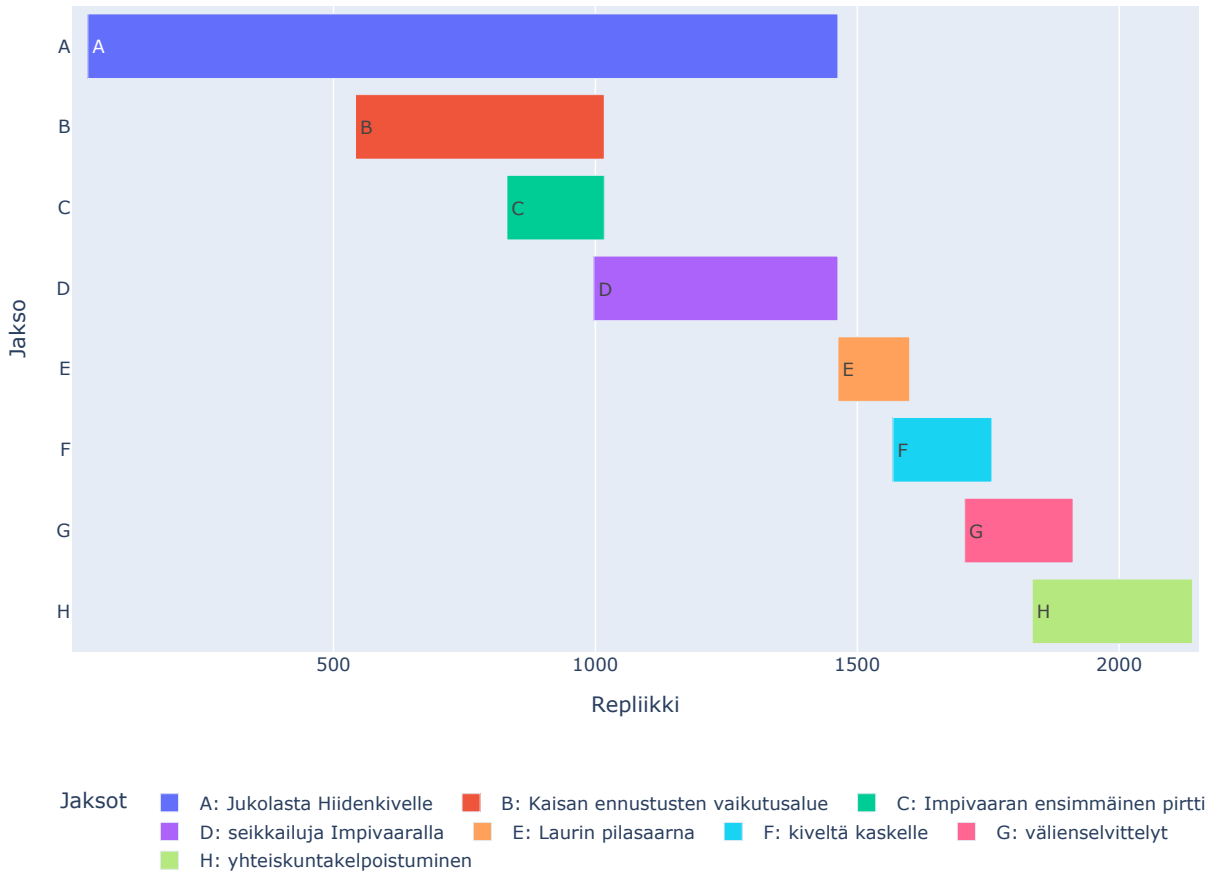
Luvussa 13 veljekset palaavat Jukolaan, tekevät sovintoa niin toukolaisten kuin Rajamäen rykmentinkin kanssa ja alkavat muun muassa Juhanin ja Männistön Venlan kihlauksen kautta solmia siteitä yhteisöön. Veljekset ”[– –] alottavat näin paluumatkansa järjestyneen yhteiskunnan keskuuteen”, ja ”[– –] heidän harharetkensä päättyvät miltei idylliseen onneen ja rauhaan” (Tarkiainen 1910, 56).

## 4.3 Kokonaiskuva

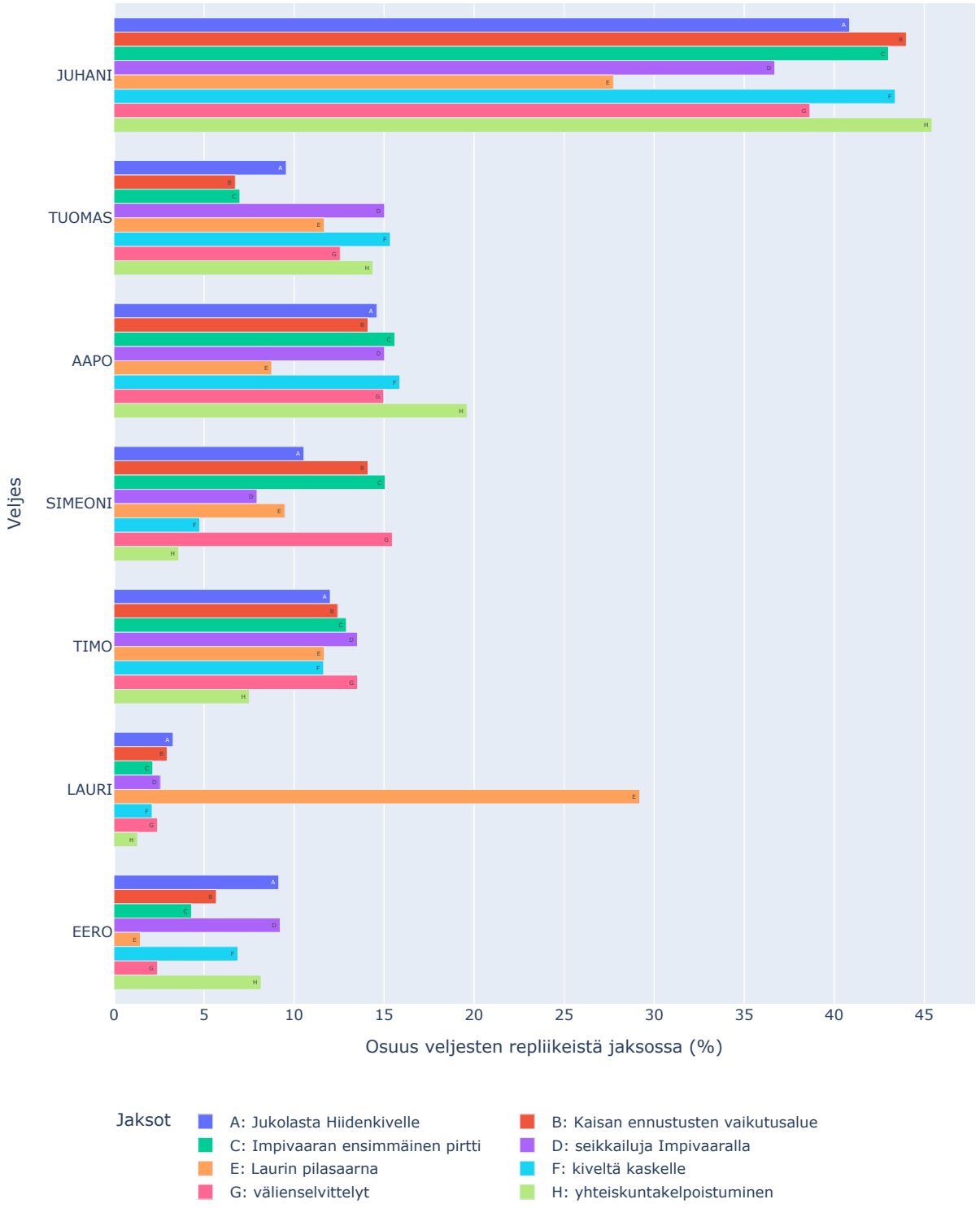
Edellä on täsmentänyt kuvissa 3 ja 4 esitetystä jaksokartasta silmämääräisesti tunnistettujen jaksojen rajat täsmällisiksi lähilukemisen avulla. Kuvassa 5 on vielä esitetty lopulliset jaksot kiteytyksineen aikajanan muodossa. Aikajanan ulkopuolella jäävät teoksen alun lapsuudenkohtaus ja teoksen lopetusluku 14, jotka toki olisi mahdollista nimetä myös omiksi jaksoikseen.

Jaksot ovat osin päällekkäisiä. Erityisesti jakso A on pitkä ja pitää sisällään kolme muuta jaksoa. Voidaankin pohtia, olisi mielekkäämpää katkaista jakso A jakson B aloituskohtaan ja muutenkin edellyttää, etteivät jaksot saisi olla päällekkäisiä. Toisaalta tällainen vaatimus absoluuttisen selvärajaisista jaksoista tuntuu liian tiukalta, kun tutkimuskohteena on kaunokirjallinen teos, johon jo lajinsa kautta kuuluu monitulkintaisuus. Kuvasta 4 nähdään, että jakson A jatkuminen aina jakson E (Lauri-jakso) alkuun asti on puhujasekvenssin mallinnuksen perusteella kohtuullisen todennäköinen hypoteesi. Sitä paitsi rinnakkaiset, eri aikatasoihin liittyvät jaksot voidaan myös tulkita osaksi teoksen todennäköisyysrakennetta (Kukkonen 2020, 20).

Samoin jakson C sisältyminen kokonaan jaksoon B herättää kysymyksen siitä, pitäisikö jaksot yhdistää tai jaksoa B typistää. Asiaa voidaan pohtia myös tarkastelemalla jälkikäteen laskettuja veljesten repliikkien suhteellisia osuuksia eri jaksoissa. Nämä osuudet on




**Kuva 5:** Aikajana luvussa 4.2 tarkennetuista *Seitsemän veljeksen* jaksoista.




**Kuva 6:** Veljesten suhteelliset osuudet eri jaksoissa.

havainnollistettu kuvassa 6. Kuvasta nähdään, että jaksojen B ja C väliset erot ovat pääosin melko pieniä verrattuna esimerkiksi näiden jaksojen ja jakson A välisiin eroihin. Juhanin, Eeron ja Laurin suhteelliset osuudet jaksossa B ovat kyllä suuremmat kuin jaksossa C, mutteivät kovin radikaalisti.

Kuvasta 6 voidaan joka tapauksessa huomata, että *Seitsemästä veljeksestä* tunnistettujen jaksojen välillä on (kenties B:tä ja C:tä lukuunottamatta) selkeitä eroja. Timon suhteellinen osuus näyttää pysyvän poikkeuksellisen vakaana jaksosta toiseen, mutta esimerkiksi Juhanin ja Eeron osuudet vaihtelevat huomattavasti jaksosta toiseen. Jakson D osalta on syytä huomata Tuomaan osuuden merkittävä kasvu ja vastaavasti Juhanin osuuden lasku, mikä tukee luvussa 4.2.4 esitettyä pohdintaa Tuomaan asemasta. Laurin suuri osuus jaksossa E on vertaansa vailla.

Vedenjakajana toimivan Laurin jakson E jälkeen jaksot näyttävät F–H jakavan teoksen jälkipuoliskon kolmeen vaiheeseen, joilla on selkeät tulkinnat osana sitä kokonaishahmotusta, että Hiidenkivi on *Seitsemän veljeksen* ratkaiseva käännekohta (Viljanen 1963, 11). Ensiksi alkaa työnteko, kun Hiidenkiveltä siirrytään – vielä pakon sanelemana – uutteran työnteon ääreen jaksossa F. Veljesten on korjattava välinsä ympäröivään yhteisöön ja yhteiskuntaan jaksossa G. Sen jälkeen alkaa lopullinen yhteiskuntakelpoistuminen ja romaani saa päätöksensä jaksossa H.

# Luku 5

# Yhteenveto

> Hänen kuntonsa kirjoituksessa ja luvunlaskussa saattoi myös hänelle paljon tointa ja askaretta, ja myöskin sisääntuloja.
>
> (SV, 391.)

Tässä tutkielmassa olen kuvannut Kukkosen esittelemät ennustavan lukijan ja todennäköisyysrakenteen käsitteet ja tuonut niihin laskennallisen näkökulman bayesiläisen tilastotieteen kautta. Olen valanut tilastomatemaattisen pohjan, joka mahdollistaa näiden käsitteiden kanssa yhteensopivien laskennallisten työkalujen laatimisen kirjallisuudentutkimuksen apuvälineiksi.

Konkreettisesti olen muodostanut Aleksis Kiven romaanista *Seitsemän veljestä* (2020/1870) sekundäärisen puhujasekvenssiaineiston ja esitellyt laskennallisen mallin, joka etsii tästä puhujasekvenssistä muutoskohtia ennustavan käsittelyn avulla. Tässä muutoskohdat ovat analogisia romaanin todennäköisyysrakenteeseen, ja laskennallinen ennustava käsittely vertautuu ennustavan lukijan toimintaan.

Lähestymistapani tuloksena olen tuottanut *Seitsemästä veljeksestä* (tarkemmin sen puhujasekvenssistä) jaksokartaksi kutsumani visuaalisen esityksen, jonka pohjalta olen lähilukemisen avulla jakanut teoksen kahdeksaan osittain päällekkäiseen jaksoon. Olen pohtinut näiden jaksojen muodostaman jaon mielekkyyttä ja kytkenyt ne teoksen sisältöön ja aiempaan *Seitsemää veljestä* koskevaan tutkimukseen.

Tutkielma yhdistää Kukkosen 2010- ja 2020-luvuilla kehittämiä kognitiivisen kirjallisuudentutkimuksen menetelmiä osin klassisiin, osin 2000-luvulla kehitettyihin matemaattisiin ja laskennallisiin aineksiin. Tutkielma laajentaa Kukkosen näkökulmaa konkreettisen laskennalliseen suuntaan, tuottaa mielekkäitä konkreettisia tuloksia ja validoi sitä kautta sekä Kukkosen näkökulmaa että laskennallisten työkalujen käyttökelpoisuutta kirjallisuudentutkimuksessa.

Edellä esitetyt huomiot muodostavat vastauksen ensimmäiseen luvussa 1.1 esittämistäni tutkimuskysymyksistä. Myös jälkimmäinen kysymys saa positiivisen vastauksen: tulosteni perusteella puhujasekvenssiin perustuva lähestymistapa voi tuottaa ainakin *Seitsemän veljeksen* tapauksessa mielekkäitä tuloksia, vaikka käytetty sekundäärinen aineisto hukkaakin valtaosan teoksen varsinaiseen tekstiin sisältyvästä informaatiosta.

Tutkielma herättää jatkotutkimuskysymyksiä useisiin eri suuntiin. *Seitsemän veljeksen* osalta olisi perusteltua selvittää, kantaisiko rikkaamman sekundäärisen aineiston käyttö hedelmää: teoksessahan esimerkiksi repliikkin *pituus* vaihtelee suuresti muun muassa henkilöhahmojen välillä (Alén ym. 2018, 119), eikä tässä tutkielmassa käytetty malli ota lainkaan huomioon teoksen ei-dialogimuotoisia osuuksia, vaikka dialogin osuus ja merkitys vaihtelevat luvusta toiseen (Kinnunen 1987, 72). Laskennallisen metodologian osalta luvuissa 4.1 ja 4.2 käsin tehty jaksokartan alustavien jaksojen tunnistaminen olisi syytä automatisoida sopivan algoritmin avulla. Yleisemmin laskennallista tutkimusta on aiemmin sovellettu vain vähän lukemisen tutkimukseen ja kognitiiviseen tutkimukseen, ja ennustavan lukijan käsite ja mallintamisen ajatus ylipäänsä avaavat kiinnostavia näköaloja.

Tutkielman perustavalaatuisin tulos onkin se, että moderniin kognitiiviseen kirjallisuudentutkimukseen perustuvasta lukijakäsityksestä on mahdollista johtaa analogisesti siirtymä laskennalliseen teosanalyysiin, ja että tällaisella lähestymistavalla voidaan tuottaa kirjallisuudentutkimuksellisesti mielekästä informaatiota kohdeteoksesta. Kiven *Seitsemän veljestä* ja sen puhujasekvenssi muodostavat hedelmällisen analyysin kohteen, mutta mikään ei estä soveltavasta vastaavaa paradigmaa *mutatis mutandis* muihinkin teoksiin. Luvussa 2.2.3 olen kuvannut yleiset lähtökohdat, joiden pohjalta voidaan konstruoida bayesiläisen ennustavan lukijan malliin sopivia sekundäärisiä aineistoja erilaisista kaunokirjallisista teoksista. Lukujen 3.3 ja 3.4 lähestymistapaa soveltamalla Adamsin ja MacKayn algoritmi voidaan mukauttaa erityyppisten aineistojen muutoskohtien tunnistamiseen, ja näin voidaan tuottaa eri ilmiöitä kuvaavia jaksokarttoja teosanalyysin tueksi.

# Viitteet

1. Käytän tutkielmassani Sakari Katajamäen toimittamaa laitosta *Seitsemän veljestä ja opas sen lukemiseen* (Kivi 2020/1870), jonka teksti pohjautuu Suomalaisen Kirjallisuuden Seuran tekeillä olevaan teokseen *Seitsemän veljestä. Kriittinen editio. I–II*, toim. Jyrki Nummi (päätoimittaja), Sakari Katajamäki, Ossi Kokko, Petri Lauerma, Juhani Niemi, Kirsi-Maria Nummila ja Pentti Paavolainen. Tutkielman laskennallisen osuuden aineisto pohjautuu Project Gutenbergin sähköisesti julkaisemaan *Seitsemän veljestä* -editioon (Kivi 2004/1870).

2. Adams ja MacKay käsittelevät artikkelissaan tapauksia, joissa tutkittavaa aikasarja-aineistoa on mielekästä mallintaa normaalijakaumalla tai Poisson-prosessina. Nämä eivät kuitenkaan sovellu puhujasekvenssin analysoimiseen. Kuten Adams ja MacKay kuitenkin toteavat artikkelinsa luvussa 2.3, mallia on mahdollista mukauttaa muunkinlaisiin aineistoihin. Tässä tapauksessa sopiva valinta on Dirichlet–multinomijakauma. Priorijakaumaksi sopii $\boldsymbol{\alpha} = (1, 1, \ldots, 1)$, joka vastaa sitä neutraalia ennakko-oletusta, että repliikit jakautuvat tasaisesti kaikille seitsemälle veljekselle. Samalla tämä priori estää nollalla jakamisen ongelman, kun malli ennustaa aina nollaa suuremman todennäköisyyden sellaisellekin veljekselle, joka ei ole esittänyt yhtään repliikkiä.

   Malli edellyttää lisäksi muutoskohtien esiintymistiheyden parametrisointia (ks. Adamsin ja MacKayn luku 2.1). Käytän kirjoittajien esimerkin mukaisesti vakiomuotoista funktiota $H(\tau) = 1/\lambda$ ja hyperparametria $\lambda = 100$. Toisin sanoen oletan, että muutoskohtia on keskimäärin sadan repliikin välein. Tämä on melko heikko oletus, koska tällöin muutoskohtien välin arvioidaan olevan noin 90 %:n todennäköisyydellä vähintään viisi ja enintään 300 repliikkiä.

3. Käyttämäni algoritmi noudattelee Adamsin ja MacKayn artikkelissa esitettyä pseudokoodimuotoista algoritmia. Python-ohjelmointikielellä tekemäni käytännön toteutus on omani, mutta sitä ovat inspiroineet Gregory Gundersenin (2019, 2020) blogikirjoitukset. Tiedossani ei ole, että algoritmia olisi aiemmin toteutettu Dirichlet–multinomijakauman kanssa.

4. Matemaattinen syy on pohjimmiltaan se, ettei nollalla jakaminen ole mahdollista. Kuvittelemalla, että jokainen veljeksistä olisi esittänyt yhden repliikin, voidaan laskea, että kenen tahansa yksittäisen veljeksen siihenastinen suhteellinen osuus esitetyistä repliikeistä on $1/7$. Vastaava laskutoimitus ei ole mahdollinen, jos yhtään repliikkiä ei ole lausuttu, koska silloin vastaavaksi osuudeksi tulisi $0/0$.

   Nämä ”virtuaaliset” romaanin alkua edeltävät seitsemän repliikkiä voi halutessaan ajatella tiedoksi siitä, että kukin veljeksistä on olemassa tarinamaailmassa.

5. ”Veljesten nimet vanhimmasta nuorimpaan ovat: Juhani, Tuomas, Aapo, Simeoni, Timo, Lauri ja Eero” (SV, 15).

# Kirjallisuutta


Adams, Ryan Prescott & David J. C. MacKay 2007: *Bayesian Online Changepoint Detection*. arXiv:0710.3742 [stat.ML]. doi:10.48550/arXiv.0710.3742

Alén, Niklas, Sakari Katajamäki & Ossi Kokko 2018: Sano kielin, puhu mielin – seitsemän veljestä keskustelijoina. Teoksessa R. Holopainen, S. Katajamäki & O. Kokko (toim.), *Ihmissydän. Henkilöitä ja kohtaloita A. Kiven maailmoissa*. Helsinki: Ntamo, 117–137, 260–262.

Aminikhanghahi, Samaneh & Diane J. Cook 2017: A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51, 339–367. doi:10.1007/s10115-016-0987-z

Bayes, Thomas 1763: An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418. doi:10.1098/rstl.1763.0053

Bode, Katherine 2018: *A World of Fiction: Digital Collections and the Future of Literary History*. Ann Arbor: University of Michigan Press.

Clark, Andy 2013: Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204. doi:10.1017/S0140525X12000477

Culler, Jonathan 1975: *Structuralist Poetics: Structuralism, Linguistics and the Study of Literature*. London: Routledge.

Da, Nan Z. 2019: The Computational Case against Computational Literary Studies. *Critical Inquiry*, 45(3), 601–639. doi:10.1086/702594

Fish, Stanley E. 1970: Literature in the Reader: Affective Stylistics. *New Literary History* 2(1), 123–162.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari & Donald B. Rubin 2013: *Bayesian Data Analysis*. Third edition. Boca Raton, FL: CRC Press.

Gundersen, Gregory 2019: *Bayesian Online Changepoint Detection*. Blogikirjoitus, julkaistu 13.8.2019, luettu 24.4.2024. Verkko-osoite: https://gregorygundersen.com/blog/2019/08/13/bocd/

Gundersen, Gregory 2020: *Implementing Bayesian Online Changepoint Detection*. Blogikirjoitus, julkaistu 20.10.2020, luettu 24.4.2024. Verkko-osoite: https://gregorygundersen.com/blog/2020/10/20/implementing-bocd/

Hohwy, Jacob 2018: The Predictive Processing Hypothesis. Teoksessa Newen, Albert, Leon de Bruin & Shaun Gallagher (toim.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press, 129–146.

Kauppinen, Sirppa 1969: *Aleksis Kiven Seitsemän veljestä ryhmätoiminnallisen oppimistapahtuman kuvauksena*. Kasvatustieteen laudaturtutkielma. Helsingin yliopiston kasvatustieteen laitos.

Kinnunen, Aarne 1987: *Tuli, aurinko ja seitsemän veljestä. Tutkimus Aleksis Kiven romaanista*. Toinen, täydennetty painos. Helsinki: SKS.

Kinnunen, Aarne 2002: *Seitsemän veljestä ja lukemisen juonet*. Helsinki: WSOY.

Kivi, Aleksis 2004/1870: *Seitsemän veljestä*. Project Gutenberg EBook 11940, https://www.gutenberg.org/ebooks/11940

SV = Kivi, Aleksis 2020/1870: *Seitsemän veljestä ja opas sen lukemiseen*. Toim. Sakari Katajamäki. Helsinki: SKS.

Kortekallio, Kaisa & Anna Ovaska 2020: Lähilukeminen ennen ja nyt: Ruumiillisia, ympäristöllisiä ja poliittisia näkökulmia. *Avain*, 17(3), 52–69. doi:10.30665/av.95530

Koskimies, Rafael 1958: Seitsemän veljeksen rakenne. Luku teoksessa *Novellin teoria ja muita tutkielmia*. Helsinki: Otava, 39–80.

Kukkonen, Karin 2014a: Bayesian Narrative: Probability, Plot and the Shape of the Fictional World. *Anglia*, 132(4), 720–739. doi:10.1515/ang-2014-0075

Kukkonen, Karin 2014b: Presence and Prediction: The Embodied Reader's Cascades of Cognition. *Style*, 48(3), 367–384. doi:10.5325/style.48.3.0367

Kukkonen, Karin 2014c: Quixotic Reasoning: Counterfactuals, Causation and Literary Storyworlds. *Paragraph*, 37(1), 47–61. doi:10.3366/para.2014.0109

Kukkonen, Karin 2020: *Probability Designs: Literature and Predictive Processing*. New York: Oxford University Press.

Lyytikäinen, Pirjo 2004: *Vimman villityt pojat. Aleksis Kiven* Seitsemän veljeksen *laji*. Helsinki: SKS.

Moretti, Franco 2000: Conjectures on World Literature. *New Left Review*, 1, 54–68.

Moretti, Franco 2011: Network Theory, Plot Analysis. *New Left Review*, 68, 80–102.

Newen, Albert, Shaun Gallagher & Leon de Bruin 2018: 4E Cognition: Historical Roots, Key Concepts, and Central Issues. Teoksessa Newen, Albert, Leon de Bruin & Shaun Gallagher (toim.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press, 3–16.

Parente-Čapková, Viola 2021: Digitaaliset ihmistieteet kirjallisuudentutkimuksessa. *Avain*, 18(4), 86–95. doi:10.30665/av.111973

Prince, Gerald 2013/2011: Reader. *The living handbook of narratology*. Luotu 8.6.2011, muokattu 25.9.2013. Verkko-osoite: https://www-archiv.fdm.uni-hamburg.de/lhn/node/52.html (luettu 8.3.2024).

Rossi, Riikka 2022: Ääniä pohjoisesta. Fonoemotionaalinen näkökulma Timo K. Mukan romaaniin *Maa on syntinen laulu*. *Avain*, 19(1), 10–31. doi:10.30665/av.112650

Stockwell, Peter 2002: *Cognitive Poetics: An introduction*. London & New York: Routledge.

Tarkiainen, Viljo 1910: *Aleksis Kiven "Seitsemän veljestä": kirjallinen tutkimus*. Väitöskirja, Suomen Keisarillinen Aleksanterin Yliopisto. Porvoo: WSOY.

Tolonen, Mikko & Leo Lahti 2015: Aatehistoria ja digitaalisten aineistojen mahdollisuudet. *Ennen ja nyt*, 2015/2.

Viljanen, Lauri 1963: 'Seitsemän veljeksen' aines ja rakenne. *Parnasso*, 13(1), 5–12.

Zunshine, Lisa 2015: Introduction to Cognitive Literary Studies. Teoksessa Zunshine, L. (toim.), *The Oxford Handbook of Cognitive Literary Studies*. New York: Oxford University Press, 1–9.