Pari sanaa asumisesta – miten yliopisto-opiskelijoiden kirjoittamissa vaihtoraporteissa asetetaan tekstikappaleen topiikki?

This paper explores the various strategies writers use to shift from one topic to another based on a corpus of student exchange reports. The main questions addressed in the study are 1) what linguistic means students use to set a new topic and 2) how these means could be defined functionally and syntactically. The data for the study consists of 492 reports from exchange period abroad. Both qualitative and statistical approaches were adopted in order to capture the complexities of the phenomenon. The paper describes 13 separate functional groups of topic-setting strategies. The statistical analysis conducted on these groups shows that majority of them can be drawn from three simple micro-level variables. This information can be used to develop teaching and writing practises especially in L2 context but to some degree in L1 also.

Avainsanat: topiikki, kirjoittaminen, korpusmentelmät

Keywords: topic, writing, corpus-based methods

# Johdanto

Lähestyessämme 2020-lukua Suomen oppilaitoksissa ja korkeakouluissa opiskelee kasvava joukko kirjoittajia (L1 ja L2), jolla ei opintojensa alkuvaiheessa ole opinnoissa menestymiseen tarvittavia tekstitaitoja. Kaiken tasoisilla kirjoittajilla on kuitenkin saatavilla lähes rajaton tekstimaailma, jota olisi mahdollista hyödyntää tai oppia hyödyntämään oman kirjoitusprosessinsa tukena. Ns. uuden, tutkivan kirjoittajan olisi ainakin teoreettisella tasolla mahdollista tutkia, miten ja millaisia kielen rakenteita käyttäen muut kirjoittajat ovat saaneet aikaan tavoitteena olevan tekstilajin edustajia. Tekstimassojen jäsentely kirjoittamisen tarpeisiin ei kuitenkaan ole yksinkertaista, eikä siihen tässä – ainakin jossain määrin uudessa - tilanteessa ole vielä tarjolla sovelluksia (ks. Koskenniemi et al. 2012). Toistaiseksi meiltä puutuu myös aineistopohjainen, tekstin kirjoittamista silmällä pitäen jäsennetty fraasipankki, jota korkeakouluopiskelija voisi käyttää.

Tämä jo lähtökohdiltaan soveltava tutkimuksemme asettuu erikoiskielten kontekstiin, mutta lähestyy ennemminkin tekstin, kielen rakenteiden ja leksikon jakamia rajapintoja kuin tekstin kokonaisuutta tai tekstilajin muodostumista sinänsä. Akateemisen kirjoittamisen tarkasteluun liittyy usein käsitys, että tekstilajitietoisuus ja -tuntemus edesauttaa ja laajentaa kielenkäyttäjän kielellistä repertuaaria, ja tekstilajitutkimuksessa lähestymistapa on yleensä tekstin merkityksistä ja funktioista kielen rakenteita kohti suuntaava (Vuorijärvi 2013: 11–15). Yhtä lailla on kuitenkin esitetty, että kielen rakenteiden erilaisten käyttötapojen hallinta edeltäisi tekstitaitojen kehittymistä (ks. esim. Jisa & Tolchinsky 2009).

Useammin tarkastelun kohteina olleiden, hierarkkisesti rakentuneiden tekstien rinnalla (ks. esim. Juvonen 2014; Komppa 2012; Paldanius 2017; Vuorijärvi 2013) elää kuitenkin huomattava joukko tekstilajeja, joissa tekstin rakennetta voi kuvata ennemmin topiikkipohjaiseksi ja temaattisen sisällön pohjalta rakentuvaksi (Hiippala 2013; Komppa 2012; Vuorijärvi 2013 SIVUNUMEROITA NÄIHIN). Tällöin rakenne muodostuu sekventiaalisesti tai lineaarisesti peräkkäin asetetuista sisältötopiikeista, jotka kytkeytyvät tavalla tai toisella tekstin diskurssitopiikkiin eli pääaiheeseen.

Meitä kiinnostaa, millä tavalla ja mitä rakenteita käyttäen kirjoittaja asettaa tekstikappaleen topiikin eli puheenaiheen (ks. VISK, määritelmät) ja siirtyy topiikista toiseen. Tavoitteenamme on kartoittaa mahdollisimman laajasti sitä rakenteiden kirjoa, jolla eri kirjoittajat ovat asettaneet yhden, spesifin sisältötopiikin. Siksi tutkimus toteutetaan korpusvetoisesti yhdistelemällä laadullisia (osio 4.1) ja määrällisiä (osio 4.2) menetelmiä. Analyysin avulla haetaan vastausta seuraaviin tutkimuskysymyksiin:

1. Millaisia tapoja yliopisto-opiskelijat käyttävät saman topiikin asettamiseen?
2. Miten näitä tapoja voidaan määritellä toisaalta funktionaalisesti, toisaalta syntaktisesti?

Tutkimus toteutetaan korpusvetoisesti yhdistelemällä laadullisia (osio 4.1) ja määrällisiä (osio 4.2) menetelmiä. Tekstikappaleiden avausvirkkeiden holistisella tarkastelulla aineistosta löytyi 13 erilaista ryhmää, joiden analyysiä sitten syvennettiin syntaktisen jäsennyksen ja sen pohjalle rakentuneen päättelypuun avulla. Tutkimuksen aineisto koostuu opiskelijavaihdosta kirjoitetuista vaihtoraporteista.

Tutkimuksemme lähtökohta on, että kirjoittajalla voi olla muiden kielten tai tekstilajitietoisen kouluopetuksen pohjalta ymmärrystä tekstilajista, mutta hänen repertuaarissaan ei välttämättä ole sopivia kielen rakenteita, joilla pääsisi alkuun ja näin pystyisi itse kirjoittamaan tavoitemuotoisen tekstin. Tästä lähtökohdasta käsin on myös valittu tutkimuksen keskeinen analyysiyksikkö - tekstikappale - joka on toistaiseksi harvakseltaan ollut tutkimuksen kohteena (Komppako?) mutta jonka kirjoittamiseen ohjeistetaan oppikirjoissa ja kirjoittamisoppaissa. Kirjoittajan kannalta kappalejako on konkreettinen kielenyksikkö ja työkalu.

# Tekstikappale ja sen topiikki

Tekstin kokonaisrakenteen tarkastelussa ja tekstilajitutkimuksessakin lähdetään usein siitä ajatuksesta, että teksti on hierakkisesti rakentunut, ja sen osat toteuttavat kokonaisuuden kannalta erilaisia funktioita (ks. esim. Juvonen 2014a, Komppa 2012, Paldanius 2017, Vuorijärvi 2013). Englanninkielisessä kirjoittamisen opetuksen perinteessä on ollut tapana käyttää topic sentence -käsitettä kuvaamaan virkettä, jolla kirjoittaja ilmaisee kappaleen tärkeimmän ajatuksen (ks. esim. Smith 2008). Suomeksi tämä kääntyy nykyisellään ydinvirkkeeksi, joka yhdessä tukivirkkeiden kanssa muodostaa kappaleen. Tekstikappaleen ydin voi kuitenkin sijaita yhtä lailla ortografisen kappaleen alussa kuin muuallakin, eikä tekstikappaleen ensimmäinen lause tai virke läheskään aina paljasta mitään siitä, millaisessa asemassa kappale hierakkisesti on tekstin kokonaisrakenteen kannalta. (kotimaiset; Smith 2008: 85).

Joidenkin tekstien rakennetta voi kuitenkin kuvata ennemmin topiikkipohjaiseksi ja temaattisen sisällön pohjalta rakentuvaksi (Paltridge 1997; ks. myös Hiippala 2013; Komppa 2012; Vuorijärvi 2013). Tällöin rakenne muodostuu sekventiaalisesti tai lineaarisesti peräkkäin asetetuista sisältötopiikeista, jotka kytkeytyvät tavalla tai toisella tekstin diskurssitopiikkiin eli pääaiheeseen. Tutkimusaineistomme osoittautui juuri tämän tyyppisiksi teksteiksi. Topiikkipohjaisuutta on havaittu korkeakouluopiskelijoiden teksteissä laajemminkin (ks. Vuorijärvi 2013).

Tekstikappaleen ytimen tavoittelun sijaan kohdistammekin huomiomme siihen, miten kirjoittaja siirtyy tekstissään aiheesta toiseen. Nimitämme siirtymän toteuttavia virkkeitä *topiikkia asettaviksi* (vrt. Juvonen 2014).

Käyttämämme topiikin käsite pohjaa Simon Dikin funktionaalisessa kielioppiin (1989), jota vastaavanlaisessa tekstikappaleiden rakenteen analyysissa on soveltanut mm. Jesus Guijarro (2001). Dikin mukaan (mts. 266) kaikilla diskursseilla (joka Dikin määrittelemänä tarkoittaa mitä tahansa koherenttia tekstiä) on se ominaisuus, että ne kertovat jostakin. Tämä ominaisuus, jota nimitämme *aiheisuudeksi* (*aboutness*), koskee paitsi tekstiä kokonaisuutena, myös siitä eroteltuja alakokonaisuuksia, kuten kappaleita ja virkkeitä. Määrittelemmekin topiikin tämän artikkelin puitteissa siksi, mistä teksti, kappale tai virke kertoo (vrt. Lambrecht 1996: 118).

Tekstin sisällä topiikit muodostavat Dikin (mt. 271, vrt. myös Givón 1983) jaottelussa topiikkiketjuja, joiden katsomme tässä artikkelissa koostuvan *esittelytopiikeista* (Dikillä *new topic*), *esitellyistä topiikeista* (Dikillä *given topic*) ja *lohkotopiikeista* (Dikillä *subtopic*). Esittelytopiikkeja ovat tapaukset, joissa kirjoittaja tuo jonkin aiheen tekstiin ensimmäistä kertaa ja esiteltyjä topiikkeja taas sellaiset tapaukset, jotka on jo tuotu aiemmin ja joihin nyt palataan. Lohkotopiikeilla viittamme tilanteeseen, jossa kirjoittaja käyttää hyväkseen johonkin edellä esiteltyyn topiikkiin kiinteästi liittyvää assosiaatioiden joukkoa viitatakseen entiteettiin, jota itseään ei kuitenkaan vielä ole mainittu.

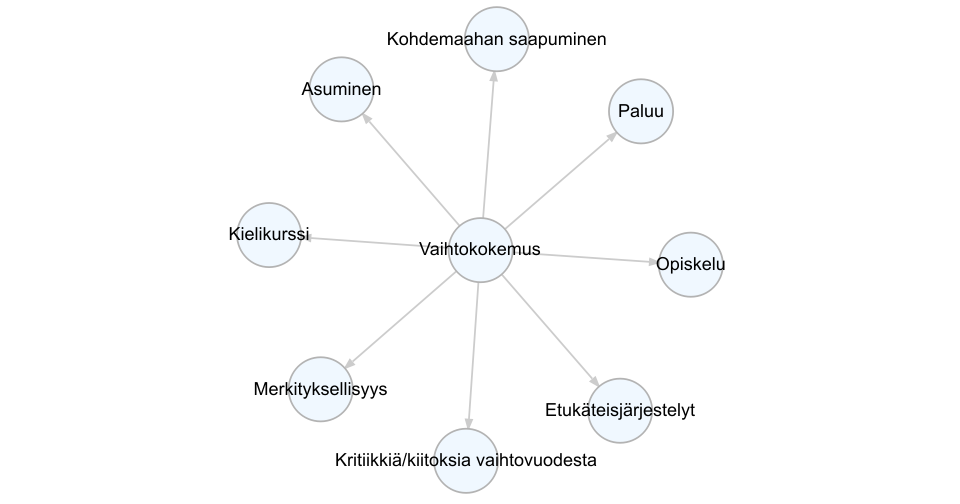
Ajatus lohkotopiikista kytkeytyy Wallace Chafen semiaktiivisuuden käsitteeseen (1987: 29), ja käytettävissä olemiseen Knud Lambrechtilla (1996: 99). Esimerkki Dikiltä (mts. 275) on lause *John gave a party last week, but the music was awful*, jossa *musiikki* on juhlien käsitteen avulla tekstiin tuotu lohkotopiikki. Kuten esimerkistäkin näkyy, lohkotopiikkien muodostumisessa keskeisiä ovat myös sanojen väliset merkityssuhteet kuten hyponymia ja meronymia (vrt. Shore 2008: 33).

Huolimatta siitä, että tekstuaalisia suhteita, joita on haasteellista ankkuroida eksplisiittiisiin tekstinosiin, on kuvattu eri viitekehyksissä eri tavoin, niiden merkitys kirjoittajalle on keskeinen.

# Tutkimusaineisto ja -metodi

Tutkimuksen aineisto koostuu yliopisto-opiskelijoiden vaihtokokemuksestaan kirjoittamista teksteistä, joita nimitämme matkaraporteiksi.[[1]](#footnote-2) Kyseinen raportti on institutionaalinen teksti, jolla on selkeä funktio: tukea vaihtojaksoa suunnittelevia opiskelijoita päätöksenteossa ja omaan tulevaan vaihtojaksoonsa valmistautumisessa. Niiden voi katsoa edustavan samaa tekstilajia, joka on korkeakoulukontekstissa kuitenkin suhteellisen yksinkertainen (ks. esim. Bhatia 2012). Raporteista koostettu korpus sisältää julkaisuluvan saaneet raportit vuosilta 2011-2017 ja kattaa kaikkiaan 492 matkaraporttia (464 744 sanetta).

Tehtävänanto tekstien laatimiseen on annettu erillisissä ohjeissa, joissa opiskelijoita pyydetään kirjoittamaan “matkakertomus, jossa kuvaillaan ja arvioidaan opiskelua ulkomaisessa yliopistossa tai harjoittelua ulkomailla”. Ohjeet sisältävät kahdeksankohtaisen listan mahdollisesti käsiteltävistä asioista (mm. *etukäteisjärjestelyt*, *asuminen* ja *merkityksellisyys*). Listaa kehotetaan tulkitsemaan toteamalla, että “seuraavat ohjeet eivät ole sitovia, mutta ne on hyvä lukea läpi ennen kirjoittamista”. Valtaosa teksteistä on silti kirjoitettu ohjetekstin ehdottaman aiherajauksen pohjalta. Tekstien rakenne on kuvion 1 kaltainen: Niiden diskurssitopiikkina on vaihtokokemus ylipäätään, ja tätä ylemmän tason topiikkia käsitellään jakamalla se alatopiikkeihin, joista tavallisimmat on lueteltu kuviossa.



KUVIO 1: Matkakertomustekstien diskurssitopiikki ja alatopiikit

Tutkimus koostui kolmesta menetelmällisesti erilaisesta vaiheesta. Kahdessa ensimmäisessä vaiheessa käytettiin laadullisia menetelmiä. Kolmas, tilastollisia menetelmiä hyödyntävä vaihe rakentui edellisten varaan ja xxx.

Ensimmäisessä vaiheessa raportit annotoitiin kappaleittain siten, että jokaiselle tekstikappaleelle merkittiin topiikki. Kaksikymmentä tekstiä käsittäneen pilotin perusteella tarkastelu rajautui *asumisesta* kertoviin kappaleisiin; se osoittautui melko selvärajaiseksi aiheeksi ja siitä kirjoitettiin lähes jokaisessa näytetekstissä. Koska tekstit analysoi kaksi eri henkilöä (kirjoittajat), näyteaineiston avulla myös testattiin, kuinka samanlaisia tulkinnat kappaleiden aiheista olivat. Asumiskappaleiden osalta arvioijien annotointi osui yhteen 45 tapauksessa 47:stä. Aineistoksi valikoituneiden tekstikappaleiden aiheisuus oli siis luotettavasti tavoitettavissa.

492 tekstistä 415:ssä oli vähintään yksi asumisesta kertova kappale. Nämä aineistoksemme valikoituneet tekstit ovat keskimäärin 945 sanaa pitkiä ja koostuvat keskimäärin 13 kappaleesta. Jotta

analysoitava otos edustaisi mahdollisimman kattavasti koko tutkimusaineistoa, mutta työmäärä pysyisi kohtuullisena, aineistosta otettiin painotettu satunnaisotanta tarkempaa analyysia varten (satunnaisotantaan liittyvistä periaatteista ks. esim. Levshina 2015: 8). Satunnaisotanta tehtiin kappaleista, joissa *topiikki-indikaattorina* (eli ensimmäisenä leksikaalisena vihjeenä) oli jokin *asua*-verbin johdos (kuten *asunto*, *asuntoasia* tai *asuminen*) tai *asua*-verbi kappaleen ensimmäisessä virkkeessä. Topiikki-indikaattorin käsitettä käytettiin operationalisoimaan tekstin sisältöön perustuvan topiikin paikkaa. Tutkimustehtävämme kannalta on keskeistä, että topiikkia asettavat virkkeet ovat osoitettavalla tavalla yhteneväisiä. Virkkeen sijainti kappaleen alussa ja leksikaalinen vihje ovat luotettavasti toistettavia keinoja mahdollisimman vertailukelpoisen otoksen valitsemiseksi.

Ennen satunnaisotantaa tekstikappaleille suoritettiin automaattinen syntaktinen ja morfologinen jäsennys sekä lemmatisointi hyödyntämällä Turun yliopistossa kehitettyä dependenssijäsennintä (Haverinen ym. 2014) ja siihen liittyviä komponentteja.

Jos tutkitaan tarkemmin niitä kategorioita, jotka syntaktinen jäsennin on merkinnyt tutkimukseen valikoituneiden kappaleiden ensimmäisille *asua*-johdoksille, saadaan taulukon 1 mukainen jakauma.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Frekvenssi | Selitys | Esimerkki | N |
| dobj | 94 | objekti | *Hankin asunnon* | 47 |
| root | 60 | predikaattiverbi | *Asuin kerrostalossa* | 30 |
| nmod | 50 | luokittelematon adverbiaali | *Asunnon suhteen* | 25 |
| nmod:gobj | 34 | nominaalisen verbimuodon objekti | *Asunnon hankkiminen* | 17 |
| nsubj | 32 | subjekti | *Asuminen järjestyi* | 16 |
| nmod:poss | 29 | omistusrakenteen kohde | *Asuntoloiden vuokrataso* | 14 |
| nsubj:cop | 21 | kopulalauseen subjekti | *Asuntola on uusi* | 10 |
| muu | 18 | esim. infinitiivitäydennykset | *Halusin asumaan paikallisten kanssa* | 9 |
| Yht. | 338 | — | — | 168 |

TAULUKKO 1. Asua-johdosten jakautuminen eri syntaktisiin kategorioihin.

Taulukon Frekvenssi-sarakkeesta nähdään, että selvästi tavallisimpia topiikki-indikaattoreita olivat objektina esiintyvät asua-johdokset. Kategoriat, joiden edustajia oli yhteensä alle kymmenen, niputettiin yhdeksi *muu*-kategoriaksi. Varsinainen satunnaisotanta tehtiin taulukossa esitettyjen kategorioiden perusteella siten, että jokaisesta taulukossa luetellusta ryhmästä (mukaan lukien *muu*-ryhmä) valittiin satunnaisesti ryhmän kokonaisfrekvenssin puolikasta vastaava määrä tapauksia (ks. taulukon viimeinen sarake).

Satunnaiset 168 tapausta muodostat tutkimuksen 2. vaiheen aineiston.

Tässä ryhmittelimme virkkeet siten, että yhden ryhmän muodostivat aina ne tavat asettaa topiikki, jotka sisällöltään ja esitystavaltaan muistuttavat toisiaan. Jokaista topiikkia asettavaa virkettä tutkittiin ensin laajemmassa tekstiyhteydessään, minkä jälkeen sitä verrattiin muihin virkkeisiin ja määriteltiin, oliko kyseessä jokin jo muodostettuun joukkoon kuuluva tapaus vai ennemminkin muista erottuva tapausjoukko. Prosessin kuluessa muodostuneista ryhmistä eksplikoitiin piirteet, jotka erottavat ne muista ryhmistä.

Kolmannessa vaiheessa näitä ryhmiä tarkasteltiin tilastollisesti ehdollisten päätöksentekopuiden avulla.

# Tulokset

Edellä kuvatun analyysin tuloksena tutkimusaineiston tekstit jaoteltiin 13 ryhmään, joita osiossa 4.1. luonnehditaan yhden edustavan esimerkin avulla. Kuten metodin kuvauksen yhteydessä todettiin, tässä kuvatut eri topiikin asettamistapoja edustavat ryhmät on muodostettu systemaattisesti sijoittamalla toisaan muistuttavat tapaukset samaan kategoriaan. Ryhmien väliset rajat eivät luonnollisestikaan ole suoraviivaisia, ja jaottelun olisi voinut tehdä muillakin tavoilla. Eri ryhmät ovat lisäksi enemmän tai vähemmän läheisessä yhteydessä toisiinsa, ja näitä linkkejä on tässä esitetyssä kuvauksessa pyritty osoittamaan. Vaikka varsinainen ryhmäjako perustuu laadulliseen samankaltaisuuteen, tulkitaan osiossa 4.2. analyysin tulosta eli itse jaottelua tilastollisin menetelmin.

## Kolmetoista tapaa asettaa topiikki

Asetimme johdannossa tutkimuksen tavoitteeksi kuvauksen siitä, millä eri tavoilla kirjoittajat ovat asettaneet yhden ja saman topiikin. Edellä kuvatussa tutkimuksen ensimmäisessä vaiheessa rajasimme tarkastelumme niihin tapauksiin, joissa vihje topiikista on annettu jo uuden kappaleen ensimmäisessä virkkeessä. Tätä kaikkiaan 168 kattanutta tekstijoukkoa tutkittaessa aineistosta erottui pääpiirteissään kolmetoista erilaista tapausryhmää, jotka on esitetty taulukossa 2. Tapausten samankaltaisuutta määriteltäessä kiinnitettiin huomiota tapausten kontekstiin koko tekstin tasolla. Olennaisiksi ryhmiä erottaviksi piirteiksi muodostuivat se, millaisessa suhteessa aloitusvirkkeessä esitetty topiikki on aiempaan tekstiin, millaisesta näkökulmasta käsin kirjoittaja lähestyy tekstinsä aihetta ja toisaalta se, miten aikaa ja paikkaa on käytetty topiikin ankkuroimiseen. Suhdetta aiempaan tekstiin on taulukossa 2 esitetyssä analyysissä käsitteellistetty osiossa 2 määriteltyjen topiikin alalajien (esittelytopiikin, esitellyn topiikin ja lohkotopiikin) avulla.

Taulukon 2 ryhmät on esitetty koon mukaisessa järjestyksestä suurimmasta pienempään. Taulukon ensimmäisessä sarakkeessa on ilmaistu ryhmän koko sekä suuruusjärjestyksen mukainen numero, jonka perusteella ryhmiin viitataan osion 4.2. tilastollisessa analyysissa. Toinen sarake sisältää yhden esimerkkivirkkeen ja kolmanteen on koostettu ne piirteet, joiden perusteella kukin tapausjoukko eroteltiin omaksi ryhmäkseen.

TAULUKKO 2. Esimerkit laadullisen analyysin tuloksena syntyneistä ryhmistä

|  |  |  |
| --- | --- | --- |
| Ryhmä | Esimerkkivirke | Ryhmää yhdistävät tekijät |
| 1 (31 kpl) | *Asuntoni sain vuokrattua Apartiksen( paikallinen TOAS) kautta.* | Topiikki-indikaattorilla on lohkotopiikin funktio siitä huolimatta, ettei näissä tapauksissa täyty Dikin (1989: 275) kuvailema ehto siitä, että varsinainen ylätopiikki olisi erikseen esitelty lukijalle. Kirjoittaja operoi niin selkeästi tehtävänannossa ehdotetun rakenteen puitteissa, että hän olettaa lukijan automaattisesti jakavan käsityksen siitä, mitä aiheita tekstin diskurssitopiikkiin, vaihtokokemukseen, liittyvien asioiden joukko pitää sisällään. |
| 2 (21 kpl) | *Vaihtoni aikana asuin Grandmontin asuntolassa.* | Indikaattorilla on esittelytopiikin funktio. Virkkeiden pääasiallisena sisältönä informaatio siitä, missä kirjoittaja on vaihtonsa aikana asunut. |
| 3 (21kpl) | *Olin löytänyt asunnon ensimmäiseksi kuukaudeksi Airbnb:n kautta ennen Italiaan saapumista.* | Myös tässä ryhmässä indikaattorilla on esittelytopiikin funktio. Nämä virkkeet ovat ajallisesti muita monitasoisempia: lukija johdatetaan kappaleen aiheeseen kuvaamalla jotakin, mikä on tapahtunut ennen sitä ajanjaksoa, jolle varsinainen topiikki sijoittuu. Kappaleen alussa luodaan jännite, joka myöhemmässä kappaleessa puretaan (vrt. ryhmä 12). |
| 4 (16kpl) | *Asuminen, tai siis lähinnä asunnon hankkiminen, oli varmaankin vaihdon hankalin juttu.* | Näissä tapauksissa kirjoittaja luonnehtii henkilökohtaista kokemustaan asumisesta: onnistumisia, epäonnistumisia, sujuvuutta (vrt. ryhmä 8, jossa keskitytään yleisen tason havaintoihin). Kuten ryhmässä 1, myös tässä indikaattorin funktiona on lohkotopiikki, mutta asumisen yhteys vaihtokokemukseen on monessa tapauksessa eksplisiittisesti ilmipantu. |
| 5 (14 kpl) | *Asuntoloista valitsin Julius Raab Heimin vanhojen matkakertomusten perusteella.* | Ryhmän 5 tapauksissa kirjoittaja on pohjustanut asumisaihetta jo edeltävässä tekstissä. Esimerkiksi tähän valittua virkettä edeltää kappale, jossa kirjoittaja esittelee erilaisia asumisvaihtoehtoja. Näissä tapauksissa topiikki-indikaattorit edustavat siis funktioltaan esiteltyjä topiikkeja. |
| 6 (12 kpl) | *Ulkomaisille opiskelijoille tarkoitettu asuntola on vain kolme vuotta vanha ja täten melko moderni.* | Asumistopiikki esitellään lähtemällä liikkeelle konkreettisesti itse asunnon tai asuntolan luonnehtimisesta. Funktioltaan näiden tapausten indikaattorit ovat esittelytopiikkeja, joskin hyppäys edellisestä aiheesta on tyypillisesti rajumpi kuin esimerkiksi ryhmässä 2. Tämä selittää sitä, että monissa ryhmän 6 tapauksessa asumisesta kertovaa kappaletta edeltää väliotsikko *Asuminen*. |
| 7 (10 kpl) | *Itselläni kävi tuuri asunnon suhteen löytäessäni edullisen huoneen Chelseasta aivan Manhattanin keskeltä.* | Ryhmän 7 ominaispiirteenä on, että niissä kirjoittaja vertaa itseään muihin vaihdossa olleisiin. Asumistopiikki tuodaan tekstiin usein lohkotopiikin funktiolla ja siihen liittyy ryhmän 1 tavoin oletus siitä, että lukija jakaa kirjoittajan mielessä olevan listan vaihtokokemukseen liittyvistä aiheista. |
| 8 (10 kpl) | *Asuminen Tokiossa on aika kallista.* | Nämä tapaukset muistuttavat ryhmää 4, mutta erona on, että ryhmässä 8 kirjoittaja jakaa lukijalle maailmantietoa eikä niinkään kuvaa omia kokemuksiaan. Kuten ryhmä 4, myös nämä tapaukset ovat funktioltaan lohkotopiikkeja, mutta lukijaan ei kohdistu vastaavaa oletusta saman assosiaatiojoukon jakamisesta, sillä nämä virkkeet ankkuroidaan yleensä siihen fyysiseen paikkaan, jossa vaihto tapahtui. |
| 9 (9 kpl) | *Suurlähetystön henkilökunta tarjosi resurssiensa puitteissa apuaan mm. maahantuloon, käytännön järjestelyihin sekä asunnon löytämiseen liittyvissä asioissa.* | Ryhmä 9 näyttää, että yleistys kirjoittajan oman kokemuspiirin ulkopuolelle voidaan tehdä myös toisella tavalla. Kyseessä on ryhmien 2,3 ja 6 tapaan esittelytopiikki, mutta itse esittely tapahtuu hyödyntämällä jotakin tekstin kannalta keskeistä, kirjoittajasta erillistä toimijaa. |
| 10 (9 kpl) | *Lensin Wieniin sunnuntaina ja lento-kentältä suuntasin suoraan asunnolleni.* | Tyyliltään kertovia tapauksia, joissa topiikki-indikaattorin funktiona on esittelytopiikki. |
| 11 (6 kpl) | *Sitten pari sanaa asumisesta.* | Kirjoittaja tuo asumisaiheen tekstiin esittelytopiikkina käyttämällä jotain eksplisiittistä metatekstuaalista ilmaisutapaa. |
| 12 (5 kpl) | *Saapuessani Varsovaan minulla oli onneksi jo asunto tiedossa.* | Ryhmien 3 ja 10 välimaastossa olevia esittelytopiikin sisältäviä tapauksia. Siinä missä ryhmässä 3 esitetään kappaleen alussa avoin tilanne ja myöhemmässä kappaleessa tilanteen ratkaisu, aloittaa ryhmä 12 ratkaisusta (tässä virkkeessä siitä, että saavuttaessa asunto oli jo löydetty) ja käyttää lopun kappaleen takaumamaisesti sen kuvailuun, miten ratkaisuun päästiin |
| 13 (4kpl) | *Lähes kaikki Bratislavan kauppakorkeakoulun vaihto-opiskelijat asuivat Ekonom-nimisessä dormitoryssa.* | Lähdetään liikkeelle siitä, mitä *yleensä* tehdään tai miten *suurin osa* tekee. Eroaa kuitenkin ryhmästä 7 siinä, että kirjoittaja ei aseta itseään tämän yleisen toimintatavan vastakohdaksi, vaan päinvastoin samaistuu kuvaamaansa joukkoon. Topiikki-indikaattorilla esittelytopiikin funktio. |

## Tilastollinen tarkastelu

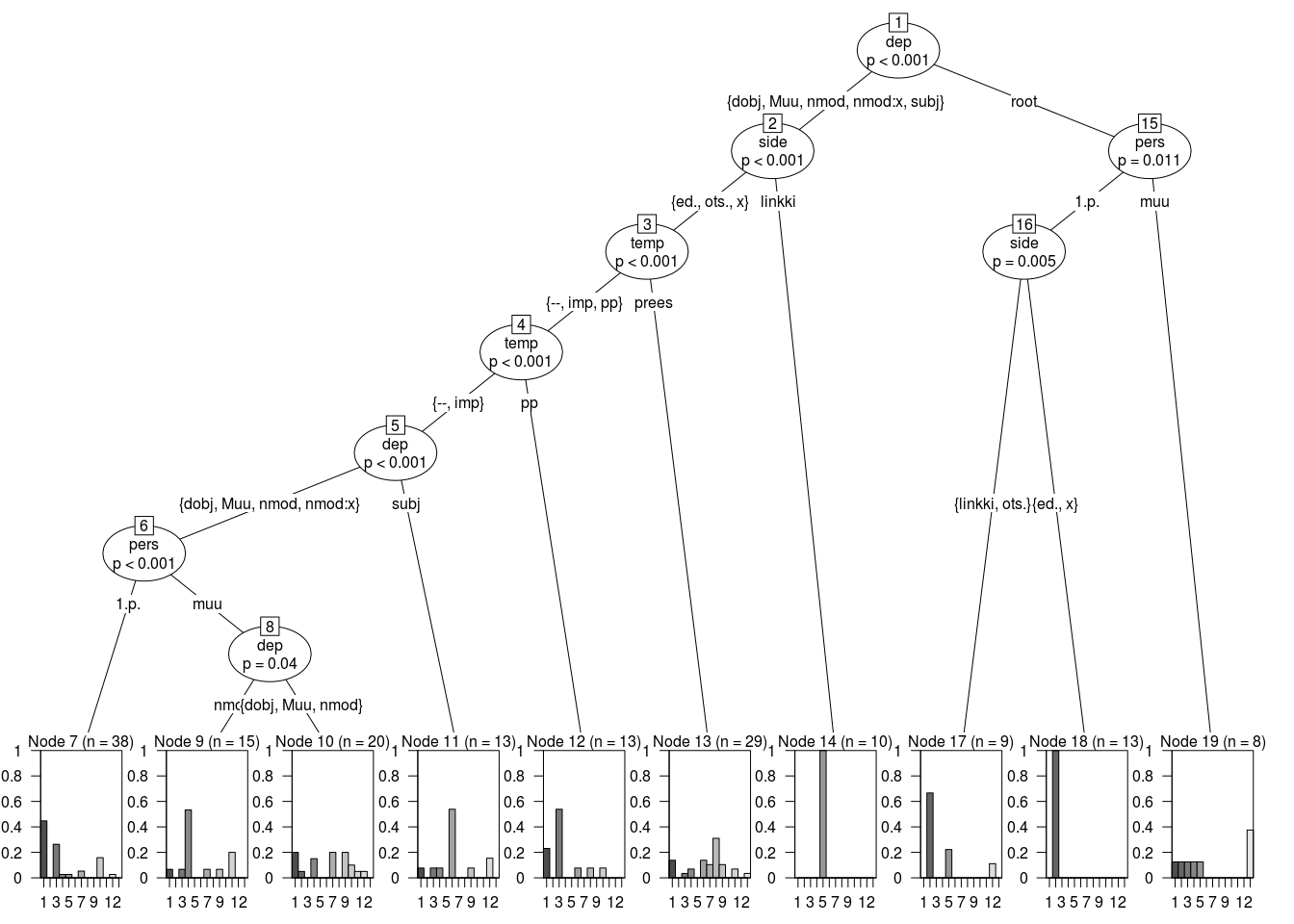
Taulukossa 2 esitimme tavan ryhmitellä asumisesta kertovat tekstikappaleet sen mukaan, miten kirjoittaja toteuttaa siirtymän aiheesta toiseen. Seuraavassa sovellamme ryhmäjakoon tilastollisia menetelmiä, tarkemmin ottaen ehdollisia päätöksentekopuita (conditional inference trees).

Ehdolliset päätöksentekopuut ovat regressiotyyppinen luokittelumenetelmä, jossa yhden tai useamman selittävän muuttujan assosiaatiota selitettävään muuttujaan (termeistä ks. Ketokivi 2015: 131) tarkastellaan sarjana tutkimusaineistoa jaottelevia päätöksiä (Levshina 2015: 291). Meidän tutkimukseemme sovellettuna tämä tarkoittaa sen selvittämistä, mitkä aineiston 168 tapauksen ominaisuudet (selittävät muuttujat) parhaiten ennustavat sen, mihin kolmestatoista ryhmästä (selitettävä muuttuja) tapaus on luokiteltu. Käytössämme ovat taulukon 3 mukaiset selittävät muuttujat.

TAULUKKO 3: Tilastollisen analyysin selittävät muuttujat

|  |  |
| --- | --- |
| Muuttujan nimi ja lyhenne | Muuttujan arvot |
| Koneellisen jäsentimen topiikki-indikaattorille asettama syntaktinen rooli (dep) | Arvot esitetty taulukossa 1, kuitenkin lavennettuna siten, että arvot *nsubj* ja *nsubj:cop* yhdistettiin yhdeksi subjektin kategoriaksi (*subj*) ja arvot *nmod:poss* sekä *nmod:gobj* yhdeksi *nmod:x*-kategoriaksi, jolla viitattiin kaikkiin kompleksisiin nominaalirakenteisiin (*asunnon hankinta*, *asumisen järjestäminen*). |
| Topiikki-indikaattorin sijainti virkkeessä (sij) | Kaksi arvoa a) Tapaukset, joissa topiikki-indikaattori on koko virkkeen ensimmäisenä. b) Muut tapaukset. |
| lauseen predikaattiverbin persoona (pers) | Kaksi arvoa: a) Tapaukset, joissa predikaattiverbi on ensimmäisessä persoonassa ja b) Muut tapaukset. |
| Side muuhun tekstiin (side) | Neljä arvoa: a) Tapaukset, joissa topiikki asetetaan viittaamalla selkeästi johonkin edellä muodostettuun linkkiin (*linkki*) b) Tapaukset, joissa asumistopiikkia on pohjustettu edellisessä kappaleessa, muttei niin selvästi, että kyseessä olisi linkkitapaus (*edel.*) c) Tapaukset, joissa asumiskappaletta edeltää väliotsikko (*ots*) d) Tapaukset, joissa mitään edellä kuvatuista siteistä ei ole (*x*). |
| Topiikki-indikaattorin morfologinen rakenne (morf) | Kolme arvoa: a) finiittiverbit b) infintiivimuodot (mukaan lukien minen-johdokset) c) substantiivit. |
| Aikamuoto (tense) | Neljä arvoa: a) preesens, imperfekti, pluskvamperfekti/perferkti[[2]](#footnote-3), ellipsi (–). |

Tässä käytetty päätöksentekopuu rakennettiin R-ohjelmiston *party*-kirjastoon kuuluvaa *ctree*-funktiota (Hothorn ym. 2006) käyttäen.[[3]](#footnote-4) Analyysin tulos on esitetty kuviossa 2.



KUVIO 2: Topiikkia asettavien ryhmien rakenne luokittelupuun avulla tarkasteltuna

Kuviossa 2 tutkimusaineisto on jaettu kymmeneen ryhmään (kuvion alimmalla rivillä esitetyt pylväskaaviot) edellä kuvattujen selittävien muuttujien perusteella. Kuvio esittää selittävien muuttujien vaikutuksen sarjana päätöksiä (kuvion ympyränmuotoiset noodit), jotka on numeroitu yhdestä 19:ään, esimerkiksi: *Jos jonkin tapauksen dependenssirooli on predikaatti (noodi 1) ja pääverbin persoona ensimmäinen (noodi 15) ja jos lisäksi tapausta ei edellä väliotsikko, kyseessä on aina ryhmään 2 kuuluva tapaus (noodi 18)*. Jos taas dependenssirooli on predikaatti (noodi 1), joka ei kuitenkaan esiinny ensimmäisessä persoonassa, on kyseessä useimmiten ryhmän 13 tapaus, vaikka muutkin vaihtoehdot ovat mahdollisia. Jokaisen kuvion haarauman kohdalle on myös merkitty p-arvo, joiden avulla voidaan arvioida kunkin haarauman merkitsevyyttä (Levshina 2015: 292). Korkein p-arvo kuviossa 3 on noodilla 15, joka sekin on vain noin yhden prosentin luokkaa, mikä tarkoittaa että kuviossa esitetyt jaot ovat kaiken kaikkiaan merkitseviä (p-arvoista ks. esim. Baayen 2012: 69; Levshina 2015: 10).

Oleellista tässä esitetyssä päätöksentekopuussa on, että vaikka itse luokittelu ryhmiin tehtiin funktionaalisesti kirjoittajien tulkinnan perusteella, on ryhmiä mahdollista kuvailla melko pitkälle myös rakenteellisten ominaisuuksien valossa. Karkeasti ottaen voidaan todeta, että suurin osa laadullisen analyysin tuloksena laadituista kolmestatoista ryhmästä pystytään päättelemään hyödyntämällä kolmea[[4]](#footnote-5) yksinkertaista rakenteellisen mikrotason muuttujaa: verbin persoonaa, aikamuotoa ja indikaattorisanan dependenssiroolia. Taulukossa 4 on tiivistetty näiden muuttujien vaikutus topiikkia asettavan virkkeen päätymiseen johonkin laadullisessa analyysissa määritellyistä ryhmistä.

TAULUKKO 4: Kvantitatiivisesti pääteltävissä olevat ryhmät

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Verbin persoona | Verbin aikamuoto | Indikaattorisanan dependenssirooli | ryhmä | luokittelupuun loppunoodit |
| *1.* | *imperfekti* | *ei predikaatti* | 1 | 7 |
| *1.* | – | *predikaatti* | 2 | 17 ja 18 |
| – | *pl.perf/perf.* | – | 3 | 12 |
| *muu kuin 1.* | *imperfekti* | *osa kompleksista lauseketta* | 4 | 9 |
| – | *imperfekti* | *subjekti* | 6 | 11 |
| – | *preesens* | *ei predikaatti* | 8 | 13 |
| *muu kuin 1.* | – | *predikaatti* | 13 | 19 |
| *muu kuin 1.* | *imperfekti* | *ei predikaatti, ei subjekti, ei osana kompleksista lauseketta* | 7 ja 9 | 10 |

Taulukon mukaan ryhmä 1 voidaan rakenteiden tasolla määritellä siten, että siinä on nominaalinen topiikki-indikaattori, jonka pääverbi esiintyy ensimmäisessä persoonassa ja imperfektissä. Ryhmän 4 puolestaan erottaa muista se, että asumisaiheeseen viitataan siinä jollakin kompleksisella lausekkeella, kuten *asunnon hankinta* tai *asumisen järjestäminen*. Ryhmä 3 taas voidaan määritellä ennen kaikkea aikamuodon pohjalta: siihen kuuluvat tapaukset, joissa verbi on pluskvamperfektissä. Ryhmien 10, 11 ja 12 osalta ei tässä esitetyn päätöksentekopuun perusteella voida esittää selkeitä piirteitä, jotka johtaisivat tapauksen sijoittamiseen juuri näihin ryhmiin. Lisäksi on huomattava, että esimerkin sijoittaminen ryhmään 5 on perustunut täysin artikkelin kirjoittajien tulkintaan: se, että tulevaa topiikkia on eksplisiittisesti pohjustettu aiemmassa tekstissä, ei ole pääteltävissä siitä, minkälaisia rakenteita topiikkia asettavassa virkkeessä on käytetty.

# Päätelmät

Tarkastelimme tässä tutkimuksessa tekstikappaleen topiikin muodostumista korpusvetoisesti, kahdella menetelmällä. Tavoitteenamme oli selvittää, millaisia tapoja yliopisto-opiskelijat käyttävät matkaraporteissaan saman topiikin asettamiseen ja miten näitä tapoja voidaan määritellä toisaalta funktionaalisesti, toisaalta syntaktisesti.

Tutkimuksen aineistona oli yliopisto-opiskelijoiden kirjoittamista matkakertomuksista koostettu korpus, ja osa artikkelin löydöksistä liittyykin eittämättä myös tekstilajille ominaisiin ilmaisutapoihin, vaikka tutkimuksen keskiössä olivat virkkeet ja niiden väliset suhteet. Käyttämämme aineisto ja menetelmät eivät tarjoa vastauksia siihen, miksi kirjoittajat ovat päätyneet ehkä hieman luettelomaisestikin kirjoittamaan ohjeessa annetuista aiheista sen sijaan, että olisivat kirjoittaneet hierarkkisesti rakentuneen tekstin.

Laadullisen analyysin tuloksena erotimme 13 erilaista tapaa toteuttaa kappaleenalkuinen siirtymä topiikista toiseen. Pääpiirteissään näiden 13 ryhmän eroavaisuudet palautuvat siihen, millaisessa suhteessa aloitusvirkkeessä esitetty topiikki on aiempaan tekstiin, millaisesta näkökulmasta käsin kirjoittaja lähestyy tekstinsä aihetta ja siihen, miten aikaa ja paikkaa on käytetty topiikin ankkuroimiseen. Tilastollisessa rakenteiden tason analyysissa monet laadullisen analyysin tuloksena laadituista ryhmistä pystyttiin lopulta päättelemään hyödyntämällä kolmea yksinkertaista muuttujaa: predikaatin persoonaa ja aikamuotoa sekä indikaattorisanan dependenssiroolia.

Esimerkiksi tekstilajeja tai argumentointia käsitellessään opettajan tai oppimateriaalin tekijän on usein turvauduttava melko abstrakteihin komponentteihin ja määritelmiin. Yhtenä tämän tutkimuksen tuloksena ovatkin funktionaalisten kuvausten lisäksi löydetyt syntaktiset tunnuspiirteet, jotka tarjoavat vaihtoehtoisen lähestymistavan kappaleiden rakentamisen opettamiseen ja omaksumiseen. Konkreettisiin kielenaineksiin viittaavat ohjeet ja mallit voivat olla lähestyttävämpiä kuin abstraktit kuvaukset siitä, minkälaisia kappaleiden tulisi olla. Tässä tuotetut ryhmittelyt voisivat toimia myös lähtölaukauksena myös erilaisia todellisista teksteistä poimittuja mallirakenteita esittelevälle fraasipankille (vrt. Chitez ym. 2015). Juuri tällaiset materiaalit istuvat hyvin ajatukseen uudesta kirjoittajasta, joka ei niinkään kysy, miten asiat pitäisi tehdä, vaan etsii malleja siitä, millä eri tavoilla on jo tehty.

# Kirjallisuus

Baayen, R. H. 2012. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge university press.

Bhatia, V. K. 2012. Interdiscursivity in Academic Genres. Teoksessa C. Benkenkotter, V. K. Bhatia, & M. Gotti (toim.) *Insights into academic genres*. Linguistic insights 160. Bern: Peter Lang, 47–65.

Chafe, W. 1987. Cognitive constraints on information flow. Teoksessa R. S. Tomlin (toim.) *Coherence and Grounding in Discourse : Outcome of a Symposium, Eugene, Oregon, June 1984.* Typological Studies in Language 11. Amsterdam: John Benjamins, 21–25.

Chitez, M., C. Rapp & O. Kruse 2015. Corpus-supported academic writing: how can technology help? Teoksessa F. Helm L. Bradley & S. Thouësny. (toim.) *Critical CALL-Proceedings of the 2015 EUROCALL Conference, Padova, Italy,*. Dublin, Ireland: Research-publishing.net.

Dik, S. C. 1989. *The Theory of Functional Grammar, Part I: The Structure of the Clause*. Functional grammar series 9. Dordrecht: Foris Publications.

Guijarro, A. J. M. 2001. Topicality chains in two discourse genres. *Estudios Ingleses de la Universidad Complutense* (9), 103–128. <http://revistas.ucm.es/index.php/EIUC/article/view/EIUC0101110103A>.

Haverinen, K., J. Nyblom, T. Viljanen, V. Laippala, S. Kohonen, A. Missilä, S. Ojala, T. Salakoski & F. Ginter 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation* *48* (3), 1–39.

Hiippala, T. 2013. *Modelling the structure of a multimodal artefact*. Helsinki: Helsingin yliopisto. [URN:ISBN:978-952-10-9427-9](urn:ISBN:978-952-10-9427-9).

Hothorn, T., K. Hornik & A. Zeileis 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* *15* (3), 651–674.

Jisa, H. & L. Tolchinsky 2009. Developing a personalized stance through linguistic means in typologically different languages. Written expository discourse. *Written Language & Literacy* *12* (1), 1–25.

Juvonen, R. 2014. *Kirjoitelma ja tekijän ääni: kehystämisen yhdyslauseet suomenkielisen ylioppilasaineen dialogisuuden hallinnassa*. Helsinki: Helsingin yliopisto. [URN:ISBN:978-951-51-0264-5](urn:ISBN:978-951-51-0264-5).

Ketokivi, M. 2015. *Tilastollinen päättely ja tieteellinen argumentointi*. Helsinki: Gaudeamus.

Komppa, J. 2012. *Retorisen rakenteen teoria suomi toisena kielenä-ylioppilaskokeen kirjoitelman kokonaisrakenteen ja kappalejaon tarkastelussa*. Helsinki: Helsingin yliopisto. [URN:ISBN:978-952-10-8164-4](urn:ISBN:978-952-10-8164-4).

Koskenniemi, K., K. Lindén, L. Carlson, M. Vainio, A. Arppe, M. Lennes, H. Westerlund, M. Hyvärinen, I. Bartis, P. Nuolijärvi & A. Piehl 2012. *Suomen kieli digitaalisella aikakaudella – The Finnish Language in the Digital Age*. Springer. <http://www.meta-net.eu/whitepapers/volumes/finnish>.

Lambrecht, K. 1996. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge studies in linguistics 71. Cambridge: Cambridge university press.

Levshina, N. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Paldanius, H. 2017. Historian esseevastauksen funktionaalinen rakenne. Teoksessa S. Latomaa, E. Luukka, & N. Lilja (toim.) *AFinLAn vuosikirja 2017*. Suomen soveltavan kielitieteen yhdistyksen julkaisuja 75. Jyväskylä: Suomen soveltavan kielitieteen yhdistys.

Paltridge, B. 1997. *Genre, frames and writing in research settings*. Amsterdam: John Benjamins.

Shore, S. 2008. Lauseiden tekstuaalisesta jäsennyksestä. *Virittäjä* *112* (1), 24–65.

Smith, C. G. 2008. Braddock revisited: The frequency and placement of topic sentences in academic writing. *The Reading Matrix* *8* (1), 78–95.

Vuorijärvi, A. 2013. *Tekstilaji ja yhteisö: Ammattikorkeakoulun opinnäytetyön diskussio tekstinä*. Helsinki: Helsingin yliopisto. [URN:ISBN:978-952-10-8928-2](urn:ISBN:978-952-10-8928-2).

1. Aineiston käyttöön on saatu lupa kyseiseltä yliopistolta, ja artikkelin esimerkkivirkkeet on pseudonymisoitu kirjoittajien yksityisyyden suojaamiseksi. [↑](#footnote-ref-2)
2. Perfektitapauksia aineistossa oli vain yksi ja sekin konditionaalissa, joten molemmat suomen liittoaikamuodot yhdistettiin samaan kategoriaan. [↑](#footnote-ref-3)
3. Analyysin tuottanut R-koodi nähtävillä GitHub-palvelussa osoitteesta [anonymisoitu]. [↑](#footnote-ref-4)
4. Päätöksentekopuun tuottaneessa algoritmissa indikaattorin sijainti osoittautui hieman yllättäenkin merkityksettömäksi, samoin kuin morfologinen rakenne. [↑](#footnote-ref-5)