JYU DISSERTATIONS 783

Taina Mylläri

Kompleksisuus osana suomenoppijan kielitaitoa

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF HUMANITIES AND
SOCIAL SCIENCES

JYU DISSERTATIONS 783

Taina Mylläri

Kompleksisuus osana suomenoppijan kielitaitoa

Esitetään Jyväskylän yliopiston humanistis-yhteiskuntatieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston vanhassa juhlasalissa S212
toukokuun 25. päivänä 2024 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities and Social Sciences of the University of Jyväskylä,
in building Seminarium, old festival hall S212, on May 25, 2024, at 12 o'clock.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2024

Editors
Jarmo Harri Jantunen
Department of Language and Communication Studies, University of Jyväskylä
Päivi Vuorio
Open Science Centre, University of Jyväskylä

ISBN 978-952-86-0149-4 (PDF)
URN:ISBN:978-952-86-0149-4
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-952-86-0149-4

ABSTRACT

Mylläri, Taina
Complexity in Finnish learner language
Jyväskylä: University of Jyväskylä, 2024, 65 p. + original articles
(JYU Dissertations
ISSN 2489-9003; 783)
ISBN 978-952-86-0149-4 (PDF)

This dissertation focuses on syntactic complexity in texts written by second language learners of Finnish. It aims to explore the changes in syntactic complexity between CEFR proficiency levels in learner Finnish. By providing empirical evidence from a morphologically rich language, this dissertation also aims to contribute to the discussion on learner language complexity. In this study, complexity in learner Finnish is first analysed with seven frequently used quantitative measures of syntactic complexity. The reliability of the production units used in these measures and the applicability of the measures to learner Finnish are tested. Then, the focus is shifted to the use of conjunctions and multi-verb constructions in the learners' texts. The data of the study are drawn from the University of Jyväskylä Cefling project corpus. The data contain formal and informal messages together with argumentative texts written by adult and adolescent learners, and they cover CEFR levels A1 to C2. The results of the study are reported in four articles. They show that although syntactic complexity at different proficiency levels is by no means similar, the quantitative measures often do not differentiate between proficiency levels. This may partly be explained by qualitative differences in the units used for measuring complexity, variance within proficiency levels and differences between task types. The findings also show differences in complexity between adult and adolescent learners' texts. This supports the view that the relation between learner language complexity and proficiency cannot be assumed to be straight-forward. The findings also resonate with the view that quantitative measures only gauge some facets of complexity. The results suggest that the study of complexity could benefit from a more qualitative approach and from looking at syntactic, morphological and lexical complexity together instead of treating them as separate dimensions of learner language complexity.

Keywords: Finnish as a second language, syntactic complexity, CEFR

TIIVISTELMÄ

Mylläri, Taina
Kompleksisuus osana suomenoppijan kielitaitoa
Jyväskylä: Jyväskylän yliopisto, 2024, 65 s. + alkuperäiset artikkelit
(JYU Dissertations
ISSN 2489-9003; 783)
ISBN 978-952-86-0149-4 (PDF)

Tässä väitöstutkimuksessa tarkastellaan suomi toisena kielenä -tekstien syntaktista kompleksisuutta ja sen muuttumista suhteessa Eurooppalaisen viitekehyksen taitotasoihin. Tutkimuksen tavoitteena on selvittää, miten kompleksisuutta ja siinä tapahtuvia muutoksia voi mitata ja miten oppijansuomen syntaktisen kompleksisuuden tarkastelu voi täydentää kuvaa oppijankielen kompleksisuudesta. Oppijansuomen kompleksisuutta tarkastellaan seitsemän syntaktisen kompleksisuuden tutkimuksessa yleisesti käytetyn määrällisen mittarin avulla, minkä lisäksi kompleksisuutta eri taitotasoilla kartoitetaan teksteissä esiintyvien konjunktioiden ja moniverbisten konstruktioiden avulla. Tutkimuksen aineisto on koottu Jyväskylän yliopiston Cefling-hankkeen suomi toisena kielenä -aineistosta, ja se koostuu aikuisten ja yläkouluikäisten kirjoittamista muodollisista ja epämuodollisista viesteistä sekä mielipideteksteistä, jotka yhdessä kattavat viitekehyksen taitotasot A1–C2. Tutkimuksen tulokset on raportoitu neljässä artikkelissa. Tulosten mukaan oppijansuomen taitotasojen välillä on eroja syntaktisessa kompleksisuudessa, mutta määrällisten mittareiden kyky erotella etenkin peräkkäisiä taitotasoja on oppijansuomessa heikko. Tähän vaikuttaa osaltaan se, että taitotasojen sisällä ja tehtävätyyppien välillä on vaihtelua syntaktisessa kompleksisuudessa. Lisäksi tässä tutkimuksessa tarkastellut kielenpiirteet kehittyvät taitotasolta toiselle eri tavoin aikuisten ja yläkoululaisten aineistossa. Tulokset tukevatkin aiempia havaintoja siitä, että oppijankielen kompleksisuuteen vaikuttavat myös muut tekijät kuin kielitaidon taso. Lisäksi tulokset vahvistavat näkemystä, jonka mukaan määrälliset mittarit tavoittavat oppijankielen kompleksisuuden vain osittain. Tutkimus osoittaa, että oppijansuomen kompleksisuudesta piirtyy monipuolisempi kuva, kun syntaktista, morfologista ja leksikaalista kompleksisuutta tarkastellaan yhdessä.

Avainsanat: suomi toisena kielenä, syntaktinen kompleksisuus, Eurooppalainen viitekehys

Author

Taina Mylläri
Department of Language and Communication Studies
University of Jyväskylä
Email taina.m.myllari@jyu.fi
ORCID https://orcid.org/0000-0003-1669-5559

Supervisors

Professor Jarmo Harri Jantunen
Department of Language and Communication Studies
University of Jyväskylä

Professor, Emerita Maisa Martin
Department of Language and Communication Studies
University of Jyväskylä

Reviewers

Professor Scott Jarvis
Northern Arizona University

Lecturer Yrjö Lauranto (Title of Docent)
Charles University, Prague, and Finnish National Agency for Education

Opponent

Professor Scott Jarvis
Northern Arizona University

ESIPUHE

Tämä väitöskirja on lopultakin valmis. On tullut aika kiittää ja kumartaa. Vaikka väitöskirjan kirjoittaminen on välillä yksinäistä puuhaa, tämäkään väitöskirja ei olisi valmistunut, jos se olisi pitänyt tehdä yksin.

Ihan ensin haluan kiittää ohjaajiani Maisa Martinia ja Jarmo Harri Jantusta. Ilman teitä en olisi nyt tässä. Kiitos tuesta, kannustuksesta ja kaikesta, mitä olen teiltä oppinut. Parempia ohjaajia en voisi kuvitella. Kiitokset myös työni esitarkastajille Yrjö Laurannolle ja Scott Jarvisille, joiden asiantuntevat kommentit auttoivat minua työn viimeistelyssä. Scott Jarvisia kiitän lisäksi siitä, että hän on lupautunut tämän työn vastaväittäjäksi.

Kaikkia tämän työn valmistumiseen vaikuttaneita on mahdotonta luetella tässä. Silti jokainen heistä ansaitsee kiitokset. Erityisesti haluan kiittää Jyväskylän yliopiston suomi toisena kielenä -tohtoriseminaarin ja myöhemmin suomen kielen tohtoriseminaarin vetäjiä Minna Sunia ja Esa Lehtistä sekä kaikkia seminaarilaisia niin saamastani palautteesta kuin siitä, että olen saanut kuulua joukkoon. Kiitos myös kaikille niille, joiden kanssa olen saanut tehdä töitä Jyväskylän yliopistossa, Opetushallituksessa, Comeniuksen yliopistossa Bratislavassa ja Vilnan yliopistossa tämän väitöskirjan kirjoittamisen aikana. Jyväskylän yliopistoa kiitän myös työni rahoittamisesta apurahoin sekä mahdollisuudesta työskennellä väitöskirjatutkijana. Työskentelyapurahasta haluan kiittää myös Suomalaista Konkordia-liittoa ja Emil Aaltosen säätiötä. Ilman mahdollisuutta tehdä tutkimusta myös kokopäiväisesti tämä väitöskirja tuskin olisi valmistunut.

Erikseen haluan vielä kiittää Nina Reimania kaikista kompleksisuudesta ja sen tutkimisesta etenkin työn alkuvaiheessa käydyistä inspiroivista keskusteluista. Ideasta koodata aineisto XML-muotoon kiitän Jussi Piitulaista Helsingin yliopistosta. Filip Ginter, Veronika Laippala ja Jenna Kanerva Turun yliopistosta opastivat minua kädestä pitäen aineiston koodauksessa suureksi avuksi olleen jäsentimen käytössä, mistä iso kiitos. Miikka Silfverbergiä kiitän siitä Python-skriptistä, joka ei lopulta tullut käyttöön tässä tutkimuksessa mutta joka innosti minua opettelemaan Python-ohjelmointia. Eleanor Underwoodia kiitän työni englanninkielisten osien asiantuntevasta ja kannustavasta kielenhuollosta sekä mahdollisuudesta oppia lisää englannin kielestä.

Lisäksi veljeni Lauri ansaitsee erityiskiitoksen opastuksesta Python-ohjelmoinnin maailmaan ja veljeni Juha avusta tilastollisten menetelmien kanssa. Ystävääni Minnaa kiitän kaikista käytännön neuvoista ja pitkästä ystävyydestä. Lisäksi haluan kiittää niin veljiäni ja Minnaa kuin Mikkoa ja kaikkia muita läheisiäni siitä, että olette elämässäni ja että olette myös jakamassa sen, että tämä väitöskirja on, lopultakin, valmis.

Vilnassa huhtikuussa 2024
Taina Mylläri

TAULUKOT

SISÄLLYS

ALKUPERÄISET JULKAISUT

Mylläri, Taina 2020. Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? Journal of the European Second Language Association 4 (1), 13–23.

Mylläri, Taina 2020. Measuring syntactic complexity in learner Finnish. Apples – Journal of Applied Language Studies 14 (2), 67–92.

Mylläri, Taina 2022. Konjunktiot ja syntaktinen kompleksisuus: konjunktioiden käyttö suomi toisena kielenä -teksteissä eri taitotasoilla. Puhe ja kieli 42 (3), 175–200.

Mylläri, Taina 2023. Moniverbiset konstruktiot ja oppijansuomen kompleksisuus kielitaidon eri tasoilla. Lähivõrdlusi. Lähivertailuja 33, 152–180.

1 JOHDANTO

1.1 Aluksi

Mitä kielitaito on? Miten se kehittyy? Miten kielitaitoa ja sen tasoa voi mitata? Tässä väitöskirjassa etsitään vastauksia näihin kysymyksiin tarkastelemalla oppijansuomen syntaktista kompleksisuutta suhteessa Eurooppalaisen viitekehyksen (EVK, 2003) taitotasoihin. Tavoitteena on kartoittaa, millaista oppijansuomen kompleksisuus eri taitotasoilla on, miten se muuttuu kielitaidon edistyessä ja miten määrälliset syntaktisen kompleksisuuden mittarit tavoittavat oppijansuomen eri taitotasoilla näkyvän kielitaidon kehittymisen.

Kielitaidon arviointiin käytetään Suomessa yleisesti Eurooppalaista viitekehystä (2003), jossa keskiössä on kommunikatiivinen kielitaito. Viitekehyksessä kielitaito jaetaan taitotasoihin sen perusteella, miten kielenkäyttäjä selviytyy erilaisista viestintätilanteista. Eurooppalaisen viitekehyksen toiminnallisiin taitotasokuvauksiin perustuvaa arviointikriteeristöä käytetään sellaisenaan tai soveltaen niin perusopetuksen kieltenopetuksen kuin aikuisille tarkoitettujen Yleisten kielitutkintojen arvioinnissa (Huhta & Hilden 2016: 20). Eri taitotasoille tyypillisiä kielellisiä piirteitä ei Eurooppalaisessa viitekehyksessä kuitenkaan juuri kuvata (esim. Alanen, Huhta & Tarnanen 2010: 33; Martin 2013b). Kielen rakenteiden ja eri piirteiden kehitys jää taka-alalle. Kielen oppimisen ja opettamisen kannalta on kuitenkin merkityksellistä, miten kielenpiirteitä opitaan. Kielenpiirteiden kehitystä suhteessa kielitaidon taitotasoihin on tutkittu muun muassa Jyväskylän yliopiston Cefling-hankkeessa, jonka yhtenä tavoitteena on ollut yhdistää toisen kielen oppimisen tutkimusta ja kielitaidon arviointia (Martin ym. 2010: 58). Tämä väitöskirja ei ole osa Cefling-hanketta mutta kulkee sen jalanjäljissä ja pyrkii selvittämään syntaktisen kompleksisuuden ja Eurooppalaisen viitekehyksen taitotasojen välistä yhteyttä.

Yksi tapa seurata kielenpiirteiden kehittymistä on tarkastella oppijankielen kompleksisuutta (engl. *complexity*), tarkkuutta (engl. *accuracy*) ja sujuvuutta (engl.

fluency) sekä niissä tapahtuvia muutoksia (ks. esim. Housen, Kuiken & Vedder 2012; Michel 2017; Pallotti 2021). Yksinkertaisimmillaan tarkkuus voidaan määritellä tuotetun kielen virheettömyydeksi, sujuvuus kielen tuottamisen helppoudeksi ja kompleksisuus oppijankielen laajuudeksi, taidokkuudeksi ja monimuotoisuudeksi (Michel 2017: 50). Vaikka näitä kolmea oppijankielen ulottuvuutta on tutkittu vuosikymmeniä, etenkään oppijankielen kompleksisuuden määritelmästä tai parhaista mittaustavoista ei edelleenkään ole yksimielisyyttä (Housen ym. 2019: 4; Michel 2017: 64). Lisäksi aiemmat tutkimustulokset ovat epäyhtenäisiä (Housen ym. 2019: 9), mikä voi johtua eroista sekä siinä, miten kompleksisuus on eri tutkimuksissa määritelty, että siinä, miten kompleksisuutta on mitattu (esim. Ortega 2012). Tässä tutkimuksessa keskitytään oppijansuomen kompleksisuuteen, jota tarkastellaan sekä vakiintuneiden syntaktisen kompleksisuuden määrällisten mittareiden avulla että kahdesta näitä täydentävästä näkökulmasta.

1.2 Tutkimuksen tavoitteet ja tutkimuskysymykset

Tämän tutkimuksen tavoitteena on selvittää, millaista kompleksisuus oppijansuomessa on kielitaidon eri taitotasoilla ja miten oppijansuomen kompleksisuutta voi mitata. Oppijankielen kompleksisuutta eri kohdekielten näkökulmasta on tutkittu vasta vähän (Housen ym. 2019: 11), vaikka aiemmat tutkimukset ovat osoittaneet, että syntaktisen kompleksisuuden määrälliset mittarit voivat tuottaa erilaisia tuloksia eri kohdekielissä (Gyllstad ym. 2014; Kuiken & Vedder 2019). Oppijansuomen kieliopillista kompleksisuutta on aiemmin tutkittu tämän aineiston kanssa päällekkäistä aineistoa käyttäen lähinnä yksittäisten kielenpiirteiden näkökulmasta (esim. Martin ym. 2010; Reiman 2011b; Seilonen 2013) sekä pienemmissä pitkittäisaineistoissa erilaisten määrällisten mittareiden avulla (esim. Spoelman & Verspoor 2010; Tilma 2014).

Tässä tutkimuksessa tarkastelun kohteeksi on valittu oppijansuomi, jonka kompleksisuutta kartoitetaan ensin tyypillisten oppijankielen syntaktisen kompleksisuuden määrällisten mittareiden avulla. Kuvaa oppijansuomen kompleksisuudesta täydennetään tarkastelemalla tutkimusaineiston teksteissä esiintyviä konjunktioita ja moniverbisiä konstruktioita. Väitöskirjan **tutkimuskysymys 1** on: *Miten oppijansuomen syntaktinen kompleksisuus kehittyy suhteessa Eurooppalaisen viitekehyksen taitotasoihin?*

Syntaktisen kompleksisuuden tutkimuksen määrällisiä mittareita on lukuisia, mutta niiden suhde kompleksisuuden määritelmiin ei aina ole selvä (Michel 2017: 64). Kompleksisuuden määritelmistä ja mittaamisesta käydäänkin jatkuvaa keskustelua (ks. esim. Housen ym. 2019). Määrällisiä mittareita on arvosteltu siitä, että ne ovat osittain ristiriidassa kompleksisuuden määritelmien kanssa ja että ne tavoittavat vain osan oppijankielen kompleksisuuden kehityksestä (De Clercq & Housen 2017; Ortega 2012; Pallotti 2015; Paquot 2019.) Esimerkiksi Bernardini ja Granfeldt (2019: 213) ovat myös kyseenalaistaneet käsityksen oppijankielen

kompleksisuuden mittareiden universaaliudesta. Tästä keskustelusta nousee tämän väitöskirjan **tutkimuskysymys 2**: *Miten syntaktisen kompleksisuuden määrälliset mittarit tavoittavat oppijansuomen kompleksisuuden kehittymisen?*

Oppijankielen kompleksisuuden tutkituin osa-alue on syntaktinen kompleksisuus (De Clercq & Housen 2017: 317). Tyypillisesti tutkittavana kohdekielenä on ollut englanti, tosin viime aikoina myös joidenkin muiden kohdekielten kompleksisuutta on tutkittu (esim. Brezina & Pallotti 2019). Uudet tutkimustulokset eri kielistä voivat täydentää tai haastaa vallitsevia käsityksiä siitä, mitä oppijankielen kompleksisuus on ja miten se näkyy oppijankielessä kielitaidon eri taitotasoilla. Näihin kysymyksiin etsii vastausta **tutkimuskysymys 3**: *Millaista oppijankielen kompleksisuus on, kun sitä tarkastellaan oppijansuomen näkökulmasta?*

Tutkimusaineistona on yhteensä 1 078 tekstiä (65 466 sanaa), jotka on kerätty aikuisilta ja yläkouluikäisiltä suomenoppijoilta Jyväskylän yliopiston Cefling-hankkeessa. Aineiston tekstit ovat testi- tai luokkatilanteessa kirjoitettuja epämuodollisia viestejä, muodollisia viestejä ja mielipidetekstejä, ja ne on Cefling-hankkeessa arvioitu Eurooppalaisen viitekehyksen taitotasoille. Aikuisten aineisto kattaa Eurooppalaisen viitekehyksen taitotasot A1–C2 ja nuorten aineisto taitotasot A1–B2. Lisäksi aineistoon kuuluu 453 tekstin (19 826 sanan) vertailuaineisto, joka on kerätty suomea ensikielenä puhuvilta yläkoululaisilta.

Ensimmäiseen tutkimuskysymykseen vastataan analysoimalla tutkimusaineiston tekstejä seitsemän aiemmissa tutkimuksissa yleisesti käytetyn syntaktisen kompleksisuuden määrällisen mittarin avulla (osatutkimus 2), tarkastelemalla teksteissä eri taitotasoilla esiintyviä konjunktioita (osatutkimus 3) ja kartoittamalla aikuisten aineiston eri taitotasoilla esiintyviä moniverbisiä konstruktioita (osatutkimus 4). Toiseen tutkimuskysymykseen vastataan kartoittamalla määrällisissä mittareissa käytettyjen kielenyksiköiden rajaamista oppijansuomen aineistossa (osatutkimus 1), arvioimalla määrällisten mittareiden kykyä erotella oppijansuomen taitotasoja (osatutkimus 2) sekä tarkastelemalla sellaisia oppijansuomen syntaktisen kompleksisuuden muutoksia, jotka eivät välttämättä näy tyypillisissä määrällisissä mittareissa (osatutkimukset 3 ja 4). Kolmanteen tutkimuskysymykseen vastataan tarkastelemalla osatutkimusten keskeisistä tuloksista muodostuvaa kokonaiskuvaa.

Tämän väitöskirjan rakenne on seuraava. Luvussa 2 esittelen lyhyesti tutkimukseni lähtökohdat, minkä jälkeen käsittelen ensin oppijankielen kompleksisuutta, sitten tässä tutkimuksessa keskeistä syntaktisen kompleksisuuden mittaamista ja lopuksi aiemmissa tutkimuksissa kertynyttä tietoa oppijansuomen kompleksisuudesta. Luvussa 3 kuvaan tutkimuksen aineiston, korpuksen annotointiprosessin sekä aineiston analyysissa sekä tilastollisessa kuvaamisessa ja vertailussa käytetyt menetelmät. Luvussa 4 esittelen artikkeleina julkaistut osatutkimukset, minkä jälkeen kokoan vastaukset tämän väitöskirjan tutkimuskysymyksiin luvussa 5. Väitöskirjan viimeisessä luvussa pohdin tulosten merkitystä ja esittelen tutkimuksessani esiin nousseita mahdollisia jatkotutkimusaiheita.

2 TUTKIMUKSEN TEOREETTINEN VIITEKEHYS

2.1 Oppijankielestä tutkimuskohteena

Oppijankieltä voidaan tarkastella yksilön hallitsemana kielisysteeminä tai käyttöympäristössä muotoutuvana vuorovaikutukseen osallistumisena (esim. Larsen-Freeman 2010). Oppijankielen tutkimuksen kohteena voivatkin olla niin oppijankielessä esiintyvät piirteet kuin kommunikatiivinen kompetenssi (Ellis & Barhuizen 2005: 362). Tässä tutkimuksessa läsnä ovat molemmat lähestymistavat. Tutkimusaineiston tekstit on sijoitettu Eurooppalaisen viitekehyksen taitotasoille viitekehyksen funktionaalisten taitotasokriteerien perusteella, ja aineiston oppijansuomen kielenpiirteitä tarkastellaan yhden näkökulman, oppijankielen kompleksisuuden, kautta.

Suomessa kielitaidon arviointiin käytetään yleisesti Eurooppalaista viitekehystä. Viitekehyksessä kielitaito nähdään kielenoppijan kykynä selvitä erilaisista viestintätilanteista. Vaikka viitekehystä käytetään myös kielitaidon arvioinnissa (esim. Huhta & Hilden 2016: 20), sen taitotasoja ei ole tarkoitettu lineaariseksi mitta-asteikoksi (EVK 2003: 10). Viitekehyksen taitotasot A1 ja A2 muodostavat yhdessä perustason, taitotasot B1 ja B2 itsenäisen kielenkäyttäjän tason ja taitotasot C1 ja C2 taitavan kielenkäyttäjän tason (mts. 47). Viitekehystä täydentää ja taustoittaa CEFR Companion Volume (CEFR Companion 2020). Tässä tutkimuksessa nojataan kuitenkin alkuperäiseen viitekehykseen, koska se on ollut käytössä tämän tutkimuksen aineiston arvioinnissa.

Oppijankielen kielenpiirteiden kehittymistä voidaan tarkastella sujuvuuden, tarkkuuden ja kompleksisuuden käsitteiden avulla. Tarkkuudella viitataan oppijankielen virheettömyyteen ja sitä tarkastellaan suhteessa kohdekielen normeihin (Housen & Kuiken 2009: 462–463). Sujuvuutta voidaan etenkin arkipuheessa käyttää kielitaidon synonyymina, mutta toisen kielen oppimisen kontekstissa sillä usein tarkoitetaan jonkinlaista kielen tuottamisen helppoutta (mts. 463).

Kaikkein vähiten yksimielisyyttä on kompleksisuuden määritelmästä. Oppijankielen kompleksisuudella voidaan toisen kielen oppimisen tutkimuksessa tarkoittaa monia eri asioita, mikä voi johtua ainakin osittain englannin kielen sanan *complexity* monitulkintaisuudesta (esim. Housen ym. 2019: 9). Kompleksisuuteen ja sen mittaamiseen palataan kahdessa seuraavassa alaluvussa. Englanninkielisessä tutkimuksessa tarkkuuden, sujuvuuden ja kompleksisuuden muodostamaan kolmikkoon viitataan usein myös lyhenteellä CAF (*Complexity, Accuracy, Fluency*). Näiden ilmiöiden suhde kielitaidon taitotasoon ei ole yksiselitteinen. Vaikka kielitaidon kehityksen voi ajatella näkyvän näiden kaikkien kolmen osa-alueen lisääntymisenä ja intuitiivisesti on helppo ajatella, että etenkin tarkkuuden ja sujuvuuden kohdalla enemmän on parempaa, näin ei välttämättä ole (ks. esim. Pallotti 2009: 597–599). Ilmaus voi myös olla hyvinkin kompleksinen, tarkka ja sujuva mutta ei silti välttämättä käyttötilanteeseen sopiva (Michel 2017: 62).

Kompleksisuus, tarkkuus ja sujuvuus eivät ole ainoa tapa lähestyä kielenpiirteiden kehittymistä. Oppijansuomen taidon kehittymisen seuraamiseen on Cefling-hankkeessa kehitetty niin sanottu DEMfad-malli, joka on saanut alkunsa tarpeesta luoda CAF-triadia ja etenkin kompleksisuutta paremmin kirjoitetun oppijansuomen analyysiin soveltuva lähestymistapa. DEMfad-mallissa tarkastellaan opittavan kielenpiirteen tai konstruktion (*Domain*) kehittymistä sen ensiesiintymisestä (*Emergence*) kyseisen ilmiön hallintaan (*Mastery*). Kehitystä seurataan konstruktion käyttötaajuuden (*frequency*), tarkkuuden (*accuracy*) ja distribuution tai variaation (*distribution*) avulla. Mallissa kieltä tarkastellaan konstruktioina, joiden katsotaan ilmaantuvan oppijankieleen silloin, kun ensimmäinen esiintymä on tunnistettavissa. Konstruktion oppiminen katsotaan alkavaksi sen ensimmäisestä tunnistettavasta esiintymästä, ja tarkasteltavan konstruktion hallinta katsotaan saavutetuksi, kun 80 % esiintymistä on kohdekielen käytänteiden mukaisia. Distribuutiota voidaan tarkastella konstruktion laajenemisena ja muuntumisena. (Franceschina ym. 2006; Martin 2022: 83–85; Martin ym. 2010: 58–62; Mustonen 2015: 84–89; Seilonen 2013: 29–32.) Vaikka tässä väitöskirjassa DEMfad-mallia ei käytetä analyysin työkaluna, siltä lainataan kaksi lähtöoletusta. Ensimmäinen liittyy käsitykseen kielenpiirteiden oppimisesta. Kielenpiirteen oppimisella voidaan tarkoittaa joko piirteen ilmaantumista oppijankieleen tai sen kohdekielen mukaista käyttöä (Ellis 2012: 8). Tässä tutkimuksessa kielenpiirteen ilmaantuminen oppijankieleen katsotaan DEMfad-mallin tapaan ensimmäiseksi osoitukseksi sen oppimisesta (ks. Martin ym. 2010: 59–60). Lisäksi myös tässä tutkimuksessa joitakin kielen rakenteita tarkastellaan konstruktioina, vaikka syntaktisen kompleksisuuden määrällisessä tutkimuksessa perusyksikköinä käytetään tyypillisesti sanoja, lauseita, T-yksiköitä tai virkkeitä.

2.2 Kompleksisuus osana toisen kielen taitoa

Oppijankielen kompleksisuuden erilaisia määritelmiä yhdistää ajatus kompleksisuudesta oppijankielen monipuolisuutena, vaihteluna tai kehittyneisyytenä (Ellis & Barkhuizen 2005: 139; Ortega 2015: 86). Oppijankielen kompleksisuutta voidaan tarkastella siihen kuuluvien muotojen tai rakenteiden monipuolisuutena ja kehittyneisyytenä (Ortega 2003: 492) tai oppijankielen sisältämien komponenttien ja niiden välisten yhteyksien määränä ja monipuolisuutena (esim. Bulté & Housen 2012: 24; Pallotti 2015: 120). Määritelmien eroihin vaikuttaa se, että oppijankielen kompleksisuutta on tarkasteltu eri näkökulmista (ks. Pallotti 2015: 118). Ortega (2012) jakaa oppijankielen kompleksisuuden tutkimukset kolmeen ryhmään sen mukaan, mistä näkökulmasta kompleksisuutta on niissä lähestytty. Tyypillisesti kompleksisuutta on tarkasteltu suhteessa kielitaidon kehitykseen, suhteessa tehtävän vaatimuksiin tai suhteessa kielitaidon tasoon. Näkökulma vaikuttaa myös tutkimuksen tavoitteisiin. Kun kompleksisuutta tarkastellaan suhteessa kielitaidon kehitykseen, tavoitteena on yleensä löytää yhteys kehitysvaiheen ja kompleksisuuden välillä. Kompleksisuuden tarkastelussa suhteessa tehtävän vaatimuksiin tavoitteena on tyypillisesti löytää yhteys tehtävän kognitiivisen kompleksisuuden ja oppijankielen kompleksisuuden välillä. Lisäksi kompleksisuutta on tutkittu suhteessa kielitaidon tasoon, jolloin tavoitteena on löytää sellaisia kompleksisuuden piirteitä, joiden avulla taitotaso voidaan määrittää. (Mts. 128–130.) Tässä väitöskirjassa oppijansuomen kompleksisuutta tarkastellaan suhteessa kielitaidon tasoon aineistossa, jonka tekstit on arvioitu Eurooppalaisen viitekehyksen eri taitotasoille.

Myös typologisessa kielentutkimuksessa kompleksisuutta voidaan tarkastella kielisysteemin näkökulmasta absoluuttisena kompleksisuutena (Miestamo 2008) tai kielenkäyttäjän näkökulmasta relatiivisena kompleksisuutena (Kusters 2008). Eroa on myös siinä, mitä kompleksisuudella tarkoitetaan. Kompleksisuudella voidaan viitata teoreettiseen käsitteeseen, jota käytetään kielen kuvaamiseen, tai empiiriseen ilmiöön, jota pyritään selittämään teorian avulla (Kusters 2008: 4). Tässä tutkimuksessa kompleksisuutta lähestytään empiirisenä ilmiönä, joka nähdään kielisysteemin ominaisuutena eli absoluuttisena tai lingvistisenä kompleksisuutena ja jota tarkastellaan kielenoppijoiden kirjoittamissa teksteissä esiintyvien kielenpiirteiden avulla. Tarkastelussa ei oteta kantaa kielenpiirteiden mahdolliseen vaikeuteen eikä yksilötason oppimisprosessiin.

Tyypillisesti oppijankielen kompleksisuutta lähestytään moniulotteisena ilmiönä. Bulté & Housen (2012) jakavat sen leksikaaliseen ja kieliopilliseen kompleksisuuteen, joista jälkimmäisen he jakavat edelleen syntaktiseen ja morfologiseen kompleksisuuteen. Näistä eniten on tutkittu syntaktista kompleksisuutta (Brezina & Pallotti 2019: 100; Bulté & Housen 2012; De Clercq & Housen 2017: 317; Lu 2017: 496). Leksikaalinen, syntaktinen ja morfologinen kompleksisuus eivät kuitenkaan ole toisistaan irrallisia ilmiöitä. Esimerkiksi sanavaraston laajuus voi vaikuttaa syntaksiin, kun oppija kompensoi epätarkkaa tai yksinkertaista sanaa tavalla, joka syntaktisen kompleksisuuden mittareissa näkyy suurempana

kompleksisuutena, esimerkiksi käyttämällä yleissanaa ja selittämällä sitä tarkentavalla relatiivilauseella (De Clercq & Housen 2017: 330; Lambert & Nakamura 2019: 251). Tämän tutkimuksen neljännessä osatutkimuksessa näitä oppijankielen kompleksisuuden ulottuvuuksia tarkastellaan yhdessä tutkimalla moniverbisiä rakenteita konstruktioina.

Norris & Ortega (2009) ovat esittäneet, että myös syntaktinen kompleksisuus on moniulotteinen ilmiö. Tyypillisesti oppijankielen tutkimuksessa keskeisessä asemassa on ollut T-yksikkö eli päälauseen ja sille alisteisten lauseiden muodostama kokonaisuus, jota on käytetty myös ensikielen kehityksen tutkimuksessa (Foster, Tonkyn & Wiggelsworth 2000; Ortega 2003; Wolfe-Quintero ym. 1998), mutta sen käyttöön oppijankielen tarkastelussa on suhtauduttu myös epäillen (Bardovi-Harlig 1992; Polio 2012; Rimmer 2006). Samoin kuin ensikielen myös toisen kielen taidon voidaan ajatella kehittyvän yksinkertaisista ilmauksista kohti yhä kompleksisempaa ilmaisua, kun kielitaito kehittyy, minkä ajatellaan syntaktisessa kompleksisuudessa näkyvän niin, että alkeistasolla tyypillistä on lauseiden rinnastaminen, keskitasolla sivulauseiden käyttö, ylimmillä taitotasoilla lausekkeiden piteneminen ja monipuolistuminen (esim. Kuiken & Vedder 2019; Norris & Ortega 2009; Wolfe-Quintero ym. 1998). Syntaktinen kompleksisuus voi siis ilmetä kielitaidon eri tasoilla eri tavoin, mikä tulee ottaa huomioon myös syntaktisen kompleksisuuden mittaamisessa (esim. Norris & Ortega 2009; Ellis & Barkhuizen 2005: 155). Tämän väitöskirjan toisessa osatutkimuksessa käytetyt syntaktisen kompleksisuuden määrälliset mittarit on valittu niin, että mukana on niin yleisen kompleksisuuden, alisteisuuden, rinnasteisuuden kuin lausetason yleisen kompleksisuuden vakiintuneita mittareita.

Kielitaidon taso ei ole ainoa kompleksisuuteen vaikuttava tekijä. Esimerkiksi tekstilaji voi vaikuttaa oppijankielen kompleksisuuteen (esim. Michel 2017: 54, 60), ja myös puheen ja kirjoitetun kielen kompleksisuus voi olla erilaista (esim. Biber, Gray & Poonpon 2011; Biber, Gray & Staples 2016). Lintusen ja Mäkilän (2014) mukaan tämä voi johtua niin puheen ja kirjoittamisen välisistä eroista kuin niistä mittareista, joilla syntaktista kompleksisuutta analysoidaan. Eroja on myös siinä, millaisia tuloksia mittarit tuottavat eri kohdekielissä. Eurooppalaisen viitekehyksen taitotasojen ja syntaktisen kompleksisuuden suhdetta tutkineet Gyllstad, Granfeldt, Bernardini ja Källkvist (2014) totesivat, että taitotasoille A1–B2 sijoitetuissa englannin, ranskan ja italian kielen oppijoiden teksteissä T-yksikön sanamäärä ja lauseen sanamäärä poikkesivat tilastollisesti merkitsevästi A-tason ja B-tason välillä, kun kohdekielinä olivat englanti ja ranska, mutta eivät silloin, kun kohdekielenä oli italia. Samassa aineistossa syntaktista kompleksisuutta kieliopillisten kategorioiden avulla tarkastelleet Bernardini ja Granfeldt (2019) totesivat kuitenkin B-tasolle arvioidut tekstit tilastollisesti merkitsevästi kompleksisemmiksi kuin A-tasolle sijoitetut tekstit sekä silloin, kun kaikkia kieliä tarkasteltiin yhdessä, että silloin, kun kieliä tarkasteltiin erikseen. Myös Kuiken ja Vedder (2019) löysivät syntaktisen kompleksisuuden mittareissa eroja kielten välillä taitotasoista A2–B1 koostuneessa oppijankielen aineistossa, jossa kohdekielinä olivat hollanti, italia ja espanja. Aiemmat tutkimukset ovat osoittaneet myös, että ensikieli voi vaikuttaa toisen kielen kompleksisuuteen (esim. Khushik & Huhta

2010; Lu & Ai 2015). Lisäksi eri kohdekielten välillä voi olla eroja siinä, mitä pidetään kompleksisena tai mitä pidetään osoituksena hyvästä kieli- tai kirjoitustaidosta. Esimerkiksi suomen kielen konjunktioalkuisten sivulauseiden soveltuvuus oppijansuomen kompleksisuuden mittaamiseen voidaan kyseenalaistaa, sillä ne eivät lauseen aloittavaa konjunktiota lukuun ottamatta poikkea syntaktisesti tai morfologisesti päälauseista eivätkä runsaasti sivulauseita sisältävät virkkeet välttämättä ole tyylillisesti tavoiteltavia suomeksi kirjoitettaessa (Martin ym. 2010: 61).

Syntaktisen kompleksisuuden ja kieli- tai kirjoitustaidon välinen suhde ei olekaan yksiselitteinen. Syntaktisen kompleksisuuden runsaus ei aina ole osoitus hyvästä tai edistyneestä kieli- tai kirjoitustaidosta (esim. Crossley & McNamara 2014; Michel 2017: 53–54; Ortega 2003; Taguchi, Crawford & Wetzel 2013: 426). Toisaalta kompleksisten ajatusten ilmaisemiseen ei välttämättä tarvita kompleksista syntaksia, vaan harjaantuneet kirjoittajat saattavat käyttää aloittelijoita yksinkertaisempaa ilmaisua välittäessään kompleksisia ajatuksia (Lambert & Kormos 2014: 612). Kompleksisuuden määrään voivat lisäksi vaikuttaa kirjoittajan tyylivalinnat, ja tilanteista ja henkilökohtaista vaihtelua kompleksisuudessa voidaan löytää myös ensikielisestä aineistosta (Pallotti 2009). Joissakin viestintätilanteissa toisen kielen oppijat voivat myös käyttää ensikielisiä puhujia kompleksisempaa syntaksia, ja oppimisen edetessä oppijankielen kompleksisuus voi alkaa muistuttaa kohdekielen ensikielisten käyttäjien kompleksisuutta niin, että se joissakin tehtävissä lisääntyy ja joissakin vähenee, toisin sanoen kielitaidon kehittyminen voi joskus näkyä myös kompleksisuuden vähenemisenä (mts. 597–598).

2.3 Oppijankielen syntaktisen kompleksisuuden mittaaminen

Syntaktista kompleksisuutta mitataan tyypillisesti määrällisillä mittareilla, joiden avulla tarkastellaan lauseiden, virkkeiden tai T-yksiköiden sisältämää sanamäärää tai niiden suhdetta toisiinsa. Mittareita on kehitetty lukuisia, ja niiden suuri määrä voi vaikeuttaa sekä mittareiden valintaa että eri tutkimuksissa saatujen tulosten vertaamista (Ellis & Barkhuizen 2005: 163; Michel 2017: 64). Oppijankielen kompleksisuuden tutkimusta on arvosteltu myös siitä, ettei tutkimuksissa käytettyjen mittareiden yhteys kompleksisuuden määritelmään tai teoreettisiin taustaoletuksiin aina ole selvä (Bulté & Housen 2012; Norris & Ortega 2009). Yleisimmin käytettyihin mittareihin kuuluvat T-yksikön keskipituus (*MLT* tai *MLTU* eli *Mean Length of T-unit*) ja lauseen keskipituus (*MLC* eli *Mean Length of Clause*), joissa molemmissa pituutta mitataan sanamäärän avulla. Usein käytettyjä ovat myös T-yksiköiden sisältämien lauseiden määrä (*C/T* tai *C/TU* eli *Mean Number of Clauses per T-unit*) tai sivulauseiden osuus kaikista lauseista (*DC/C* eli *Mean Number of Dependent Clauses per Clause*). Lisäksi syntaktista kompleksisuutta on mitattu edistyneeseen kielitaitoon liitettyjen kielenpiirteiden, kuten passiivin, esiintymistaajuuden avulla. (Kuiken & Vedder 2019; laajempi katsaus

eri mittareista ks. esim. Bulté & Housen 2012; Norris & Ortega 2009; Ortega 2003; Wolfe-Quintero ym. 1998).

Nykyään syntaktista kompleksisuutta lähestytään usein Norrisin ja Ortegan (2009) suosituksen mukaisesti useiden rinnakkaisten mittareiden avulla moniulotteisena ilmiönä, joka voi näkyä kielitaidon eri tasoilla eri tavoin. Yleistä syntaktista kompleksisuutta mitataan tyypillisesti virkkeen tai T-yksikön sanamäärän avulla (Bulté & Housen 2012; Norris & Ortega 2009). T-yksikön pituutta on pidetty sekä sujuvuuden mittarina (ks. Wolfe-Quintero ym. 1998) että kompleksisuuden mittarina, eikä tutkijoiden keskuudessa ole yksimielisyyttä siitä, kumpaa sen avulla mitataan (Ortega 2012: 141). Molempia on pidetty hyvinä mittareina, sillä etenkin keskitasolla niiden ja kielitaidon kehittymisen välillä on useissa tutkimuksissa havaittu yhteys (ks. Ortega 2003; Wolfe-Quintero ym. 1998). On kuitenkin mahdollista, että eri tutkimuksissa havaittu lineaarinen kasvu on tyypillistä vain joillekin kielitaidon tasoille tai kehitysvaiheille ja että tietyn tason jälkeen kasvu tasaantuu (Bulté & Housen 2012: 37). Näiden kahden mittarin lisäksi yleistä syntaktista kompleksisuutta voidaan mitata myös virkkeen sisältämien lauseiden määrällä (*C/S* eli *Mean Number of Clauses per Sentence*), joka on mukana esimerkiksi Lun (2010, 2011) kehittämässä syntaktisen kompleksisuuden automatisoidussa analysointisovelluksessa. Nämä kaikki kolme yleisen kompleksisuuden mittaria ovat mukana tämän väitöskirjan osatutkimuksessa, jossa tarkastellaan määrällisten mittareiden soveltumista oppijansuomen taitotason määrittämiseen.

Syntaktisen kompleksisuuden tutkimuksessa suosituimpia ovat olleet alisteisuuteen (engl. *subordination*) keskittyvät mittarit, kuten T-yksikön sisältämien lauseiden määrä ja sivulauseiden osuus kaikista lauseista (esim. Bulté & Housen 2012). Molempien on tutkimuksissa havaittu kasvavan lineaarisesti kielitaidon kehittyessä, mutta muutokset eivät aina ole olleet tilastollisesti merkitseviä (esim. Wolfe-Quintero ym. 1998). Vaikka alisteisiin lauseisiin perustuvia mittareita usein pidetään hyvinä syntaktisen kompleksisuuden mittareina etenkin kielitaidon keskitasolla (esim. Norris & Ortega 2009), alisteisuuden keskeistä asemaa syntaktisen kompleksisuuden mittaamisessa on myös arvosteltu (esim. Lambert & Kormos 2014: 608–609). Myös Bulté ja Housen (2012: 37) ovat nostaneet esiin sen, että alisteisuuteen perustuvat mittarit kiinnittävät huomiota vain yhteen kompleksisuuden puoleen jättäen sivuun sen muut ulottuvuudet, kuten lauseiden rinnastamisen ja lauseketason kompleksisuuden. Biber, Gray ja Poonpon (2011: 29) ovat huomauttaneet, että englannin kieltä koskevien korpustutkimusten mukaan sivulauseet ovat tyypillisempiä puheessa kuin akateemisessa tekstissä ja että akateemisten tekstien kompleksisuuden mittaamiseen tarvitaan erilaisia mittareita kuin puhutun kielen kompleksisuuden mittaamiseen. Oppijankielen fraseologista kompleksisuutta tutkinut Paquot (2019: 134) puolestaan totesi, ettei tutkimuksessa käytetyissä alisteisuuteen perustuvissa syntaktisen kompleksisuuden mittareissa ollut systemaattista kehitystä tai tilastollisesti merkitseviä eroja taitotasojen B2 ja C2 välillä edistyneiden englanninoppijoiden teksteistä koostuneessa tutkimusaineistossa. Paquotin mukaan onkin mahdollista,

että etenkin ylemmillä taitotasoilla kielitaidon kehitys voi näkyä enemmän fraseologisessa kuin syntaktisessa tai leksikaalisessa kompleksisuudessa ja että edistyneessä oppijankielessä fraseologinen kompleksisuus vaikuttaa olennaiselta osalta toisella kielellä kirjoittamisen taitoa (mts. 139, 141). Koska alisteisuuteen perustuvilla mittareilla on kuitenkin ollut keskeinen asema oppijankielen syntaktisen kompleksisuuden tutkimuksessa, myös tässä tutkimuksessa on mukana kaksi lauseiden alisteisuuteen perustuvaa mittaria.

Lauseiden välisten alistussuhteiden lisäksi kompleksisuutta voidaan mitata rinnastamiseen keskittyvillä mittareilla. Rinnastamista voidaan tarkastella esimerkiksi virkkeen sisältämien T-yksiköiden määrän (*T/S* tai *TU/S* eli *Mean Number of T-units per Sentence*) ja Bardovi-Harligin (1992) kehittämän koordinaatioindeksin (*CI* eli *Coordination Index*) avulla. Vaikka rinnastamista mittaavaa virkkeen sisältämien T-yksiköiden määrää ei perinteisesti ole pidetty toisen kielen oppimisen näkökulmasta hyödyllisenä mittarina (Wolfe-Quintero ym. 1998), Norris ja Ortega (2009) suosittelevat sen käyttämistä etenkin kielitaidon alemmilla tasoilla, joilla rinnasteiset suhteet lauseiden välillä ovat heidän mielestään tyypillisempiä kuin alisteiset suhteet. T-yksiköiden määrä virkkeessä on mittari, joka tavoittaa vain päälauseiden rinnastamisen. Sivulauseiden väliset rinnastussuhteet ja lausekkeiden rinnastaminen lauseen sisällä jäävät sen avulla havaitsematta. Esimerkiksi Lu (2010: 491) on kuitenkin havainnut, että kielitaidon edistyessä T-yksiköt voivat pidetä myös rinnasteisten lausekkeiden, ei vain sivulauseiden, määrän lisääntyessä. Yksi tapa tarkastella rinnastamista laajemmin on tutkia, miten rinnastuskonjunktioita käytetään oppijankielessä, kuten oppijansaksaa tutkinut Vyatkina (2012, 2013) on tehnyt. Konjunktioiden käyttöä osana oppijankielen syntaktista kompleksisuutta ovat tutkineet myös esimerkiksi Benevento ja Storch (2011) ranskanoppijoiden teksteissä ja Grant ja Ginther (2000) sekä Taguchi, Crawford ja Wetzel (2013) englanninoppijoiden teksteissä. Oppijansuomen konjunktioita ovat aiemmin tarkastelleet oppijansuomen sanastoa tutkineet Määttä (2012) ja Honko (2013) sekä edistyneen oppijansuomen kielenpiirteitä tarkastellut Ivaska (2015). Myös tämän väitöskirjan yhdessä osatutkimuksessa rinnastamista samoin kuin lauseiden välisiä alistussuhteita tarkastellaan aineiston teksteissä esiintyvien konjunktioiden avulla.

Edellä kuvatut syntaktisen kompleksisuuden määrälliset mittarit keskittyvät lausetta laajempiin kielenyksiköihin. Norrisin ja Ortegan (2009) mukaan etenkin ylemmillä taitotasoilla syntaktisen kompleksisuuden mittaamisessa tulee ottaa huomioon myös lauseen sisällä tapahtuvat muutokset, joita voidaan yleisellä tasolla mitata lauseen sanamäärän (MLC) avulla. Lauseiden piteneminen on aiempien tutkimustulosten valossa yhteydessä kielitaidon kehittymiseen: ylemmillä taitotasoilla käytetyt lauseet ovat tyypillisesti pitempiä kuin alemmilla taitotasoilla käytetyt (ks. esim. Lu 2011; Ortega 2003). Lauseen sanamäärää käytetään yhtenä syntaktisen kompleksisuuden mittarina myös tässä väitöskirjassa.

Oppijankielen kompleksisuuden tutkimuksessa on usein käytetty vain yhtä tai kahta mittaria, jotka ovat saattaneet keskittyä samaan kompleksisuuden osa-alueeseen (Bulté & Housen 2012: 34). Lisäksi tutkimusaineistot ovat usein olleet varsin pieniä, mihin on osaltaan saattanut vaikuttaa aineiston käsittelyn työläys

(esim. Lu 2011: 475). Tästä syystä syntaktisen kompleksisuuden mittaamiseen on kehitetty automatisoituja sovelluksia, joiden avulla laajoistakin aineistoista voidaan mitata monia erilaisia kielenpiirteitä ja näin saada yhdellä kertaa mittaustuloksia kompleksisuuden eri osa-alueista. Tällaisia ovat esimerkiksi Biber Tagger (Biber ym. 1999), Coh-Metrix (McNamara ym. 2014) ja L2 Syntactic Complexity Analyzer (Lu 2010, 2011), joita voidaan käyttää englanti toisena kielenä -aineistojen syntaktisen kompleksisuuden analysointiin.

Määrälliset mittarit tavoittavat kuitenkin vain osan oppijankielen erilaisten piirteiden kehityksestä. Esimerkiksi yleisen kompleksisuuden mittarina pidetty T-yksiköiden keskipituus sanoina (MLTU) havaitsee kyllä T-yksiköiden sanamäärän muutokset mutta ei T-yksiköiden välisiä laadullisia eroja. Saman sanamäärän sisältävät T-yksiköt voivat olla rakenteeltaan hyvin erilaisia (Biber ym. 2011: 14; Kyle & Crossley 2018: 334; Rimmer 2006: 507). Määrä ei myöskään välttämättä ole osoitus laadusta. Kuten Ellis ja Barkhuizen (2005: 155) huomauttavat, alisteinen sivulause on alisteinen silloinkin, kun se ei ole kohdekielen mukainen. Määrällisten mittareiden luotettavuuteen vaikuttaa myös se, että määrällisiä mittareita varten oppijankieli on segmentoitava mittareissa käytettyihin yksiköihin. Oppijankieli ei kuitenkaan etenkään alemmilla taitotasoilla aina ole yksiselitteisesti segmentoitavissa näihin kielenyksiköihin, on kyse sitten puhutusta (esim. Foster ym. 2000) tai kirjoitetusta kielestä (esim. Martin 2013a). Lisäksi oppijankielen vaihteluun kuuluu se, että siinä voi esiintyä rinnakkain sekä kohdekielen mukaisia että kohdekielestä poikkeavia muotoja (Ellis 2012: 117).

Toinen syntaktisen kompleksisuuden määrällisiin mittareihin liittyvä ongelma on, että ne keskittyvät usein vain johonkin tiettyyn oppijankielen kompleksisuuden ulottuvuuteen (Bulté & Housen 2012). Vaikka monimuotoisuus tai monipuolisuus on usein mukana syntaktisen kompleksisuuden määritelmissä ja vaikka diversiteetti otetaan huomioon leksikaalisen kompleksisuuden (ks. Jarvis 2013) ja morfologisen kompleksisuuden (esim. Brezina & Pallotti 2017) tutkimuksessa, syntaktisen kompleksisuuden tutkimuksessa monimuotoisuus on jäänyt vähälle huomiolle (De Clercq & Housen 2017: 317). Yksi ratkaisu on tarkastella oppijankielen kompleksisuutta määrällisten mittareiden lisäksi laadullisesti. Esimerkiksi oppijankielen konstruktioiden tarkastelu voi nostaa näkyväksi sellaisia kompleksisuuden kehityskulkuja, joita määrälliset mittarit eivät tavoita (Reiman 2011b). Siksi tässä tutkimuksessa oppijansuomen syntaktista kompleksisuutta tarkastellaan perinteisten määrällisten mittareiden ja teksteissä esiintyvien konjunktioiden lisäksi myös moniverbisten konstruktioiden avulla.

2.4 Oppijansuomen kompleksisuus ja sen kehitys

Oppijansuomen kompleksisuutta on tähän mennessä tarkasteltu lähinnä joko pienehköissä pitkittäisaineistoissa tai yksittäisten kielenpiirteiden näkökulmasta. Pitkittäisaineistossa oppijansuomen sujuvuuden kehitystä tutkinut Alisaari

(2016) on tarkastellut T-yksikön sanamäärän kehittymistä 32 oppijan kirjoittamissa kertomuksissa taitotasolla A2. Tulosten mukaan T-yksiköiden sanamäärissä ei tapahtunut merkittäviä muutoksia neljän viikon seuranta-aikana (mts. 44).

Spoelman ja Verspoor (2010) ovat tutkineet oppijansuomen kompleksisuutta ja tarkkuutta dynaamisten systeemien teorian (DST) näkökulmasta. Tutkimusaineistona oli 54 yhdeltä suomenoppijalta kolmen ensimmäisen opiskeluvuoden aikana kerättyä tekstinäytettä. Kompleksisuutta tarkasteltiin kolmella tasolla: sanatasolla sanan sisältämien morfeemien määrän avulla, lauseketasolla substantiivilausekkeiden sisältämien sanojen määrällä ja virketasolla jakamalla virkkeet yksinkertaisiin virkkeisiin (*simple sentence*), yhdyslauseisiin (*compound sentence*), kompleksisiin virkkeisiin (*complex sentence*) ja kompleksisiin yhdyslauseisiin (*compound-complex sentence*). Tulosten mukaan näillä mittareilla mitattu kompleksisuus lisääntyi mutta kasvu ei ollut lineaarista. (Spoelman & Verspoor 2010.)

Oppijansuomen syntaktista ja morfologista kompleksisuutta on tarkastellut Tilma (2014), joka tutki suomea toisena ja vieraana kielenä opiskelevien suomen kielen kompleksisuuden ja tarkkuuden kehitystä. Tilman aineistona oli kahdeksalta opiskelijalta yhdeksän kuukauden aikana kerätyt tekstinäytteet. Syntaktisen kompleksisuuden mittarina Tilma käytti samoja lause- ja virketyyppejä kuin Spoelman ja Verspoor (2010), minkä lisäksi hän tarkasteli virkkeen ja lauseen keskipituutta morfeemeina sekä eri sijamuotojen käyttöä. Morfologisen kompleksisuuden mittareina toimivat sanojen keskipituus morfeemeina ja menneen ajan aikamuotojen käyttö. Tilman mukaan parhaina mittareina toimivat lauseen ja virkkeen morfeemimäärä, jotka molemmat kasvoivat oppimisen edetessä, vaikka mittareiden ja ryhmätason kehityksen välillä ei ollut tilastollisesti merkitsevää korrelaatiota. (Tilma 2014.)

Ensikielisten lasten kielen kompleksisuutta on tutkittu Scarborough'n (1990) kehittämän produktiivisen syntaksin indeksin (IPSyn) avulla. Indeksin suomen kieleen kehitetyn version (Nieminen ja Torvelainen 2003) soveltumisesta alle kouluikäisten puhutun oppijansuomen aineistoon on raportoinut Suni (2006). Sunin mukaan indeksin pistemäärien kasvun ja suomenkielisessä hoidossa vietetyn ajan pituuden välillä oli odotuksenmukainen yhteys ja indeksi sopi hyvin alle kouluikäisten S2-puhujien morfosyntaktisen kompleksisuuden tarkasteluun (mts. 436).

Morfosyntaktinen kompleksisuus voi myös olla ensikielisten ja suomenoppijoiden tekstejä erottava tekijä. Oppijansuomen konstruktiopiirteitä edistyneiden suomenoppijoiden teksteistä koostuvassa aineistossa tarkastellut ja niitä ensikielisten suomenpuhujien aineistoon verrannut Ivaska (2015) on todennut, että ensikielisten teksteissä on oppijansuomen aineistoa runsaammin morfosyntaktista kompleksisuutta. Erot näkyivät esimerkiksi A-infinitiivin ja VA-partisiipin käytössä. Lisäksi suomenoppijoiden ensikieli näytti vaikuttavan konjunktioiden käyttöön aineiston teksteissä. (Ivaska 2015.)

Oppijansuomen kompleksisuutta tai kompleksisina pidettyjen kielenpiirteiden esiintymistä suhteessa Eurooppalaisen viitekehyksen taitotasoihin on sel-

vitetty myös tämän tutkimuksen aineiston kanssa päällekkäisessä poikittaisaineistossa. Yksi tällainen kielenpiirre on passiivi. Impersonaalisen passiivin käyttöä Cefling-aineistossa tarkastellut Seilonen (2013) totesi passiivin ilmaantuvan sekä aikuisten että nuorten teksteihin heti taitotasolla A1. Nuorten aineistossa impersonaalinen passiivi hallittiin DEMfad-mallin mukaan jo tällä taitotasolla, aikuisten aineistossa taitotasolla B1. Passiivin käyttö lisääntyi aikuisten aineistossa tilastollisesti merkitsevästi taitotasoilla A1–C2. Nuorten aineistossa käytön lisääntyminen ei ollut tilastollisesti merkitsevää määrällisessä tarkastelussa mukana olleiden taitotasojen A1–B1 välillä. Seilonen havaitsi myös aikamuotojen ja modusten variaation lisääntyvän taitotason noustessa. Sekä aikuisten että nuorten aineistossa muodon variaatiota ja leksikaalista variaatiota oli eniten kahdella ylimmällä taitotasolla. Aikuisten aineistossa muodon variaatiota oli eniten taitotasolla C1 ja nuorten aineistossa taitotasolla A2, leksikaalista variaatiota puolestaan aikuisten aineistossa taitotasolla C2 ja nuorten aineistossa taitotasoilla B1. (Seilonen 2013.)

Oppijansuomen kompleksisuutta on sekä määrällisestä että laadullisesta näkökulmasta tarkastellut Reiman (2011b), joka on tutkinut transitiivikonstruktion kehittymistä Cefling-aineiston aikuisten (Reiman 2011a) ja nuorten (Reiman 2014) teksteissä. Reiman havaitsi, että yleensä keskitasolle tyypillisenä pidettyjä alisteisia sivulauseita käytettiin jo aineiston taitotasolla A1 ja että keskitasolla, etenkin taitotasolla B1, sivulauseiden runsaus näyttäytyi jopa liikakäyttönä. Lisäksi Reiman totesi, että transitiivikonstruktioiden laadullisen tarkastelun avulla päästään kiinni oppijankielen variaatioon ja sen kautta oppijankielen syntaktisten resurssien monipuolisuuteen ja kehittymiseen. (Reiman 2011b.)

Tässä väitöskirjassa oppijansuomen syntaktista kehitystä tarkastellaan suhteessa kielentaidon taitotasoihin A1–C2. Tarkastelussa käytetään seitsemää syntaktisen kompleksisuuden määrällistä mittaria, minkä lisäksi määrällisten mittareiden tavoittamattomiin mahdollisesti jääviä oppijansuomen syntaktisen kompleksisuuden puolia tarkastellaan aineiston teksteissä esiintyvien konjunktioiden ja moniverbisten konstruktioiden avulla.

3 TUTKIMUKSEN AINEISTO JA MENETELMÄT

3.1 Cefling-aineisto tässä tutkimuksessa

Nyt käsillä olevan tutkimuksen aineisto on koostettu suomi toisena kielenä -teksteistä, jotka on kerätty ja arvioitu Eurooppalaisen viitekehyksen mukaisille taitotasoille Jyväskylän yliopiston Cefling-hankkeessa (Cefling-hankkeesta ks. Jantunen & Pirkola 2015). Aineisto on kerätty kielen oppimisen tutkimusta varten, ja sen avulla on mahdollista tutkia erilaisten kielenpiirteiden esiintymistä Eurooppalaisen viitekehyksen eri taitotasoilla (esim. Alanen, Huhta & Tarnanen 2010). Cefling-aineistoa on käytetty muun muassa nollapersoonaisten ilmausten (Seilonen 2013), eksistentiaalilauseiden (Kajander 2013), transitiivikonstruktion (Reiman 2011a, 2011b, 2014) sekä paikallissijojen (Mustonen 2015) käyttöön keskittyneissä tutkimuksissa. Lisäksi Cefling-hankkeen aineistoa hyödyntävissä pro gradu -tutkielmissa on käsitelty esimerkiksi *olla*-verbirakenteita (Kynsijärvi 2007) sekä verbiketjuja ja niiden kehittymistä (Haapala 2008; Paavola 2008) (Cefling-hankkeen yhteydessä tehdyistä maisterintutkielmista tarkemmin esim. Martin 2022).

Cefling-aineisto sisältää kirjoitettuja tekstejä aikuisilta ja yläkouluikäisiltä suomenoppijoilta. Aikuisten aineiston tekstit on kerätty Yleisten kielitutkintojen (YKI) kirjoittamisen osakokeen suorituksista, nuorten aineisto puolestaan yläkouluikäisiltä S2-oppilailta tehtävillä, jotka on laadittu vastaamaan tehtävätyypiltään aikuisten aineiston tehtäviä (aineistosta ja sen keruusta ks. esim. Alanen ym. 2010; Martin ym. 2010; Mustonen 2015). Aikuisten aineiston tehtävät on jaettu kolmeen tyyppiin: epämuodollinen viesti, muodollinen viesti ja mielipide. Nuorten aineistossa epämuodollisissa viesteissä tehtävänantona on ollut kirjoittaa sähköposti ystävälle tai opettajalle. Muodollisen viestin tehtävänantona on ollut kirjoittaa verkkokauppaan ja valittaa lahjaksi saadusta pelistä, joka toimii huonosti. Mielipidekirjoituksen tehtävänantona on ollut kaksi vaihtoehtoista ot-

sikkoa: Kännykät pois koulusta! ja Vanhemmat saavat päättää, miten lapset käyttävät Internetiä. Nuorten aineiston keruussa käytetyt tehtävät suunniteltiin vastaamaan aikuisten aineiston tehtävänantoja, jotka eivät ole julkisia. Nuorten aineistossa kaikki kirjoittivat samoista tehtävistä, aikuisten aineistossa eri taitotasoilla tehtävät ovat poikenneet jonkin verran toisistaan. Lisäksi A-tasolla mielipidetekstit on kerätty saman tyyppisellä tehtävällä kuin muodolliset viestit: molemmissa tehtävänä on ollut antaa palautetta. Nuorilta on kerätty erillisellä tehtävänannolla myös kertomus, jota vastaavaa tehtävätyyppiä ei ole mukana aikuisten aineistossa ja joka on siksi rajattu pois tutkimusaineistosta. Cefling-hankkeen aineistoon kuuluu myös ensikielisten suomenpuhujien vertailuaineisto, joka on kerätty yläkouluikäisiltä nuorilta samoilla tehtävillä kuin S2-nuorten aineisto. Cefling-aineiston tekstit on kirjoitettu käsin ilman apuvälineitä, ja ne on tuotettu rajoitetussa ajassa, nuorten tekstit oppitunneilla ja aikuisten tekstit testitilanteessa. (Jantunen & Pirkola 2015; Mustonen 2015.)

Tässä tutkimuksessa käytetty aineisto sijoittuu Eurooppalaisen viitekehyksen taitotasoille A1–C2. Tekstit on arvioitu taitotasoille Cefling-hankkeessa (Alanen ym. 2010) tavalla, jota voidaan pitää luotettavana sekä tilastollisin että laadullisin menetelmin tarkasteltuna (Huhta ym. 2014). Aikuisten aineisto kattaa Eurooppalaisen viitekehyksen kaikki kuusi taitotasoa. Nuorten aineisto sijoittuu taitotasoille A1–B2, mutta valtaosa nuorten teksteistä sijoittuu taitotasoille A1–B1. Cefling-hankkeen vertailuaineistoa ei ole arvioitu taitotasoille (äidinkielisen aineiston taitotasoarvioinnin haasteista ks. esim. Toropainen, Härmälä & Lahtinen 2012).

Tutkimuksen aineisto on poikittainen. Sen avulla on mahdollista tarkastella kielenpiirteiden esiintymistä eri taitotasoilla tai ryhmätason kehitystrendejä suhteessa kielitaidon taitotasoon. Yksilötason kehityskulkuja ei tällaisessa aineistossa kuitenkaan voida seurata eikä ryhmätason trendeistä voi päätellä yksilötason kielenoppimisen kulkua (esim. Larsen-Freeman 2006). Näin ollen tässä tutkimuksessa ei seurata suomen kielen taidon tai oppijansuomen kompleksisuuden kehitystä suhteessa aikaan tai oppimisprosessin kulkuun vaan valittuja syntaktisen kompleksisuuden alaan kuuluvia kielenpiirteitä tarkastellaan suhteessa Eurooppalaisen viitekehyksen taitotasoihin.

Tämän tutkimuksen aineistoon on otettu kaikki aikuisten aineiston tekstit sekä näitä kolmea tehtävätyyppiä vastaavilla tehtävillä kerätyt nuorten aineiston tekstit. Aineiston teksti- ja sanamäärät on esitetty taulukossa 1. Ne poikkeavat jonkin verran aiemmissa tutkimuksissa käytetyn Cefling-aineiston teksti- ja sanamääristä. Erot johtuvat kolmesta ratkaisusta, jotka tehtiin tämän tutkimuksen alussa. Ensinnäkin tässä tutkimuksessa aineistosta on Childes-tiedostojen sijaan käytetty Word-tiedostoihin puhtaaksikirjoitettuja tekstejä. Toiseksi annotoinnin yhteydessä (tarkempi kuvaus alaluvussa 3.2) aineistosta on myös rajattu tarkastelun ulkopuolelle joukko verbittömiä ilmauksia, kuten tervehdyksiä ja yhteystietoja. Tarkastelun ulkopuolelle on rajattu myös kaksi nuorten aineiston tekstiä, joista toinen puuttuu Cefling-aineiston Word-tiedostoista ja toisen kirjoittaja- ja tehtävätiedot ovat epäselvät, sekä kaksi aikuisten aineiston tekstiä, joista toinen

sisältää vain verbittömistä ilmauksista koostuvan luettelon ja toisen teksti on kopioitu sanasta sanaan tehtävänannosta. Ensikielisten yläkoululaisten vertailuaineistosta tässä tutkimuksessa ovat mukana vain ne tehtävätyypit, joita vastaavat S2-tekstit ovat mukana aineistossa.

TAULUKKO 1 Aineiston teksti- ja sanamäärät taitotasoittain ja kirjoittajaryhmittäin

	Epämuodolliset viestit		**Muodolliset viestit**		**Mielipidetekstit**		**Yhteensä**	
	tekstit	**sanat**	**tekstit**	**sanat**	**tekstit**	**sanat**	**tekstit**	**sanat**
Aikuiset								
A1	39	1582	22	752	50	2261	111	4595
A2	39	1533	27	1494	37	2272	103	5299
B1	41	2453	42	2290	43	5142	126	9885
B2	39	2206	34	1983	35	4166	108	8355
C1	26	1528	45	3337	46	5876	117	10741
C2	14	970	58	5152	30	3879	102	10001
Nuoret								
A1	25	677	33	800	32	775	90	2312
A2	79	2879	40	1489	39	1589	158	5957
B1	64	2971	40	1980	40	2232	144	7183
B2	12	677	7	461	-	-	19	1138
Vertailuryhmä	163	6186	162	6931	128	6709	453	19826

Aikuisten aineiston tekstejä on yhteensä 667 ja kirjoittajia 481. Tekstit jakautuvat kirjoittajien kesken niin, että 40 kirjoittajalta on kolme tekstiä, 106 kirjoittajalta kaksi tekstiä ja 335 kirjoittajalta yksi teksti kultakin. Tyypillisesti useamman kuin yhden tekstin kirjoittaneiden tekstit ovat eri tehtävätyyppiä, mutta aikuisten aineistossa on 12 kirjoittajaa, joilta jokaiselta on aineistossa mukana kaksi muodollista viestiä. Nuorten aineiston 411 tekstiä on kerätty 212 kirjoittajalta. Heistä yhteensä 45 on kirjoittanut kolme tekstiä, 109 kaksi tekstiä ja 58 yhden tekstin. Nuorten vertailuaineisto koostuu 453 tekstistä yhteensä 175 kirjoittajalta, joista 115 on kirjoittanut kolme tekstiä, 48 kaksi tekstiä ja 12 yhden tekstin. Sekä S2-nuorten aineistossa että vertailuaineistossa useamman kuin yhden tekstin kirjoittaneiden tekstit edustavat eri tehtävätyyppejä.

Aineistoon kuuluvat samalta kirjoittajalta kerätyt tekstit sijoittuvat useimmiten samalle taitotasolle. Näin ei kuitenkaan etenkään nuorten aineistossa aina ole. Esimerkiksi nuorten aineistossa kolme eri tehtävää kirjoittaneiden teksteistä 16 kirjoittajan kaikki kolme tekstiä sijoittuvat samalle taitotasolle, kun taas 27 kirjoittajaa on sellaisia, joiden teksteistä yksi sijoittuu yhtä ylemmälle tai alemmalle taitotasolle kuin kaksi muuta tekstiä, ja kahden kirjoittajan teksteistä kaikki kolme on arvioitu keskenään eri taitotasoille. Kaksi tekstiä kirjoittaneista noin puolet (55) on sellaisia, joiden molemmat tekstit on arvioitu samalle taitotasolle, ja noin puolet (54) sellaisia, joiden kaksi tekstiä sijoittuvat eri taitotasoille. Aikuisten aineistossa saman kirjoittajan tekstit asettuvat useammin keskenään samalle taitotasolle: useamman kuin yhden tekstin kirjoittaneista 146 kirjoittajasta vain

16 on sellaisia, joiden kaikki aineistossa mukana olevat tekstit eivät sijoitu keskenään samalle taitotasolle.

Aineiston tekstit ovat suhteellisen lyhyitä, keskimäärin alle sata sanaa. Aikuisten aineiston mielipidetekstit ovat keskimääräiseltä pituudeltaan sata sanaa tai enemmän taitotasolta B1 alkaen. Nuorten aineiston S2-tekstien pituus kuitenkin vastaa melko hyvin yläkouluikäisiltä kerätyn vertailuaineiston tekstien pituutta. Aikuisten kirjoittamat tekstit ovat keskimäärin pitempiä kuin nuorten. Sadan sanan tekstinäytteitä ovat käyttäneet oppijansuomen tutkimuksissa Spoelman ja Verspoor (2010) sekä Tilma (2014), jotka käyttivät myös lyhyempiä tekstinäytteitä etenkin seurantajakson alussa, kun oppijoiden kirjoittamat tekstit olivat pituudeltaan alle sata sanaa.

3.2 Aineiston käsittely

Tässä tutkimuksessa oppijansuomen syntaktista kompleksisuutta tarkasteltiin sanoihin, lauseisiin sekä niistä muodostuviin laajempiin kielenyksiköihin eli virkkeisiin ja T-yksiköihin perustuvin määrällisin mittarein. Koska Cefling-aineistoon ei ole valmiiksi koodattu kaikkia näitä kielenyksiköitä, aineisto segmentointiin ja koodattiin uudelleen tutkimuksen alussa.

Cefling-aineisto on saatavissa kahdessa tiedostomuodossa, CHILDES-ympäristössä[1] käytettävinä CHAT-koodattuina[2] tiedostoina (.cha) ja Microsoft Word -tiedostoina (.doc). Cefling-hankkeen tutkimuksissa on tyypillisesti käytetty CHILDES-ympäristöä ja aineiston CHAT-tiedostoja (esim. Seilonen 2013; Mustonen 2015). Tässä tutkimuksessa aineistoa päädyttiin käsittelemään XML-tiedostoina, jotka luotiin Cefling-hankkeen Word-tiedostoista. Näin tutkimusaineistoon saatiin mukaan myös sellaiset Word-tiedostoihin tallennetut tiedot alkuperäisten käsinkirjoitettujen tekstien muotoiluista, esimerkiksi kappalejako, joita CHAT-tiedostoissa ei ollut ja joita voitiin käyttää apuna tekstien segmentoinnissa. Tutkimusaineiston muodoksi valittiin XML, sillä XML-merkintäkieli mahdollistaa sekä aineiston joustavan koodauksen että sen automatisoidun lukemisen. XML-koodauksessa käytetyt tunnisteet luotiin tämän tutkimuksen tarpeisiin eikä niitä sovitettu yhteen muiden oppijansuomen korpusten tai TEI-standardin[3] kanssa. XML-koodattujen tiedostojen rakennetta ja niissä käytettyjä tunnisteita täydennettiin osatutkimusten aikana tarpeen mukaan.

Cefling-korpuksen epämuodolliset viestit, muodolliset viestit ja mielipidetekstit muunnettiin tätä tutkimusta varten XML-tiedostoiksi. Muuntamisessa käytettiin hyväksi tekstinkäsittelyohjelman mahdollisuutta tallentaa tekstitiedostot XML-muotoon. Näin saadut tiedostot muokattiin yhdenmukaiseen XML-asuun tätä tarkoitusta varten tässä tutkimuksessa kirjoitettujen Python-skriptien

[1] CHILDES = The Child Language Data Exchange System, http://childes.talkbank.org.
[2] CHAT = Codes for Human Analysis of Transcripts. MacWhinney, Brian. 1991. The Childes Project: Tools for Analyzing Talk.
[3] TEI = Text Encoding Initiative. Text Encoding Initiative Consortium, http://www.tei-c.org.

avulla, minkä jälkeen XML-tiedostot vielä tarkistettiin ja korjattiin käsin. Samalla tarkastelun ulkopuolelle rajattiin tekstien aluista ja lopuista lause- ja virkejaon kannalta ongelmalliset ilmaukset, jotka olivat joko finiittiverbittömiä kokonaisuuksia, kuten nimimerkkejä, osoitetietoja ja tervehdyksiä, tai tehtävänannosta sanasta sanaan kopioituja otsikoita (ks. myös Foster ym. 2000: 370–371).

Word-tiedostoista kerättyjä tekstejä verrattiin CHAT-tiedostojen sisältämiin teksteihin. Näin pyrittiin varmistamaan, että XML-tiedostojen tekstien sisältö oli sama kuin aiemmissa tutkimuksissa käytettyjen CHAT-tiedostojen. Koska ensimmäisessä osatutkimuksessa verrattiin Cefling-hankkeessa tehtyä virkejakoa tässä tutkimuksessa tehtyyn, pilottiaineiston mielipidetekstien CHAT- ja Word-tiedostojen erot sanamäärissä käytiin yksitellen läpi käsin. Nuorten aineiston tekstien tarkistuksissa käytettiin apuna alkuperäisiä käsin kirjoitettuja tekstejä, aikuisten aineiston tarkistuksissa YKI-tietokantaa, josta tekstit oli poimittu, koska alkuperäisiä käsinkirjoitettuja tekstejä ei aikuisten teksteistä ollut saatavissa. Jos alkuperäistä versiota ei ollut käytettävissä, Word-tiedoston mukainen asu tekstistä jätettiin tutkimusaineistoon. Muodollisten ja epämuodollisten viestien teksteistä käytettiin aina Word-tiedostoihin tallennettuja tekstiversioita.

Tarkistuksen jälkeen aineiston tekstit segmentoitiin syntaktisen kompleksisuuden määrällisissä mittareissa tarvittaviin kielenyksiköihin. Mielipidetekstit segmentoitiin ensin käsin ja sitten Turun yliopistossa kehitetyn Turku University Finnish Depency Parser -jäsentimen[4] (Haverinen ym. 2013) avulla, ja näitä kahta segmentointitapaa verrattiin toisiinsa väitöskirjan ensimmäisessä osatutkimuksessa. Seuraavassa vaiheessa muodolliset ja epämuodolliset viestit segmentoitiin ensin saman jäsentimen avulla ja tarkistettiin sen jälkeen käsin vastaamaan mielipidetekstien segmentoinnissa käytettyjä periaatteita.

Oppijankielen segmentointi on aina osin subjektiivista, ja esimerkiksi kielenoppijan, aikuisen kielenkäyttäjän ja kielentutkijan käsitykset kielen yksiköistä saattavat poiketa toisistaan (Peters 1983). Annotointia vaikeuttaa myös se, että oppijankieli voi poiketa kohdekielen normeista (esim. Granger 2002), ja norminmukaisesta kielestä poikkeavien ilmausten annotoinnissa on aina mukana myös tulkintaa (Brunni ym. 2015; Ragheb & Dickinson 2011; Rehbein ym. 2012). Tämä ei välttämättä rajoitu oppijankieleen, vaan myös ensikielisessä aineistossa voi olla epätyypillisiä kielenyksiköitä ja lausejaon kannalta ongelmallisia ilmauksia, esimerkiksi finiittiverbittömiä virkkeitä. Oppijansuomen lause- ja virkerajojen tunnistamisen ongelmallisuutta on havainnollistanut Martin (2013a) kokeella, jossa 35 äidinkielistä suomen kielen yliopisto-opiskelijaa jakoi kolme suomenoppijoiden kirjoittamaa tekstiä virkkeisiin ja lauseisiin päätyen yhtenäisestä taustastaan huolimatta hyvin erilaisiin tuloksiin. Kiinnostavaa tuloksissa oli, että silloinkin, kun lauseiden tai virkkeiden määrä jossakin tekstissä oli kahden tai useamman segmentoijan mielestä sama, tekstistä rajatut lauseet tai virkkeet saattoivat poiketa toisistaan. Erot eivät vaikuta määrällisiin mittareihin silloin, kun

[4] Tässä tutkimuksessa käytetty jäsentimen versio on edelleen vapaasti saatavissa osoitteessa http://turkunlp.github.io/Finnish-dep-parser/. Nykyisin käytössä oleva uudempi versio ei ollut aineiston annotoinnin aikana vielä saatavissa.

tekstin sana-, lause- ja virkemäärä on sama eri segmentoinneissa. Ne osoittavat kuitenkin sen, että lauseiden, virkkeiden ja T-yksiköiden rajat eivät aina ole selkeitä ja että tekstin segmentoiminen vaatii usein myös tulkintaa.

Tässä tutkimuksessa päädyttiin segmentointiin, jossa sanat ja virkkeet rajattiin ortografisin perustein. Segmentointi aloitettiin virkkeistä, joita esimerkiksi Ellis ja Barkhuizen (2005: 147) pitävät kirjoitetussa kielessä luontevina tarkastelukohteina. Virkkeet rajattiin isojen alkukirjainten ja loppuvälimerkkien (yhden tai useamman pisteen, huutomerkin, kysymysmerkin tai näiden yhdistelmien) lisäksi myös seuraavilla ortografisilla perusteilla: luettelomerkki luettelon rivin alussa ja uuden tekstikappaleen alku tulkittiin uuden virkkeen aluksi ja edellisen virkkeen lopuksi, minkä lisäksi koko tekstin loppu tulkittiin viimeisen virkkeen lopuksi silloinkin, kun jakso ei alkanut isolla kirjaimella tai päättynyt virkkeen päättävään välimerkkiin. Jokaisen virkkeen tulkittiin sisältävän vähintään yhden päälauseen, mikä yhtäältä mahdollisti kaikkien sanojen sisällyttämisen määrällisin mittarein analysoitavaan aineistoon mutta toisaalta tuotti epätyypillisiä, finiittiverbittömiä lauseita. Lisäksi rajaus tuotti lauseita, joita voidaan pitää itsenäisinä sivulauseina (ks. Foster ym. 2000; Kalliokoski 2006), mutta samalla se tuotti vain sellaisia kirjoittajan virkkeeksi rajaamia kokonaisuuksia, joissa oli vähintään yksi T-yksikkö.

Näin saadut T-yksiköt sisälsivät siis vähintään yhden päälauseen, jolle alisteiset lauseet koodattiin kyseiselle päälauseelle alisteisiksi sivulauseiksi ja tarvittaessa näille alisteisiksi sivulauseiksi. Alun perin lause määriteltiin finiittiverbin ja sen laajennusten muodostamaksi kokonaisuudeksi, mutta finiittiverbin vaatimuksesta luovuttiin, kun pilottivaiheessa selvisi, että tällaisen määritelmän käyttö tuottaa virkkeitä, joissa ei ole yhtään lausetta. Lisäksi rajaus osoittautui ongelmalliseksi tapauksissa, joissa päälauseena toimi jokin verbitön ilmaus. Tällaisia finiittiverbittömiä lausemaisia rakenteita oli pilottiaineistossa aikuisten teksteissä kahdella alimmalla taitotasolla noin viidennes kaikista lauseiksi koodatuista jaksoista. Nämä kielenyksiköt olivat ongelmallisia myös T-yksiköiden segmentoinnin kannalta. T-yksiköiden kannalta ongelmallisia olivat lisäksi virkkeet, joissa lauseiden väliset alistus- tai rinnastussuhteet olivat tulkinnanvaraisia. Lauseiden väliset suhteet eivät kuitenkaan aina ole selviä tai yksitulkintaisia (Vilkuna 1996, 71–72). Tällaisia virkkeitä löytyi pilottiaineiston kaikilta taitotasoilta, eniten kuitenkin alimmilta taitotasoilta, ja segmentoinnin näkökulmasta ongelmallisia virkkeitä löytyi lähes 40 prosentista pilottiaineiston teksteistä.

Koska aineiston segmentoinnin aikana tehdyt tulkinnat aineiston sisältämistä, määrällisissä mittareissa välttämättömistä kielenyksiköistä vaikuttavat myös määrällisillä mittareilla saataviin tuloksiin, käytetty koodaus dokumentoitiin mahdollisimman yksityiskohtaisesti ja raportoitiin osatutkimuksessa 1. Samalla osatutkimuksen aineiston segmentointi toimi pilottina, jonka perusteella valittua mallia käytettiin, kun tutkimuksen muu aineisto segmentoitiin seuraavia osatutkimuksia varten.

Segmentoinnin jälkeen jokaisesta tekstistä laskettiin sanojen, virkkeiden sekä pää- ja sivulauseiden määrät, ja tiedot koodattiin XML-tiedostoihin. T-yksiköitä ei koodattu erikseen, vaan ne poimittiin ja niiden tunnusluvut laskettiin

pää- ja sivulauseiden koodauksen perusteella. Tästä segmentoidusta aineistosta laskettiin kaikki ensimmäisessä ja toisessa osatutkimuksessa käytetyt tunnusluvut. Kolmatta osatutkimusta varten XML-tiedostoihin koodattiin kaikki konjunktiot. Neljännessä osatutkimuksessa käytettiin kolmannen osatutkimuksen aineistosta aikuisten mielipidetekstejä, joiden virkkeisiin koodattiin käsin kaikki finiittiverbien esiintymät ja erilaiset finiittiverbistä ja infiniittisistä verbinmuodoista koostuvat moniverbiset konstruktiot.

3.3 Tutkimusmenetelmät

Tässä tutkimuksessa aineiston tekstejä on tarkasteltu sekä yksittäisinä teksteinä että taitotasoittain tekstimassana. Ensimmäisen lähestymistavan etuna on, että tekstien väliset erot ja taitotasojen sisäinen vaihtelu saadaan näkyviin selvemmin kuin vain ryhmätason keskiarvojen vertailussa. Yksittäisten konjunktioiden esiintymistä eri taitotasoilla on kuitenkin tarkasteltu tekstimassassa, jolloin saadaan yleiskuva kullekin taitotasolle sijoittuvien tekstien sisältämistä eri konjunktiolekseemeistä. Myös moniverbisiä konstruktioita on tarkasteltu taitotasoittain tekstimassassa.

Konjunktioiden ja moniverbisten konstruktioiden tarkastelussa aineiston tilastollisessa kuvaamisessa on käytetty frekvenssejä. Koska yksittäisten tekstien pituus ja kullekin taitotasolle sijoittuvien tekstien yhteissanamäärät vaihtelevat, esiintymien määrää on tarkasteltu sekä absoluuttisten että sanamäärään suhteutettujen frekvenssien avulla. Yksittäisten tekstien kohdalla esiintymät on suhteutettu sataan sanaan, taitotasojen ja kirjoittajaryhmien rinnakkaisessa tarkastelussa tuhanteen sanaan. Tuhanteen sanaan suhteutetut esiintymistaajuudet ovat verrattavissa aiemmissa Cefling-aineistolla tehdyissä tutkimuksissa saatuihin tuloksiin kielenpiirteiden esiintymistaajuudesta, tosin vertailussa on otettava huomioon, että kokonaissanamäärät eivät ole täysin yhteneväisiä (aineiston rajauksesta ks. alaluku 3.1). Absoluuttisia frekvenssejä käytettiin variaation tarkastelun yhteydessä kontrolloimaan tehtyjä tulkintoja: kun verrataan taitotasoja, joilla esiintymiä on hyvin erilainen määrä, yksi variaatiota selittävä tekijä voi olla esiintymien absoluuttinen esiintymistaajuus, sillä eri variantteja ei luonnollisesti voi olla enempää kuin esiintymiä on yhteensä.

Syntaktisen kompleksisuuden määrällisten mittareiden luotettavuuteen vaikuttaa keskeisesti aineiston segmentoinnin ja annotoinnin luotettavuus (esim. Foster ym. 2010). Koska tutkimuksessa aineistoa oli segmentoimassa vain yksi tutkija, koodauksen luotettavuutta arvioitiin kahdella tavalla. Ensinnäkin pilottiaineiston poikkeamat analysoitiin ja ongelmallisten kielenyksiköiden määrät laskettiin. Analyysissa keskityttiin sellaisiin poikkeamiin, jotka mahdollisesti vaikuttivat mittareissa käytettyjen kielenyksiköiden määriin ja pituuksiin. Tällaisia olivat virkkeiden ja lauseiden rajat, lauseiden jako pää- ja sivulauseisiin sekä predikaattiverbittömät virkkeet ja lauseet. Lisäksi pilottiaineistoon käsin tehtyä virkejakoa verrattiin Cefling-hankkeen CHAT-koodauksen yhteydessä tehtyyn virkejakoon ja tässä tutkimuksessa apuna käytetyn jäsentimen (ks. alaluku 3.2)

tuottamaan virkejakoon. Näitä kolmea verrattiin tarkkuuden (*precision*), saannin (*recall*) ja F-mitan (*F-score*) avulla. Annotoinnin luotettavuutta voidaan mitata selvittämällä, miten moni koodatuista esiintymistä on koodattu oikein (tarkkuus) ja miten moni kaikista aineiston esiintymistä on mukana koodatuissa esiintymissä (saanti), sekä laskemalla sitten F-mitta, joka on näiden kahden harmoninen keskiarvo (van Rooy 2015: 85). Tässä tutkimuksessa ei kuitenkaan ollut käytettävissä luotettavalla tavalla todennettua oikeaa segmentointia. Siksi tarkkuus, saanti ja F-mitta laskettiin Lun (2010) ja Brantsin (2000) esimerkin mukaisesti vertailtavien annotointien tuottamien keskenään identtisten virkkeiden avulla. Tarkkuus laskettiin jakamalla kahden annotointitavan tuottamien samojen virkkeiden määrä ensimmäisen annotointitavan virkkeiden kokonaismäärällä ja saanti jakamalla samojen virkkeiden määrä toisen annotointitavan virkkeiden kokonaismäärällä. Tällöin F-mitta ei kerro annotoinnin virheettömyydestä vaan eri annotointitapojen yhdenmukaisuudesta (Brants, 2000; Lu, 2010).

Kompleksisuuden määrällisten mittareiden tarkastelussa taitotasojen välisten erojen tilastolliseen vertailuun käytettiin varianssianalyysia ja sen epäparametrista vastinetta Kruskal-Wallisin testiä (ks. esim. Larson-Hall 2010: 378–380). Jos näiden testien avulla löytyi tilastollisesti merkitseviä eroja, suoritettiin pareittaiset testit. Tässä käytettiin t-testiä ja sen epäparametrista vastinetta Wilcoxonin järjestyslukujen summan testiä (tunnetaan myös nimellä Mann-Whitneyn U-testi) (ks. esim. Larson-Hall 2010: 376–378; 404). Tilastollisessa vertailussa käytettiin sekä parametrisia että epäparametrisia testejä, koska parametristen testien taustaoletukset normaalijakaumasta ja ryhmien sisäisestä varianssista täyttyivät osittain (ks. esim. Gries 2013: 215; Larson-Hall 2010: 374). Konjunktioiden esiintymistaajuuksien tilastollisissa vertailuissa käytettiin vain epäparametrisia testejä, sillä aineisto ei täyttänyt parametristen testien taustaoletuksia. Koska pareittaisia vertailuja tehtiin useita, tilastollista merkitsevyyttä arvioitaessa käytettiin Bonferroni-korjausta (ks. esim. Winter 2020: 176–177). Varianssianalyysin tulosten tilastollisen merkitsevyyden arvioimiseksi laskettiin myös efektikoko (ks. esim. Larson-Hall 2010: 117–120; Tähtinen, Laakkonen & Broberg 2020: 44–49). Tilastollista tarkastelua täydennettiin visualisoimalla aineisto laatikkojanojen avulla (ks. esim. Larson-Hall 2010: 245). Tilastollisissa testeissä käytettiin R-ohjelmiston versiota 3.4.4 ja RStudion versiota 1.1.456.

4 OSATUTKIMUSTEN ESITTELY

4.1 Oppijansuomen sanat, lauseet, virkkeet ja T-yksiköt mittayksikköinä

Ensimmäisessä osatutkimuksessa tarkasteltiin tyypillisissä syntaktisen kompleksisuuden määrällisissä mittareissa käytettyjä kielenyksiköitä eli sanoja, lauseita, virkkeitä ja T-yksiköitä sekä niiden koodaamista S2-teksteihin. Tavoitteena oli kartoittaa, miten objektiivisia ja luotettavia mittayksiköitä kirjoitetun oppijankielen sanat, virkkeet, lauseet ja T-yksiköt ovat. Osatutkimuksen tutkimuskysymys oli: Miten oppijankielen poikkeamat kohdekielen normeista vaikuttavat oppijankielen segmentointiin sanoiksi, virkkeiksi, lauseiksi ja T-yksiköiksi?

Osatutkimuksen aineistona oli mielipidetekstien osakorpus eli 241 aikuisten ja 111 yläkouluikäisten kirjoittamaa S2-tekstiä sekä 128 tekstistä koostuva yläkouluikäisten S1-vertailuaineisto. Aikuisten aineisto kattoi taitotasot A1–C2, S2-nuorten aineisto taitotasot A1–B1.

Kaikilla taitotasoilla helpoimmin rajattavissa olevat yksiköt olivat sana ja virke, jotka molemmat ovat erotettavissa tekstistä ortografisin perustein. Aineiston tekstien segmentointi tosin osoitti, että myös virkerajan merkitsemisessä oli horjuntaa etenkin taitotasolla A1, jolla vain noin puolet (aikuisten aineistossa 48 % ja nuorten aineistossa 55 %) virkkeeksi tulkituista jaksoista sekä alkoi isolla kirjaimella että päättyi pisteeseen, huutomerkkiin tai kysymysmerkkiin. Ongelmallisimpia yksiköitä olivat lause ja T-yksikkö, jotka sanojen ohella ovat syntaktisen kompleksisuuden määrällisissä mittareissa yleisimmin käytetyt kielenyksiköt. Esimerkiksi etenkin alimmilla taitotasoilla varsin yleiset predikaattiverbittömät virkkeet olivat ongelmallisia lausejaon kannalta. Tällaiset virkkeet voidaan koodata lauseeksi ja yhdestä lauseesta koostuvaksi T-yksiköksi, osaksi jotakin toista lausetta tai T-yksikköä tai ne voidaan jättää kokonaan analyysin ulkopuolelle. Valinnat vaikuttavat niihin määrällisiin mittareihin, joissa käytetään lauseita tai T-yksiköitä mittayksiköinä. T-yksiköiden koodauksessa ongelmallisia voivat olla

myös lauseiden keskinäiset suhteet, jotka eivät aina ole yksiselitteisesti tulkittavissa tekstin pintarakenteesta (ks. myös Martin 2013a).

Osatutkimuksen tulokset tukevat näkemyksiä siitä, ettei oppijankieltä ole mahdollista jakaa jäännöksettömästi sellaisiin kielenyksiköihin kuin lauseet tai T-yksiköt (ks. myös esim. Foster ym. 2000; Rimmer 2006: 508). Lausejaon ongelmallisuus on nostettu esiin myös aiemmissa Cefling-aineistoa hyödyntäneissä tutkimuksissa (Kajander 2013: 49, 86-87; Reiman 2014: 190; Seilonen 2013: 31). Jos oppijankieltä kuitenkin tarkastellaan näihin kielenyksiköihin perustuvilla mittareilla, tutkimusten luotettavuuden ja vertailtavuuden kannalta on tärkeää tehdä näkyväksi niin käytettyjen kielenyksiköiden määritelmät kuin se, miten määritelmiä on sovellettu aineiston segmentoinnissa. Pyrkimys jäännöksettömään segmentointiin voi myös johtaa siihen, että näissä kielenyksiköissä tapahtuva kehitys jää osittain huomaamatta. Myös tästä syystä oppijankielen laadullinen tarkastelu voi nostaa näkyviin uusia puolia oppijankielen syntaksin ja kompleksisuuden kehittymisestä.

4.2 Oppijansuomen syntaktinen kompleksisuus määrällisesti mitattuna

Toisessa osatutkimuksessa oppijansuomen syntaktista kompleksisuutta tarkasteltiin seitsemän määrällisen mittarin avulla. Tavoitteena oli selvittää, miten käytetyimmät syntaktisen kompleksisuuden määrälliset mittarit soveltuvat oppijansuomen tarkasteluun. Tutkimuskysymyksiä oli kaksi: 1) Miten kirjoitetun oppijansuomen syntaktinen kompleksisuus muuttuu Eurooppalaisen viitekehyksen taitotasolta toiselle? ja 2) Onko määrällisissä mittareissa tilastollisesti merkitseviä eroja taitotasojen välillä? Osatutkimuksen aineistona olivat tutkimusaineiston kaikki tekstit: epämuodolliset viestit, muodolliset viestit ja mielipidetekstit. Eri tehtävätyyppejä tarkasteltiin rinnakkain, mutta aineiston luonteen vuoksi (ks. myös luku 3.1) tehtävätyyppejä ei verrattu toisiinsa tilastollisesti. Aikuisten ja nuorten tekstejä tarkasteltiin erikseen. Taitotasojen välisiä eroja verrattiin tilastollisesti sekä parametrisilla että epäparametrisilla testeillä. Tilastolliset vertailut tehtiin tehtävätyypeittäin erikseen aikuisten ja nuorten aineistossa.

Mittareiden valinnassa pyrittiin monipuolisuuteen ja siihen, että niiden avulla voidaan tavoittaa syntaktisen kompleksisuuden eri osa-alueet (ks. Bulté & Housen 2012; Lu 2010, 2011; Norris & Ortega 2009; Ortega 2003; Wolfe-Quintero ym. 1998). Yleisen kompleksisuuden mittareiksi valittiin T-yksikön sanamäärä (MLTU), virkkeen sanamäärä (MLS) ja virkkeen lausemäärä (C/S). Alisteisina lauserakenteina näkyvän kompleksisuuden mittareina toimivat T-yksikön lausemäärä (C/TU) ja sivulauseiden osuus kaikista lauseista (DC/C). Lausetason rinnastamisen mittariksi valittiin virkkeen sisältämien T-yksiköiden määrä (TU/S) ja lauseen sisäisen kompleksisuuden mittariksi lauseen sanamäärä (MLC).

Tulosten mukaan lauseiden, virkkeiden ja T-yksiköiden sanamäärät kasvavat taitotason noustessa. Samoin lauseiden määrä virkkeessä ja T-yksikössä kasvaa aineiston teksteissä taitotason noustessa. Virkkeen sisältämien lauseiden määrän kasvu näyttäisi selittyvän sivulauseiden osuuden lisääntymisellä, sillä myös sivulauseiden osuus kaikista lauseista kasvoi ja T-yksiköiden (eli rinnasteisten päälauseiden) määrä virkkeessä väheni. Taitotasojen sisällä oli kuitenkin useimmissa mittareissa paljon variaatiota, ja joissakin mittareissa taitotasojen vertailu keskiarvojen avulla ja mediaanien avulla tuotti erilaisen kuvan ryhmätason kehitystrendeistä. Vaikka eri tehtävätyyppejä ei verrattu tilastollisesti toisiinsa, niiden tarkastelu rinnakkain viittasi siihen, että myös tehtävätyyppi voi vaikuttaa määrällisillä mittareilla tavoitettavaan syntaktiseen kompleksisuuteen. Lisäksi aikuisten ja nuorten välillä näyttäisi tulosten perusteella olevan eroja ryhmätason kehityksessä.

Ryhmien sisäinen varianssi vaikutti myös tilastolliseen analyysiin ja sen tuloksiin. Tilastollisesti merkitseviä eroja oli aikuisten aineistossa lähinnä A-tason ja sitä ylempien taitotasojen välillä, nuorten aineistossa taitotason A1 ja muiden taitotasojen välillä. Aikuisten aineistossa kaikissa käytetyissä mittareissa oli tilastollisesti merkitseviä eroja ainakin joidenkin taitotasojen välillä. Sen sijaan nuorten aineistossa tilastollisesti merkitseviä eroja löytyi vain neljällä mittarilla eivätkä taitotasojen väliset erot virkkeiden sanamäärässä, virkkeiden sisältämien T-yksiköiden määrässä tai virkkeiden lausemäärässä olleet tilastollisesti merkitseviä. Näistä kahdessa viimeisessä, virkkeen sisältämien T-yksiköiden ja virkkeen sisältämien lauseiden määrässä, ei myöskään ollut tilastollisesti merkitsevää eroa aikuisten aineiston taitotasojen A1 ja C2 välillä.

Osatutkimuksen tulosten perusteella määrällisten mittareiden kyky tavoittaa oppijansuomen kompleksisuudessa tapahtuvia muutoksia on rajallinen. Tulokset tukevat näkemyksiä, joiden mukaan kompleksisuuden ja taitotason suhteen tarkasteluun tarvitaan myös muita mittareita. Siksi seuraavassa osatutkimuksessa oppijansuomen rinnasteisuutta ja alisteisuutta osana syntaktista kompleksisuutta tarkasteltiin konjunktioiden avulla.

4.3 Konjunktiot täydentämässä kuvaa oppijansuomen syntaktisesta kompleksisuudesta

Kolmannessa osatutkimuksessa aineiston syntaktisia kytköksiä tarkasteltiin teksteissä käytettyjen konjunktioiden avulla. Osatutkimuksessa tavoitteena oli selvittää, miten paljon ja millaisia konjunktiolla merkittyjä syntaktisia kytköksiä aineiston eri taitotasoilla käytetään. Tutkimuskysymykset olivat: 1) Miten paljon konjunktioilla merkittyjä syntaktisia kytköksiä käytetään eri taitotasoilla? ja 2) Miten konjunktioiden käyttö muuttuu taitotasolta toiselle?

Osatutkimuksessa käytettiin samaa aineistoa kuin toisessa osatutkimuksessa, mutta eri tehtävätyypit (epämuodollinen viesti, muodollinen viesti, mielipideteksti) yhdistettiin. Aikuisten ja nuorten aineistoa tarkasteltiin myös tässä

osatutkimuksessa rinnakkain. Aineistosta poimittiin kaikki konjunktioiden esiintymät, ja ne koodattiin syntaktisin perustein alistus- ja rinnastuskonjunktioiksi. Lopuksi laskettiin jokaisen yksittäisen konjunktion esiintymien lukumäärä.

Konjunktioiden käytön yleisyyttä eri taitotasoilla kartoitettiin vertaamalla tilastollisesti kaikkien konjunktioiden sekä erikseen alistuskonjunktioiden ja rinnastuskonjunktioiden normalisoituja esiintymistaajuuksia eri taitotasoilla. Alisteisissa syntaktisissa kytköksissä tapahtuvia muutoksia tarkasteltiin selvittämällä, miten suuri osa aineiston sivulauseista sisältää alistuskonjunktion. Rinnasteisten syntaktisten kytkösten muuttumista puolestaan tarkasteltiin selvittämällä, mitä kielenyksiköitä rinnastuskonjunktioilla yhdistetään kielitaidon eri tasoilla. Lisäksi selvitettiin, mitä eri konjunktioita aineiston teksteissä eri taitotasoilla esiintyy.

Konjunktioiden sanamääriin suhteutettujen esiintymistaajuuksien tarkastelu osoitti, että eniten konjunktioita esiintyi taitotasolla B1. Erot kaikkien konjunktioiden määrässä olivat tilastollisesti merkitseviä aikuisten aineistossa A-tason ja B-tason sekä B-tason ja C-tason välillä, nuorten aineistossa taitotasojen A1 ja B1 välillä. Alistuskonjunktioiden esiintymistaajuuden erot olivat tilastollisesti merkitseviä vain aikuisten aineiston taitotasojen B2 ja C2 välillä. Rinnastuskonjunktioiden normalisoitujen esiintymistaajuuksien välillä oli aikuisten aineistossa tilastollisesti merkitsevä ero A-tason ja B-tason välillä, nuorten aineistossa taitotasojen A1 ja B1 välillä.

Kaikilla taitotasoilla vähintään 60 % sivulauseista sisälsi alistuskonjunktion. Muiden kuin alistuskonjunktion sisältävien sivulauseiden osuus kaikista sivulauseista kolminkertaistui aikuisten aineistossa taitotasolta A1 taitotasolle C2, jolla se oli noin kolmasosa kaikista sivulauseista. Nuorten aineistossa muiden kuin alistuskonjunktion sisältävien sivulauseiden osuus lähes puolitoistakertaistui taitotasolta A1 taitotasolle B2, jolla se oli noin viidennes kaikista sivulauseista. Rinnastuskonjunktioita puolestaan käytettiin tyypillisesti virkkeiden sisällä yhdistämään lauseita. Vain aineiston taitotasoilla C1 ja C2 hieman yli puolet rinnastuskonjunktioiden esiintymistä sijaitsi muualla kuin lause- tai virkerajalla.

Aineiston seitsemän yleisintä konjunktiota olivat käytössä jo taitotasolla A1. Aikuisten aineistossa teksteissä esiintyvien konjunktioiden valikoima kaksinkertaistui taitotasojen A2 ja B1 välillä. Nuorten aineistossa ei näkynyt vastaavaa hyppäystä taitotasojen välillä vaan käytettyjen konjunktioiden valikoima laajeni tasaisemmin taitotasolta toiselle. Sekä aikuisten että nuorten aineistossa konjunktioiden valikoima eri konjunktiolekseemeinä laskettuna oli suurimmillaan taitotasolla B1. Konjunktioiden valikoiman laajeneminen näkyi aineistossa epätasaisesti: osasta konjunktioita oli vain yksittäisiä esiintymiä koko aineistossa eikä kerran aineistoon ilmaantunut konjunktio välttämättä esiintynyt kaikilla ylemmillä taitotasoilla. Näin ollen yksittäisen konjunktion esiintymistä tai esiintymättömyyttä ei voi tulkita jonkin tietyn taitotason indikaattoriksi.

4.4 Moniverbiset konstruktiot ikkunana kieliopilliseen kompleksisuuteen

Neljännessä osatutkimuksessa oppijansuomen kompleksisuutta lähestyttiin tarkastelemalla eri taitotasoilla esiintyviä moniverbisiä konstruktioita. Osatutkimuksessa selvitettiin, millaisia moniverbisiä konstruktioita oppijansuomen taitotasoilla A1−C2 käytettiin ja millainen kuva niiden avulla saadaan oppijansuomen kompleksisuudesta ja sen muuttumisesta kielitaidon tason noustessa.

Osatutkimuksen aineistona käytettiin aikuisten S2-kirjoittajien mielipidetekstejä, joista poimittiin kaikki finiittiverbien esiintymät. Lähempään tarkasteluun valittiin konstruktiot, jotka sisälsivät finiittiverbin lisäksi vähintään yhden infiniittisen verbinmuodon, ja esiintymät luokiteltiin kolmeen rakennetyyppiin finiittimuotoista verbiä seuraavan infiniittisen muodon perusteella: finiittiverbin ja A-infinitiivin (esimerkiksi *voi tehdä*), finiittiverbin ja MA-infinitiivin (esimerkiksi *joutuu tekemään*) tai finiittiverbin ja partisiippimuodon (esimerkiksi *on tehtävä*) sisältäviin konstruktioihin. Moniverbisten konstruktioiden yleisyyttä eri taitotasoilla selvitettiin laskemalla niiden osuus kaikista finiittiverbiesiintymistä. Lisäksi tarkasteltiin konstruktioiden morfologista ja leksikaalista variaatiota sekä erilaisten infiniittisten rakenteiden yhdistämistä eri taitotasoilla.

Moniverbiset konstruktiot ilmaantuivat aineiston teksteihin heti taitotasolla A1, jolla ne olivat kuitenkin melko harvinaisia ja koostuivat tyypillisesti finiittiverbistä ja yhdestä A-infinitiivistä. Moniverbisten konstruktioiden osuus kaikista finiittiverbiesiintymistä lähes kaksinkertaistui taitotasojen A2 ja B1 välillä, minkä jälkeen kasvu tasaantui. Konstruktioiden variaatio kuitenkin jatkoi kasvuaan myös taitotason B1 jälkeen, ja ylimmillä taitotasoilla käytössä oli eniten erilaisia moniverbisiä konstruktioita. Variaatio oli sekä leksikaalista että morfologista. Lisäksi ylemmillä taitotasoilla oli eniten sellaisia moniverbisiä konstruktioita, joissa infiniittisiä muotoja yhdistetään rinnastamalla tai käyttämällä niitä toisten infiniittisten muotojen laajennuksina. Ylemmillä taitotasoilla käyttöön tulivat myös merkitykseltään eriytyneet konstruktiot ja samalla ilmausten idiomaattisuus lisääntyi.

Osa näistä muutoksista voi näkyä syntaktisen kompleksisuuden määrällisissä mittareissa. Kun konstruktiot pitenevät, myös lauseiden ja niitä laajempien kielenyksiköiden sanamäärät voivat kasvaa. Toisaalta moniverbiset konstruktiot voivat myös tiivistää ilmaisua, jolloin niissä tapahtuvat muutokset eivät välttämättä näy sana- tai lausemääriin perustuvissa syntaktisen kompleksisuuden mittareissa kielenyksiköiden pitenemisenä.

Moniverbisissä konstruktioissa leksikaalinen, morfologinen ja syntaktinen kompleksisuus kietoutuvat yhteen ja vaikuttavat toisiinsa. Osatutkimuksen tulosten valossa näiden kompleksisuuden osa-alueiden tarkastelu erillisinä ilmiöinä ei välttämättä ole perusteltua suomen kaltaisessa morfologisesti rikkaassa kielessä.

5 KOLME NÄKÖKULMAA SYNTAKTISEEN KOMPLEKSISUUTEEN

5.1 Oppijansuomen syntaktinen kompleksisuus eri taitotasoilla

Tämän tutkimuksen lähtökohtana oli kartoittaa, miten tyypilliset syntaktisen kompleksisuuden määrälliset mittarit toimivat oppijansuomen aineistossa. Tavoitteena oli selvittää, löytyykö mittareiden avulla eri taitotasoille tyypillisiä kielenpiirteitä ja erottelevatko mitatut kielenpiirteet oppijansuomen taitotasoja. Tässä alaluvussa keskitytään siihen, millaisen kuvan käytetyt määrälliset mittarit yhdessä aineiston teksteissä esiintyvien konjunktioiden ja moniverbisten konstruktioiden käytön tarkastelun kanssa muodostavat oppijansuomen syntaktisesta kompleksisuudesta Eurooppalaisen viitekehyksen eri taitotasoilla. Mittareiden kykyä erotella oppijansuomen taitotasoja käsitellään tarkemmin luvussa 5.2.

Tutkimuksen alussa syntaktista kompleksisuutta tarkasteltiin aiemmissa oppijankielen kompleksisuuden tutkimuksissa yleisesti käytettyjen määrällisten mittareiden avulla. Ne valittiin Norrisin ja Ortegan (2009: 574) suositusta seuraten siten, että mukana oli yleisen kompleksisuuden, lauseiden alisteisuuden ja rinnasteisuuden sekä lauseiden sisällä tapahtuvan laajenemisen mittareita (syntaktisen kompleksisuuden määrällisistä mittareista ks. myös luku 2.3). Tutkimuksen muissa kahdessa osatutkimuksessa keskityttiin sellaisiin muutoksiin, joita ei välttämättä saatu näkyviin määrällisillä mittareilla. Kielenyksiköiden yhdistämistä toisiinsa tarkasteltiin eri taitotasoilla esiintyvien konjunktioiden avulla, ja lauseiden sisällä tapahtuvia muutoksia kartoitettiin teksteissä esiintyvien moniverbisten konstruktioiden avulla. Näin pyrittiin saamaan näkyviin mahdollinen monipuolistuminen ja variaatio, jotka voivat jäädä määrällisten mittareiden tavoittamattomiin.

Oppijansuomen yleistä syntaktista kompleksisuutta eri taitotasoilla mitattiin kolmen määrällisen mittarin avulla. Mittareiksi valittiin virkkeen keskipituus

sanoina (MLS), T-yksikön keskipituus sanoina (MLTU) sekä virkkeen keskimääräinen lausemäärä (C/S). Näitä mittareita on tyypillisesti käytetty yleisen syntaktisen kompleksisuuden mittareina (ks. esim. Bulté & Housen 2012), ja niiden on todettu kasvavan kielitaidon kehittyessä (myös Ortega 2003; Wolfe-Quintero ym. 1998). Etenkin T-yksikön sanamäärää on perinteisesti pidetty hyvänä toisen kielen taidon kehityksen mittarina, jonka kasvulla on selvä yhteys kielitaidon edistymiseen (esim. Wolfe-Quintero ym. 1998). Nyt saadut tulokset ovat osittain samansuuntaisia aiempien tutkimusten tulosten kanssa, sillä myös oppijansuomen aineistossa sekä virkkeiden että T-yksiköiden sanamäärät kasvoivat taitotason noustessa. Sanamäärien kasvu ei kuitenkaan kaikissa tehtävätyypeissä jatkunut ylimmille taitotasoille saakka. Tämä on osittain odotuksenmukaista, sillä aiemmissa tutkimuksissakin on havaittu, että oppijankielessä T-yksikön pituuden kasvu usein tasaantuu tietyn keskipituuden jälkeen (esim. Bulté & Housen 2012: 37). On kuitenkin huomattava, että nyt käsillä olevan tutkimuksen aineistossa taitotasot eivät olleet yhtenäisiä, vaan taitotasojen sisällä tekstien välillä oli paljon vaihtelua niin T-yksiköiden kuin virkkeiden keskipituuksissa. Virkkeen sisältämien lauseiden määrän muutoksissa puolestaan ei ollut yhtä selvää trendiä.

Yleisen kompleksisuuden mittaamisessa epämuodollisia viestejä, muodollisia viestejä ja mielipidetekstejä tarkasteltiin erikseen. Kun tehtävätyyppejä tarkasteltiin rinnakkain, mittareiden tuottamissa tuloksissa havaittiin eroja tehtävätyyppien välillä. Erojen tilastollista merkitsevyyttä ei kuitenkaan testattu. Oppijansuomen syntaktisen kompleksisuuden suhde tehtävätyyppiin jää jatkotutkimuksen tehtäväksi. Aineisto tarjoaa siihen kaksi mielenkiintoista asetelmaa, joissa molemmissa on mahdollista kontrolloida kirjoittajakohtaisten taustatekijöiden vaikutusta. Koska aineiston jokainen teksti on erikseen arvioitu taitotasolle, on mahdollista tarkastella saman kirjoittajan eri tehtävätyyppejä edustavia tekstejä, jotka on arvioitu keskenään joko samalle taitotasolle tai eri taitotasoille. Molemmissa asetelmissa tutkimusaineisto olisi tämän tutkimuksen aineistoa pienempi, joten tarkasteluun sopisi laadullisempi ote, joka jo itsessään voisi tuoda näkyviin sellaisia tehtävätyyppien välisiä eroja, jotka jäävät havaitsematta tässä tutkimuksessa käytetyin mittarein.

Tässä väitöskirjatutkimuksessa aikuisten ja nuorten aineisto pidettiin erillään. Näiden kahden kirjoittajaryhmän tekstien tarkastelu rinnakkain osoitti, että määrällisin mittarein kuvattu yleinen syntaktinen kompleksisuus ja siinä näkyvät muutokset taitotasolta toiselle eivät olleet samanlaisia. Kaikkien kolmen mittarin mukaan nuorten tekstit olivat ryhmätasolla tarkasteltuna lähes kaikilla taitotasoilla hieman kompleksisempia kuin aikuisten tekstit vastaavilla taitotasoilla. Aikuisten ja nuorten tekstejä ei kuitenkaan verrattu toisiinsa tilastollisin menetelmin, joten erojen tilastolliseen merkitsevyyteen ei tässä tutkimuksessa oteta kantaa. Tulosta voi osittain selittää se, että tekstit on arvioitu taitotasoille niissä käytettyjen kielenpiirteiden sijaan kommunikatiivisin kriteerien. Aiemmat tutkimukset ovat antaneet viitteitä siitä, että suurempi kompleksisuuden määrä ei välttämättä tarkoita parempaa kielitaitoa. Esimerkiksi kielenpiirteiden ja englanti toisena kielenä -kirjoittamisen laadun välistä yhteyttä tutkineet Taguchi,

Crawford ja Wetzel (2013) totesivat, että kun yliopiston vaihto-opiskelijoiden lähtötasotestien teksteistä koostunut aineisto jaettiin pistemäärän perusteella kahteen ryhmään, vähemmän pisteitä saaneet tekstit olivat joidenkin tutkimuksessa käytettyjen mittareiden mukaan syntaktisesti kompleksisempia kuin enemmän pisteitä saaneet tekstit. Yhtenä mahdollisena selittävänä tekijänä he mainitsivat alistuskonjunktioiden liikakäytön, joka tuotti liiallista kompleksisuutta (mts. 426).

Kompleksisuuden mittaamisessa keskeisessä roolissa on kuitenkin usein ollut juuri alisteisuus (esim. Ellis & Barkhuizen 2005: 140). Sivulauseiden käyttöä on pidetty etenkin keskitasolle tyypillisenä syntaktisen kompleksisuuden piirteenä (mts. 155; Norris & Ortega 2009). Tähän tutkimukseen otettiin mukaan kaksi sivulauseiden osuutta kuvaavaa määrällistä mittaria: T-yksiköiden keskimääräinen lausemäärä (C/TU) ja sivulauseiden osuus kaikista lauseista (DC/C). Näistä ensimmäinen, T-yksikön sisältämien lauseiden määrä, kasvoi nyt tarkastellussa oppijansuomen aineistossa taitotason noustessa, mutta kasvu ei ollut yhtäjaksoista alimmilta taitotasoilta ylimmille. Kun aikuisten ja nuorten aineistoa ja eri tehtävätyyppejä tarkasteltiin rinnakkain, kirjoittajaryhmien välillä havaittiin eroja. Aikuisten aineistossa lausemäärän kasvu oli selvintä mielipideteksteissä, joissa erot A-tason ja kaikkien ylempien taitotasojen välillä todettiin tilastollisesti merkitseviksi, kun taas nuorten aineiston mielipideteksteissä ei ollut tilastollisesti merkitseviä eroja taitotasojen välillä. Tätä selittää ainakin osittain taitotason A1 suuri sisäinen vaihtelu etenkin nuorten mielipideteksteissä. Toinen alisteisuuden määrällinen mittari, sivulauseiden osuus kaikista lauseista, kehittyi varsin samalla tavalla kuin T-yksikön sisältämien lauseiden määrä, mikä on odotuksenmukaista, kun kyse on kahdesta saman syntaktisen kompleksisuuden ulottuvuuden mittarista. Tulokset eivät tue näkemyksiä alisteisuudesta erityisesti keskitasolle tyypillisenä kielellisenä keinona, sillä vaikka aikuisten aineistossa kahden alimman taitotason ja ylempien taitotasojen välillä oli tilastollisesti merkitseviä eroja, ainoat tilastollisesti merkitsevät erot keskitason ja ylimpien taitotasojen välillä olivat taitotasojen B2 ja C1 välillä epämuodollisissa viesteissä. Niissä taitotasolla B2 sekä lauseiden määrä T-yksikössä että sivulauseiden osuus kaikista lauseista olivat tilastollisesti merkitsevästi suurempia kuin taitotasoilla C1 ja B1.

Määrällisin mittarein saatuja tuloksia täydennettiin tarkastelemalla teksteissä käytettyjä alistuskonjunktioita. Kullekin taitotasoille sijoittuvista teksteistä laskettiin niiden sisältämät alistuskonjunktiot sekä alistuskonjunktion sisältävien sivulauseiden osuus kaikista sivulauseista. Tulokset osoittivat, että alistuskonjunktiot ilmaantuvat suomenoppijoiden teksteihin jo heti taitotasolla A1 ja että taitotasojen väliset erot alistuskonjunktioiden sanamäärään suhteutetuissa esiintymistaajuuksissa olivat tilastollisesti merkitseviä vain aikuisten aineistossa taitotasojen B2 ja C2 välillä. Vaikka alistamista on tyypillisesti pidetty kielenpiirteenä, joka opitaan vasta keskitasolla (esim. Ellis & Barkhuizen 2005: 155), alistuskonjunktioiden on myös muissa tutkimuksissa todettu ilmaantuvan oppijankieleen jo alkeistasolla (Määttä 2012; Vyatkina 2012, 2013). Tästä näkökulmasta on mielenkiintoista, että tämän tutkimuksen nuorten aineistossa taitotasoilla A1 ja A2 alistuskonjunktioiden suhteellinen esiintymistaajuus oli hieman suurempi

kuin rinnastuskonjunktioiden. Kaikilla muilla taitotasoilla alistuskonjunktioiden esiintymistaajuus on pienempi kuin rinnastuskonjunktioiden.

Toisaalta näkemystä sivulauseiden tärkeydestä osana etenkin keskitason oppijankielen syntaktista kompleksisuutta tukee se, että ryhmätason tarkastelussa alistuskonjunktioiden tuhanteen sanaan suhteutetut esiintymistaajuudet olivat aikuisten aineistossa suurimmillaan taitotasoilla B1 ja B2 ja nuorten aineistossa taitotasoilla A2 ja B1. Tämän kanssa linjassa on myös se, että nuorten aineistossa laajin valikoima eri alistuskonjunktiota oli käytössä samoilla taitotasoilla eli taitotasoilla A2 ja B1. Sen sijaan aikuisten aineistossa eri konjunktioiden valikoima oli laajin taitotasoilla C1 ja C2. Nuorten aineiston taitotason B2 muita taitotasoja pienempää konjunktioiden variaatiota selittää todennäköisesti tälle taitotasolle sijoitettujen tekstien pieni määrä. Taitotasolla B2 myös aineiston sanamäärä ja konjunktioiden esiintymien määrä ovat pienemmät kuin muilla taitotasoilla.

Konjunktioiden esiintymistaajuuksien lisäksi tarkasteltiin alistuskonjunktion sisältävien lauseiden osuutta kaikista sivulauseista. Tulosten perusteella kaikilla taitotasoilla suurin osa sivulauseista sisältää alistuskonjunktion, mutta muiden kuin alistuskonjunktion sisältävien sivulauseiden osuus on aikuisten aineistossa taitotasolla C2 kolminkertainen (32 %) taitotasoon A1 (10 %) verrattuna. Nuorten aineistossa muiden kuin alistuskonjunktion sisältävien sivulauseiden osuus kaikista sivulauseista noin puolitoistakertaistuu A-tasolta (14 %) taitotasolle B1 (20 %). Näiden lukujen valossa alistuskonjunktiot eivät välttämättä toimi sivulauseiden likiarvona edes alimmilla taitotasoilla. Tulokset viittaavat myös siihen, että vaikka sivulauseiden osuus kaikista lauseista ei välttämättä enää kasva tilastollisesti merkitsevästi keskitasolta ylimmille taitotasoille oppijansuomessa, ylimmillä taitotasoilla erilaisten sivulauseiden tai sivulausetyyppien kirjo on laajempi. Tässä tutkimuksessa sivulauseiden kirjoa tarkasteltiin kuitenkin varsin kapeasta näkökulmasta alistuskonjunktioiden esiintymien avulla. Lisätietoa sivulauseiden merkityksestä oppijansuomen kompleksisuudessa antaisi esimerkiksi erilaisten sivulausetyyppien käytön tarkastelu eri taitotasoilla.

Tässä tutkimuksessa oppijansuomen syntaktista kompleksisuutta mitattiin myös rinnasteisuuden avulla. Mittarina käytettiin virkkeen sisältämien T-yksiköiden määrää (TU/S). Rinnastamista on pidetty lähinnä alkeistasolle kuuluvana kompleksisuuden piirteenä, jonka merkitys vähenee, kun kielitaidon taso nousee (esim. Norris & Ortega 2009). Tämän tutkimuksen tulokset tukevat tätä näkemystä vain osittain. Virkkeen sisältämien T-yksiköiden määrä väheni ryhmätasolla yhtäjaksoisesti alimmalta taitotasolta ylimmälle vain nuorten aineiston epämuodollisissa viesteissä, kun ryhmätason kehitystä tarkasteltiin keskiarvojen avulla. Aikuisten aineistossa virkkeen sisältämien T-yksiköiden määrä oli suurin keskitasolla kaikissa tehtävätyypeissä. Koska käytetty määrällinen mittari (TU/S) tavoittaa vain virkkeessä olevien rinnasteisten päälauseiden määrän, rinnastamista tarkasteltiin myös aineistossa esiintyvien rinnastuskonjunktioiden avulla. Rinnastamisen yleisyyttä tarkasteltiin ryhmätasolla rinnastuskonjunktioiden absoluuttisen ja tuhanteen sanaan suhteutetun esiintymistaajuuden avulla, ja taito-

tasoja verrattiin tilastollisesti tekstikohtaisten sataan sanaan suhteutettujen esiintymistaajuuksien avulla. Lisäksi rinnastamisen monipuolisuutta tarkasteltiin selvittämällä, käytetäänkö rinnastuskonjunktioita lauseiden vai muiden kielenyksiköiden yhdistämiseen ja mitä eri rinnastuskonjunktioita eri taitotasoilla esiintyy.

Tyypillisesti rinnastuskonjunktioita käytettiin aineiston teksteissä lauseiden yhdistämiseen, ja suurin osa aineiston rinnastuskonjunktioista esiintyikin lauseen alussa. Aikuisten aineiston taitotasot C1 ja C2 poikkesivat muusta aineistosta siten, että niillä noin puolet rinnastuskonjunktioiden esiintymistä oli muualla kuin lause- tai virkerajalla. Lausekkeiden rinnastamista pidetäänkin tyypillisenä nimenomaan ylemmille kielitaidon tasoille (Norris ja Ortega 2009). Ryhmätasolla rinnastuskonjunktioiden absoluuttinen ja suhteellinen esiintymistaajuus olivat suurimmillaan taitotasolla B1. Taitotasojen väliset erot rinnastuskonjunktioiden suhteellisessa esiintymistaajuudessa olivat aikuisten aineistossa tilastollisesti merkitseviä A-tason ja B-tason välillä, nuorten aineistossa taitotasojen A1 ja B1 välillä. Sen sijaan tekstikohtainen rinnastuskonjunktioiden normalisoitu esiintymistaajuus ei vaikuttanut erottavan alimpia ja ylimpiä taitotasoja toisistaan. Nuorten aineiston osalta tulokseen on kuitenkin voinut vaikuttaa myös taitotasolle B2 arvioitujen tekstien vähyys. Eniten eri rinnastuskonjunktioita esiintyi aikuisten aineistossa taitotasoilla B1 ja B2, nuorten aineistossa taitotasolla B1. Etenkin taitotason B1 tuloksia voi osittain selittää se, että kyseisellä taitotasolla myös rinnastuskonjunktioiden absoluuttinen esiintymistaajuus oli suurimmillaan.

Rinnastuskonjunktiolla alkavia lauseita ei tässä tutkimuksessa jaettu erikseen pää- ja sivulauseisiin, mutta aineistoesimerkkien valossa rinnastuskonjunktioilla yhdistetään oppijansuomen eri taitotasoilla sekä pää- että sivulauseita. Keskenään rinnasteiset sivulauseet voivat vaikuttaa virkkeiden ja T-yksiköiden sanamäärään. Siten ne voivat myös näkyä määrällisissä mittareissa, jotka kuvaavat virkkeiden ja T-yksiköiden sana- tai lausemäärää tai sivulauseiden osuutta kaikista lauseista. Keskenään rinnasteiset sivulauseet jäävät kuitenkin tässä tutkimuksessa rinnastamisen mittarina käytetyn virkkeen sisältämien T-yksiköiden määrän tavoittamattomiin. Viimeksi mainitun mittarin tavoittamattomiin jää myös rinnastaminen lauseiden sisällä. Lauseen osien rinnastaminen voi sen sijaan näkyä virkkeiden, T-yksiköiden ja lauseiden sanamäärässä ja siten myös lauseiden yleistä kompleksisuutta mittaavassa lauseen sanamäärässä (MLC).

Oppijankielen syntaktisen kompleksisuuden ajatellaan usein näkyvän etenkin ylimmillä taitotasoilla juuri lauseiden sisällä, ei niinkään lauseiden välisissä suhteissa, ja sitä voidaan yleisellä tasolla mitata juuri lauseiden sisältämien sanojen määrän avulla (Norris & Ortega 2009). Tämän tutkimuksen tulosten mukaan oppijansuomen lauseet pitenevät lähes yhtäjaksoisesti alimmalta taitotasolta ylimmälle sekä aikuisten että nuorten aineistossa lähes kaikissa tehtävätyypeissä. Aikuisten aineistossa erot lauseiden pituudessa olivat tilastollisesti merkitseviä etenkin, kun kahta alinta taitotasoa verrattiin ylempiin taitotasoihin. Kun taitotasoja A1 ja A2 verrattiin muodollisissa viesteissä taitotasoihin B2–C2 ja mielipideteksteissä taitotasoihin B1–C2 sekä epämuodollisissa viesteissä C-tasoon ja taitotasoa A1 taitotasoon B2, erot olivat tilastollisesti merkitseviä. Myös B-tason ja C-tason välillä oli tilastollisesti merkitseviä eroja sekä epämuodollisissa että

muodollisissa viesteissä. Nuorten aineistossa erot taitotasojen välillä olivat tilastollisesti merkitseviä tälläkin mittarilla vain, kun taitotasoa A1 verrattiin taitotasoon B1 muodollisissa viesteissä ja mielipideteksteissä. Myös tässä mittarissa näkyy se, että määrällisesti mitattu kompleksisuus voi kehittyä suhteessa taitotasoon eri tavoin eri tehtävätyypeissä ja että näin mitattu kompleksisuus ja siinä tapahtuvat muutokset eivät välttämättä ole samanlaisia aikuisten ja nuorten aineistossa.

Lauseiden sisäistä kompleksisuutta tarkasteltiin lauseiden sanamäärän lisäksi moniverbisten konstruktioiden avulla aikuisten aineiston mielipideteksteissä. Tarkastelu osoitti, että ensimmäiset moniverbiset konstruktiot ilmaantuivat aineiston teksteihin jo taitotasolla A1 ja että ne alkoivat yleistyä taitotasolla B1, jolla myös niiden leksikaalinen variaatio oli selvästi suurempaa kuin A-tasolla. Taitotason B1 jälkeen moniverbisten konstruktioiden yleisyys ei enää juuri muuttunut. Sen sijaan sekä leksikaalinen että morfologinen variaatio lisääntyivät, ja ylimmillä taitotasoilla oli käytössä eniten erilaisia konstruktioita. Tulokset ovat samansuuntaisia kuin esimerkiksi Cefling-aineiston modaalisia ja nesessiivisiä verbiketjuja tutkineen Seilosen (2013) tulokset, joiden mukaan modaalisia ja nesessiivisiä verbiketjuja käytetään jo alimmilla taitotasoilla ja ylemmillä taitotasoilla on käytössä myös merkitykseltään eriytyneitä konstruktioita. Oppijansuomen erilaisia moniverbisiä rakenteita tutkineet Haapala (2008), Kynsijärvi (2008), Paavola (2008) ja Puhakka (2010) ovat myös todenneet sekä moniverbisten rakenteiden esiintymistaajuuden että niissä esiintyvien verbien kirjon kasvavan taitotasolta toisella ja konstruktioiden rakenteellisen variaation lisääntyvän taitotason noustessa.

5.2 Syntaktinen kompleksisuus oppijansuomen taitotason indikaattorina

Tämän väitöskirjan yhtenä tavoitteena oli selvittää, miten perinteiset syntaktisen kompleksisuuden määrälliset mittarit erottelevat oppijansuomen taitotasoja. Käytetyssä tutkimusaineistossa syntaktisen kompleksisuuden määrälliset mittarit tuottivat tilastollisesti merkitseviä eroja lähinnä vain silloin, kun alimpia taitotasoja, aikuisten aineistossa taitotasoja A1 ja A2 ja nuorten aineistossa taitotasoa A1, verrattiin ylempiin taitotasoihin. Mittareiden kyky erotella muita taitotasoja toisistaan oli selvästi heikompi. Lisäksi mittarit tuottivat erilaisia tuloksia syntaktisen kompleksisuuden ja taitotason suhteesta eri tehtävätyypeissä. Aikuisten aineistosta ja nuorten aineistosta saadut tulokset poikkesivat myös joiltain osin toisistaan.

Tutkituista seitsemästä syntaktisen kompleksisuuden mittarista vain neljällä tavoitettiin nuorten aineistossa tilastollisesti merkitseviä eroja taitotasojen välillä, kun taas aikuisten aineistossa eroja löytyi vähintään yhdessä tehtävätyypissä kaikilla käytetyillä seitsemällä syntaktisen kompleksisuuden määrällisellä

mittarilla. Osatutkimuksen 2 tulosten perusteella vaikuttaisi siltä, että perinteisistä määrällisistä mittareista parhaiten taitotasoja erottelee nuorten aineistossa T-yksikön keskipituus sanoina (MLTU), aikuisten aineistossa puolestaan lauseen keskipituus sanoina (MLC). Tulosta voi osin selittää se, että nuorten aineistossa ei ollut lainkaan tekstejä C-tasolta, jolloin ylin taitotaso oli mielipideteksteissä B1 ja viesteissä B2 eli nuorten aineistossa ylintä taitotasoa edustavat Eurooppalaisen viitekehyksen keskitason tekstit. Nämä molemmat syntaktisen kompleksisuuden määrälliset mittarit ovat kuitenkin siinä mielessä ongelmallisia, että lauseet ja T-yksiköt ovat etenkin alimmilla taitotasoilla segmentoinnin näkökulmasta osittain tulkinnanvaraisia. Lauseiden ja T-yksiköiden laadulliset erot eivät kuitenkaan välttämättä tule näkyviin, kun eri taitotasoja tarkastellaan syntaktisen kompleksisuuden määrällisillä mittareilla.

Tehtävätyyppien tarkastelu rinnakkain osoitti, että kompleksisuuden määrälliset mittarit eivät kaikissa aineiston tehtävätyypeissä tuota samanlaisia tuloksia vaan syntaktisessa kompleksisuudessa ja sen kehityksessä on näillä mittareilla tarkasteltuna eroja tehtävätyyppien välillä. Vaikka eri tehtävätyyppejä ei verrattu toisiinsa tilastollisesti, tulosten voi tulkita tukevan näkemyksiä, joiden mukaan tehtävätyyppi vaikuttaa oppijankielen syntaktiseen kompleksisuuteen. Myös aiemmissa tutkimuksissa on havaittu, että kielitaidon tason lisäksi tekstilaji tai tehtävä voivat vaikuttaa oppijankielen kompleksisuuteen (ks. esim. Michel 2017: 54, 60). Näin ollen mittareiden kyky erotella taitotasoja voi vaihdella myös sen mukaan, millaisista teksteistä tarkasteltava aineisto koostuu.

Tuloksissa oli havaittavissa eroja myös, kun aikuisten ja nuorten aineistoja tarkasteltiin rinnakkain. Aiemmat tutkimukset kielenpiirteiden kehittymisestä Cefling-aineiston teksteissä ovat osoittaneet, että transitiivikonstruktion (Reiman 2011a, 2014), nollapersoonaisten ilmausten (Seilonen 2013), eksistentiaalilauseiden (Kajander 2013) sekä paikan ja tilan ilmausten (Mustonen 2015) kehityksessä suhteessa taitotasoon on eroa aikuisten ja nuorten aineistojen välillä. Tämän tutkimuksen tulosten perusteella myös syntaktinen kompleksisuus suhteessa taitotasoon on erilaista aikuisten ja nuorten aineistoissa. Tämä tulee ottaa huomioon, jos oppijansuomen taitotasoa halutaan mitata syntaktisen kompleksisuuden avulla.

Kolmas erilaisten taustatekijöiden vaikutusta tukeva havainto on monilla taitotasoilla todettu suuri taitotason sisäinen variaatio. Tämä näkyi myös siinä, että mittareiden kuvaamat ryhmätason kehitystrendit eivät välttämättä olleet samanlaisia keskiarvoilla ja mediaaneilla tarkasteltuna. Taitotasojen sisäinen vaihtelu oli nuorten aineistossa tyypillisesti suurempaa kuin aikuisten aineistossa, tosin tässäkin oli jonkin verran vaihtelua eri mittareissa sekä tehtävätyypeittäin ja taitotasoittain. Yksi taitotasojen sisäiseen variaatioon vaikuttava tekijä voi olla kirjoittajien taustojen heterogeenisuus. Esimerkiksi eri ensikieliä on aineiston taustatietojen mukaan yli 20 (Jantunen & Pirkola 2015: 93; tarkemmin kirjoittajien taustatiedoista ks. Mustonen 2015: 72–80). Tässä tutkimuksessa muut taustatekijät kuin tehtävätyyppi ja kirjoittajan kuuluminen joko aikuisten tai nuorten ryhmään oli kuitenkin rajattu tarkastelun ulkopuolelle, koska tavoitteena oli löytää

yleisesti eri taitotasoja yhdistäviä ja erottavia kielenpiirteitä. Tehty rajaus ei kuitenkaan sulje pois sitä mahdollisuutta, että ensikielen kaltaiset taustamuuttujat ovat vaikuttaneet tekstien kompleksisuuteen eri taitotasoilla ja sitä kautta myös ryhmien sisäiseen yhtenäisyyteen. Taitotasojen sisäinen varianssi puolestaan on voinut vaikuttaa siihen, miten tässä tutkimuksessa käytetyt tilastolliset menetelmät tavoittivat taitotasojen väliset erot kompleksisuudessa.

Taitotasojen suuri sisäinen varianssi on voinut vaikuttaa tilastollisten analyysien tuloksiin. Niihin on osaltaan voinut vaikuttaa myös se, ettei poikkeavia havaintoja poistettu aineistosta ennen tilastollisia testejä. Tämä saattoi heikentää tilastollisten testien kykyä havaita eroja ryhmien välillä. Ilman poikkeavia havaintoja mahdolliset ryhmätason kehityskulut olisivat siis voineet tulla eri tavalla näkyviin. Koska tässä tutkimuksessa kartoitettiin valittujen määrällisten mittareiden soveltuvuutta taitotason mittaamiseen, myös poikkeavia havaintoja pidettiin osana taitotasolla mahdollista kielenpiirteiden kirjoa ja ne pidettiin mukana tilastollisessa analyysissa. Poistamisen sijaan poikkeavien arvojen vaikutusta pyrittiin kontrolloimaan käyttämällä parametristen testien rinnalla myös epäparametrisia testejä, joihin ryhmien sisäinen hajonta vaikuttaa vähemmän. Tehdyt valinnat ovat silti saattaneet vaikuttaa taitotasojen välisten tilastollisesti merkitsevien erojen vähyyteen. Lisäksi määrällisiä mittareita tarkasteltiin yksittäin. Niiden tarkastelu yhdessä monimuuttujaisilla tilastollisilla menetelmillä olisi voinut antaa toisenlaisen kuvan syntaktisen kompleksisuuden suhteesta kielitaidon taitotasoon. Tämä jää kuitenkin jatkotutkimuksen tehtäväksi.

Tilastollisten testien lisäksi aineistoa tarkasteltiin visuaalisesti laatikkojanojen avulla. Tarkastelu osoitti, että taitotasot olivat osittain päällekkäisiä. Mitä enemmän eri taitotasoilta saaduissa mittaustuloksissa on päällekkäisyyttä, sen vaikeampaa on määritellä taitotasoille tyypillisiä vaihteluvälejä, joiden avulla yksittäisiä tekstejä voitaisiin sijoittaa jollekin tietylle taitotasolle. Taitotasojen päällekkäisyys ei koske vain tämän tutkimuksen aineistoa. Jo T-yksikön käyttöä oppijankielen taidon indikaattorina tarkastellut Gaies (1980) on raportoinut, että useiden tutkimusten mukaan T-yksikköön perustuvat mittarit erottelevat osaamistasoja huonommin oppijankielessä kuin ensikielisessä aineistossa, koska oppijankielen aineistossa vierekkäisillä taitotasolla on yleensä paljon päällekkäisyyttä.

Tilastollisesti merkitsevien erojen puutteeseen on voinut vaikuttaa myös aineiston luonne. Aineiston tekstit ovat melko lyhyitä. Tästä syystä tuloksia ei voi varauksetta yleistää koskemaan oppijansuomen ylimpiä taitotasoja. Myös aineiston tehtävätyypit saattoivat vaikuttaa tuloksiin. Cefling-aineisto on tarkoitettu kielenpiirteiden ja Eurooppalaisen viitekehyksen taitotasojen välisen yhteyden tutkimiseen erilaisten kielenpiirteiden avulla, ei nimenomaan syntaktisen kompleksisuuden tutkimukseen. On todennäköistä, että esimerkiksi akateemisissa esseissä edistyneiden oppijoiden kielessä olisi voinut näkyä erilaisia piirteitä kuin esimerkiksi epämuodollisissa viesteissä. Myös nuorten aineistosta saatujen tulosten yleistettävyyteen on syytä suhtautua tietyllä varauksella etenkin taitotason B2 tulosten osalta, sillä taitotasolla B2 on vain pieni määrä tekstejä ja ne kaikki ovat tehtävätyypiltään muodollisia tai epämuodollisia viestejä.

Määrällisten mittareiden toimivuuteen kielitaidon indikaattorina vaikuttaa myös se, ettei oppijankielen segmentointi näissä mittareissa käytettyihin kielenyksiköihin ole ongelmatonta (esim. Foster ym. 2000; Rimmer 2006: 508). Sanoihin, lauseisiin, virkkeisiin ja lauseiden välisiin suhteisiin perustuvien mittareiden tarkkuuteen ja luotettavuuteen vaikuttaa se, miten aineisto on segmentoitu. Oppijankielen annotointi ja segmentointi on aina osittain tulkintaa (Brunni ym. 2015; Ragheb & Dickinson 2011; Rehbein ym. 2012). Oppijansuomi ei ole poikkeus tästä (Martin 2013a). Siksi segmentointiprosessista ja aineiston annotoijien yhdenmukaisuudesta raportointia pidetään tärkeänä syntaktista kompleksisuutta määrällisesti mittaavissa tutkimuksissa (Ortega 2012: 140). Tässä tutkimuksessa annotoinnin on kuitenkin tehnyt vain yksi henkilö. Useamman annotoijan käyttö olisi mahdollistanut neuvottelun segmentoinnin kannalta ongelmallisista kohdista ja johtanut kenties joihinkin erilaisiin ratkaisuihin.

Toisaalta useamman annoitoijan käyttökään ei olisi muuttanut sitä, että oppijankieli sisältää segmentoinnin näkökulmasta ongelmallisia ilmauksia. Määrällisten tulosten tulkinnassa onkin otettava huomioon se, etteivät mittareissa käytetyt kielenyksiköt eri taitotasoilla ole keskenään täysin vertailukelpoisia. Etenkin alimmilla taitotasoilla tekstit sisältävät paljon epätyypillisiä lauseita, jotka eroavat monella tavalla ylempien taitotasojen lauseista. Laadulliset erot jäävät määrällisillä mittareilla havaitsematta silloin, kun erot eivät vaikuta lauseiden sanamäärään tai siihen, tulkitaanko lauseet pää- vai sivulauseiksi.

Oppijankielen kompleksisuuden soveltuvuuteen kielitaidon taitotason indikaattoriksi voi vaikuttaa myös se, ettei kompleksisuus aina ole kommunikoinnin kannalta välttämätöntä (De Clercq & Housen 2017: 317). Näin ollen taitotason nousu ei aina näy kompleksisuuden lisääntymisenä. Etenkin harjaantuneet kirjoittajat voivat myös mukauttaa viestintäänsä ja säädellä käyttämänsä kielen kompleksisuutta (Lambert & Kormos 2014: 612). Eri tilanteissa odotuksenmukainen kompleksisuus voikin olla erilaista (esim. Pallotti 2009). Koska tämän tutkimuksen aineistossa eri tehtävätyyppien tekstit tulivat osin eri kirjoittajilta, tehtävätyyppejä ei verrattu tilastollisesti toisiinsa. Siten nyt saatujen tulosten perusteella ei ole mahdollista erottaa toisistaan tehtävätyypin, kirjoittajien tekemien tyyli- ja rekisterivalintojen tai muiden mahdollisten taustatekijöiden vaikutusta tekstien syntaktiseen kompleksisuuteen. Nyt havaitut erot yhtäältä tehtävätyyppien ja toisaalta aikuisten ja nuorten aineistojen välillä sekä taitotasojen sisäinen variaatio kuitenkin viittaavat siihen, että tällaista vaihtelua on myös oppijansuomessa.

5.3 Täydentyvä kuva oppijankielen kompleksisuudesta

Koska perinteiset syntaktisen kompleksisuuden määrälliset mittarit vaikuttivat soveltuvan melko huonosti oppijansuomen taitotasojen erottelemiseen, niiden lisäksi aineistoa tarkasteltiin konjunktioiden ja moniverbisten konstruktioiden avulla. Tarkastelu osoitti, että oppijansuomessa syntaktisen kompleksisuuden alaan kuuluvissa kielenpiirteissä tapahtuu myös sellaista kehitystä, joka ei näy

perinteisissä määrällisissä mittareissa. Määrällisten mittareiden kyvyn tavoittaa vain osa oppijansuomen kompleksisuuden kehityksestä ovat aiemmin nostaneet esiin esimerkiksi Martin, Mustonen, Reiman ja Seilonen (2010) sekä Reiman (2011b). Laadullisempaa tutkimusotetta oppijankielen kompleksisuuden tutkimukseen ovat peräänkuuluttaneet myös Larsen-Freeman (2006, 2009) sekä Rimmer (2006, 2009).

Tämän tutkimuksen tulokset osoittavat myös, että jotkin vasta myöhemmin oppijankieleen ilmaantuviksi ajatellut kielenpiirteet ilmaantuvat oppijansuomeen jo kielitaidon perustasolla. Yksi tällainen kielenpiirre on alisteisten lauseiden käyttö. Oppijansuomeen sivulauseet ilmaantuvat jo taitotasolla A1. Samoin moniverbisiä konstruktioita esiintyy suomenoppijoiden teksteissä jo taitotasolla A1, vaikka sekä esiintymiä että variaatiota on vähän. Vastaavan havainnon kompleksisena pidettyjen kielenpiirteiden varhaisesta ilmaantumisesta oppijansuomeen on tehnyt Seilonen (2013), joka totesi, että usein edistyneenä kielenpiirteenä pidetty passiivi ilmaantuu oppijoiden teksteihin jo ensimmäisellä taitotasolla ja että impersonaalisen passiivin käyttö myös hallitaan DEMfad-mallin kriteerien mukaan nuorten aineistossa jo taitotasolla A1.

Oppijankielen syntaktista kompleksisuutta on tutkittu eniten englannin kielessä (esim. Brezina & Pallotti 2019: 100). Samat mittarit eivät kuitenkaan aina tuota samoja tuloksia eri kielissä (Gyllstad ym. 2014; Kuiken & Vedder 2019), eivätkä syntaktisen kompleksisuuden mittarit toimi samalla tavalla kaikissa kielissä (Bernardini & Granfeldt 2019: 213). Tässä tutkimuksessa saatujen tulosten perusteella perinteiset määrälliset syntaktisen kompleksisuuden mittarit eivät ole sillä tavalla yhteydessä oppijansuomen taitotasoihin, että niiden avulla voitaisiin mitata suomenoppijan kielitaitoa. Yksi syy tähän voi olla se, että suomi morfologisesti rikkaana kielenä poikkeaa englannista, jonka tutkimukseen mittarit on alun perin kehitetty. Siksi oppijankielen kompleksisuuden mittaamiseen saatetaan tarvita toisenlaisia mittareita. Tilma (2014) esimerkiksi on todennut, että suomen kaltaisessa morfologisesti rikkaassa kielessä morfeemit voivat olla määrällisissä mittareissa sanoja informatiivisempia mittayksiköitä. Myös suomen kieleen sopeutettu produktiivisen syntaksin indeksi (Nieminen & Torvelainen 2003) saattaisi soveltua etenkin oppimisen alkuvaiheessa myös oppijansuomen rakenteiden kehityksen analysointiin (Suni 2006: 473). Niemisen (2007) mukaan produktiivisen syntaksin indeksin käyttäminen perinteisten määrällisten mittareiden rinnalla toi suomea ensikielenä puhuvien lasten kielen kompleksisuuden kehityksen moniulotteisuuden näkyviin paremmin kuin esimerkiksi pelkkä ilmausten keskipituuden (*Mean Length of Utterance*) tarkastelu.

Perinteisten syntaktisen kompleksisuuden mittareiden sijasta oppijansuomen taitoa ja sen kehittymistä voitaisiin myös tarkastella esimerkiksi DEMfad-mallin (Franceschina ym. 2006) avulla. Oppijansuomen tarkastelu konstruktioina sanojen, lauseiden, virkkeiden ja T-yksiköiden sijaan näyttää tarjoavan määrällisiä mittareita tarkemman kuvan oppijansuomen kompleksisuudesta ja sen kehittymisestä (Reiman 2011b). Tätä näkemystä tukevat myös tämän väitöskirjan havainnot moniverbisten konstruktioiden esiintymisestä oppijansuomessa. Sa-

malla havainnot viittaavat siihen, että suomen kaltaisessa morfologisesti rikkaassa kielessä syntaktinen ja morfologinen kompleksisuus voivat kietoutua toisiinsa niin, ettei niiden tarkastelu erillisinä osa-alueina ole välttämättä perusteltua.

Vaikka tämän väitöskirjan tutkimusote oli määrällinen, saadut tulokset havainnollistavat sitä, että jotkin oppijankielen syntaktisen kompleksisuuden puolet jäävät tyypillisten määrällisten mittareiden ulottumattomiin. Mittareiden tavoittamattomiin jäävä oppijankielen diversiteetin lisääntyminen näkyi tässä tutkimuksessa aineiston teksteissä esiintyvien konjunktioiden varannon kasvuna sekä moniverbisten konstruktioiden leksikaalisen ja morfologisen variaation lisääntymisenä. Jos oppijankielen kompleksisuutta tarkastellaan sen sisältämien kompleksiseksi miellettyjen kielenpiirteiden esiintymisen sijaan oppijankielen sisältämien erilaisten ilmaisukeinojen määränä ja monipuolisuutena, tässä tutkimuksessa havaittuja muutoksia konjunktioiden käytössä ja moniverbisissä konstruktioissa voidaan pitää merkkinä kompleksisuuden lisääntymisestä.

6 LOPUKSI

Myös tämän tutkimuksen tulosten valossa kielitaidon taitotason ja oppijankielen kompleksisuuden suhde on monimutkainen. Kun oppijansuomen syntaktista kompleksisuutta mitataan kompleksisena pidettyjen kielenpiirteiden avulla, enemmän ei välttämättä tarkoita ylempää kielitaidon tasoa. Kompleksisuuden kohdalla määrä ei aina olekaan samaa kuin laatu (esim. Michel 2017: 62; Pallotti 2009: 597). Kompleksisena pidettyjen kielenpiirteiden kehitys ei myöskään oppijansuomessa ole lineaarista tai jatku yhdenmukaisesti alimmilta taitotasoilta ylimmille. Epälineaarisuutta voidaan pitää oppijankielen kehitykselle tyypillisenä, etenkin jos kieltä tarkastellaan dynaamisena systeeminä (Larsen-Freeman 2006). Oppijankielen kehityksen epälineaarisuus voi näkyä myös niin, että yhdessä vaiheessa jo osatut muodot jäävät pois jossakin myöhemmässä vaiheessa ja ilmestyvät sitten uudelleen oppijankieleen (Lightbown & Spada 2013: 189).

Tämän tutkimuksen tulosten mukaan oppijansuomessa esiintyy jo taitotasolla A1 kielenpiirteitä, joiden yleensä on katsottu ilmaantuvan oppijankieleen vasta myöhemmin. Epälineaarinen kehitys ja joidenkin syntaktisesti kompleksisina pidettyjen rakenteiden varhainen ilmaantuminen oppijansuomeen voivat osaltaan selittää sitä, etteivät syntaktisen kompleksisuuden määrälliset mittarit näytä soveltuvan kirjoitetun oppijansuomen taitotason mittaamiseen.

Tässä väitöskirjassa oppijankielen kompleksisuutta on tarkasteltu suhteessa kielitaidon taitotasoon. Nyt saatuja tuloksia ei voi yleistää koskemaan oppijansuomen oppimisprosessia ja kompleksisuuden roolia suomen kielen taidon kehittymisessä, koska tutkimuksessa on käytetty poikittaisaineistoa, jolla ei voida päästä käsiksi tällaisiin kehityskulkuihin. Vaikka syntaktinen kompleksisuus näiden tulosten valossa toimii huonosti oppijansuomen taitotason indikaattorina, se voi tarjota arvokkaan näkökulman kielenoppimisen seuraamiseen. Ryhmä- ja yksilötason oppimisprosessien tutkiminen pitkittäisaineistossa, esimerkiksi Topling-aineistossa, antaisi tietoa oppijansuomen kompleksisuudesta suhteessa kielitaidon kehittymiseen.

Tutkimusaineisto koostui Eurooppalaisen viitekehyksen taitotasoille arvioiduista teksteistä. Kielitaidon taso oli siis määritelty kommunikatiivisen kompe-

tenssin perusteella. Syntaktista kompleksisuutta puolestaan tarkasteltiin määrällisesti eri kielenpiirteiden avulla. Kompleksisuus ei kuitenkaan ole välttämättä kielitaitoa (esim. Crossley & McNamara 2014; Ortega 2003), eikä nyt käytetyillä mittareilla näkyvä kompleksisuuden lisääntyminen välttämättä tarkoita parempaa suomen kielen taitoa. Eurooppalaisessa viitekehyksessä taitotason kriteereissä on keskeistä se, miten kielenoppija selviää erilaisista viestintätilanteista. Syntaktinen kompleksisuus ei kaikissa tilanteissa ole kommunikaation kannalta välttämätöntä (De Clercq & Housen 2017: 317; Lambert & Kormos 2014: 612). Oppijankielen kompleksisuus ei välttämättä olekaan ilmiö, joka odotuksenmukaisesti lisääntyy jatkuvasti kielitaidon edistyessä (Pallotti 2009: 598).

Tämän tutkimuksen tulokset tukevat näkemyksiä, joiden mukaan oppijankielen kompleksisuuteen vaikuttavat muutkin tekijät kuin kielitaidon taitotaso. Esimerkiksi tehtävän tai tekstilajin (ks. esim. Michel 2017: 54, 60) ja ensikielen (Khushik & Huhta 2020; Lu & Ai 2015) on todettu vaikuttavan määrällisesti mitattuun syntaktiseen kompleksisuuteen. Erilaisten henkilökohtaisten taustatekijöiden vaikutus oppijansuomen kompleksisuuteen ja sen kehittymiseen oli rajattu tarkastelun ulkopuolelle, mutta niiden tutkiminen jatkossa voisi antaa arvokasta lisätietoa kompleksisuuden ja kielitaidon tason suhteesta.

Tilanteisen ja henkilökohtaisen vaihtelun lisäksi määrällisesti mitattuun oppijankielen syntaktiseen kompleksisuuteen eri taitotasoilla voi vaikuttaa myös kohdekieli (Bernardini & Granfeldt 2019: 213; Gyllstad ym. 2014; Kuiken & Vedder 2019). Tässä tutkimuksessa kohdekielenä oli suomi. Morfologisesti rikkaana kielenä se poikkeaa englannista, jonka syntaktista kompleksisuutta on tutkittu eniten. Tämän tutkimuksen tulokset tukevat näkemyksiä, joiden mukaan syntaktinen kompleksisuus on vain osa oppijankielen kieliopillista kompleksisuutta. Morfologialtaan erilaisissa kielissä syntaktinen ja morfologinen kompleksisuus voivat tuottaa erikseen tarkasteltuna erilaisen kuvan kompleksisuudesta, sillä samaa tietoa ei kielessä välttämättä koodata sekä morfologian että syntaksin tasolla (Juola 2008: 106). Siksi suomen kaltaisessa morfologisesti rikkaassa kielessä kompleksisuuden morfologinen ulottuvuus tulisi ottaa huomioon syntaktisen rinnalla.

Oppijansuomen kompleksisuutta tarkasteltiin myös kielellisten keinojen monipuolisuutena konjunktioiden ja moniverbisten konstruktioiden avulla. Tulosten valossa vaikuttaa siltä, että kielellisten keinojen monipuolisuus ja kyky käyttää erilaisia kielellisiä keinoja eri konteksteissa lisääntyvät kielitaidon tasolta toiselle edettäessä. Sujuvuuden ja tarkkuuden ohella kielitaito näyttäisi siis olevan kykyä hyödyntää kielessä olevaa erilaisten keinojen valikoimaa kuhunkin tilanteeseen sopivalla tavalla. Miten kielenoppijan tai -käyttäjän hallussa olevaa varantoa ja sen laajuutta voi mitata yksittäisistä tekstinäytteistä tai oppijankielen korpuksista, jää kuitenkin jatkotutkimuksen selvitettäväksi. Oppijankielen analysointi konstruktioiden avulla näyttäisi kuitenkin mahdollistavan leksikaalisen, morfologisen ja syntaktisen kompleksisuuden tarkastelun yhdessä. Samalla konstruktioiden tarkastelussa esimerkiksi DEMfad-mallin avulla kompleksisuuden tutkimukseen tulisi mukaan myös laadullinen ulottuvuus. Konstruktionäkökulman mukainen muodon, merkityksen ja käytön tarkasteleminen yhdessä voisi

tarjota lisätietoa kompleksisuudesta ja sen kehityksestä suhteessa sekä kommunikatiivisin perustein määritettyyn kielitaidon taitotasoon että suhteessa yksilölliseen ja tilanteiseen vaihteluun. Laadullisempi tutkimusote mahdollistaisi myös oppijankielen kompleksisuuden tarkastelun erilaisten kielellisten ilmaisukeinojen monipuolisuutena ja kielenoppijan taitona hyödyntää kohdekielen tarjoamaa erilaisten ilmausten varantoa.

SUMMARY

Theoretical background of the study

When analysing learner language proficiency, the focus may be on learners' communicative competence or on linguistic features of the learner language. In the Common European Framework of Reference for Languages (CEFR 2001), learner language proficiency is divided into six levels, each described by the learner's ability to communicate efficiently in different situations. As the focus is on communicative competence, linguistic features typical of each level are mostly outside the scope of the CEFR scales. This has created a need to study the relation between the CEFR scales and the development of features of the learner language (e.g. Bartning, Martin &Vedder 2010). One of the aims of this dissertation is to meet this need by exploring syntactic complexity in written learner Finnish at the proficiency levels of CEFR.

The development of linguistic features in a second language (L2) can be studied by analysing the complexity, accuracy and fluency of the learner language (e.g. Housen, Kuiken & Vedder 2012; Michel 2017; Pallotti 2021; Wolfe-Quintero, Inagaki & Kim 1998). In the field of second language learning, there are multiple ways to define and operationalise the construct of complexity (e.g. Housen et al. 2019; Pallotti 2021). What the different definitions of learner language complexity seem to share is the idea of the range, variety and sophistication of learner language (Ellis & Barkhuizen 2005: 139; Ortega 2015: 86). The focus of analysis can be on the forms and structures of learner language (e.g. Ortega 2003: 492) or on the different components and the connections between them in learner language (e.g. Bulté & Housen 2012: 24). In this dissertation, both of these views are explored in L2 Finnish, first by applying traditional quantitative measures of syntactic complexity to the data and then by taking a closer look at the use of subordinating and coordinating conjunctions as well as multi-verb constructions in the L2 Finnish data.

According to Ortega (2012), previous studies have approached learner language complexity as an indicator of development, as an indicator of proficiency, or in relation to cognitive complexity. When complexity is studied in relation to language proficiency, the aim is often to find complexity measures that would be reliable indicators of proficiency (ibid. 128–129). Typically, these measures are quantitative. When analysing written learner language, among the most frequently used measures are Mean Length of T-unit (MLT or MLTU), Mean Length of Clause (MLC), Mean Number of Clauses per T-unit (C/T or C/TU), and Mean Number of Dependent Clauses per Clause (DC/C) (see e.g. Bulté & Housen 2012; Norris & Ortega 2009; Ortega 2003; Wolfe-Quintero et al. 1998 for extensive reviews of complexity measures). Three of these measures can be seen as measures tapping into subordination. However, Norris and Ortega (2009) have suggested also including measures of coordination and measures of elaboration at the phrasal level, as coordination could be a more useful measure at beginner levels and phrasal-level elaboration at advanced levels. Following

Norris and Ortega (2009), complexity has often been analysed more recently using sets of measures gauging these different dimensions of complexity. This also applies to the present study.

Although quantitative measures can be considered objective, they have also been subject to criticism on various grounds. One cause for criticism is the lack of clarity in the relation between the measures and the construct of complexity (e.g. Bulté & Housen 2012). Another is the focus on syntactic complexity and the relatively little attention given to other aspects of complexity, such as morphological complexity (e.g. Brezina & Pallotti 2019). However, it is not clear if syntactic, morphological, and lexical complexity are separate dimensions of complexity. For example, the use of more complex syntax can compensate for a simple lexicon (e.g. Lambert & Nakamura 2019: 251). In this dissertation, complexity is also explored by focusing on multi-verb constructions in L2 Finnish. In this approach, syntactic, morphological and lexical complexity can be studied together.

Previous research suggests that the relation between complexity and proficiency is not straightforward. In addition to proficiency, factors such as task type or genre (Michel 2017: 54. 60), the target language (e.g. Bernardini & Granfeldt 2019; Kuiken & Vedder 2019) or the learner's L1 background (e.g. Khusik & Huhta 2010) can also affect syntactic complexity. Also, an increase in complexity does not always equal an increase in proficiency. When complexity is measured as the frequency of linguistic features considered complex, e.g. subordination, high complexity scores do not necessarily indicate high proficiency (e.g. Crossley & McNamara 2014). Lambert and Kormos (2014: 612) also point out that expert writers can in some contexts use less complex language than less experienced writers. Pallotti (2009) notes that complexity can also be affected by the language user's choices and preferences, whether they are first or second language speakers.

The relation between complexity and proficiency in L2 Finnish has previously been studied using small longitudinal datasets and by analysing the development of certain linguistic structures. In the longitudinal studies, quantitative measures have been applied to written L2 Finnish texts. Alisaari (2016), in her study of L2 fluency, measured mean length of T-unit in written L2 Finnish data at CERF level A2 and found no significant change during a four-week course. Spoelman and Verspoor (2010), who studied L2 Finnish complexity in a DST framework using a longitudinal data set collected from one learner over a period of three years, found that the development of complexity was not linear over time. In her study tapping both syntactic and morphological complexity in L2 Finnish, Tilma (2014) measured the average length of sentence and clause in morphemes instead of words. She found that both measures grew over time but found no statistically significant correlation between the measures and development. She concluded that average length of clause in morphemes seemed the best measure of syntactic complexity in L2 Finnish. (Tilma 2014.)

The relation between some linguistic features and the CEFR levels in L2 Finnish has also been studied previously. In studies using the Cefling project data,

from which the data of the current study are also drawn, growth in frequency, variation and accuracy across the CEFR proficiency levels has been reported in the use of indirect references (Seilonen 2013), in existential sentences (Kajander 2013), and in transitive constructions (Reiman 2011a, 2011b, 2014).

The research questions

The focus of this dissertation is on syntactic complexity in written L2 Finnish across CEFR proficiency levels. On the one hand, the current study follows in the footsteps of the the University of Jyväskylä Cefling project in exploring the relation between the development of linguistic features and the CEFR proficiency levels. On the other hand, it contributes to the study of complexity, accuracy, and fluency in learner language by providing empirical evidence from a less studied target language. The aim is to trace the development of complexity in L2 Finnish across CEFR proficiency levels, to test the suitability of quantitative measures of complexity in L2 Finnish, and to view the construct of complexity from the perspective of a morphologically rich target language.

This study aims to answer the following research questions:

RQ1: How does syntactic complexity in L2 Finnish develop across CEFR proficiency levels?

RQ2: How well do the quantitative measures of syntactic complexity gauge the development of complexity in L2 Finnish?

RQ3: How does L2 Finnish complexity compare with L2 complexity as it is currently understood?

Data and methods

The data were drawn from a corpus of 667 texts (48,876 tokens) from 481 adult L2 Finnish learners and 411 texts (16,590 tokens) from 212 adolescent L2 Finnish learners. For comparative data, a set of 453 texts (19,826 tokens) from 175 L1 Finnish adolescent writers was included. The adults' texts come from the National Certificates of Language Proficiency examination database, and the adolescents' texts were collected at school from learners in school years 7, 8 and 9 using similar writing tasks. All the writing tasks were completed in a limited time without the use of any aids. (Alanen, Huhta & Tarnanen 2010; Martin et al. 2010.) The task types included in the study presented here are informal messages (e.g. an email to a friend), formal messages (e.g. a complaint to an online store), and argumentative texts (e.g. a text expressing an opinion on a given topic, such as the use of mobile phones at school). The narrative texts collected from the adolescent writers in the Cefling project were excluded since no narrative texts had been collected from the adult writers. Each L2 text in the data had been assessed at a CEFR proficiency level in the Cefling project (e.g. Alanen et al. 2014). The adults' texts cover all the CEFR levels, from A1 to C2, and the texts by L2 adolescent writers cover levels A1 to B2, although there is a limited number of

texts at B2. The L1 texts were not assessed to a proficiency level; instead, they were arranged according to school year.

In this study, the texts were manually segmented into words, clauses, sentences, and T-units. To compare the process of segmenting L2 and L1 Finnish data, the L1 data were also segmented. Additionally, all subordinating and coordinating conjunctions were coded in the data. In the L2 adult argumentative texts, all the finite verbs were coded as such, and finite verbs followed by one or more non-finite verb forms were coded as multi-verb constructions. All the coding was performed by one researcher.

To analyse syntactic complexity in the L2 Finnish texts, seven quantitative measures were used. They were applied to the adult learners' texts and to the adolescent learners' texts separately. The three task types were also treated as separate data subsets. To test the measures' ability to differentiate between CEFR proficiency levels in both writer groups and in all the task types, ANOVA and the Kruskal-Wallis test together with t-tests and the Wilcoxon rank sum test (also known as the Mann-Whitney U test) with Bonferroni corrections were used. In analysing the use of subordinating and coordinating conjunctions, the statistical significance of differences between the CEFR proficiency levels was tested using the Kruskal-Wallis and the Wilcoxon rank sum tests. Again, Bonferroni corrections were used in the pair-wise comparisons. For all the statistical tests, R version 3.4.4 and RStudio version 1.1.456 were used.

Main findings

In this dissertation, seven quantitative measures of syntactic complexity were applied to the data. For overall complexity, three measures were used: mean length of sentence, mean length of T-unit and mean number of clauses per sentence. Of these, the first two were observed to grow with an increase in proficiency, but there was no clear pattern in the mean number of clauses per sentence. However, the results varied according to task type, and there were differences between the adult and adolescent writers' data.

The role of coordination in L2 Finnish complexity was tested with one quantitative measure, i.e. mean number of T-units per sentence. Although the number was expected to decrease with increasing proficiency, this was true only for one task type in the data from the adolescent writers. In the data from the adult writers, the mean number of T-units in a sentence peaked at the intermediate levels. When the coordinating conjunctions used in the texts were analysed, it was discovered that they were mainly used to combine clauses. However, in the adult learners' texts at CEFR levels C1 and C2, half of all occurrences of coordinating conjunctions were found within clauses, not between them.

To gauge the role of subordination in L2 Finnish complexity, two quantitative measures were used. Both measures, mean number of clauses per T-unit and mean number of dependent clauses per clause, showed growth when

proficiency increased, but neither of the measures grew linearly from the lowest proficiency levels to the highest, and there were differences both between the task types and between the two writer groups (i.e. adults and adolescents). The study of subordinating conjunctions revealed that at the lower proficiency levels, the adolescent writers' texts contained more subordinating conjunctions than coordinating conjunctions. At all other proficiency levels, coordinating conjunctions were more frequent than subordinating conjunctions. The results also indicate that subordinating conjunctions cannot be considered reliable indicators of subordination. While the majority of subordinate clauses contained a subordinating conjunction, this was not the case with one third of the subordinate clauses at level C2 in the adult writers' data and with one fifth of the subordinate clauses at level B1 in the adolescent writers' data. When focusing on the different conjunctions used at each proficiency level, it was discovered that the range of conjunctions used was the widest at levels A2 and B1 for the adolescent learners and at C1 and C2 for the adult learners.

Additionally, syntactic complexity was measured using a quantitative measure of clause length (mean number of words per clause). Also this measure grew with the increase in proficiency from the lowest proficiency levels to the highest. However, not all the differences between proficiency levels were statistically significant, and there were differences in the results between the three task types and between the adults' and adolescents' groups. When the focus was on the multi-verb constructions in the adult writers' argumentative texts, the results showed that multi-verb constructions were already present at the lowest proficiency levels. According to the results, there was no significant growth in the length of these constructions, but lexical and morphological variation grew and the constructions became more idiomatic with increasing proficiency.

The statistical tests used in this study indicated that quantitative measures have a limited ability to differentiate between proficiency levels. While all seven quantitative measures of syntactic complexity showed statistically significant differences between at least two proficiency levels in the adult L2 data, only four of the measures did so in the adolescent L2 data. Additionally, not all the measures were able to differentiate between adjacent proficiency levels. The results show that quantitative measures can yield different results according to task type or the age of the writers. These results indicate that the quantitative measures used in the current study cannot be considered reliable measures of learner Finnish proficiency. However, the limitations of this study should be kept in mind when interpreting these results. In the first place, segmenting learner language into the production units used in the quantitative measures can be challenging (e.g. Foster et al. 2000; Martin 2013a), and in the current study, the data were segmented by one person only. Secondly, the statistical differences were tested using ANOVA and Kruskal-Wallis tests. No outliers were excluded from the statistical analyses. Choosing different statistical tests or cleaning the data could have yielded different results. However, visualising the data with box plots supports the results of the statistical tests as there was a lot of overlap

between proficiency levels and a lot of variation within many of the proficiency levels.

In line with the findings of Seilonen (2013), Kajander (2013) and Reiman (2011a, 2014), there seem to be differences in linguistic features and their development between adult and adolescent L2 Finnish learners. These results suggest that it may not be feasible to use the same measures for both adult and younger L2 Finnish learners.

The results suggest that measuring syntactic complexity as the presence, absence or frequency of coordination or subordination, or as the mean number of words in a given production unit, does not give a reliable indication of learner Finnish proficiency. On the one hand, this may depend on the way complexity was measured in this study. When learner language is segmented into the production units used in quantitative measures, the segments identified can, especially at the lower proficiency levels, be subject to interpretation. The nature of learner language also leads to qualitative differences that are not visible in quantitative measures. The mean length of clauses, T-units or sentences is no indication of whether or not the units are well-formed or are appropriate to the context in which they are used. On the other hand, the results may indicate that defining complexity as the increasing use of syntactic features considered to be complex is not a valid approach for all languages. According to the findings of this study, changes in morphology and lexicon may also play a part in the increase in complexity in a morphologically rich target language.

The results also support earlier findings that the relation between complexity and proficiency is not straightforward: increasing proficiency is not always reflected in quantitative measures as increasing complexity. In this regard, the results support earlier findings that there may be several other factors affecting complexity, such as the task used to elicit the samples or the age of the writer. Although the reasons behind within-group variation were not analysed in this study, the variation was considerable, especially at the intermediate levels, and a closer look at this within-group variance would make an interesting topic for future research.

Conclusion

In this dissertation, learner language complexity and the relation between complexity and proficiency were approached from two angles. First, complexity was operationalised as the frequency of certain linguistic features, such as the mean number of clauses per T-unit, and the data were analysed using quantitative measures. Statistical tests were used to find differences between the CEFR proficiency levels. Second, complexity was explored as variety in the range of the components in L2 Finnish at each CEFR level and as connections between these components. Here, the focus was on the use of subordinating and coordinating conjunctions and of multi-verb constructions. In the study of the

verb constructions, syntactic complexity in learner Finnish was observed to be closely connected to both morphological and lexical complexity.

In light of the findings of this study, analysing learner language complexity from a more qualitative viewpoint and studying complexity rather as the variety of linguistic means used than as the presence (or absence) of certain linguistic features might be a more fruitful approach to understanding learner language complexity. For languages with a rich morphology, studying syntactic, morphological and lexical complexity together could also be a valuable approach.

LÄHTEET

Alanen, R., Huhta, A. & Tarnanen, M. 2010. Designing and assessing L2 writing tasks across CEFR proficiency levels. Teoksessa I. Bartning, M. Martin & I. Vedder (toim.) Communicative proficiency and linguistic development: Intersections between SLA and language testing research. EUROSLA Monograph Series 1. European Second Language Association, 21–56. http://eurosla.org/monographs/EM01/EM01home.html

Alisaari, J. 2016. Songs and poems in the second language classroom. The hidden potential of singing for developing writing fluency. Annales Universitaris Turkuensis, Series B Humaniora 426. Turku: Turun yliopisto.

Bardovi-Harlig, K. 1992. A second look at T-unit analysis: Reconsidering the sentence, Tesol Quarterly 26 (2), 390–395. https://doi.org/10.2307/3587016

Bartning, I, Martin, M. & Vedder, I. (toim.) 2010. Communicative proficiency and linguistic development: Intersections between SLA and language testing research. EUROSLA Monograph Series 1. European Second Language Association. http://eurosla.org/monographs/EM01/EM01home.html

Benevento, C. & Storch, N. 2011. Investigating writing development in secondary school learners of French. Assessing Writing, 16 (2), 97–110. https://doi.org/10.1016/j.asw.2011.02.001

Bernardini, P. & Granfeldt, J. 2019. On crosslinguistic variation and measures of linguistic complexity in learner texts: Italian, French and English. International Journal of Applied Linguistics, 29 (2), 211–232. https://doi.org/10.1111/ijal.12257

Biber, D., Gray, B. & Poonpon, K. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?, TESOL Quarterly 45 (1), 5–35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Gray, B. & Staples, S. 2016. Predicting patterns of grammatical complexity across language exam task ypes and proficiency levels. Applied Linguistics, 37 (5), 639–668. https://doi.org/10.1093/applin/amu059

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. Longman grammar of spoken and written English. London: Longman.

Brants, T. 2000. Inter-Annotator agreement for a German newspaper corpus. LREC. http://www.lrec-conf.org

Brezina, V. & Pallotti, P. 2019: Morphological complexity in written L2 texts. Second Language Research 35 (1), 99–119. https://doi.org/10.1177/0267658316643125

Brunni, S., Lehto, L., Jantunen, J. & Airaksinen, V. 2015. How to annotate morphologically rich learner language. Principles, problems and solutions. Bergen Language and Linguistics Studies 6, 133–152. https://doi.org/10.15845/bells.v6i0.812

Bulté, B. & Housen, A. 2012. Defining and operationalising L2 complexity. Teoksessa A. Housen, V. Kuiken & I. Vedder (toim.) Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA. Amsterdam and Philadelphia: John Benjamins Publishing Company, 21–46.

CEFR = Council of Europe. 2001. Common European framework of reference for languages: learning, teaching, assessment. https://www.coe.int/lang-cefr

CEFR Companion = Council of Europe. 2020. Common European framework of reference for languages: learning, teaching, assessment. Companion volume. https://www.coe.int/lang-cefr

Crossley, S. & McNamara, D. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners, Journal of Second Language Writing 26, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

De Clercq, B. & Housen, A. 2017. A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. The Modern Language Journal, 101 (2), 315–334. https://doi.org/10.1111/modl.12396

Ellis, R. 2012. The study of second language acquisition. Second edition. Oxford: Oxford University Press.

Ellis, R. & Barkhuizen, G. 2005. Analysing learner language. Oxford: Oxford University Press.

EVK 2003 = Eurooppalainen viitekehys 2003. Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys. Helsinki: WSOY.

Foster, P., Tonkyn A. & Wigglesworth G. 2000. Measuring spoken language: a unit for all reasons, Applied Linguistics 21 (3), 354–375. https://doi.org/10.1093/applin/21.3.354

Franceschina F., Alanen, R., Huhta A. & Martin M. 2006. A progress report on the CEFLING project. Paper presented at SLATE Workshop, 1.–2. December 2006, Amsterdam.

Gaies S. 1980. T-Unit analysis in second language research: Applications, problems and limitations. TESOL Quarterly 14 (1), 53–60. https://doi.org/10.2307/3586808

Granger, S. 2002. A bird's-eye view of learner corpus research. Teoksessa S. Granger, J. Hung & S. Petch-Tyson (toim.) Computer learner corpora, second language acquisition and foreign language teaching. Amsterdam: Benjamins, 3–33.

Grant, L. & Ginther, A. 2000. Using computer-tagged linguistic features to describe L2 writing differences. Journal of Second Language Writing, 9 (2), 123–145.

Gries, S. T. 2013. Statistics for linguistics with R: A practical introduction. (2. uudistettu painos). Berlin: De Gruyter Mouton.

Gyllstad, H., Granfeldt, J., Bernardini, P. & Källkvist, M. 2014. Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. Teoksessa L. Roberts, I. Vedder & J.H. Hulstijn (toim.) Eurosla Yearbook 14. Amsterdam: John Benjamins, 1–30 .

Haapala, T. 2008. Finiittiverbeistä verbiketjuihin: verbiytimien kompleksistuminen S2-oppijoiden kielessä. Pro gradu -tutkielma. Tampereen yliopisto.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T. & Ginter, F. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. Language Resources and Evaluation 48(3), 493–531. https://doi.org/10.1007/s10579-013-9244-1

Honko, M. 2013. Alakouluikäisen leksikaalinen tieto ja taito: Toisen sukupolven suomi ja S1-verrokit. Acta Universitatis Tamperensis 1865. Tampere: Tampereen yliopisto.

Housen, A., De Clercq, B., Kuiken, F. & Vedder I. 2019. Multiple approaches to complexity in second language research. Second Language Research 35 (1), 3–21. https://doi.org/10.1177/0267658318809765

Housen, A. & Kuiken, V. 2009. Complexity, Accuracy, and Fluency in Second Language Acquisition. Applied Linguistics 30 (4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, V. & Vedder, I. (toim.) 2012. Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA. Amsterdam: John Benjamins Publishing Company.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M. & Hirvelä, T. 2014. Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. Language Testing 31 (3), 307–328. https://doi.org/10.1177/0265532214526176

Huhta, A. & Hilden, R. 2016. Kielitutkinnot ja muu laajamittainen kielitaidon arviointi Suomessa. Teoksessa A. Huhta & R. Hildén (toim.) Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2016 (9), 3–26.

Ivaska, I. 2015. Edistyneen oppijansuomen konstruktiopiirteitä korpusvetoisesti: avainrakenneanalyysi. Annales Universitatis Turkuensis, Series C Scripta Lingua Fennica Edita 409. Turku: Turun yliopisto.

Jantunen, J. & Pirkola, S. 2015. Oppijansuomen sähköiset tutkimusaineistot. Nykytilanne. Virittäjä, 119 (1), 88–103.

Jarvis, S. 2013. Capturing the diversity in lexical diversity. Language Learning, 63 (1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Juola, P. 2008. Assessing linguistic complexity. Teoksessa M. Miestamo, K. Sinnemäki & F. Karlsson (toim.) Language complexity: Typology, contact, change. Amsterdam: Benjamins, 89–108.

Kajander, M. 2013. Suomen eksistentiaalilause toisen kielen oppimisen polulla. Jyväskylä studies in humanities 220. Jyväskylä: Jyväskylän yliopisto.

Kalliokoski, J. 2006. Virke, dialogisuus ja argumentaatio: irralliset sivulauseet ja toisella kielellä kirjoittaminen. Teoksessa T. Nordlund, T. Onikki-Rantajääskö, T. Suutari & H. Forsberg (toim.) Kohtauspaikkana kieli: näkökulmia persoonaan, muutoksiin ja valintoihin. Helsinki: Suomalaisen Kirjallisuuden Seura, 212–231.

Khushik, G. & Huhta, A. 2020. Investigating syntactic complexity in EFL learners' writing across Common European framework of reference levels A1, A2, and B1. Applied Linguistics 41 (4), 506–532. https://doi.org/10.1093/applin/amy064

Kuiken, F. & Vedder, I. 2019. Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. International Journal of Applied Linguistics, Special issue, 192–210. https://doi.org/10.1111/ijal.12256

Kusters, W. 2008. Complexity in linguistic theory, language learning and language change. Teoksessa M. Miestamo, K. Sinnemäki & F. Karlsson (toim.) Language complexity: Typology, contact, change. Amsterdam: Benjamins, 3–22.

Kyle, K. & Crossley S. 2018. Measuring syntactic complexity in L2 writing using fine grained clausal and phrasal indices. The Modern Language Journal, 102(2), 333–349. https://doi.org/10.1111/modl.12468

Kynsijärvi, T. 2007. *Se johtuu siitä, että minulla oli muistinmenetys*. *Olla*-verbirakenteiden kehkeytyminen oppijankielessä. Pro gradu -tutkielma. Jyväskylän yliopisto.

Lambert, C. & Kormos, J. 2014. Complexity, accuracy and fluency in task-based research: Toward more developmentally based measures of second language acquisition. Applied Linguistics 35 (5), 607–614. https://doi.org/10.1093/applin/amu047

Lambert, C. & Nakamura, S. 2019. Proficiency related variation in syntactic complexity: A study of English L1 and L2 oral descriptive discourse. International Journal of Applied Linguistics 29, 248–264. https://doi.org/10.1111/ijal.12224

Larsen-Freeman, D. 2006. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. Applied Linguistics 27 (4), 590–619. https://doi.org/10.1093/applin/aml029

Larsen-Freeman, D. 2009. Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. Applied Linguistics, 30(4), 579–589. https://doi.org/10.1093/applin/amp043

Larsen-Freeman, D. 2010: Having and doing: Learning from a complexity theory perspective. Teoksessa P. Seedhouse, S. Walsh& C. Jenks (toim.) Conceptualising 'learning' in applied linguistics. Basingstoke: Palgrave Macmillan, 52–68.

Larson-Hall, J. 2010. A guide to doing statistics in second language research using SPSS. New York: Routledge.

Lightbown, P. & Spada, N. 2013. How Languages are Learned. Fourth edition. Oxford: Oxford University Press.
Lintunen, P. & Mäkilä, M. 2014. Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? Research in Language 12 (4), 377–399. https://doi.org/10.1515/rela-2015-0005
Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics 15 (4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu
Lu, X. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL Quarterly 45 (1), 36–62. https://doi.org/10.5054/tq.2011.240859
Lu, X. 2017. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. Language Testing 34 (4), 493–511. https://doi.org/10.1177/0265532217710675
Lu, X. & Ai, H. 2015.Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. Journal of Second Language Writing 29, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003
Martin, M. 2022. Käyttötaajuus ja tarkkuus toisen kielen kehityksessä. Teoksessa L. Aarikka, K. Priiki & I. Ivaska (toim.) Soveltavan kielitieteilijän sormenjälkiä etsimässä: Kielen rakenteet ja niiden käyttäjät Kirsti Siitosen tutkimuksellisina kiintopisteinä. In search of the fingerprints of an applied linguist: Linguistic structures and their users as scholarly focal points in Kirsti Siitonen's career. AFinLA-teema 14, 81–102.
Martin, M. 2013a. Sentences and clauses as complexity measures in second language writing: a segmentation experiment. Teoksessa M. Järventausta & M. Pantermöller (toim.) Finnische Sprache, Literatur und Kultur im deutschsprachigen Raum – Suomen kieli, kirjallisuus ja kulttuuri saksankielisellä alueella. Greifswald: Veröffentlichungen der Societas Uralo-Altaica, 185–198.
Martin, M. 2013b. The complex simple – a problematic adjective in the CEFR writing scales. Nordand 8 (2), 63–85
Martin, M., Mustonen, S., Reiman, N. & Seilonen, M. 2010. On becoming an independent user. Teoksessa I. Bartning, M. Martin & I. Vedder (toim.) Communicative proficiency and linguistic development: Intersections between SLA and language testing research. EUROSLA Monograph Series 1. European Second Language Association, 57–80. http://eurosla.org/monographs/EM01/EM01home.html
McNamara, D., Graesser, A., McCarthy, P. & Cai, Z. 2014. Automated evaluation of text and discourse with Coh-Metrix. Cambridge: Cambridge University Press.
Michel, M. 2017. Complexity, accuracy and fluency in L2 production. Teoksessa S. Loewen & S. Masatoshi (toim.) Routledge handbook of instructed second language acquisition. New York: Routledge, 50–68.

Miestamo, M. 2008. Grammatical complexity in a cross-linguistic perspective. Teoksessa M. Miestamo, K. Sinnemäki & F. Karlsson (toim.) Language complexity: Typology, contact, change. Amsterdam: Benjamins, 23–42.
Mustonen, S. 2015. Käytössä kehittyvä kieli. Paikat ja tilat suomi toisena kielenä -oppijoiden teksteissä. Jyväskylä studies in humanities 255. Jyväskylä: Jyväskylän yliopisto.
Määttä, T. 2012. Oppikirjan sanaston vaikutuksesta ruotsinkielisten alkeistason suomenoppijoiden kirjallisiin tuotoksiin. Lähivõrdlusi. Lähivertailuja 22, 188–218.
Nieminen, L. 2007. A complex case: a morphosyntactic approach to complexity in early child language. Jyväskylä studies in humanities 72. Jyväskylä: Jyväskylän yliopisto.
Nieminen, L. & Torvelainen, P. 2003. Produktiivisen syntaksin indeksi – suomenkielinen versio. Puhe ja kieli (3), 119–132.
Norris, J. M. & Ortega, L. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Applied Linguistics 30 (4), 555–578. https://doi.org/10.1093/applin/amp044
Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics 24 (4), 492–518. https://doi.org/10.1093/applin/24.4.492
Ortega, L. 2012. Interlanguage complexity. A construct in search of theoretical renewal. Teoksessa B. Kortmann & B. Szmrecsanyi (toim.) Linguistic Complexity: Second Language Acquisition, Indigenization, Contact. Berlin: De Gruyter, 127–155.
Ortega, L. 2015. Syntactic complexity in L2 writing: Progress and expansion. Journal of Second Language Writing 29, 82–94. http://doi.org/10.1016/j.jslw.2015.06.008
Paavola, V. 2008. Haluaisitko menna muunkansa kalastaman? Verbiketjujen kehkeytyminen suomi toisena kielenä -oppijoiden kielessä. Pro gradu -tutkielma. Jyväskylän yliopisto.
Pallotti, G. 2009. CAF: Defining, Refining and Differentiating Constructs. Applied Linguistics 30 (4), 590–601. https://doi.org/10.1093/applin/amp045
Pallotti, G. 2015. A simple view of linguistic complexity, Second Language Research 31 (1), 117–134. https://doi.org/10.1177/0267658314536435
Pallotti, G. 2021. Measuring Complexity, Accuracy, and Fluency (CAF). Teoksessa P. Winke & T. Brunfaut (toim.) The Routledge Handbook of Second Language Acquisition and Language Testing. Routledge: New York, 201–210.
Paquot, M. 2019. The phraseological dimension in interlanguage complexity research. Second Language Research 35 (1), 121–145. https://doi.org/10.1177/0267658317694221
Peters, A. 1983: The units of language acquisition. Cambridge: Cambridge University Press.

Polio, C. 2012. How to research second language writing. Teoksessa A. Mackey & S. Gass (toim.) Research methods in second language acquisition. A practical guide. Chichester: Wiley-Blackwell, 139–157.
Puhakka, Martta 2010. "*Sit se meni ja tuli hetken päästä takas*": verbit *mennä* ja *tulla* suomi toisena kielenä -oppijoiden teksteissä. Pro gradu -tutkielma. Jyväskylän yliopisto.
Ragheb, M. & Dickinson, M. 2011. Avoiding the comparative fallacy in the annotation of learner corpora. Teoksessa G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. P. Botana & E. Rhoades (toim.) Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions. Somerville, MA: Cascadilla Proceedings Project, 114–124. http://www.lingref.com/cpp/slrf/2010/paper2620.pdf
Rehbein, I., Hirschmann, H., Lüdeling, A. & Reznicek, M. 2012. Better tags give better trees – or do they?, Linguistic Issues in Language Technology 7 (10), 1–18. https://journals.linguisticsociety.org
Reiman, N. 2011a. Transitiivikonstruktio ikkunana syntaksin kehitykseen: infiniittiset rakenteet ja passiivi taidon indikaattoreina S2-oppijoiden teksteissä. Teoksessa E. Lehtinen, S. Aaltonen, M. Koskela, E. Nevasaari & M. Skog-Södersved (toim.) AFinLAe 3, 142–157. http://ojs.tsv.fi/index.php/afinla/issue/view/694
Reiman, N. 2011b. Two faces of complexity: structural measures and diversity of constructions. Nordand 6 (2), 9–23.
Reiman, N. 2014. Yläkoulun S2-oppilaiden transitiivi-ilmausten käyttö Eurooppalaisen viitekehyksen taitotasoilla. Lähivõrdlusi. Lähivertailuja 24, 183–220. https://doi.org/10.5128/LV24.07
Rimmer, W. 2006. Measuring grammatical complexity: the Gordian knot. Language Testing 23 (4), 497–519. https://doi.org/10.1191/0265532206lt339oa
Rimmer, W. 2009. Can what counts in complexity be counted? University of Reading: Language Studies Working Papers, 1, 25–34.
Scarborough, Hollis S. 1990. Index of productive syntax. Applied Psycholinguistics 11 (1), 1–22. https://doi.org/10.1017/S0142716400008262
Seilonen, M. 2013. Epäsuora henkilöön viittaaminen oppijansuomessa. Jyväskylä Studies in Humanities 197. Jyväskylä: Jyväskylän yliopisto.
Spoelman, M. & Verspoor, M. 2010. Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. Applied Linguistics 31 (4), 532–553. https://doi.org/10.1093/applin/amq001
Suni, M. 2006. Lasten toisen kielen kehitys IPSyn-kokeilun valossa. Teoksessa H. Tommola & A. Pajunen (toim.) XXXII Kielitieteen päivät Tampereella 19.–20.5.2005. Valikoima pidettyihin esitelmiin pohjautuvista artikkeleista. Tampere: Tampereen yliopisto, 427–439.

Taguchi, N., Crawford, W. & Wetzel, D. 2013. What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. TESOL Quarterly 47 (2), 420–430. https://doi.org/10.1002/tesq.91

Tilma, C. 2014. The dynamics of foreign versus second language development in Finnish writing. Jyväskylä studies in humanities 233. Jyväskylä: Jyväskylän yliopisto.

Toropainen, O., Härmälä, M. & Lahtinen, S. 2012. Kaksi asteikkoa, kaksi eri tilannetta: äidinkielellä ja vieraalla kielellä kirjoitettujen tekstien kriteeripohjaisen arvioinnin haasteita. AFinLA-e: Soveltavan kielitieteen tutkimuksia 4, 60–79. http://journal.fi/afinla/article/view/7038

Tähtinen, J., Laakkonen, E. & Broberg, M. 2020. Tilastollisen aineiston käsittelyn ja tulkinnan perusteita. Turun yliopiston kasvatustieteiden tiedekunnan julkaisuja C: 22. (2. uudistettu painos). Turku: Turun yliopiston kasvatustieteiden laitos.

van Rooy, B. 2015. Annotating learner corpora. Teoksessa S. Granger, G. Gilquin & F. Meunier (toim.) The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press, 79–106.

Vilkuna, M. 1996. Suomen lauseopin perusteet. Helsinki: Edita.

Vyatkina, N. 2012. The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. The Modern Language Journal 96 (4), 576–598. https://doi.org/10.1111/j.1540-4781.2012.01401.x

Vyatkina, N. 2013. Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. The Modern Language Journal 97 (S1), 11–30. https://doi.org/10.1111/j.1540-4781.2012.01421.x

Winter, B. 2020. Statistics for linguists: An introduction using R. New York: Routledge.

Wolfe-Quintero, K., Inagaki, S. & Kim, H. 1998. Second language development in writing: Measures of fluency, accuracy, and complexity. Technical report No. 1. Honolulu: Second Language Teaching and Curriculum Center.

ORIGINAL PAPERS

I

WORDS, CLAUSES, SENTENCES, AND T-UNITS IN LEARNER LANGUAGE: PRECISE AND OBJECTIVE UNITS OF MEASURE?

by

Mylläri, Taina 2020

Journal of the European Second Language Association 4 (1), 13–23.

https://doi.org/10.22599/jesla.63

JESLA | Journal of the European Second Language Association

Mylläri, T. (2020). Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association,* 4(1), 13–23. DOI: https://doi.org/10.22599/jesla.63

RESEARCH

Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure?

Taina Mylläri

In research on learner language complexity, accuracy and fluency (CAF), syntactic complexity is often studied with quantitative measures based on words, clauses, sentences, and T-units. The findings have been mixed, but segmenting learner language into these units of measure has seldom been problematised, even if the need for accurate coding is well known. The present study explores words, clauses, sentences, and T-units as production units in written learner language using a corpus of 352 L2 Finnish texts (28,813 words). The results illustrate how written learner language can be hard to fit into the production unit categories, which are essential for the most frequently used quantitative measures of syntactic complexity. On the one hand, the results support calls to include explicit definitions of the units of measure when reporting findings obtained with these quantitative measures. On the other hand, they align with calls to introduce new measures to better gauge the changes in learner language syntax as it develops with increasing language proficiency.

Keywords: Common reference levels; Complexity; Learner Finnish; Learner writing; Segmentation

1. Introduction

When second-language (L2) learning is analysed in terms of complexity, accuracy, and fluency, complexity is often quantified using measures that are based on the length of clauses, sentences, and T-units, or on the relation of these production units to each other (e.g., Bulté & Housen, 2012; Pallotti, 2015; Wolfe-Quintero et al., 1998). These measures require the consistent and reliable segmenting of learner language, but the possible effects of inconsistencies in coding learner language have seldom been discussed (e.g., Byrnes et al., 2010, p. 169).

Learner language does not always fit neatly into the categories used in these quantitative measures of complexity. Deviations from the target language norms are a challenge for annotation (e.g., Granger, 2002), and there can be several interpretations of the intended target form (e.g., Brunni et al., 2015; Ragheb & Dickinson, 2011; Rehbein et al., 2012). These challenges affect the segmenting of learner language into clauses, sentences, and T-units, especially on lower proficiency levels, when learner language can be fragmented and elliptic in both its oral (e.g., Foster et al., 2000) and written forms (e.g., Martin, 2013). The ambiguity of clause and sentence boundaries in written learner language is illustrated by Martin's (2013) segmenting experiment, in which a group of 35 university students of Finnish segmented three learner Finnish texts into clauses and sentences. The results showed variation in the numbers of both sentences and clauses, and even when two students arrived at the same number of clauses or sentences, the production units identified were not necessarily identical (Martin, 2013).

Differences in the numbers of production units are likely to lead to different results when complexity is measured using these units. Segmenting learner language into clauses, sentences, and T-units may also affect the quantitative measures that have typically been used to measure the syntactic complexity of written learner language, as among the most frequently used measures have been mean length of sentence, mean length of clause, mean length of T-unit, mean number of clauses per T-unit, mean number of T-units per sentence, and mean number of dependent clauses per clause (e.g., Ortega, 2003).

The present study seeks to explore how objective and reliable words, sentences, clauses, and T-units are as units of measure in written learner language. This is done by taking a close look at the segments that cause difficulties in splitting the data into these production units. The research question is: How do deviations from target language norms affect the segmenting of written learner language into words, sentences, clauses, and T-units? To answer this question, a corpus of written learner Finnish texts from different proficiency levels, from beginners to advanced, was segmented into these production units, and the segments not fitting into these categories were analysed. While the results are in part language specific, the problems are not limited to learner Finnish: Similar problems arise with other languages too.

Department of Language and Communication Studies, University of Jyväskylä, FI

taina.myllari@jyu.fi

2. Word, sentence, clause, and T-unit as production units

When words, clauses, sentences, and T-units are used as units of measure, they need to be identified in the data and their frequency of occurrence needs to be counted. These units can, however, be defined in more ways than one. In this section, words, clauses, sentences, and T-units are discussed in relation to their use in measuring syntactic complexity.

2.1. Word

One way to measure complexity is to calculate the mean length of a given production unit in words (e.g., Bulté & Housen, 2012). In many languages, a word can be defined as an orthographic unit separated from other text units by a blank space or by punctuation. While this simple definition is not suitable for all languages and it may overlook some linguistic features of words and differences between languages (e.g., Booij, 2012), it can in many cases be considered a reasonable way of defining a word in written language (Haspelmath, 2011, p. 69). It also makes automated word counts easy in languages in which words are separated by blank spaces.

This simple definition of a word seems reasonable within a study or within a language, but some language-specific conventions or orthographic rules, such as those concerning compound words, may cause differences in word count. When the number of words is based on orthography, elements in compound words are each counted as one word if they are separated from other elements by a blank space. This way of counting seems suitable for the present study, as compound words in Finnish normally consist of two or more words spelled as one orthographic unit (e.g., *ruokapöytä* for *ruoka+pöytä* 'food' + 'table') 'dining/dinner table'. It may, however, cause problems in languages with different orthographic conventions. Additionally, errors in orthography with compound words made by both L2 and first-language (L1) writers, such as *iso äiti* for *isoäiti* 'grandmother' or *jokapäivä* for *joka päivä* 'every day', may affect the word count.

Another possible source of differences in the length of a clause, sentence, or T-unit in words are differences in morphology. In morphologically rich languages, some syntactic information may be encoded within a single word, as illustrated in example (1). Such differences, and their impact on word count, should be taken into consideration if the length of a given syntactic unit in words is compared across languages.

(1)	talo-ssa=ni	luk-isi-t=ko
	house-INESS=POSS.1SG	read-COND.2SG=Q
	'in my house'	'would you read'

Some less-frequently occurring elements in written texts may also affect the word count. These include abbreviations pointing to multiple words (e.g, *jne* for *ja niin edelleen* 'and so on'), orthographic units containing hyphens or slashes, and word-like units containing or consisting of other characters than letters of the alphabet, such as expressions of quantity written with numbers (e.g., *1–2*), or amounts specified with a combination of a number and a unit of measurement (e.g., *12 tuntia* '12 hours'; *11 tuntia* '11 hours').[1]

2.2. Clause

Some of the most widely used measures of syntactic complexity involve counting the number of clauses per given unit (Pallotti, 2015) and mean length of clause in words (e.g., Ortega, 2003). Although grammars offer relatively clear definitions of a clause, in reality texts, both in L1 and L2, contain segments that do not fit these descriptions. Nevertheless, these segments should also somehow be acknowledged and included in analyses of complexity.

In studies on syntactic complexity in learner language, especially in learner English, a clause has typically been defined as a production unit containing either a subject and a finite verb or a subject and a finite or non-finite verb form (e.g., Lu, 2011, p. 44; Wolfe-Quintero et al., 1998, p. 70). When measuring syntactic complexity, infinitive forms in verb clusters can be considered to either belong to a verb construction within one clause or to form non-finite dependent clauses (e.g., Pallotti, 2015). In Finnish, structures with a non-finite verb form are typically considered verb phrases rather than clauses (Hakulinen et al., 2004, pp. 488–489; Vilkuna, 2003, pp. 14–15). Regarding the measures of complexity, coding verb clusters to belong to one clause or to more clauses has an impact on the mean length of clause, as well as on the number of clauses (Bulté & Housen, 2012). This decision also affects the number of dependent clauses and thus any ratios in which the number of dependent clauses is used.

In the above definitions of a clause, a subject is also considered a mandatory element. While this requirement suits non-null-subject languages, such as English, it is not practical for null-subject languages or partial null-subject languages, such as Finnish. In a quantitative study of Finnish syntax, Hakulinen et al. (1996) conclude that an overt subject cannot be considered a mandatory element of a clause in Finnish, because in their data, consisting of factual prose such as newspaper articles, more than 30% of the clauses did not have an overt subject (Hakulinen & Karlsson, 1980). There are several linguistic features contributing to this. In Finnish, it is possible to incorporate the first- and second-person subject in the verb form, leaving out the corresponding pronoun. Hence, for example, 'I say' can be expressed either with two words (*minä sanon*) or one word (*sanon*). There are also clause types that do not allow an overt subject. These types include all clauses in the passive voice (Hakulinen et al., 2004, p. 1245; Karlsson, 2015, p. 200) and some clauses containing meteorological expressions (e.g., *Satoi.* rain-PAST-3SG 'It was raining.') or causative verbs (e.g., *Minua pelottaa.* me-OBJ frighten-PRS-3SG 'I feel frightened.') (e.g., Karlsson, 2015, p. 81; for more detail, see Hakulinen et al., 2004, pp. 856–862, 1286). Such differences between languages need to be considered when defining a clause.

2.3. Sentence

In segmenting written language, the sentence can be considered "the obvious unit" (Ellis & Barkhuizen, 2005, p. 147). A sentence is usually defined as an orthographic unit beginning with a capital letter and ending with appropriate punctuation. These indicators of sentence boundaries are marked by the writer, but in some texts, the use of punctuation and capital letters may be inconsistent. These inconsistencies may be caused by problems in writing in the target language or by problems in writing in general.

The unsystematic use of punctuation can sometimes create sentences without a verb (as in example (2)) or an apparent independent clause (see example (3)). Considering this kind of punctuation intentional or erroneous affects the number of sentences and the kind of elements they consist of.

(2) Saa syödä purukumia tunnilla ellei se
can eat chewing.gum in.class unless.not it
häiritse. muita.
disturbs others.
'You/One can eat chewing gum in class unless it disturbs. others.' (F-010, adolescent A1)

(3) Oppilaat eivät sais ottaa kännyköitä kouluun
pupils not should take mobiles to.school
mukaan. Koska ne häiritse tunneilla.
along because they disturb in.classes
'Pupils should not take mobile phones to school. Because they disturb the class.' (F-733, adolescent B1)

Not all sentences without a verb or an independent clause result from errors in punctuation. For example, newspaper headlines, interactive elements such as greetings, and certain idiomatic expressions can be punctuated as sentences even when they do not contain a grammatically complete clause (e.g., Biber et al., 1999, pp. 224–225; Leech & Svartvik, 2002, p. 262). This also applies to Finnish. According to standard Finnish grammar, the minimal length of a sentence is one word, and this word does not need to be a verb (Hakulinen et al., 2004, p. 827).

There are also sentences that contain only clauses or structures that are traditionally not considered independent. For example, Foster et al. (2000) raise the question of the dependence or independence of adverbial clauses beginning with the conjunction *because* but lacking an apparent main clause. In written Finnish, sentences containing only clauses that begin with a subordinator can be found in both L1 and L2 writers' texts (Kalliokoski, 2006). In Finnish, there are also sentences that contain only infinitive verb forms (Visapää, 2008).

Sentences containing grammatically incomplete clauses or lacking an independent clause present a challenge to coding learner language and to the quantitative measures of complexity. Annotating these sentences to contain at least one clause or zero clauses affects all measures in which the number of clauses is used. Similarly, coding these sentences to contain at least one independent clause or only dependent clauses also affects measures relying on the number of dependent or independent clauses.

2.4. T-unit

The T-unit, first introduced by Hunt in 1965 in the L1 context, has gained ground in L2 research, but it has also been the target of some criticism (Bardovi-Harlig, 1992; Biber et al., 2011; Crossley & McNamara, 2014). There are several definitions of the T-unit. Most often it refers to one independent clause and any dependent clauses attached to it, although there has been variation in the inclusion or exclusion of fragments and in the counting of elements across sentence boundaries (e.g., Foster et al., 2000, pp. 360–363). In measuring syntactic complexity, the T-unit is among the most popular production units (Foster et al., 2000; Ortega, 2003; Wolfe-Quintero et al., 1998).

However, the relationship between clauses can sometimes be ambiguous, which makes it hard to determine whether a clause is coordinated or subordinated (Lieko, 1992, pp. 29–31; Quirk et al., 1972, pp. 795–796). Additionally, it is not always clear which independent clause is the main clause of a given dependent clause (as in example (4)), where it is not clear which of the independent clauses functions as the main clause for the clause beginning with *jos* 'if'.

(4) jos kotona on kiire, valmistan ruokaa, ja
if at.home is hurry I.make food and
huomasin että ei ole maitoa, menen
I.noticed that no is milk I.go
lähikauppaan.
to.corner.shop
'if it's busy at home, I cook, and I noticed that there is no milk, I go to the corner shop.' (F-253, adult A2)

Nevertheless, distinguishing between the two and identifying the dependency relationships are essential when using the T-unit as a unit of measure.

3. Design of the study

In the present study, a corpus of written learner Finnish and a comparative set of L1 Finnish adolescent writers' texts were split into words, sentences, clauses, and T-units to create a corpus for measuring syntactic complexity in learner Finnish with the frequently used quantitative measures. To find the production units, a set of definitions, described in Section 4, was used, and segments not fitting into these categories were examined. The focus was on problematic segments that could lead to different interpretations of the number of the relevant production units (i.e., words, sentences, independent clauses, and dependent clauses). The problematic segments were analysed qualitatively and quantified by counting their frequency. The aim was to identify the key challenges and evaluate their significance.

3.1. The data

The data in the present study comprise 352 learner Finnish (L2) texts (28,813 words) and 128 native Finnish (L1) texts (7,049 words) from the Cefling project corpus,[2] which contains texts elicited by means of communicative writing tasks. The Cefling corpus was collected for L2 research by selecting L2 Finnish adult learner texts from the National Certificates of Language Proficiency examination database and by collecting texts from adolescent L2 Finnish learners and L1 writers in school years 7 to 9 (age 12 to 16) with matching tasks (Martin et al., 2010). For the present study, the argumentative texts from the Cefling corpus were used.

To facilitate research into the development of different linguistic features in relation to language proficiency, all the L2 Finnish texts were assessed and placed according to the proficiency levels of the Common European Framework of Reference (CEFR, Council of Europe, 2001) by a team of trained raters in the Cefling project. Each text was rated by three raters using scales based on the CEFR (Alanen et al., 2010). The reliability of the ratings has been shown by both quantitative and qualitative analysis (for more detail, see Huhta et al., 2014). The adult learners' texts cover CEFR proficiency levels A1 to C2, and the adolescent learners' argumentative texts cover levels A1 to B1.

In the present study, segments that were copied word by word from the task prompts or contained only verbless greetings, pseudonyms, or contact information were considered echo responses and interactional elements, and they were not included in the analysis (cf. Foster et al., 2000). This led to the exclusion of 328 segments (961 words). The remaining text in the Cefling project Microsoft Word files was organised into a project corpus (**Table 1**).

To enable comparisons between language learners and native speakers, the L2 and L1 data were kept separate. To observe differences between learner age groups and between proficiency levels, the L2 data were separated into two groups, referred to in this study as adult learners and adolescent learners, and arranged according to the assessed proficiency level. Similarly, the L1 data were organised into three subgroups based on the school year of the participants.[3]

3.2. Analysis of the data

To answer the research question, the data were coded as words, sentences, clauses, and T-units. Segments not complying with the definitions and thus not fitting into these categories were analysed linguistically, and the frequency of such segments was calculated. On the sentence level, the focus was on irregularities in sentence marking which could affect the number of clauses, sentences, and T-units. On the clause level, the focus was on segments that could affect the number of clauses or their status as independent or dependent. If the problematic segments were not considered to affect the number of production units or the division of clauses into independent and dependent, they were outside the scope of this study.

Because there was only one annotator and a high number of problematic segments were found during coding, the sentence-level segmentation was compared with two other segmentations of the same data. The segmentation in the Cefling project CHAT files was one of those used. During the Cefling project, the texts were divided into sentences by seven native Finnish-speaking graduate students pursuing their Master's degree in Finnish language. If a sentence could not be clearly identified, the students were instructed to divide the text into clauses or, if the clause boundaries were also ambiguous, to group the text into clauses around the finite verbs (Cefling project, unpublished instructions). In the Cefling project, problematic segments were discussed but no inter-annotator agreement was counted or reported. The second segmentation used the open-source dependency parsing pipeline for Finnish developed by the University of Turku natural language processing (NLP) group.[4] The Finnish Dependency Parser is a statistical parser based on open-source NLP tools and trained on the Turku Dependency Treebank, whose system of annotation is a

Table 1: Amount and distribution of data across different writer groups.

CEFR level/ school year	Adult		Adolescent		Native		Total	
	texts	words	texts	words	texts	words	texts	words
A1	50	2,261	32	775	–	–	82	3,036
A2	37	2,272	39	1,589	–	–	76	3,861
B1	43	5,142	40	2,232	–	–	83	7,374
B2	35	4,166	–	–	–	–	35	4,166
C1	46	5,876	–	–	–	–	46	5,876
C2	30	3,879	–	–	–	–	30	3,879
Year 7	–	–	–	–	55	2,902	55	2,902
Year 8	–	–	–	–	50	2,831	50	2,831
Year 9	–	–	–	–	23	976	23	976
Total	241	23,596	111	4,596	128	6,709	480	34,901

Finnish-specific adaptation of the Stanford Dependency scheme (Haverinen et al., 2014).

To evaluate the reliability of the sentence-level segmentation, the three segmentations were compared using precision, recall, and F-score, which is the harmonic mean of the two. None of the segmentations was used as a gold standard annotation but instead, precision and recall were counted following Lu (2010) and Brants (2000) by dividing the number of segments identical in both the compared sets by the total number of sentences in the first set (precision) and in the second set (recall). In this kind of comparison setup, precision, recall, and F-score are considered to reflect agreement between annotations, the F-score being considered the most informative of the three (Brants, 2000; Lu, 2010).

4. Results

4.1. Words

In the present study, a word was defined as an orthographic unit containing alpha-numeric characters and separated from other units by a blank space, punctuation, or other orthographic marker, such as the beginning or the end of a line or a paragraph.

During the sentence-level comparisons, the orthography of each word in the two manual segmentations was checked and aligned to eliminate inconsistencies due to typing errors or differences in typing conventions between the file formats. Any discrepancies were resolved, when possible, based on the hand-written originals (adolescent learners and L1 writers) or the original database files (adult learners), and otherwise based on the transcription in the Word files. This resulted in identical word counts in the two manual segmentations.

In the automatically segmented data, there were four words more in the adult learner data and two words more in the L1 data than in the manual segmentations. The differences were caused by non-alphabetic characters within a word, such as quotation marks or a colon connecting a letter and a case ending. There were no differences in the word count in the adolescent learner data.

4.2. Sentences

A sentence was initially defined as an orthographic unit beginning with a capital letter and ending with a full stop, question mark, exclamation mark, or any combination of these. However, the requirement of initial capitalisation was discarded during segmenting because in some texts all the writing was originally in block capitals, or random block capitals were used within words. Consequently, segments such as those in example (2) were also coded to contain two sentences. The requirement of punctuation at the end of a sentence was also re-evaluated, and other orthographic markers, for example the organisation of text into items on a bulleted or numbered list, were considered to be indicators of sentence boundaries, as some texts were partly or completely organised as lists (as in example (5)).

(5) Minä olin syömässä ravintolassa Helsingissa, minä nähnyt 3 huonoa asiaa ja 1 hyvä asia
1/- ruokaa on hyvää.
2/- paljon ihmiset, ei riita paikkalla,
3/- He puhuvat kovaa
4/- ravintolassa tosi kuuma.
'I was eating at a restaurant in Helsinki, I seen 3 bad things and 1 good thing
1/- food is good.
2/- a lot of people, no quarrel at place,
3/- They speak loudly
4/- at the restaurant really hot.' (F-1012, adult A1)

In example (5), which is a short text from the lowest proficiency level, there is only one sentence indicated with both initial capitalisation and punctuation at the end. After careful consideration of such cases, the working definition of a sentence was changed, and the end of a whole text, a text paragraph, or a list item in a bulleted or numbered list were also defined as ending a sentence, regardless of the punctuation.

To evaluate the effect of the changes in the definition of a sentence, the sentence-level segmentation was compared to the original definition, and sentences not falling within the original definition were divided into two categories: Those ending with standard punctuation but not beginning with a capital letter, and those having no standard punctuation at the end (**Table 2**). The comparison showed that with proficiency level A1, only around half of the sentences conformed to the original definition of a sentence. Inconsistencies in punctuation were more frequent in the learner texts than in the L1 texts, where they were rare. These results should not, however, be interpreted as a straightforward relationship between the use of punctuation and L2 proficiency, as the inconsistent use of punctuation may have been caused by difficulties in writing in general, not necessarily difficulties in writing in a L2.

As for the actual number of sentences, there were only small differences in the numbers found in the different segmentations, and agreement between the segmentations was high, 90% to 99%, except in the adolescent learner data, where it was 85% and 88% on levels A1 and A2 in the comparison of the two manual segmentations (**Table 3**). The high agreement indicates that the sentences found were mainly identical.

The Cefling project segmentation contains the highest number of sentences in all the writer groups, which is in line with the instructions to split the text into clauses if the sentence boundaries were unclear. The parsed texts were found to contain the smallest number of sentences in all the writer groups. According to Haverinen et al. (2014), the parser makes its decisions based on dependencies and does not follow any separately given rules for sentence splitting.

These results seem to suggest that the working definition used in the present study could provide reliable enough criteria for identifying a sentence. It seems that

Table 2: Number and percentage of sentences with initial capitalisation and standard punctuation, sentences ending with standard punctuation but lacking initial capitalisation, and sentences not ending with standard punctuation.

Writer group	Initial capital and standard punctuation		Standard punctuation but no initial capital		No standard punctuation		Total sentences
	n	%	n	%	n	%	n
Adult A1	154	48	60	19	108	34	322
Adult A2	234	63	29	8	109	29	372
Adult B1	414	88	30	6	25	5	469
Adult B2	378	95	6	2	12	3	396
Adult C1	537	98	4	1	9	2	550
Adult C2	367	96	3	1	12	3	382
Adolescent A1	48	55	15	17	24	28	87
Adolescent A2	118	86	13	9	6	4	137
Adolescent B1	182	93	8	4	5	3	195
L1 year 7	268	93	8	3	11	4	287
L1 year 8	254	92	10	4	12	4	276
L1 year 9	104	94	2	2	5	5	111

Table 3: Number of sentences in each segmentation and the results of the sentence-level comparisons.

Writer group	Number of sentences in different segmentations			Number of identical sentences		Agreement between segmentations (F-score)	
	Present study	CHAT files	Parsed texts	Present study and CHAT files	Present study and parsed texts	Present study vs. CHAT files	Present study vs. parsed texts
Adult A1	322	337	308	308	296	0.93	0.94
Adult A2	372	375	371	364	368	0.97	0.99
Adult B1	469	488	450	452	430	0.94	0.94
Adult B2	396	405	385	387	371	0.97	0.95
Adult C1	550	554	545	546	530	0.99	0.97
Adult C2	382	388	379	377	367	0.98	0.96
Adolescent A1	87	97	84	78	81	0.85	0.95
Adolescent A2	137	145	131	124	123	0.88	0.92
Adolescent B1	195	201	192	185	183	0.93	0.95
L1 year 7	287	290	280	285	259	0.99	0.91
L1 year 8	276	277	268	273	244	0.99	0.90
L1 year 9	111	112	106	110	101	0.99	0.93

the absence of an initial capital letter can be ignored. Further, the end of a list item in a bulleted or numbered list, the end of a text paragraph and the end of the whole text could be considered indicators of a sentence ending, even if none of these markers are included in the standard definition of a sentence.

4.3. Clauses

A clause was defined as a segment within a sentence containing a finite verb and all its arguments and adjuncts. As Finnish is considered a partial null-subject language, a subject was not required. Following the definition in Hakulinen et al. (2004, pp. 827–828), a finite verb was deemed to be a mandatory element in a clause, and non-finite verbs were considered to be part of a verb phrase within a clause clustered around a finite verb, although in some studies (e.g., Hakulinen et al., 1996) or descriptions of Finnish grammar (e.g., Karlsson, 2015) also some structures clustered around non-finite verb forms have been considered clauses. As the texts were first split into sentences, and this segmenting was considered reasonably reliable, it was decided to look for clauses within sentences.

However, splitting the data into clauses proved to be problematic. In the first place, not all sentences contained a grammatical clause. In some sentences, especially with the lower proficiency levels, verbs could be completely missing or determining the presence or absence of finite verbs could require interpretation. Some of these verbless sentences were created by punctuation that seemed to split a grammatical clause into two sentences (as in example (2)). Others, especially among the higher proficiency levels, seemed to be stylistically motivated and to intentionally lack a finite verb (see example (6)). With some of these sentences, context was needed in order to choose between several interpretations (as in example (7)), in which the words *soitin* (musical_instrument.NOM or call.PAST.1SG) and *vasta* 'just' could have more than one meaning and could be labelled as more than one part of speech: The word *vasta* could also be a misspelled form of *vasta-a* (answer.PRS.3SG or answer.INF). Additionally, there were sentences containing only non-finite verb forms, such as infinitives (example (8)), participles, or a negation verb without the main verb.[5]

(6) Ensimmäinen työpäivä ja hetkessä se onkin ohi. Sitten viikko ja kuukausi.
'First day of work and suddenly it's over already. Then a week and a month.' (F-816, adult C1)

(7) Sitten sinä vasta puhelin
then you just/answer phone
soitin.
musical.instrument/I.called
'Then you answer the ringing phone.' (A possible interpretation) (F-249, adolescent A1)

(8) Kävel-lä luontossa, katso-a kauniita paikkoja,
walk-INF in.nature look-INF beautiful places
nautti-a meren- tai järven vettä.
enjoy-INF of.sea- or of.lake water
'To walk in nature, look at beautiful places, enjoy the sea or lake water.' (F-659, adult B1)

It was also problematic because in sentences with more than one finite verb, it was not always clear how many clauses the finite verbs should be divided into. As in example (9), there could be two finite verbs (i.e., *ei saa* 'may not' and *saavat* 'may'), but it was not clear if there were two clauses.

(9) ei saa lapset saa-vat ol-la kauan
not get[PRS.3SG] children get-PRS.3PL be-INF long
nettissä
on.the.web
'may not children may be on the internet for a long time.' (F-018, adolescent A1)

Thirdly, coordinators and subordinate conjunctions were sometimes used to connect segments that did not fall within the definition of a clause. As coordinators can be used to connect both clauses and phrases, segments without a finite verb could be interpreted as phrases coordinated with an element in the preceding clause. Another interpretation could be, as in example (10), that there are two coordinated clauses of which the latter is elliptic: The word *kielettyä* 'forbidden' could be interpreted as an adjective coordinated with *sallittua* 'allowed' in the preceding clause or as an elliptic clause *mutta [että kännykän pitely on] koulussa kielettyä* 'but [that holding a mobile is] at school forbidden'.

(10) toivomme että, kännykän pitely on sallittua,
we.hope that a.mobile holding is allowed
mutta koulussa kielettyä.
but at.school forbidden
'we hope that, holding a mobile is allowed, but at school forbidden.' (F-736, adolescent A2)

Regarding the use of subordinate conjunctions, this could create dependent clauses without a grammatical main clause (as in example (11)) or elements beginning with a subordinator but not containing a verb (see example (12)). We will return to this issue when exploring the T-units in the data.

(11) iso ongelma jos se tapahtuu talvella.
big problem if it happens in.winter
'a big problem if it happens in the winter.' (F-657, adult B1)

(12) Alaastella ei saa otta mukaan kouluun,
in.primary.school not get take with to.school
koska sellaiset säännöt.
because such rules
'In primary school, it is not allowed to bring to school because such rules.' (F-200, adolescent A1)

To evaluate the frequency and significance of these problems, the number of sentences without a finite verb was counted. These sentences were found on all proficiency levels, and also in the L1 texts (**Table 4**), although they were most common on the lower proficiency levels in the adult learner data. Other sentences considered problematic were counted after coding the T-units into the data.

Four possible solutions to these clause-level annotation problems were considered. The first of these was to include only sentences containing grammatical clauses. While this decision would solve the problems of clause-level coding of sentences with no finite verb, it would not solve the issues related to the number of clauses within those sentences in which there was a finite verb. It would also mean excluding one fifth of the sentences in the adult learners' texts on the two lowest proficiency levels. Secondly, consideration was given to the possibility of counting the number of clauses based on the number of finite verbs present in the texts (e.g., Verspoor et al., 2017). Although this would provide a solution to the problem of counting the number of clauses within the sentences containing at least one finite verb form, it would be affected by sentences not

Table 4: Number and percentage of sentences containing at least one finite verb, no verb, or at least one non-finite or ambiguous verb form.

Writer group	Finite verb		No verb		Other		Total sentences
	n	%	n	%	n	%	n
Adult A1	256	80	53	16	13	4	322
Adult A2	300	81	67	18	5	1	372
Adult B1	440	94	26	6	3	1	469
Adult B2	366	92	23	6	7	2	396
Adult C1	525	95	21	4	4	1	550
Adult C2	354	93	25	7	3	1	382
Adolescent A1	78	90	4	5	5	6	87
Adolescent A2	136	99	0	0	1	1	137
Adolescent B1	191	98	3	2	1	1	195
L1 year 7	264	92	19	7	4	1	287
L1 year 8	262	95	10	4	4	1	276
L1 year 9	105	95	6	5	0	0	11

containing any finite verbs. The third possible solution was to introduce a new production unit, similar to the sub-clausal element suggested by Foster et al. (2000) for analysing spoken language. While this solution would address issues related to labelling segments without a finite verb, it would introduce two new issues. On the one hand, it would mean that the exact boundaries of these units would become important if one wanted to measure their length or the clause length in words, because all words in these new units would need to be excluded from the word count of the clauses. On the other hand, it would create a need to introduce new measures in which these new units were included. Otherwise, it could entail excluding these new units and their content from the analysis. The fourth solution to the clause-level annotation problems was to also consider segments such as the grammatically incomplete clauses in examples (11) and (12) as attempted clauses and, therefore, to code them as clauses. While this solution would make it possible to include all the data in the analysis with the quantitative measures, it would create segments labelled clauses that do not fall within the original definition, in which a finite verb was required. We will return to this issue in Section 5.

4.4. T-units

A T-unit was defined as a production unit within a sentence consisting of one main clause and all the subordinate clauses connected to it directly or via another subordinate clause. In applying this definition to the data, problems similar to those in segmenting the data into clauses were encountered. First, the use of punctuation created segments in which there seemed to be a sentence boundary within a T-unit, as in example (3). Second, some dependent clauses had a grammatically incomplete clause as their main clause, as in example (11), and some segments beginning with a subordinator were not complete clauses, as in example (12).

Another type of sentence without an apparent main clause was also encountered. In the data, there were sentences that consisted of two clauses, one starting with a subordinator (e.g., *koska* 'because') and the other with a coordinating conjunction (e.g., *tai* 'or'), as in example (13). There were also sentences in which a clause starting with a subordinator seemed to be the main clause of the other clause or clauses in the sentences, as in example (14), in which the clause *Jos ajattelen* 'If I think' seems to be the main clause of two indirect questions rather than a subordinate clause of either of them. With this kind of sentence, analysis of the context is needed to determine the relationship between the clauses.

(13) Koska he eivät saisi olla kauan, tai he
because they not should be for.long or they
eivät saisi surffata nettissä.
not should surf in.net
'Because they should not be for long, or they should not surf the web.' (F-062, adolescent A2)

(14) Jos ajattelen missä Suomi geograafisesti
if I_think where Finland geographically
sijaitsee ja mitä luonnolla on meille
is.located and what nature has us
tarjottavana?
to.offer
'If I think where Finland is geographically located and what nature has to offer us?' (F-420, adult B2)

Sentences containing problems with either the number of clauses or their status as an independent or dependent clause were counted. These sentences were encountered throughout the data on all proficiency levels as well as in the L1 texts. Problematic sentences were more frequent in the adolescent learners' texts (between 22% on level A2

and 9% on B1) than in the adult learners' texts (between 13% on level A1 and 5% on C2), and the problems were not limited to the lower proficiency levels or to isolated texts. Rather, examples were spread across the data, and there was at least one problematic sentence in 40% or more of the L2 texts. There were fewer problematic sentences in the L1 data, but at least one such sentence could be found in 32% of the year 8 students' texts.

To resolve these issues, the use of the sentence as a superordinate unit was reconsidered, as some of the problems could have been solved by coding T-units across perceived sentence boundaries. This would, however, have led to treating some punctuation as erroneous, or ignoring it, which would be problematic, given that in writing, the boundaries of production units cannot be indicated by pauses or intonation, as they can in spoken language. Two other issues to be addressed were the coding of grammatically incomplete clauses or sub-clausal units, and their status as independent or dependent. These problems could have been solved by using an alternative production unit instead of the T-unit, namely the AS-unit, introduced by Foster et al. (2000) for analysing spoken language. While this solution would have acknowledged the sub-clausal units and their role in the superordinate units, it would also have disregarded the sentence boundaries the writer had marked with punctuation.

5. Discussion

When measuring learner language complexity with quantitative measures based on production units such as words, clauses, sentences, and T-units, it is important to split the data into these units reliably and consistently (e.g., Ellis & Barkhuizen, 2005; Pallotti, 2015). Nevertheless, as the results of this study show, learner language texts cannot always be divided into the aforementioned production units without making exceptions or leaving loose ends. In other words, as Rimmer (2006, p. 508) points out, authentic language does not always fit "into neat pigeon holes". It is therefore important to explicitly define the production units used and to make visible the exceptions allowed or the amount of data omitted. This information should always be included when reporting research findings.

In the present study, a sentence was defined as a segment indicated by the writer with punctuation or other orthographic means. As it was marked by the writer, a sentence was considered relevant also to the writer (cf. Peters, 1983). Therefore, it was selected as the superordinate unit (cf. Bardovi-Harlig, 1992; Ellis & Barkhuizen, 2005), and all the texts were first segmented into sentences, which were then split into clauses. In the clause-level annotation, clause boundaries and information on coordination and subordination, including information about the main clause of each dependent clause, were annotated where possible. Unclear cases were analysed and the number of sentences in which they occurred was counted. All of the words were annotated as belonging to a sentence and all sentences were annotated to contain a minimum of one independent clause (and thus also at least one T-unit), even when the sentence did not contain a finite verb or when it began with a subordinator. While these decisions led to segments not falling within the definition of the intended production units, they ensured that all the data were included in every annotation level and that they would be included in quantitative measures of syntactic complexity in future studies using this corpus.

These solutions leave room for criticism. They do, however, resonate with earlier findings of the difficulty of fitting learner language into these production unit categories (e.g., Foster et al., 2000; Rimmer, 2006), and they seem to suggest that reliance on production units that are not necessarily found in learner language could be one of the reasons behind inconsistencies in the results that have been obtained using these measures (e.g., Housen et al., 2019; Ortega, 2003; Wolfe-Quintero et al., 1998). In light of the results and the findings of other studies, three different solutions could be considered. One is forcing learner language into the categories used in quantitative measures, as was done in this study. Another is introducing new units of measure for quantitative research, as, for example, Foster et al. (2000) have done. A third solution is to analyse learner language from a more qualitative perspective and, for example, look for qualitative changes and development in selected linguistic features, as has been done by Reiman (2011) in a study on the development of transitive constructions in written learner Finnish.

There are a number of limitations to this study. The data were split into the production units by one person only. It was therefore impossible to negotiate problematic segments and calculate inter-coder agreement. Comparing the sentence-level results with two other segmentations revealed, however, only minor differences between segmentations in identifying words and sentences, which suggests that the sentence-level coding could be considered reliable enough. On the clause level, the problematic segments and their frequency of occurrence were based on the interpretations of one annotator; another annotator could have made different decisions and arrived at different results. While high inter-annotator agreement enhances the reliability of coding, having more annotators would not have eliminated the need to interpret parts of learner language, to adjust the definitions of production units used, or both.

The target language in this study was Finnish, a morphologically rich language, and it is possible that some of the ambiguities are language-specific. The data used in this study come from a heterogeneous group of learners with different proficiency levels. Some of the segmenting difficulties, such as those related to unsystematic use of punctuation, may also be related to the nature of the data. These issues, nonetheless, should be taken into account when making comparisons between studies within one language or studies on different target languages.

6. Conclusion

The level of detail in learner language coding and in reporting the process naturally depends on the aims and the research questions of each individual study. Nevertheless, segments that are problematic for coding in the data and their potential effect on result, should always

be acknowledged. This is essential for accumulating evidence on the development of complexity and for comparability across studies.

Segments which are problematic for coding could also be seen as potential sources of new information, and they could prove to be worth studying in more detail if a more qualitative approach to investigating complexity was adopted. Analysing the actual structures used by learners instead of forcing all learner language into predefined production unit categories could give new insights into the development of learner language and its complexity.

Notes

[1] The standard Finnish spelling is to separate the number and the unit.

[2] CEFLING = Linguistic basis of the Common European framework for L2 English and L2 Finnish (http://www.jyu.fi/hytk/fi/laitokset/kivi/tutkimus/hankkeet/paattyneet-tutkimushankkeet/cefling).

[3] For challenges in using the same rating scales for L1 and L2 texts, see, for example, Toropainen et al. (2012).

[4] It is available under an open licence at http://turkunlp.github.io/Finnish-dep-parser/. For this study, the branch 'master' updated May 9, 2016 was used.

[5] In Finnish, negation is expressed not with an invariable negation word but with a negation verb (e.g., Karlsson 2015: 82) that agrees with the subject in person and is followed by the main verb (e.g., Lue-n. read-PRS-1SG 'I am reading.', E-n lue. NEG-1SG read 'I am not reading.').

Acknowledgements

I would like to thank Maisa Martin, Jarmo Jantunen, the two anonymous reviewers and the editorial team for their valuable comments.

References

Alanen, R., Huhta, A., & **Tarnanen, M.** (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 21–56). EUROSLA Monographs series, 1. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2), 390–395. DOI: https://doi.org/10.2307/3587016

Biber, D., Gray, B., & **Poonpon, K.** (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35. DOI: https://doi.org/10.5054/tq.2011.244483

Biber, D., Johansson, S., Leech, G., Conrad, S., & **Finegan, E.** (1999). *Longman grammar of spoken and written English*. London: Longman.

Booij, G. (2012). *The grammar of words.* Oxford: Oxford University Press.

Brants, T. (2000). Inter-Annotator agreement for a German newspaper corpus. *LREC*. Retrieved from http://www.lrec-conf.org

Brunni, S., Lehto, L., Jantunen, J., & **Airaksinen, V.** (2015). How to annotate morphologically rich learner language. Principles, problems and solutions. *Bergen Language and Linguistics Studies, 6*, 133–152. DOI: https://doi.org/10.15845/bells.v6i0.812

Bulté, B., & **Housen, A.** (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam and Philadelphia: John Benjamins. DOI: https://doi.org/10.1075/lllt.32.02bul

Byrnes, H, Maxim, H., & **Norris, J.** (2010). Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment. *The Modern Language Journal, 94*(Supplement), 1–235. Retrieved from http://www.jstor.org/stable/40985261. DOI: https://doi.org/10.1111/j.1540-4781.2010.01139.x

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Retrieved from https://rm.coe.int/1680459f97

Crossley, S., & **McNamara, D.** (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26,* 66–79. DOI: https://doi.org/10.1016/j.jslw.2014.09.006

Ellis, R., & **Barkhuizen, G.** (2005). *Analysing learner language.* Oxford: Oxford University Press.

Foster, P., Tonkyn A., & **Wigglesworth G.** (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354–375. DOI: https://doi.org/10.1093/applin/21.3.354

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/lllt.6.04gra

Hakulinen, A., & Karlsson, F. (1980). Finnish syntax in text: Methodology and some results of a quantitative study. *Nordic Journal of Linguistics, 3*(2), 93–129. DOI: https://doi.org/10.1017/S0332586500000536

Hakulinen, A., Karlsson, F., & **Vilkuna, M.** (1996). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus* (2nd ed.). Helsinki: University of Helsinki.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., & **Alho, I.** (2004). *Iso suomen kielioppi.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica, 45*(1), 31–80. DOI: https://doi.org/10.1515/flin.2011.002

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., & **Ginter, F.** (2014). Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation, 48*(3), 493–531. DOI: https://doi.org/10.1007/s10579-013-9244-1

Housen, A., De Clercq, B., Kuiken, F., & **Vedder I.** (2019). Multiple approaches to complexity in second language research, *Second Language Research, 35*(1), 3–21. DOI: https://doi.org/10.1177/0267658318809765

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & **Hirvelä, T.** (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3), 307–328. DOI: https://doi.org/10.1177/0265532214526176

Kalliokoski, J. (2006). Virke, dialogisuus ja argumentaatio: irralliset sivulauseet ja toisella kielellä kirjoittaminen. In T. Nordlund, T. Onikki-Rantajääskö, T. Suutari, & H. Forsberg (Eds.), *Kohtauspaikkana kieli: näkökulmia persoonaan, muutoksiin ja valintoihin* (pp. 212–231). Helsinki: Suomalaisen Kirjallisuuden Seura.

Karlsson, F. (2015). *Finnish: an essential grammar.* London: Routledge.

Leech, G., & **Svartvik, J.** (2002). *A communicative grammar of English* (3rd ed). London: Longman.

Lieko, A. (1992). *The develoment of complex sentences. A case study of Finnish.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. DOI: https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36–62. DOI: https://doi.org/10.5054/tq.2011.240859

Martin, M. (2013). Sentences and clauses as complexity measures in second language writing: a segmentation experiment. In M. Järventausta, & M. Pantermöller (Eds.), *Finnische Sprache, Literatur und Kultur im deutschsprachigen Raum – Suomen kieli, kirjallisuus ja kulttuuri saksankielisellä alueella* (pp. 185–198). Greifswald: Veröffentlichungen der Societas Uralo-Altaica.

Martin, M., Mustonen, S., Reiman, N., & **Seilonen, M.** (2010). On becoming an independent user. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 57–80). EUROSLA Monographs series, 1. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518. DOI: https://doi.org/10.1093/applin/24.4.492

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117–134. DOI: https://doi.org/10.1177/0267658314536435

Peters, A. (1983). *The units of language acquisition.* Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., & **Svartvik, J.** (1972). *A grammar of contemporary English.* London: Longman.

Ragheb, M., & **Dickinson, M.** (2011). Avoiding the comparative fallacy in the annotation of learner corpora. In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. P. Botana, & E. Rhoades (Eds.), *Selected proceedings of the 2010 Second Language Research Forum: Reconsidering SLA research, dimensions, and directions* (pp. 114–124). Somerville, MA: Cascadilla Proceedings Project. Retrieved from http://www.lingref.com/cpp/slrf/2010/paper2620.pdf

Rehbein, I., Hirschmann, H., Lüdeling, A., & **Reznicek, M.** (2012). Better tags give better trees – or do they? *Linguistic Issues in Language Technology, 7*(10), 1–18. Retrieved from https://journals.linguisticsociety.org

Reiman, N. (2011). Two faces of complexity: Structural measures and diversity of constructions. *Nordand, 6*(2), 9–23.

Rimmer, W. (2006). Measuring grammatical complexity: The Gordian knot. *Language Testing, 23*(4), 497–519. DOI: https://doi.org/10.1191/0265532206lt339oa

Toropainen, O, Härmälä, M., & **Lahtinen, S.** (2012). Kaksi asteikkoa, kaksi eri tilannetta: äidinkielellä ja vieraalla kielellä kirjoitettujen tekstien kriteeripohjaisen arvioinnin haasteita. *AFinLA-e: Soveltavan kielitieteen tutkimuksia 4*, 60–79. Retrieved from http://journal.fi/afinla/article/view/7038

Verspoor, M., Lowie, W., Chan, H., & **Vahtrick, L.** (2017). Linguistic complexity in second language development: variability and variation at advanced stages. *Recherches en didactique des langues et des cultures, 14*(1), 1–27. DOI: https://doi.org/10.4000/rdlc.1450

Vilkuna, M. (2003). *Suomen lauseopin perusteet.* Helsinki: Edita.

Visapää, L. (2008). *Infinitiivi ja sen infinittiisyys: tutkimus suomen kielen itsenäisistä A-infinitiivikonstruktioista.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Wolfe-Quintero, K., Inagaki, S., & **Kim, H.** (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity. Technical report No. 1.* Honolulu: Second Language Teaching and Curriculum Center.

How to cite this article: Mylläri, T. (2020). Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association*, 4(1), 13–23. DOI: https://doi.org/10.22599/jesla.63

Submitted: 22 February 2020 **Accepted:** 20 July 2020 **Published:** 07 August 2020

WHITE ROSE UNIVERSITY PRESS Universities of Leeds, Sheffield & York

Journal of the European Second Language Association is a peer-reviewed open access journal published by White Rose University Press.

OPEN ACCESS

II

MEASURING SYNTACTIC COMPLEXITY IN LEARNER FINNISH

by

Mylläri, Taina 2020

Apples – Journal of Applied Language Studies 14 (2), 67–92.

https://doi.org/10.47862/apples.99134

Apples – Journal of Applied Language Studies
Vol. 14, 2, 2020, 67–92

Measuring syntactic complexity in learner Finnish

Taina Mylläri, University of Jyväskylä

In the study of complexity, accuracy and fluency (CAF), syntactic complexity can be measured by a multitude of measures. Traditionally, the measures are quantitative and they use production units such as words, clauses, T-units, and sentences. Despite the vast number of measures available, many studies have used only one or two of them, or parallel ones tapping the same component of complexity. The present study explores syntactic complexity using seven frequently used quantitative complexity measures to gauge different facets of complexity in written learner Finnish. The data of the study consist of texts written by adult and adolescent language learners, and they cover proficiency levels from beginner (A1) to advanced learner (C2) in the Common European Framework of Reference (CEFR). According to the results, changes in the measures are not linear from one proficiency level to the next. The results also show that while all the selected measures catch some statistically significant differences between proficiency levels in adult language learner texts, only four measures do so in adolescent language learner texts. The results also suggest that the measures are sensitive to task type.

Keywords: L2 writing, complexity, learner Finnish

1 Introduction

Learner language complexity can be defined as "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003, p. 492). Following Norris and Ortega (2009), complexity is most often considered a multi-faceted concept, and there is a multitude of quantitative measures of syntactic complexity available. Among the most popular are length-based measures, such as mean length of T-unit and mean length of clause, measures based on subordination, such as mean number of clauses per T-unit or mean number of dependent clauses per clause, and measures based on features considered sophisticated (Bulté & Housen, 2012; Norris & Ortega, 2009; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). Nevertheless, many studies use only one or two measures, or they use measures tapping the same aspect of complexity (Bulté & Housen, 2012, p. 34) or small datasets (Lu, 2011). This makes comparisons between studies difficult (e.g. Ellis & Barkhuizen, 2005; Pallotti, 2015).

According to a research synthesis by Ortega (2003), the most frequent measures of syntactic complexity have been mean length of clause (MLC), mean length of sentence (MLS), mean length of T-unit (MLTU), mean number of T-units per

Corresponding author's email: taina.myllari@jyu.fi
eISSN: 1457-9863
Publisher: University of Jyväskylä, Language Campus

https://apples.journal.fi
https://doi.org/10.47862/apples.99134

sentence (TU/S), mean number of clauses per T-unit (C/TU), and mean number of dependent clauses per clause (DC/C). Previous studies using these measures have yielded inconsistent results on the development of complexity across time or proficiency levels (e.g. Housen, De Clercq, Kuiken, & Vedder, 2019; Ortega, 2003; Wolfe-Quintero et al., 1998). Recent studies also indicate that there are differences in the development of complexity between languages (e.g. Gyllstad, Granfeldt, Bernardini, & Källkvist, 2014; Kuiken & Vedder, 2019).

In the present study, these measures are applied to written learner Finnish, and syntactic complexity is measured across the proficiency levels of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). The aim is on the one hand to test how traditional quantitative measures gauge syntactic complexity on different proficiency levels in a language that is structurally different from those more frequently studied, and on the other hand to examine if these measures could be used to indicate learner proficiency in Finnish. Seven quantitative measures, chosen on the basis of prior research on other languages, are used to tap different dimensions of syntactic complexity. The research questions are:

RQ1. How does syntactic complexity in written learner Finnish develop across CEFR proficiency levels when measured quantitatively?
RQ2. How well do the quantitative measures used in this study differentiate CEFR proficiency levels in written learner Finnish?

To answer the research questions, a pseudo-longitudinal learner Finnish corpus of texts written by second language (L2) learners with various backgrounds is analysed. Two different age groups of learners, i.e. adults and adolescents (12 to 16 years of age), are included to test if the measures yield the same results in both age groups. The adult L2 learners' texts cover all CEFR levels, from A1 to C2, and the adolescent L2 learners' texts cover levels A1 to B1/B2. The texts have been elicited with communicative tasks and divided into three types, according to task: informal messages, formal messages, and argumentative texts. These three types are analysed separately to see if there are differences in the results between the text types.

The theoretical background and measures used in the present study are introduced in Section 2. The data and methods are introduced in Section 3. The results for each measure are reported in Section 4. The article ends with discussion and conclusions in Section 5.

2 Measuring syntactic complexity

Previous studies have provided inconsistent results on the development of complexity (e.g. Housen et al., 2019). This may depend on the multitude of ways complexity has been operationalised and measured in the studies (Housen et al., 2019; Housen, Kuiken, & Vedder, 2012; Norris & Ortega, 2009; Pallotti, 2015), but differences in the settings of the studies (e.g. Ortega, 2003), individual variation in learner language (e.g. Larsen-Freeman, 2006), the non-linear development of language as a complex system (e.g. Larsen-Freeman, 2009), or complexity being manifested in different ways at different developmental stages (Housen et al., 2019; Norris & Ortega, 2009) may also partly explain the differences. Some of the inconsistencies may also be caused by typological differences between languages (Housen et al. 2019; Bernardini & Granfeldt, 2019).

One possible source of the mixed results are also the challenges involved in coding learner language. The most frequently used quantitative measures of syntactic complexity rely on production units such as clauses, sentences, and T-units, but these units can be ambiguous in both spoken and written learner language (e.g. Foster, Tonkyn, & Wiggelsworth, 2000; Martin, 2013). Differences in the definitions used in segmenting the data (e.g. Bulté & Housen, 2012; Wolfe-Quintero et al., 1998) or different interpretations made by annotators (e.g. Lu, 2010; Martin, 2013) may lead to differences in the number or length of the production units used. Such differences can affect the quantitative measures. For example, counting segments containing a non-finite verb form as either a dependent clause or as a part of a verb structure within a clause is likely to affect the number of words per clause (e.g. Bulté & Housen, 2012, pp. 39–40), with the number of dependent clauses and the total number of clauses then also affecting all the measures using these production units as units of measure (see also Mylläri, 2020).

Norris and Ortega (2009) suggest that syntactic complexity should be studied as a multi-faceted construct, and that different dimensions of complexity–overall complexity, complexity via subordination, subclausal complexity and, especially on lower proficiency levels, complexity via coordination–should be taken into account when measuring it. In addition to measures based on length or ratios of production units, complexity has been measured using the frequency of linguistic features that are considered sophisticated (e.g. Wolfe-Quintero et al., 1998).

Overall syntactic complexity, or general syntactic complexity, is typically measured by calculating mean length of sentence or T-unit in words (Bulté & Housen, 2012; Norris & Ortega, 2009). The results of studies using length-based measures have been mixed. In the research synthesis of Wolfe-Quintero et al. (1998), where sentence length (W/S) and T-unit length (W/T) were considered measures of fluency, they were both found to grow linearly with proficiency. According to Wolfe-Quintero et al. (1998), sentence length was found to correlate with proficiency in all the 10 studies, and T-unit length in 28 of the 40 studies included in the synthesis. Bulté and Housen (2012), however, point out that while length-based measures may show linear increase at some phase of development, they may well plateau at some level, in much the same way as L1 development of mean length of utterance has been shown to do.

Measures based on subordination have been among the measures most widely used in studies on syntactic complexity (e.g. Bulté & Housen, 2012). According to Wolfe-Quintero et al. (1998), clauses per T-unit (C/TU) and dependent clauses per clause (DC/C) are good indicators of proficiency as they seem to grow linearly with proficiency, although only some of the studies in their research synthesis found a correlation between the measures and proficiency. The relevance of subordination in measuring syntactic complexity especially in writing has later been questioned. Biber, Gray, and Poonpon (2011) argue that subordination is more typical of spoken language than of academic texts. Bulté and Housen (2012) note that subordination ratios only gauge one type of complexity and ignore others, such as clausal coordination or complexity at the phrasal level. Martin, Mustonen, Reiman, and Seilonen (2010) have also questioned using subordination as an indicator of learner Finnish complexity since, in Finnish, there are no apparent syntactic or morphological differences between subordinate clauses and main clauses, with the exception of relative clauses.

Measures based on subordination also overlook coordination as a part of complexity (e.g. Bardovi-Harlig, 1992). This could partly be explained by coordination being often associated with lower proficiency levels, as development is generally thought to proceed from coordination at beginning levels to subordination at intermediate levels and phrasal-level elaboration at advanced levels (e.g. Norris & Ortega, 2009). While Wolfe-Quintero et al. (1998) conclude that the coordination ratio of T-units per sentence (TU/S) has not been shown to be useful in L2 research, Norris and Ortega (2009) suggest that measures of coordination should also be included, especially in studies using data on lower proficiency levels.

A measure taking both coordination and subordination into account is number of clauses per sentence (C/S). Wolfe-Quintero et al. (1998) found that this measure had been used in only one study, and in that case the growth in the measure had been statistically significant for only a part of the study. Nevertheless, Lu (2010, 2011) has included C/S in the 14 measures in his L2 Syntactic Complexity Analyzer and labelled it a measure of sentence complexity.

Although mean length of clause (MLC) in terms of number of words can be considered a length-based measure, it is more often regarded as a measure of clausal or phrasal complexity than overall complexity, for the reason that it shows lengthening within a clause, thus indicating elaboration on the phrasal level (e.g. Norris & Ortega, 2009). Previous studies have shown MLC to correlate with proficiency: learners on the higher proficiency levels tend to use longer clauses than those on lower proficiency levels (e.g. Lu, 2011; Ortega, 2003). Wolfe-Quintero et al. (1998) considered clause length (W/C) to be a measure of fluency, and they found that it grew linearly with proficiency in all nine studies included in the synthesis, although the correlation was not statistically significant in all the studies. A similar measure, i.e. mean number of finite verbs per total number of words, has also been found to develop linearly over time (Verspoor, Lowie, Chan, & Vahtrick, 2017). Mean number of finite verbs per total number of words is the same as mean length of clause provided that each clause in the data contains a finite verb and all the words in the word count belong to a clause.

Syntactic complexity in L2 Finnish has so far been studied using small datasets or by analysing the development of some specific structures. Alisaari (2016), who used mean length of T-unit (MLTU) as a measure of fluency in written learner Finnish on CEFR level A2, found no significant development in MLTU in narrative texts written by 32 learners at the beginning and at the end of a four-week course. Tilma (2014) studied the development of complexity and accuracy in foreign and second language Finnish using written data collected from eight students over a period of nine months. Among the measures used in her study were average length of sentence and average length of clause in morphemes, both of which she found increased over time, although she found no statistically significant correlation between the indices and development on the group level. She concluded that the best measure of syntactic complexity in learner Finnish was average length of clause in morphemes. Spoelman and Verspoor (2010) studied learner Finnish complexity and accuracy in a DST framework. Using a longitudinal data set of 54 writing samples collected from one L2 Finnish learner over a period of three years, they found a non-linear increase in complexity, including the sentence complexity ratio, which was based on the average number of dependent clauses per text.

The development of linguistic features in relation to an increase in proficiency from one CEFR level to the next in written learner Finnish has been studied using the CEFLING project data, from which the data of the current study are also drawn. Seilonen (2013) studied the use of indirect references, Kajander (2013) the use of existential sentences, and Reiman (2011a, 2011b, 2014) transitive constructions. All of these linguistic features showed growth in frequency, variation and accuracy across proficiency levels. There were, however, differences in the use of these linguistic features between the adult and adolescent language learners and between task types. The results indicate that there is a leap in frequency between levels A2 and B1 in the adult learner data, whereas in the adolescent learner data a similar difference can already be found between levels A1 and A2. (Kajander, 2013; Reiman, 2014; Seilonen, 2013.)

To test the usability of the frequently used quantitative measures in written learner Finnish, the following seven measures were selected to tap the different dimensions of syntactic complexity (Table 1). To tap overall or general complexity, mean length of sentence (MLS) and mean length of T-unit (MLTU) were used. Following Lu (2010, 2011, 2017), mean number of clauses per sentence (C/S) was also calculated as a measure of overall sentence complexity. For complexity via subordination, two measures, mean number of clauses per T-unit (C/TU) and mean number of dependent clauses per clause (DC/C), were used. Complexity via coordination was measured with the mean number of T-units per sentence (TU/S). Sub-clausal complexity was measured with mean length of clause (MLC).

Table 1. Complexity measures used in the present study.

Label	Measure	Formula
MLS	Mean length of sentence	Total number of words / total number of sentences
MLTU	Mean length of T-unit	Total number of words / total number of T-units
MLC	Mean length of clause	Total number of words / total number of clauses
TU/S	Mean number of T-units per sentence	Total number of independent clauses / total number of sentences
C/S	Mean number of clauses per sentence	Total number of clauses / total number of sentences
C/TU	Mean number of clauses per T-unit	Total number of clauses / total number of independent clauses
DC/C	Mean number of dependent clauses per clause	Total number of dependent clauses / total number of clauses

Six of the measures, i.e. MLS, MLTU, MLC, TU/S, C/TU, and DC/C, are among the most frequently used in research on L2 complexity, according to Ortega (2003).

3 Data and methods

3.1 Data

The data used in the present study are drawn from the pseudo-longitudinal corpus of the Jyväskylä University CEFLING project and they comprise 667 texts (48,876 tokens) from adult L2 Finnish learners and 411 texts (16,590 tokens) from adolescent L2 learners. In the CEFLING project, the adult L2 learners' texts were selected from the National Certificates of Language Proficiency examination database, and the adolescent L2 learners' texts were collected from pupils in school years 7, 8 and 9 (between 12 and 16 years of age), together with similar texts from their native Finnish (L1) counterparts. The texts were elicited through communicative writing tasks, and they have been arranged into groups according to the type of task: informal messages (e.g. an email to a friend), formal messages (e.g. a complaint to an online store), and argumentative texts (e.g. a text expressing an opinion on a given topic, such as the use of mobile phones at school). The adolescent L2 learners and L1 writers also wrote a narrative text. The participants had a limited time in which to complete the writing tasks, and the use of aids, such as dictionaries, was not allowed. (Alanen, Huhta, & Tarnanen, 2010; Jantunen & Pirkola, 2015; Martin et al., 2010.) The L2 messages and argumentative texts are used in the present study.

In the CEFLING project, the L2 texts were assessed on the proficiency levels of the Common European Framework of Reference (CEFR), using scales based on the framework. The assessment was done by a team of trained raters, and each text was rated by three raters. (Alanen et al., 2010.) The reliability of the assessment has been shown by quantitative and qualitative analysis (Huhta, Alanen, Tarnanen, Martin, & Hirvelä, 2014). There are adult learner texts on all the CEFR proficiency levels, from A1 to C2. For the adolescent learners, the proficiency levels range from A1 to B1 in the argumentative texts and to B2 in the informal and formal messages. However, there are only a few adolescent L2 learner texts at level B2.

When annotating the data for the present study, echoes of task prompts and segments consisting of verbless greetings or contact information, such as email or street addresses and phone numbers, were excluded from the analysis (cf. Foster et al., 2000, pp. 370–371). Additionally, four whole texts were excluded during annotation: two adult L2 learners' texts containing only verbless greetings or list items, and two adolescent L2 learners' texts with inconsistencies in task type or writer identification information.

For the analysis, the L2 texts were organised according to the CEFR level. The two age groups of L2 learners, adult and adolescent learners, were studied separately, because earlier studies (Kajander, 2013; Reiman, 2014; Seilonen, 2013) using the same data have shown developmental differences between adult and adolescent learners. To control for task or genre effect (see e.g. Michel, 2017), the three task types were kept separate. The number of texts and words in the data are presented in Table 2.

Table 2. The number of texts and words and the average length of texts in words.

	Informal messages			Formal messages			Argumentative texts		
	Texts	**Words**	**Average length**	**Texts**	**Words**	**Average length**	**Texts**	**Words**	**Average length**
L2 adult learners									
A1	39	1,582	40.56	22	752	34.18	50	2,261	45.22
A2	39	1,533	39.31	27	1,494	55.33	37	2,272	61.41
B1	41	2,453	59.83	42	2,290	54.52	43	5,142	119.58
B2	39	2,206	56.56	34	1,983	58.32	35	4,166	119.03
C1	26	1,528	58.77	45	3,337	74.16	46	5,876	127.74
C2	14	970	69.29	58	5,152	88.83	30	3,879	129.30
Total	**198**	**10,272**	**51.88**	**228**	**15,008**	**65.82**	**241**	**23,596**	**97.91**
L2 adolescent learners									
A1	25	677	27.08	33	860	26.06	32	775	24.22
A2	79	2,879	36.44	40	1,489	37.23	39	1,589	40.74
B1	64	2,971	46.42	40	1,980	49.50	40	2,232	55.80
B2	12	677	56.42	7	461	65.86	-	-	-
Total	**180**	**7,204**	**40.02**	**120**	**4,790**	**39.92**	**111**	**4,596**	**41.41**

The data come from 868 different writers. Of the 481 adult learners, 40 wrote three texts each, 106 wrote two texts each, and 335 one text each. Of the 212 adolescent learners, 45 wrote three texts each, 109 two texts each, and 58 one text each. In the CEFLING project, each text was placed on a proficiency level independently (e.g. Martin et al., 2010).

3.2 Production units and segmenting the data

For the argumentative texts, a manually segmented corpus from an earlier study (Mylläri, 2020) was used. The informal and formal messages were first split into sentences, and each sentence was then split into clauses using the clause-splitting feature of the Finnish Dependency Parser (Haverinen et al., 2014)[1]. The segmentation was manually checked by the author to ensure that it was in line with the guidelines described below, and exceptions were considered case by case (see also Mylläri, 2020).

Words and sentences were segmented based on orthography. A segment was counted as a word if it contained at least one letter or number, or a symbol such as the euro sign (€), and if it was separated from other text by a space or other orthographic indicator, such as punctuation. This definition of word was considered to be feasible for this study, as in Finnish there are no articles and only a few prepositions, and compound words are generally written as one orthographic unit (e.g. *olohuone* 'a/the sitting room, *olohuoneessa* 'in a/the sitting room').

A sentence was defined as an orthographic unit ending with proper punctuation or, in the absence of punctuation, with an end-of-line character. Each sentence was annotated to contain at least one independent clause. This was applied also to sentences containing only one clause starting with a subordinator (see also Foster et al., pp. 2000: 336; Kalliokoski, 2006) and to sentences not containing a finite verb.

A clause was defined as a segment within a sentence clustered around a finite verb (cf. Lu, 2010; Wolfe-Quintero et al, 1998, p. 123). Clauses beginning with a subordinator or the surface-level ellipsis of a subordinator and having a main clause within the same sentence were annotated as dependent clauses. All other clauses were labelled as independent. Two clauses concatenated without any connectors were assumed to be coordinated with each other. Three types of exception were allowed in order to include all the words in the analysis and to maintain possibly intended subordination, even if there was no finite verb in the main clause or in the subordinate clause. First, as mentioned above, sentences not containing a finite verb were considered to contain at least one clause. Second, segments not containing a finite verb but functioning as a main clause to at least one subordinate clause within the same sentence, and not being subordinated to or coordinated with another clause, were considered independent clauses. Third, when there was a main clause within the same sentence, segments beginning with a subordinator were counted as subordinate clauses even if they did not contain a finite verb.

A T-unit was defined as consisting of one independent clause together with all the dependent clauses that are either directly or indirectly connected to it within the same sentence.

3.3 Statistical methods

All seven measures were calculated for each text. The results were rounded to two decimal places. Group mean and median, standard deviation, and interquartile range were calculated for each proficiency level for all the task types (informal message, formal message, argumentative text) separately for the adult and adolescent learners. All the texts in the data were included in the analysis, and outliers were considered to be occurrences of individual variation and therefore included in the calculations and statistical analyses. The descriptive statistics are presented in Tables A1–A7 in Appendix A.

Before making the statistical comparisons between proficiency levels, the data were visualised using boxplots and Q-Q plots. Since sample sizes varied, two tests were used to test the normality of distribution. The Shapiro-Wilk test was used for samples of 50 texts or fewer and the Kolmogorov-Smirnov test for samples larger than 50 texts. Both tests indicated violations of normality (68% of the samples of 50 texts or fewer and 24% of the samples over 50 texts). Homogeneity of variance was therefore tested with the Fligner-Killeen test (e.g. Gries, 2013, p. 229), which showed that the assumption was violated in around 31% of the comparisons.

Because of the differences in sample size and violations of assumptions of normality of distribution and homogeneity of variance, both parametric and non-parametric tests were used to test for differences between proficiency levels. For omnibus tests, one-way ANOVA and the Kruskal-Wallis test were used. For both tests, the cut-off point for statistical significance was set at $p < .05$. For effect size, adjusted R-squared in ANOVA was used with the following guidelines: > .01 small, > .06 medium, and > .14 large (see e.g. Larson-Hall 2010: 119). For complexity measures with statistically significant differences in ANOVA or the Kruskal-Wallis test, t-test and the Wilcoxon rank sum test (also known as the Man Whitney U-test) were conducted as pairwise tests of independent samples with Bonferroni corrections.

All statistical tests were done with R version 3.4.4 using RStudio version 1.1.456.

4 Results

Detailed results for each measure are presented below. Because of the nature of the data, both means and medians are used to describe the pseudo-longitudinal development trends. To address RQ1, results across proficiency levels are illustrated by boxplots where the group means are also shown. The two learner groups are presented together for each type of task to provide a visual comparison of the differences and similarities between the writer groups and the text types. The numeric values for means, standard deviations, medians, and interquartile ranges for each measure can be found in the tables in Appendix A. To address RQ2, the results of statistical comparisons are summarised in the text, and the post-hoc test results are visualised by tables indicating statistically significant differences between proficiency levels. Section 4.8 provides a summary of the results.

4.1 Mean length of sentence (MLS)

The average length of sentence measured with MLS grows across the proficiency levels, but the increase in length is continuous from the lowest level to the highest only in the formal messages when measured with group means, and in the adolescent learners' argumentative texts when measured with group medians (Figure 1).

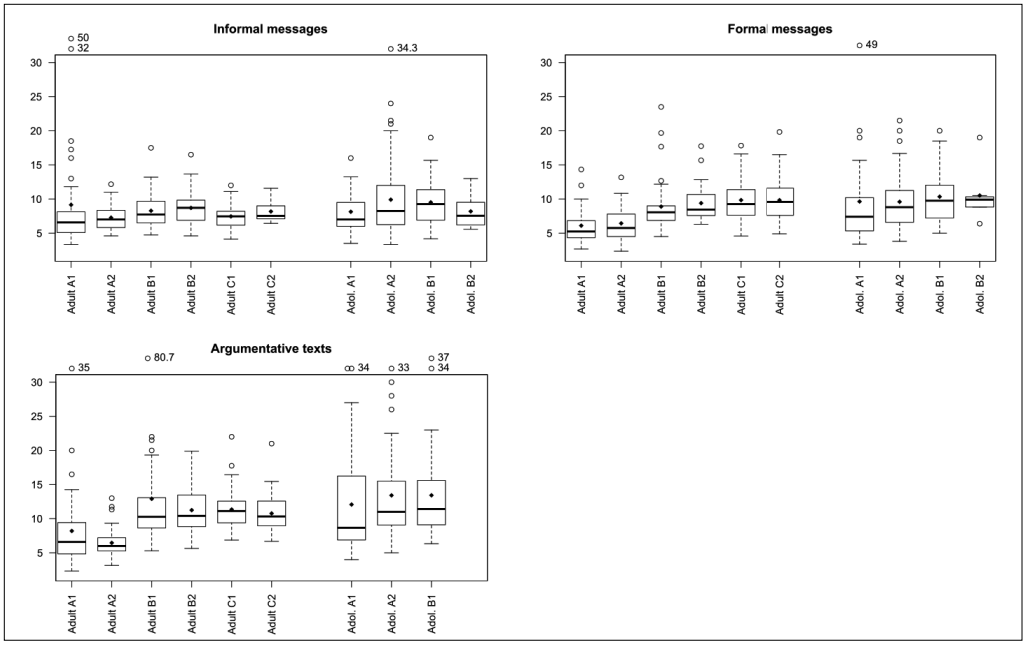


Figure 1. MLS in informal messages, formal messages, and argumentative texts.

According to both one-way ANOVA and the Kruskal-Wallis test, there are statistically significant differences in MLS in the adult learners' formal messages ($F(5,222) = 9.93$, $p < .001$, $R^2_{Adj} = .16$; $Chi\ squared = 53.53$, $p = < .001$, $df = 5$) and argumentative texts ($F(5,235) = 6.75$, $p < .001$, $R^2_{Adj} = .11$; $Chi\ squared = 75.96$, $p = < .001$, $df = 5$). According to the Kruskal-Wallis test, there are statistically

significant differences also in adult learners' informal messages (*Chi squared = 12.42, p = .029, df = 5*). The differences between the proficiency levels are not statistically significant in the adolescent learners' data according to either test.

In the adult learners' informal messages, there are no statistically significant differences in the post hoc tests even if the Kruskal-Wallis test result indicates between-group differences in MLS. Table 3 presents the levels between which there is a statistically significant difference according to the post hoc tests.

Table 3. The statistically significant between-level differences in MLS according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1							✓	✓	✓	✓		✓	✓	✓	✓
Adult A2							✓	✓	✓	✓		✓	✓	✓	✓
Adult B1															
Adult B2															
Adult C1															

4.2 Mean length of T-unit (MLTU)

A growing trend of mean length of T-unit (MLTU) is found in the formal messages and in the argumentative texts, where the growth is continuous across all the proficiency levels according to both group means and medians in the formal messages of the adolescent learners, according to group means in the L2 adult learners' formal messages, and according to group medians in the L2 adolescent learners' argumentative texts (Figure 2).

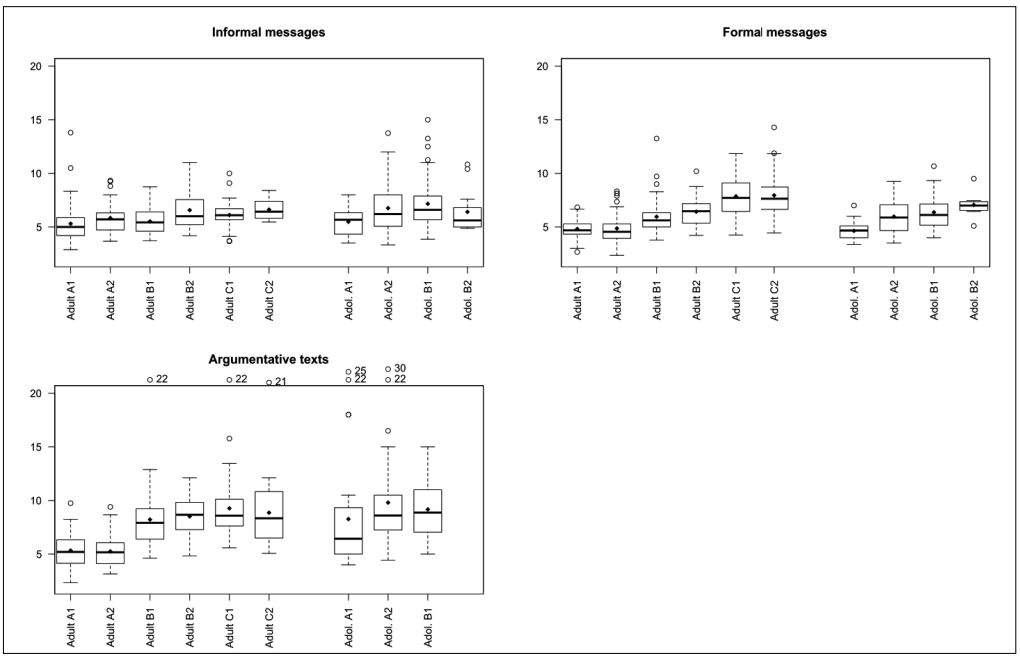


Figure 2. MLTU in informal messages, formal messages, and argumentative texts.

There are statistically significant differences between proficiency levels in MLTU according to both ANOVA and the Kruskal-Wallis test in all the task types in the adult learner data: informal messages ($F(5,192) = 3.73$, $p = .003$, $R^2_{Adj} = .06$; *Chi squared* $= 25.41$, $p < .001$, $df = 5$), formal messages ($F(5,222) = 23.95$, $p < .001$, $R^2_{Adj} = .34$; *Chi squared* $= 90.96$, $p < .001$, $df = 5$), and argumentative texts ($F(5,235) = 23.91$, $p < .001$, $R^2_{Adj} = .32$; *Chi squared* $= 108.53$, $p < .001$, $df = 5$). In the adolescent learner data there are statistically significant between-level differences according to both tests in the informal ($F(3,176) = 3.81$, $p < .001$, $R^2_{Adj} = .04$; *Chi squared* $= 12.48$, $p = .006$, $df = 3$) and formal ($F(3,116) = 23.29$, $p < .001$, $R^2_{Adj} = .23$; *Chi squared* $= 33.65$, $p < .001$, $df = 3$) messages. There are also statistically significant differences between proficiency levels in the adolescent learners' argumentative texts according to the Kruskal-Wallis test result (*Chi squared* $= 10.26$, $p = .006$, $df = 2$). The between-level differences which are statistically significant according to the post hoc tests are shown in Table 4.

Table 4. The statistically significant between-level differences in MLTU according to parametric ($✓_t$), non-parametric ($✓_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1			$✓_U$		✓		✓	✓	✓	✓		✓	✓	✓	✓
Adult A2							$✓_U$	✓	✓	✓		✓	✓	✓	✓
Adult B1			$✓_t$		✓				✓	✓					
Adult B2									✓	✓					
Adult C1															
Adol. A1	$✓_t$	✓				✓	✓	✓			$✓_U$	$✓_U$			
Adol. A2															
Adol. B1															

4.3 Mean length of clause (MLC)

There is growth in the average length of clauses measured with MLC in all three task types (Figure 3). In the informal messages, group means and medians both become higher from the lowest to the highest proficiency level in the adult learners' texts. In the formal messages, the same pattern can be found in the adult learner data (group means) and in the adolescent learner data (group means and medians). In the argumentative texts, both the group means and the medians grow in the adolescent learner texts.

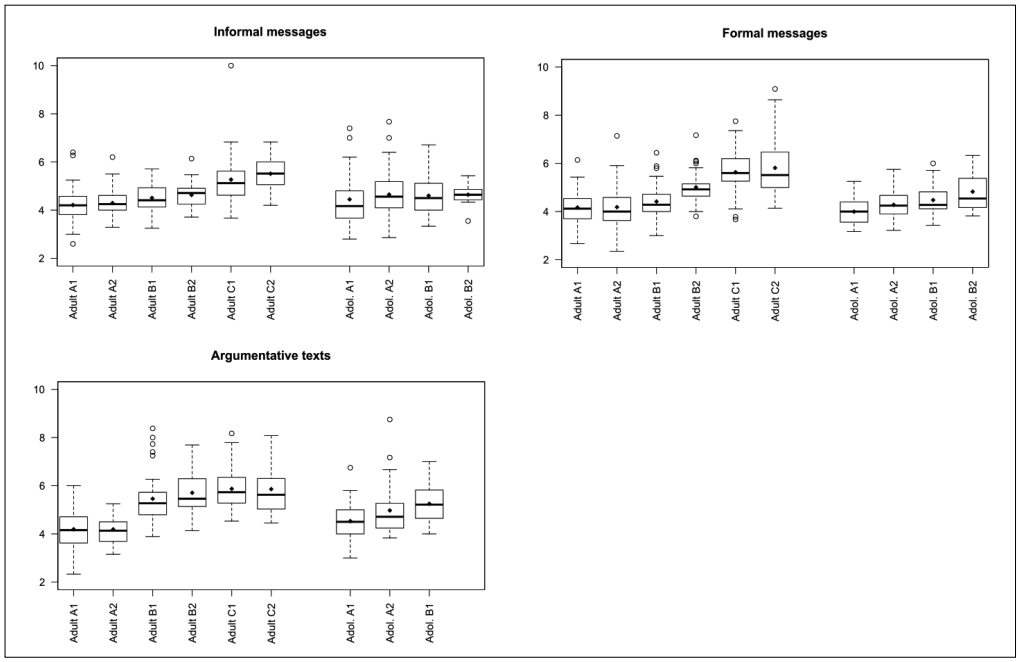


Figure 3. MLC in informal messages, formal messages, and argumentative texts.

There are statistically significant differences according to both ANOVA and the Kruskal-Wallis test between proficiency levels in MLC in the L2 adult learner informal messages ($F(5,192) = 11.95$, $p < .001$, $R^2_{Adj} = .22$; *Chi squared* $= 46.95$, $p < .001$, $df = 5$), formal messages ($F(5,222) = 26.52$, $p < .001$, $R^2_{Adj} = .36$; *Chi squared* $= 95.97$, $p < .001$, $df = 5$), and argumentative texts ($F(5,235) = 35.48$, $p < .001$, $R^2_{Adj} = .42$; *Chi squared* $= 119.07$, $p < .001$, $df = 5$), and also in the L2 adolescent learner formal messages ($F(3,116) = 6.11$, $p < .001$, $R^2_{Adj} = .11$; *Chi squared* $= 14.66$, $p = .002$, $df = 3$) and argumentative texts ($F(2,108) = 6.24$, $p = .003$, $R^2_{Adj} = .09$; *Chi squared* $= 13.22$, $p = .001$, $df = 2$). The post hoc test results are summarised in Table 5.

Table 5. The statistically significant between-level differences in MLC according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1			✓$_U$	✓	✓			✓	✓	✓		✓	✓	✓	✓
Adult A2				✓	✓			✓	✓	✓		✓	✓	✓	✓
Adult B1				✓$_U$	✓			✓	✓	✓					
Adult B2					✓				✓	✓					
Adult C1															
Adol. A1							✓					✓			
Adol. A2															
Adol. B1															

4.4 Mean number of T-units per sentence (TU/S)

The patterns of change are mixed for TU/S (Figure 4). The group means and medians grow smaller after the lowest proficiency level, but the change continues from the lowest proficiency level to the highest only in the L2 adolescent learners' informal messages (group means) and formal messages (group medians).

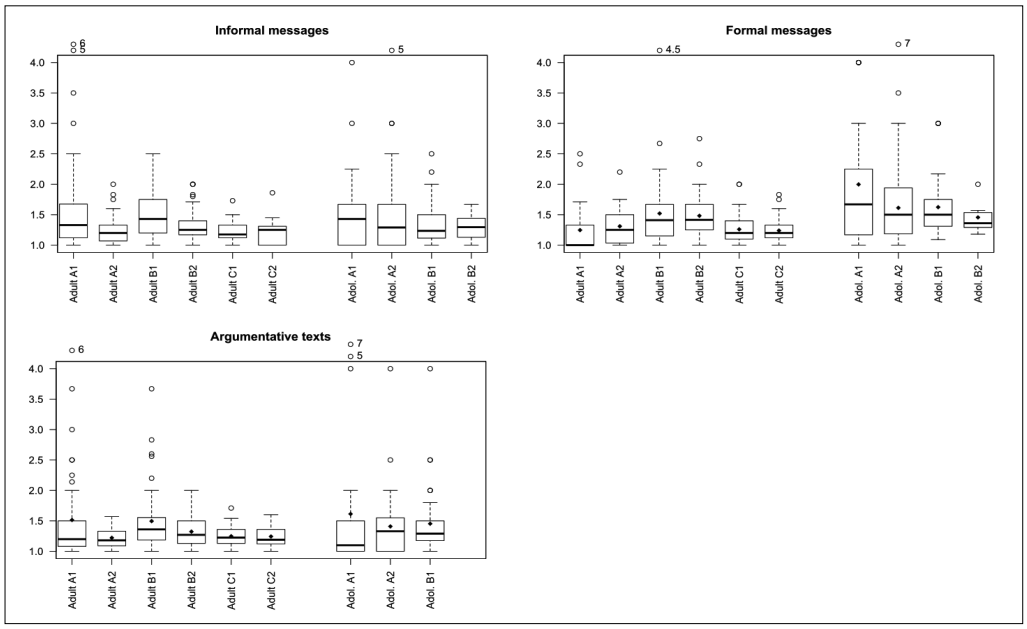


Figure 4. TU/S in informal messages, formal messages, and argumentative texts.

In the adult learners' texts, there are statistically significant differences between proficiency levels in TU/S according to both ANOVA and the Kruskal-Wallis test in the informal ($F(5,192) = 3.55$, $p = .004$, $R^2_{Adj} = .06$; $Chi\ squared = 16.25$, $p < .001$, $df = 5$) and formal ($F(5,222) = 4.62$, $p < .001$, $R^2_{Adj} = .07$; $Chi\ squared = 26.26$, $p < .001$, $df = 5$) messages and according to ANOVA in the argumentative texts ($F(5,235) = 3.24$, $p = .008$, $R^2_{Adj} = .04$). No statistically significant differences between proficiency levels were found in the adolescent learners' texts. The statistically significant between-level differences indicated by the post hoc tests are summarised in Table 6.

Table 6. The statistically significant between-level differences in TU/S according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1								✓$_U$							
Adult A2		✓										✓$_t$			
Adult B1				✓											
Adult B2									✓	✓					
Adult C1															

4.5 Mean number of clauses per sentence (C/S)

The patterns for C/S are mixed (Figure 5). The changes between proficiency levels are not linear, but there are more clauses per sentence on the highest proficiency level compared to the lowest according to group medians in the informal messages and in the argumentative texts in both the adult and adolescent learner data, and according to the group means in the formal messages written by the adult learners. In the adolescent learners' formal messages, the group mean is lower on level B2 than on level A1, and the group median is the same on levels A1 and B2.

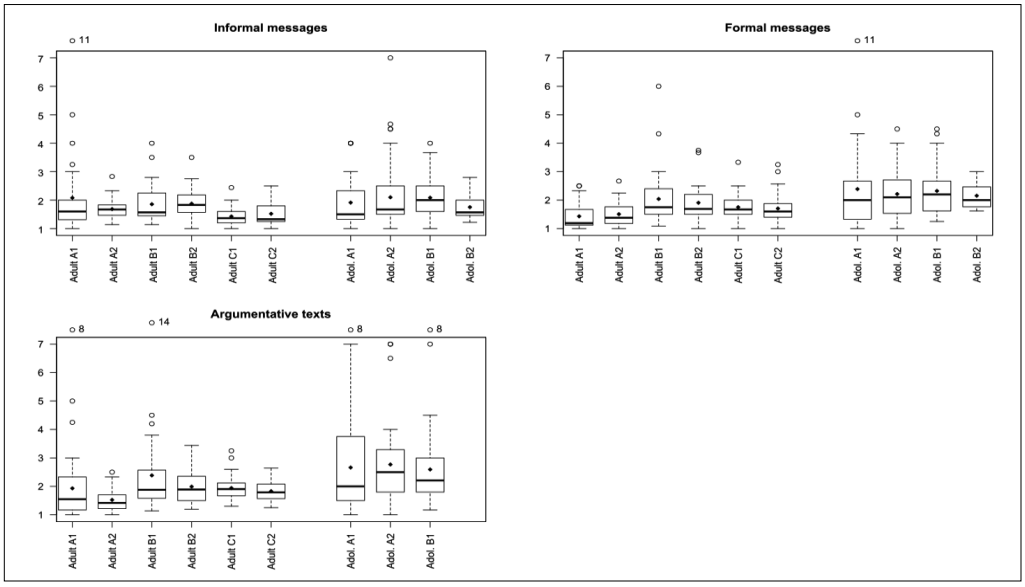


Figure 5. C/S in informal messages, formal messages, and argumentative texts.

There are statistically significant differences in C/S between proficiency levels in L2 adult learner informal messages ($F(5,192) = 2.33$, $p = .044$, $R^2_{Adj} = .03$; *Chi squared* $= 20.32$, $p < .001$, $df = 5$), formal messages ($F(5,222) = 5.03$, $p < .001$, $R^2_{Adj} = .08$; *Chi squared* $= 25.92$, $p < .001$, $df = 5$), and argumentative texts ($F(5,235) = 2.74$, $p = .020$, $R^2_{Adj} = .04$; *Chi squared* $= 27.51$, $p < .001$, $df = 5$) according to both ANOVA and the Kruskal-Wallis test. There are no statistically significant differences between proficiency levels in the L2 adolescent learner texts. The post hoc test results vary according to task type (Table 7).

Table 7. The statistically significant between-level differences in C/S according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1							✓	✓	✓$_U$						
Adult A2				✓			✓	✓$_U$				✓$_U$	✓	✓	✓
Adult B1				✓											
Adult B2				✓											
Adult C1															

4.6 Mean number of clauses per T-unit (C/TU)

For mean number of clauses per T-unit (C/TU), the patterns are varied, and both the group mean and median indicate growth in the number of clauses between the lowest and highest proficiency levels only in the adult learners' formal messages and argumentative texts and in the adolescent learners' formal messages, where the group means grow continuously from one proficiency level to the next (Figure 6).

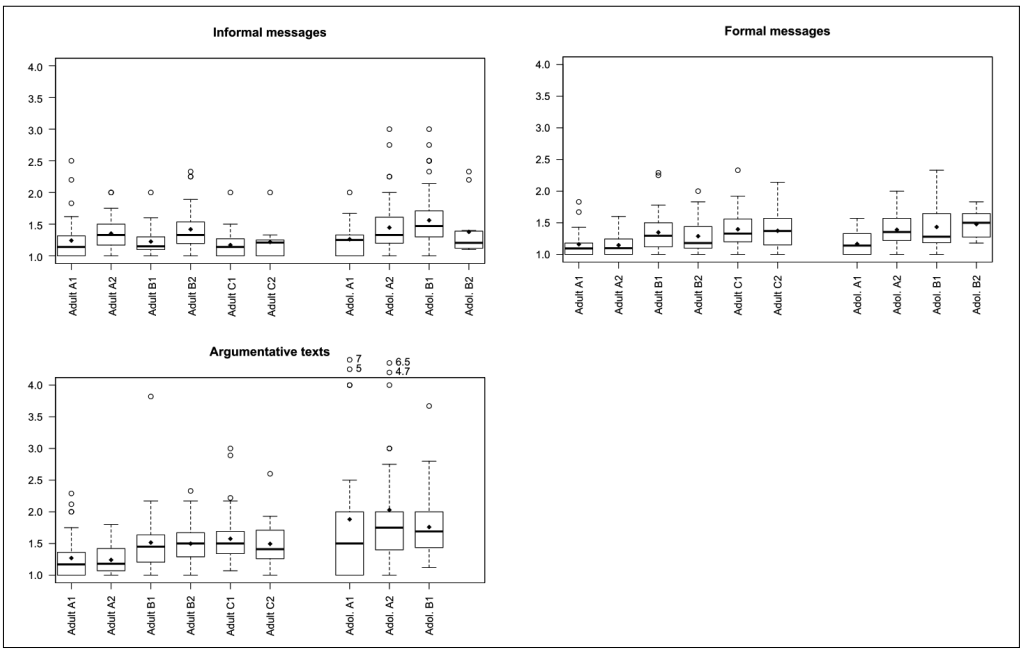


Figure 6. C/TU in informal messages, formal messages, and argumentative texts.

The differences between proficiency levels in C/TU are statistically significant according to both ANOVA and the Kruskal-Wallis test in the informal messages of both the adult ($F(5,192) = 3.98$, $p = .002$, $R^2_{Adj} = .07$; $Chi\ squared = 26.04$, $p < .001$, $df = 5$) and the adolescent learners ($F(3,176) = 3.78$, $p = .012$, $R^2_{Adj} = .04$; $Chi\ squared = 13.55$, $p = .004$, $df = 3$), in the formal messages of both the adult ($F(5,222) = 5.66$, $p < .001$, $R^2_{Adj} = .09$; $Chi\ squared = 34.13$, $p < .001$, $df = 5$) and adolescent learners ($F(5,235) = 6.69$, $p < .001$, $R^2_{Adj} = .11$; $Chi\ squared = 44.73$, $p < .001$, $df = 5$) and in the adult learners' argumentative texts ($F(5,235) = 6.69$, $p < .001$, $R^2_{Adj} = .11$; $Chi\ squared = 44.73$, $p < .001$, $df = 5$). Table 8 shows the between-level differences that are statistically significant according to the post hoc tests.

Table 8. The statistically significant between-level differences in C/TU according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					**Formal messages**					**Argumentative texts**				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1			✓$_U$				✓$_U$		✓	✓		✓$_U$	✓	✓	✓$_U$
Adult A2							✓		✓	✓		✓	✓	✓	✓

Adult B1			✓												
Adult B2				✓											
Adult C1															
Adol. A1	$✓_t$	✓				✓	✓	$✓_U$							
Adol. A2															
Adol. B1															

4.7 Mean number of dependent clauses per clause (DC/C)

The group means and medians of DC/C show that there are more dependent clauses per clause on the highest proficiency level than on the lowest in all task types, with the exception of the group medians of the adolescent learners' informal messages. In the adolescent learners' formal messages, the growth in group means is continuous from level A1 to B2 (Figure 7).

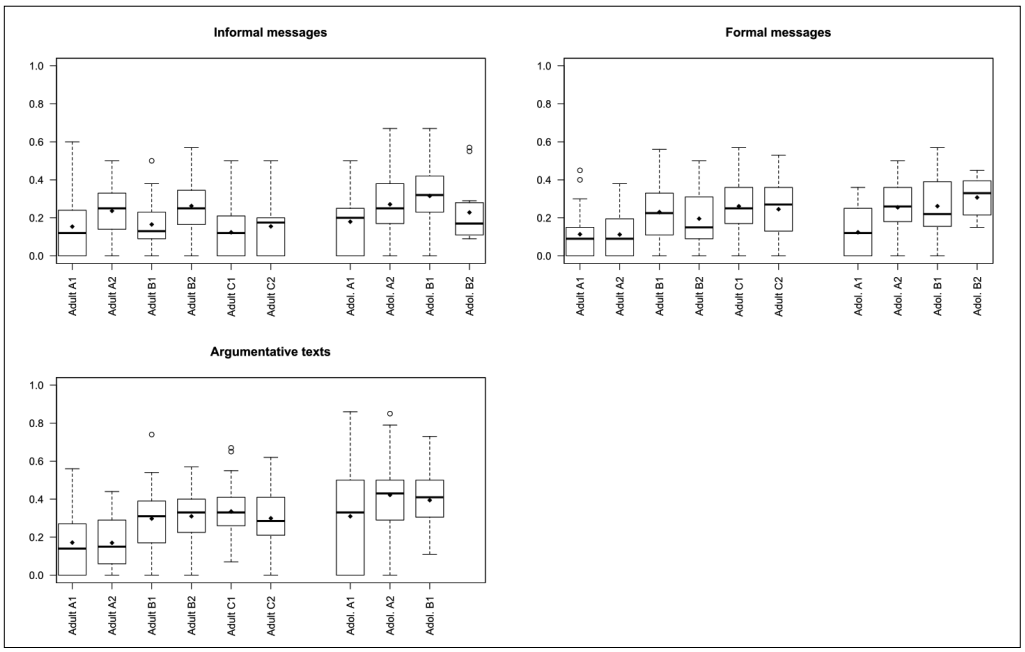


Figure 7. DC/C in informal messages, formal messages, and argumentative texts.

In DC/C, there are statistically significant differences between proficiency levels according to both ANOVA and the Kruskal-Wallis test in the informal messages of both the adult ($F(5,192) = 5.23$, $p < .001$, $R^2_{Adj} = .10$; $Chi\ squared = 25.83$, $p < .001$, $df = 5$) and adolescent learners ($F(3,176) = 4.62$, $p = .004$, $R^2_{Adj} = .06$; $Chi\ squared = 13.62$, $p = .004$, $df = 3$), in the formal messages of both the adult ($F(5,222) = 7.47$, $p < .001$, $R^2_{Adj} = .12$; $Chi\ squared = 33.94$, $p < .001$, $df = 5$) and adolescent learners ($F(2,116) = 7.61$, $p < .001$, $R^2_{Adj} = .14$; $Chi\ squared = 19.93$, $p < .001$, $df = 3$), and in the adult learners' argumentative texts ($F(5,235) = 10.93$, $p < .001$, $R^2_{Adj} = .17$; $Chi\ squared = 44.87$, $p < .001$, $df = 5$). Table 9 presents the levels between which there is a statistically significant difference according to the post hoc tests.

Table 9. The statistically significant between-level differences in DC/C according to parametric (✓$_t$), non-parametric (✓$_U$) or both (✓) post hoc tests.

	Informal messages					Formal messages					Argumentative texts				
	A2	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**	**A2**	**B1**	**B2**	**C1**	**C2**
Adult A1			✓				✓		✓	✓		✓	✓	✓	✓
Adult A2				✓			✓		✓	✓		✓	✓	✓	✓
Adult B1			✓												
Adult B2				✓											
Adult C1															
Adol. A1		✓				✓	✓	✓							
Adol. A2															
Adol. B1															

4.8 Summary of results

The measures yield different results for adult and adolescent learners and also for the different task types. In the adult learner data, there are statistically significant differences between proficiency levels in all the seven measures, while in the L2 adolescent learner data, only four of the measures, i.e. MLTU, MLC, C/TU, and DC/C, show statistically significant differences between proficiency levels. In the adult learner data, there are no statistically significant between-level differences in MLS in the informal messages. In the adolescent learner data, MLTU is the only measure showing statistically significant between-level differences in all three task types.

Developmental trends in group means may differ from those in group medians. There is growth in group means between the lowest and highest proficiency levels in MLTU, MLC and DC/C in all three task types in both the adult and adolescent learner data. The same kind of growth is found in group medians in MLS and MLC. When comparing either group means or group medians, at least one of them is higher on the highest proficiency level than on the lowest also in C/TU, and is lower in TU/S in all task types for both the adult and adolescent learners. None of the measures indicating growth reach the highest values at the highest proficiency level in all three task types, and the change in TU/S is continuous from one proficiency level to the next only in the group means of the adolescent learners' informal messages, where the between-level differences are not statistically significant.

Most measures show statistically significant differences between levels A1 and C2 in the adult learner data, although the differences between levels A1 and C2 are not statistically significant in TU/S and C/S in any of the task types and in MLS, C/TU and DC/C they are statistically significant only in the formal messages and argumentative texts. In the adolescent learner data, there are statistically significant differences between level A1 and the highest proficiency level in MLTU, C/TU and DC/C in the formal messages, and in MLTU and MLC in the argumentative texts.

There is variation in the measures' ability to gauge differences between adjacent proficiency levels. In the adult learner data, none of the measures show

statistically significant differences between levels A1 and A2 or between levels C1 and C2. In the adolescent learner data, all the statistically significant differences are between level A1 and the levels above it.

The effect sizes for all measures showing statistically significant between-level differences in one-way ANOVA are at least small (> .01) when measured with adjusted R squared. The effect size is large (> .14) in the L2 adult learner data for MLC in all task types, for MLTU and DC/C in the formal messages and argumentative texts, and for MLS in the formal messages. In the L2 adolescent learner data, the effect size is large for MLTU and DC/C in the formal messages. The effect size is medium (>. 06) in the L2 adult learner data for C/TU in all task types, for TU/S in the informal and formal messages, for MLS in the argumentative texts, for MLTU and DC/C in the informal messages, and for C/S in the formal messages. In the L2 adolescent learner data, effect size is medium for MLC in the formal messages and argumentative texts and for C/TU in the formal messages. For the measures showing no statistically significant differences between proficiency levels, the effect size is small for TU/S in the L2 adolescent learner formal messages. For the remaining measures, the effect size is less than small.

5 Discussion and conclusions

In the present study, syntactic complexity and seven quantitative measures were studied in relation to the CEFR proficiency levels in cross-sectional data of learner Finnish. Both measures for overall complexity (MLS and MLTU), the measure for sub-clausal complexity (MLC), and the two subordination measures (C/TU and DC/C) grow from the lowest proficiency level to the highest, and the coordination-based measure (TU/S) diminishes from the lowest proficiency level to the highest even if the changes are not linear. The growing trend is in line with earlier research findings, such as those included in the synthesis of Wolfe-Quintero et al. (1998). The general reduction in coordination from level A1 to the highest proficiency level is in line with the notion of clausal coordination being more typical of lower proficiency levels (e.g. Norris & Ortega, 2009).

There are statistically significant differences between proficiency levels only for some of the measures and only between some proficiency levels. Indeed, the measures often have overlapping values when proficiency levels are compared. In the adult learner data, there are differences between the beginners and advanced learners in most measures, but the intermediate learners' results overlap with the beginners' or advanced learners' results, or both, in all the measures. In the adolescent learner data, there are overlapping results in all the measures.

Regarding the first research question, *How does syntactic complexity in written learner Finnish develop across CEFR proficiency levels when measured quantitatively*, the results suggest that the development is different in the adult learner data and in the adolescent learner data. The results show that there are statistically significant between-level differences in all the measures in the adult learner data, but in only four measures in the adolescent learner data. There is also more within-group variation in many of the measures in the adolescent learner data than in the adult learner data. These findings suggest that adult and adolescent

L2 Finnish learners use some syntactic features differently in their writing. The measures also indicate different patterns of development in the two age groups: in the adult learner data, the statistically significant differences are typically found when levels A1 and A2 are compared to the higher proficiency levels, whereas in the adolescent learner data, the differences are between level A1 and the other proficiency levels. Similar differences between adult and adolescent learners have been found in the use of existential sentences (Kajander, 2013), indirect references (Seilonen, 2013), and transitive constructions (Reiman, 2014) in the same data.

The answer to the second research question, *How well do the quantitative measures used in this study differentiate CEFR proficiency levels in written learner Finnish*, is mixed. While most of the measures do differentiate the lowest proficiency levels from the highest, most of them do not differentiate the intermediate learner levels from other levels or differentiate between adjacent proficiency levels. In this regard, mean length of clause (MLC) seems the best measure of syntactic complexity for adult L2 learner Finnish, as it develops quite linearly and it is also able to differentiate the intermediate levels from the levels both below and above. For adolescent L2 learner Finnish, mean length of T-unit (MLTU) seems the best measure, as it is able to differentiate level A1 from most of the levels above it in all task types, even if the increase in MLTU is not linear in all task types. The results also suggest that the measures are sensitive to task type, which is in line with findings from other studies (see e.g. Michel, 2017). In the present study, there are statistically significant differences between proficiency levels in all seven measures in the adult learners' formal messages and argumentative texts, but only in six measures in the informal messages. In the adolescent learner data, only four measures show statistically significant differences between proficiency levels; four of them in the formal messages, three in the informal messages, and two in the argumentative texts. The formal messages also seem to show statistically significant differences between more proficiency levels than do the other two task types. Differences between the task types in the CEFLING corpus have also been found by Seilonen (2013), Kajander (2013), and Reiman (2014).

There are a number of limitations to this study. First, the data were segmented into the production units by one annotator only. As learner language contains deviations from the norms, annotating learner language always involves some level of interpretation of the intended forms (e.g. Brunni, Lehto, Jantunen, & Airaksinen, 2015; Granger, 2002). Segmenting the data used in this study into clauses, sentences, and T-units is no exception (Martin, 2013; Mylläri, 2020). Using more annotators and negotiating the segmentation could result in different numbers of clauses, sentences, and T-units in some texts, and this could affect the measures.

Second, the texts used in this study are relatively short. This can partly be explained by typological features of Finnish, such as the lack of articles and the limited use of prepositions, which affect the word count. On average, the length of the texts written by the adolescent learners corresponds to that of the texts written by their L1 counterparts in the CEFLING project. The adult learners' texts are generally longer than those by the adolescent learners, and the adult learners' argumentative texts reach the length of 100 words or more at level B1. This corresponds to the length of the 100-word random samples used in the studies of Spoelman and Verspoor (2010) and Tilma (2014), who also used

shorter samples at the beginning, when the learner texts did not reach 100 words.

Third, there would be grounds for criticising the statistical analysis of the data. There is a varying amount of individual variation in the data and there are outliers in many of the groups that were compared. Not excluding the outliers from the statistical analysis could have had an impact on the parametric tests used in the study. The effect was partly controlled by using both parametric and non-parametric tests. Also, the results for adolescent learner proficiency level B2 should be interpreted with caution since there is only a limited number of texts on that level. Therefore the statistical significance or insignificance of the results should not be interpreted as straightforward evidence of the measures' general ability or inability to gauge differences between proficiency levels.

The present study suggests some interesting topics for future research. A closer analysis of the differences between adult and adolescent L2 learners, as well as of the differences between the texts written by the adolescent L2 learners and the corresponding L1 writers, could be worthwhile. Another topic for future research could be the correlation between the measures (cf. Lu, 2017), as they may prove more powerful indicators of proficiency together than individually. Also, the present study focused on syntactic complexity in cross-sectional data and in relation to proficiency. The results therefore cannot be interpreted as reflecting the measures' value as indicators of development, which should be studied using longitudinal data.

The results of this study support calls for new ways of exploring complexity, especially in morphologically rich languages. For learner Finnish, Reiman (2011b) has argued for a more qualitative approach, and Tilma (2014) has used morphemes instead of words in length-based quantitative measures. Although measures of complexity cannot be validated by simply showing increase across time or proficiency (Bulté and Housen, 2012), and increasing complexity does not necessarily mean increasing proficiency, as pointed out by Ortega (2003) and Pallotti (2009), there is a need for new means to gauge complexity across proficiency levels if syntactic complexity is going to be used to measure learner language proficiency.

Endnote

[1] The parsing pipeline and the clause splitting feature are available under an open licence at http://turkunlp.github.io/Finnish-dep-parser/. A version available on July 29, 2018 was used.

References

Alanen, R., Huhta, A., & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 21–56). EUROSLA Monographs series 1.

Alisaari, J. (2016). *Songs and poems in the second language classroom. The hidden potential of singing for developing writing fluency.* Annales Universitaris Turkuensis, Series B Humaniora 426. Turku: University of Turku.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *Tesol Quarterly, 26(2)*, 390–395.

Bernardini, P., & Granfeldt, J. (2019). On crosslinguistic variation and measures of linguistic complexity in learner texts: Italian, French and English. *International Journal of Applied Linguistics, Special issue*, 211–232.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly, 45(1)*, 5–35.

Brunni, S., Lehto, L., Jantunen, J., & Airaksinen, V. (2015). How to annotate morphologically rich learner language. Principles, problems and solutions. *Bergen Language and Linguistics Studies 6*, 133–152.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins Publishing Company.

Council of Europe. (2001). *Common European framework of reference for: learning, teaching, assessment*. Retrieved from https://rm.coe.int/1680459f97

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language.* Oxford: Oxford University Press.

Foster, P., Tonkyn A., & Wigglesworth G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics 21(3)*, 354–375.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: Benjamins.

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction*. (2nd rev. ed.). Berlin: De Gruyter Mouton.

Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. In L. Roberts, I. Vedder, & J.H. Hulstijn (Eds.), *Eurosla Yearbook 14* (pp. 1–30). Amsterdam: John Benjamins.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., & Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation 48*, 493–531.

Housen, A., De Clercq, B., Kuiken, F., & Vedder I. (2019). Multiple approaches to complexity in second language research, *Second Language Research, 35(1)*, 3–21.

Housen, A., Kuiken, V., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA.* Amsterdam: John Benjamins Publishing Company.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study – Validating the rating process across tasks, scales and languages. *Language Testing, 31(3)*, 307–328.

Jantunen, J., & Pirkola, S. (2015). Oppijansuomen sähköiset tutkimusaineistot. Nykytilanne [Electronic corpora of learner Finnish. Current situation]. *Virittäjä, 119(1),* 88–103 .

Kajander, M. (2013). *Suomen eksistentiaalilause toisen kielen oppimisen polulla* [Paths of learning Finnish existential sentences]. Jyväskylä studies in humanities 220. Jyväskylä: University of Jyväskylä.

Kalliokoski, J. (2006). Virke, dialogisuus ja argumentaatio: irralliset sivulauseet ja toisella kielellä kirjoittaminen [Sentence, dialogue and argumentation: stand-alone subordinate clauses and second language writing]. In T. Nordlund, T. Onikki-Rantajääskö, T. Suutari, & H. Forsberg (Eds.), *Kohtauspaikkana kieli: näkökulmia persoonaan, muutoksiin ja valintoihin* [Language as a meeting place: Perspectives on person, changes and choices] (pp. 212–231). Helsinki: Suomalaisen Kirjallisuuden Seura.

Kuiken, F. & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish, *International Journal of Applied Linguistics, Special issue,* 192–210.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27(4)*, 590–619.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30(4)*, 579–589.

Larson-Hall J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15(4)*, 474–496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development, *TESOL Quarterly, 45(1)*, 36–62.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing, 34(4)*, 493–511.

Martin, M. (2013). Sentences and clauses as complexity measures in second language writing: a segmentation experiment. In M. Järventausta, & M. Pantermöller (Eds.), *Finnische Sprache, Literatur und Kultur im deutschsprachigen Raum – Suomen kieli, kirjallisuus ja kulttuuri saksankielisellä alueella* (pp. 185–198). Greifswald: Veröffentlichungen der Societas Uralo-Altaica.

Martin, M., Mustonen, S., Reiman, N., & Seilonen, M. (2010). On becoming an independent user. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 57–79). EUROSLA Monographs series 1.

Michel, M. C. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen, & S. Masatoshi (Eds.), R*outledge handbook of instructed second language acquisition* (pp. 50–68). New York: Routledge.

Mylläri, T. (2020). Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association, 4(1),* 13–23.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30(4)*, 555–578.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24(4)*, 492–518.

Pallotti, G. (2009). CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics, 30(4)*, 590–601.

Pallotti, G. (2015). A simple view of linguistic complexity, *Second Language Research, 31(1)*, 117–134.

Reiman, N. (2011a). Transitiivikonstruktio ikkunana syntaksin kehitykseen: infiniittiset rakenteet ja passiivi taidon indikaattoreina S2-oppijoiden teksteissä [The transitive construction as a window into syntax development: Infinite structures and passive as indicators of proficiency in F2 students' texts]. In E. Lehtinen, S. Aaltonen, M. Koskela, E. Nevasaari, & M. Skog-Södersved (Eds.), *AFinLAe 3* (pp. 142–157). Retrieved from http://ojs.tsv.fi/index.php/afinla/issue/view/694

Reiman, N. (2011b): Two faces of complexity: structural measures and diversity of constructions. *Nordand, 6(2)*, 9–23.

Reiman, N. (2014). Yläkoulun S2-oppilaiden transitiivi-ilmausten käyttö Eurooppalaisen viitekehyksen taitotasoilla [Lower secondary school L2 Finnish students' use of transitive expressions at the CEFR levels]. *Lähivõrdlusi. Lähivertailuja 24*, 183–220.

Seilonen, M. (2013). *Epäsuora henkilöön viittaaminen oppijansuomessa* [Indirect references in Finnish learner language] Jyväskylä Studies in Humanities 197. Jyväskylä: University of Jyväskylä.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics, 31(4)*, 532–553.

Tilma, C. (2014). *The dynamics of foreign versus second language development in Finnish writing*. Jyväskylä studies in humanities 233. Jyväskylä: University of Jyväskylä.

Verspoor, M., Lowie, W., Chan, H., & Vahtrick, L. (2017). Linguistic complexity in second language development: variability and variation at advanced stages. *Recherches en didactique des langues et des cultures, 14(1)*, 1–27

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity. Technical report No. 1.* Honolulu: Second Language Teaching and Curriculum Center.

Appendices

Appendix A.

Table A1. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, mean length of sentence (MLS).

MLS	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	9.14	8.55	6.57	3.04	22	6.10	2.86	5.24	2.39	50	8.20	5.44	6.59	4.54
Adult A2	39	7.27	1.93	7.00	2.52	27	6.43	2.68	5.75	3.29	37	6.45	2.20	6.00	1.93
Adult B1	41	8.28	2.66	7.73	3.17	42	8.89	3.78	8.06	2.13	43	12.91	11.30	10.27	4.44
Adult B2	39	8.70	2.58	8.71	2.95	34	9.40	2.59	8.45	2.98	35	11.25	3.27	10.41	4.63
Adult C1	26	7.45	1.96	7.46	1.90	45	9.84	2.75	9.25	3.76	46	11.35	2.78	11.12	3.11
Adult C2	14	8.17	1.58	7.53	1.65	58	9.83	2.79	9.58	3.82	30	10.77	3.02	10.32	3.40
Adol. A1	25	8.13	3.37	7.00	3.50	33	9.63	8.22	7.40	4.87	32	12.07	8.46	8.67	8.44
Adol. A2	79	9.90	5.40	8.25	5.75	40	9.59	4.19	8.80	4.50	39	13.41	6.73	11.00	6.44
Adol. B1	64	9.51	3.17	9.25	4.36	40	10.34	3.78	9.75	4.53	40	13.42	6.71	11.42	6.15
Adol. B2	12	8.19	2.56	7.55	2.83	7	10.51	3.99	9.91	1.50	-	-	-	-	-

Table A2. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, mean length of T-unit (MLTU).

MLTU	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	5.31	2.07	5.00	1.67	22	4.82	1.14	4.69	0.93	50	5.33	1.61	5.20	2.11
Adult A2	39	5.84	1.49	5.71	1.59	27	4.86	1.59	4.55	1.35	37	5.25	1.47	5.17	1.95
Adult B1	41	5.52	1.15	5.43	1.80	42	5.95	1.69	5.62	1.29	43	8.23	2.87	7.91	2.83
Adult B2	39	6.57	1.67	6.00	2.35	34	6.43	1.37	6.48	1.78	35	8.52	1.93	8.67	2.55
Adult C1	26	6.12	1.43	6.10	1.01	45	7.86	1.81	7.71	2.66	46	9.26	2.86	8.59	2.41
Adult C2	14	6.62	0.97	6.43	1.54	58	7.95	2.00	7.64	2.09	30	8.86	3.16	8.34	4.32
Adol. A1	25	5.48	1.19	5.67	2.00	33	4.64	0.81	4.67	1.10	32	8.27	5.26	6.44	2.45
Adol. A2	79	6.76	2.22	6.20	2.94	40	5.97	1.43	5.89	2.37	39	9.80	4.75	8.60	2.94
Adol. B1	64	7.16	2.31	6.60	2.11	40	6.37	1.59	6.12	1.92	40	9.17	2.54	8.87	2.75
Adol. B2	12	6.41	2.11	5.61	1.35	7	7.06	1.33	7.00	0.83	-	-	-	-	-

Table A3. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, mean length of clause (MLC).

MLC	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	4.21	0.77	4.20	0.76	22	4.17	0.80	4.12	0.82	50	4.19	0.76	4.16	1.08
Adult A2	39	4.30	0.62	4.25	0.61	27	4.18	0.98	4.00	0.96	37	4.19	0.61	4.13	0.81
Adult B1	41	4.51	0.62	4.41	0.80	42	4.42	0.71	4.29	0.72	43	5.46	1.03	5.27	0.93
Adult B2	39	4.63	0.50	4.71	0.66	34	5.01	0.68	4.92	0.48	35	5.71	0.90	5.46	1.16
Adult C1	26	5.27	1.22	5.12	0.96	45	5.63	0.84	5.60	0.94	46	5.87	0.86	5.73	1.04
Adult C2	14	5.51	0.72	5.52	0.86	58	5.81	1.09	5.52	1.42	30	5.86	1.04	5.63	1.26
Adol. A2	79	4.65	0.85	4.56	1.09	40	4.28	0.59	4.25	0.77	39	4.97	1.01	4.71	1.04
Adol. B1	64	4.60	0.78	4.50	1.09	40	4.48	0.56	4.28	0.66	40	5.25	0.81	5.22	1.15
Adol. B2	12	4.64	0.49	4.64	0.36	7	4.83	0.91	4.54	1.22	-	-	-	-	-

Table A4. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, T-units per sentence (TU/S).

TU/S	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	1.67	1.06	1.33	0.56	22	1.25	0.43	1.00	0.30	50	1.51	0.86	1.20	0.40
Adult A2	39	1.26	0.25	1.20	0.26	27	1.31	0.31	1.25	0.47	37	1.22	0.17	1.18	0.24
Adult B1	41	1.50	0.37	1.43	0.55	42	1.52	0.61	1.41	0.50	43	1.50	0.54	1.36	0.37
Adult B2	39	1.33	0.28	1.25	0.23	34	1.48	0.35	1.42	0.40	35	1.32	0.26	1.27	0.37
Adult C1	26	1.22	0.19	1.18	0.21	45	1.26	0.25	1.20	0.30	46	1.25	0.17	1.23	0.23
Adult C2	14	1.25	0.23	1.25	0.28	58	1.24	0.19	1.20	0.20	30	1.24	0.16	1.19	0.24
Adol. A1	25	1.52	0.72	1.43	0.67	33	2.00	1.26	1.67	1.08	32	1.61	1.31	1.10	0.50
Adol. A2	79	1.45	0.61	1.29	0.67	40	1.61	0.61	1.50	0.72	39	1.41	0.55	1.33	0.55
Adol. B1	64	1.35	0.34	1.24	0.38	40	1.62	0.47	1.50	0.43	40	1.45	0.55	1.29	0.32
Adol. B2	12	1.29	0.22	1.30	0.30	7	1.46	0.28	1.36	0.25	-	-	-	-	-

Table A5. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, clauses per sentence (C/S).

C/S	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	2.08	1.69	1.60	0.69	22	1.43	0.49	1.19	0.54	50	1.93	1.21	1.55	1.12
Adult A2	39	1.69	0.34	1.67	0.36	27	1.51	0.43	1.38	0.59	37	1.52	0.40	1.42	0.48
Adult B1	41	1.86	0.63	1.57	0.80	42	2.04	0.89	1.75	0.87	43	2.39	1.97	1.88	0.99
Adult B2	39	1.88	0.50	1.83	0.62	34	1.91	0.59	1.69	0.68	35	1.99	0.58	1.89	0.86
Adult C1	26	1.43	0.32	1.37	0.39	45	1.75	0.44	1.67	0.50	46	1.94	0.42	1.91	0.44
Adult C2	14	1.52	0.43	1.33	0.46	58	1.71	0.44	1.60	0.47	30	1.83	0.35	1.79	0.48
Adol. A1	25	1.91	0.94	1.50	1.00	33	2.39	1.87	2.00	1.34	32	2.66	1.86	2.00	2.13
Adol. A2	79	2.10	1.01	1.67	1.00	40	2.21	0.83	2.10	1.14	39	2.77	1.45	2.50	1.49
Adol. B1	64	2.08	0.67	2.00	0.90	40	2.32	0.85	2.20	1.03	40	2.59	1.37	2.21	1.20
Adol. B2	12	1.75	0.46	1.57	0.42	7	2.15	0.53	2.00	0.70	-	-	-	-	-

Table A6. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, clauses per T-unit (C/TU).

C/TU	Informal message					Formal message					Argumentative text				
	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR	n	M	SD	Mdn	IQR
Adult A1	39	1.24	0.33	1.14	0.32	22	1.16	0.22	1.10	0.18	50	1.27	0.33	1.17	0.35
Adult A2	39	1.35	0.25	1.33	0.33	27	1.15	0.16	1.10	0.25	37	1.24	0.23	1.18	0.35
Adult B1	41	1.23	0.20	1.15	0.20	42	1.35	0.29	1.30	0.37	43	1.51	0.47	1.45	0.43
Adult B2	39	1.42	0.33	1.33	0.34	34	1.29	0.26	1.18	0.33	35	1.50	0.29	1.50	0.38
Adult C1	26	1.17	0.22	1.14	0.26	45	1.40	0.27	1.33	0.36	46	1.57	0.39	1.50	0.34
Adult C2	14	1.22	0.25	1.21	0.23	58	1.37	0.27	1.37	0.40	30	1.49	0.34	1.41	0.44
Adol. A1	25	1.26	0.25	1.25	0.33	33	1.17	0.18	1.14	0.33	32	1.88	1.34	1.50	1.00
Adol. A2	79	1.45	0.38	1.33	0.41	40	1.39	0.26	1.36	0.35	39	2.03	1.06	1.75	0.60
Adol. B1	64	1.56	0.44	1.47	0.41	40	1.43	0.37	1.28	0.44	40	1.76	0.49	1.69	0.56
Adol. B2	12	1.38	0.43	1.21	0.27	7	1.48	0.24	1.50	0.37	-	-	-	-	-

Table A7. Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) by writer group and task type, dependent clauses per clause (DC/C).

DC/C	Informal message					Formal message					Argumentative text				
	n	**M**	**SD**	**Mdn**	**IQR**	**n**	**M**	**SD**	**Mdn**	**IQR**	**n**	**M**	**SD**	**Mdn**	**IQR**
Adult A1	39	0.15	0.16	0.12	0.24	22	0.11	0.13	0.09	0.15	50	0.17	0.17	0.14	0.27
Adult A2	39	0.24	0.13	0.25	0.19	27	0.11	0.11	0.09	0.20	37	0.17	0.14	0.15	0.23
Adult B1	41	0.17	0.12	0.13	0.14	42	0.23	0.14	0.23	0.21	43	0.30	0.15	0.31	0.22
Adult B2	39	0.26	0.14	0.25	0.18	34	0.20	0.14	0.15	0.21	35	0.31	0.13	0.33	0.18
Adult C1	26	0.12	0.12	0.12	0.20	45	0.26	0.13	0.25	0.19	46	0.34	0.13	0.33	0.15
Adult C2	14	0.16	0.14	0.18	0.18	58	0.25	0.14	0.27	0.22	30	0.30	0.14	0.29	0.20
Adol. A1	25	0.18	0.15	0.20	0.25	33	0.12	0.12	0.12	0.25	32	0.31	0.27	0.33	0.50
Adol. A2	79	0.27	0.16	0.25	0.21	40	0.26	0.14	0.26	0.18	39	0.42	0.20	0.43	0.21
Adol. B1	64	0.32	0.17	0.32	0.19	40	0.26	0.17	0.22	0.22	40	0.39	0.14	0.41	0.19
Adol. B2	12	0.23	0.17	0.17	0.17	7	0.31	0.11	0.33	0.18	-	-	-	-	-

Received September 18, 2019
Revision received April 25, 2020
Accepted June 26, 2020

III

KONJUNKTIOT JA SYNTAKTINEN KOMPLEKSISUUS: KONJUNKTIOIDEN KÄYTTÖ SUOMI TOISENA KIELENÄ -TEKSTEISSÄ ERI TAITOTASOILLA

by

Mylläri, Taina 2022

Puhe ja kieli 42 (3), 175–200.

https://doi.org/10.23997/pk.126146

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

KONJUNKTIOT JA SYNTAKTINEN KOMPLEKSISUUS: KONJUNKTIOIDEN KÄYTTÖ SUOMI TOISENA KIELENÄ -TEKSTEISSÄ ERI TAITOTASOILLA

Taina Mylläri, Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto

Kielitaidon kehittymistä voidaan tutkia mittaamalla oppijankielen sujuvuutta, tarkkuutta ja kompleksisuutta. Kompleksisuus voidaan määritellä oppijankielen monipuolisuudeksi ja kehittyneisyydeksi, mutta syntaktista kompleksisuutta on kansainvälisessä toisen kielen oppimisen tutkimuksessa tyypillisesti mitattu alisteisiin lauserakenteisiin perustuvilla kvantitatiivisilla mittareilla. Tässä artikkelissa syntaktisen kompleksisuuden kehitystä tutkitaan tarkastelemalla rinnastus- ja alistuskonjunktioiden käyttöä suomi toisena kielenä -teksteissä Eurooppalaisen viitekehyksen eri taitotasoilla. Aineistona käytetään Jyväskylän yliopiston Cefling-hankkeen oppijansuomen korpuksen tekstejä. Rinnasteisten ja alisteisten rakenteiden osuutta syntaktisen kompleksisuuden kehityksessä selvitetään tarkastelemalla rinnastus- ja alistuskonjunktioiden käyttöä suhteessa kielitaidon tasoon. Syntaktista kompleksisuutta oppijankielen monipuolisuutena selvitetään tarkastelemalla, mitä eri konjunktioita eri taitotasoilla käytetään. Tulokset kyseenalaistavat käsityksen, että sivulauseet ilmaantuvat oppijankieleen vasta keskitasolla, sillä ne osoittavat, että suomi toisena kielenä -teksteissä käytetään sekä rinnastus- että alistuskonjunktioita jo taitotasolla A1. Lisäksi tulokset osoittavat, että konjunktioiden käyttö saattaa kehittyä eri tavalla aikuisten ja yläkouluikäisten teksteissä.

Avainsanat: konjunktioiden käyttö, suomi toisena kielenä, syntaktinen kompleksisuus

1 JOHDANTO

Toisen kielen taitoa ja sen kehittymistä voidaan tutkia esimerkiksi oppijankielen kielenpiirteiden muuttumisen tai toisen kielen käyttäjän kommunikatiivisten taitojen kehittymisen näkökulmasta. Molemmissa lähestymistavoissa kielitaidon ajatellaan usein kehittyvän yksinkertaisista ja irrallisista ilmauksista monimutkaisempiin kokonaisuuksiin. Kun toisen kielen taidon kehitystä tarkastellaan kielenpiirteissä tapahtuvina muutoksina, apuna voidaan käyttää kompleksisuuden, tarkkuuden ja sujuvuuden käsitteitä (esim.

Kirjoittajan yhteystiedot:
Taina Mylläri
taina.myllari@jyu.fi

Ellis & Barkhuizen, 2005; Housen, Vedder & Kuiken, 2012). Näistä syntaktisen kompleksisuuden katsotaan usein kehittyvän niin, että alkeistasolla yksittäisiä merkityksiä ja lauseita aletaan yhdistää rinnastamalla, keskitasolla siirrytään rinnastamisesta sivulauseiden ja muiden alisteisten rakenteiden käyttöön ja ylimmillä taitotasoilla lausekkeet pitenevät ja monipuolistuvat (esim. Kuiken & Vedder, 2019; Norris & Ortega, 2009).

Kielitaitoa kommunikatiivisesta näkökulmasta lähestyvässä Eurooppalaisessa viitekehyksessä (EVK, 2003) kielen osaaminen jaetaan taitotasoihin sen perusteella, miten kielenkäyttäjä selviytyy erilaisista viestintätilanteista, ja taitotasoille tyypilliset kielelliset piirteet jäävät vähälle huomiolle (esim. Alanen, Huhta & Tarnanen, 2010; Martin, 2013). Konjunktioiden käyttö on yksi niistä harvoista kielenpiirteistä, jotka mainitaan EVK:n taitotasokuvauksissa: kirjoittamisen taitotasoasteikon mukaan alkeistasolla A1 kielenkäyttäjä osaa kirjoittaa ”yksinkertaisia, irrallisia ilmauksia ja lauseita” ja seuraavalla tasolla (A2) hän osaa yhdistää niitä ”tavanomaisilla sidesanoilla, kuten sanoilla ’ja’, ’mutta’, ’koska’” (EVK 2003: 96). Aiempien tutkimustulosten mukaan sekä rinnastus- että alistuskonjunktioita kuitenkin käytetään jo toisen kielen oppimisen alkuvaiheessa (esim. Määttä, 2012; Vyatkina, 2012), eikä esimerkiksi rinnasteisten päälauseiden määrä virkkeessä tai sivulauseiden osuus kaikista lauseista välttämättä ole taitotasolla A1 merkittävästi pienempi kuin taitotasolla A2 (Mylläri, 2020a).

Kompleksisuutta voidaan tarkastella kielisysteemin sisältämien osien ja niiden välisten kytkösten määränä ja monipuolisuutena (esim. Bulté & Housen, 2012). Tässä artikkelissa yhtä kompleksisuuden osa-aluetta, syntaktista kompleksisuutta, lähestytään tarkastelemalla konjunktioilla merkittyjä syntaktisia kytköksiä. Tavoitteena on täydentää aiemmassa tutkimuksessa (Mylläri, 2020a) saatua tietoa oppijansuomen syntaktisesta kompleksisuudesta EVK:n eri taitotasoilla. Tutkimuskysymykset ovat:

1. Miten paljon konjunktiolla merkittyjä syntaktisia kytköksiä käytetään eri taitotasoilla?
2. Miten konjunktioiden käyttö muuttuu taitotasolta toiselle?

Tutkimusaineisto koostuu aikuisten ja yläkouluikäisten kirjoittamista suomi toisena kielenä (S2) -teksteistä, jotka edustavat kolmea eri tehtävätyyppiä, ja se kattaa kielitaidon taitotasot A1–C2. Konjunktiolla merkittyjen syntaktisten kytkösten määrää ja yleisyyttä kartoitetaan laskemalla konjunktioiden esiintymistaajuudet eri taitotasoilla. Konjunktioiden käytön muutoksia kartoitetaan selvittämällä, millä taitotasolla yksittäiset konjunktiot ilmaantuvat ja yleistyvät S2-teksteissä, millaisia syntaktisia kielenyksiköitä rinnastuskonjunktioiden avulla yhdistetään ja miten suuri osa sivulauseista sisältää alistuskonjunktion. Artikkelissa keskitytään konjunktiolla merkittyjen syntaktisten kytkösten kuvaamiseen ryhmätasolla taitotasoittain. Konjunktioita ja niiden käyttöä käsitellään kieliopin ilmiönä.

2 SYNTAKTINEN KOMPLEKSISUUS OSANA KIELITAITOA

Toisen kielen oppimisen tutkimuksessa syntaktisen kompleksisuuden määritelmien lähtökohtana on oppijankielen rakenteiden monipuolisuus, vaihtelu tai kehittyneisyys

(esim. Ellis & Barkhuizen, 2005; Housen, De Clercq, Kuiken & Vedder, 2019; Ortega, 2003). Tutkijoiden keskuudessa ei kuitenkaan ole yksimielisyyttä siitä, miten kompleksisuus näkyy kielellisessä tuotoksessa, mikä on sen suhde kielitaidon tasoon tai oppimisprosessin kestoon ja miten sitä parhaiten mitataan (esim. Housen ym., 2019; Ortega, 2003; Pallotti, 2009, 2015; Wolfe-Quintero, Inagaki & Kim, 1998). Tyypillisesti oppijankielen syntaktista kompleksisuutta on mitattu lauseisiin tai T-yksikköihin eli yhden päälauseen ja siihen liittyvien sivulauseiden muodostamaan kokonaisuuteen perustuvilla määrällisillä mittareilla, kuten T-yksikön pituus sanoina tai lauseina ja sivulauseiden osuus kaikista lauseista, joita on kuitenkin arvosteltu siitä, että ne yksinkertaistavat syntaktista kompleksisuutta keskittymällä lähinnä alisteisuuteen (engl. *subordination*) (esim. Bulté & Housen, 2012). Ne jättävät huomiotta esimerkiksi rinnastamisen (esim. Bardovi-Harlig, 1992) ja muut kuin lausetason yhdistämiskeinot (esim. De Clercq & Housen, 2017). Lisäksi mittareita on arvosteltu siitä, että ne tavoittavat vain kielenyksiköiden määrissä tapahtuvat muutokset eivätkä muutosten syitä, sillä keskenään hyvin erilaiset virkkeet tai T-yksiköt voivat tuottaa täysin samat määrälliset tulokset (esim. Biber, Gray & Poonpon, 2011; Rimmer, 2009).

Nykyään syntaktista kompleksisuutta tutkitaankin yleensä moniulotteisena ilmiönä, jota tarkastellaan useilla rinnakkaisilla mittareilla (esim. Housen ym., 2019). Myös konjunktioita on tarkasteltu muiden syntaktisen kompleksisuuden mittareiden rinnalla. Alkeistason saksanoppijoiden rinnastus- ja alistuskonjunktioiden käyttöä tutkinut Vyatkina (2012, 2013) totesi, että ryhmätasolla rinnastuskonjunktioiden käyttö väheni ja alistuskonjunktioiden käyttö lisääntyi kielitaidon kehittyessä mutta yksilötasolla kehitys saattoi poiketa ryhmätason kehityksestä. Sanoja ja lauseita kytkeviä ilmauksia yhtenä syntaktisen kompleksisuuden mittarina käyttäneet Benevento ja Storch (2011) puolestaan havaitsivat, että näiden ilmausten määrä ranskanoppijoiden teksteissä ei välttämättä kehity samalla tavalla kuin lauseiden määrä T-yksikössä: kolmella mittauskerralla kerätyissä teksteissä lauseiden määrä T-yksikössä oli suurimmillaan eri mittauskerralla kuin kielenyksiköiden kytkemiseen käytettyjen ilmausten määrä[1]. Grant ja Ginther (2000) tutkivat englanti toisena kielenä -tekstien kielellisten piirteiden suhdetta kielitaidon arviointiin ja havaitsivat, että paremman arvosanan saaneissa teksteissä käytettiin enemmän konjunktioita ja sivulauseita kuin alemman arvosanan saaneissa. Taguchi, Crawford ja Wetzel (2013) puolestaan tutkivat yliopistotason vaihto-opiskelijoiden englannin kielen lähtötasotestien tekstejä ja havaitsivat, ettei alisteisuus välttämättä toimi kirjoittamisen taidon mittarina edistyneemmillä taitotasoilla, sillä pistemäärän perusteella kahteen ryhmään jaetuista teksteistä enemmän konjunktioita sisälsivät alemman pistemäärän saaneet tekstit, jotka myös joidenkin muiden tutkimuksessa käytettyjen syntaktisen kompleksisuuden mittareiden mukaan olivat hieman kompleksisempia kuin enemmän pisteitä saaneet tekstit. He toteavatkin, ettei lausetason kompleksisuus tai pelkkä alisteisten rakenteiden mittaaminen välttämättä anna oikeaa kuvaa kielitaidosta kaikissa konteksteissa ja ettei sivulauseiden

1 Tarkastelussa oli konjunktioiden lisäksi konnektiiveja, tosin konjunktiot *mais* (mutta) ja *parce que* (koska) Benevento ja Storch rajasivat pois niiden esiintymien suuren määrän vuoksi.

suuri määrä aina tee tekstistä parempaa vaan voi tuottaa liiallista kompleksisuutta, mikä saattaa hankaloittaa tekstin ymmärtämistä ja alentaa sen saamaa arvosanaa.

Syntaktisen kompleksisuuden ja kieli- tai kirjoitustaidon suhde ei olekaan yksiselitteinen, eikä kompleksisuuden lisääntyminen välttämättä merkitse parempaa kieli- tai kirjoitustaitoa (esim. Crossley & McNamara, 2014; Ortega, 2003; Pallotti, 2009; ensikielen osalta ks. esim. Jagaiah, Olinghouse & Kearns, 2020; Pajunen & Vainio, 2021a, 2021b). Esimerkiksi Lambert ja Kormos (2014) ovat todenneet, että harjaantuneet kirjoittajat saattavat käyttää aloittelijoita yksinkertaisempaa ilmaisua välittäessään monimutkaisia ajatuksia.

Aiemmissa kompleksisuuden tutkimuksissa on havaittu, että kompleksisuus voi kielitaidon eri tasoilla tai typologisesti erilaisissa kielissä ilmetä eri tavoin (esim. Bernardini & Granfeldt, 2019; Housen ym., 2019). Eroja voi olla siinä, miten kompleksisuus jonkin tietyn mittarin mukaan kehittyy (esim. Gyllstad, Granfeldt, Bernardini & Källkvist, 2014; Kuiken & Vedder, 2019) tai mitä pidetään kompleksisena. Esimerkiksi suomen kielen konjunktioalkuiset sivulauseet tai epäsuorat kysymykset eivät välttämättä ole syntaksin näkökulmasta päälauseita kompleksisempia (Martin, Mustonen, Reiman & Seilonen, 2010). Alisteisuus ei myöskään ilmene kaikissa kielissä samalla tavalla (esim. Herlin, Visapää & Kalliokoski, 2014), eivätkä alisteiset rakenteet välttämättä muodosta yhtenäistä syntaktista kategoriaa (Cristofaro, 2014). Myös lähtökieli voi vaikuttaa kohdekielen syntaktiseen kompleksisuuteen (esim. Khushik & Huhta, 2020). Tässä artikkelissa keskitytään ryhmätason kehityskulkuihin ja ensikielen tai muiden aiemmin opittujen kielten mahdollinen vaikutus on rajattu tutkimuksen ulkopuolelle.

Oppijansuomen syntaktista kompleksisuutta on tutkittu erilaisten määrällisten mittareiden avulla. Alisaari (2016) tutki taitotason A2 suomenoppijoiden T-yksiköiden sanamäärää[2] eikä löytänyt siinä tilastollisesti merkitsevää kehitystä neljän viikon kurssin aikana. Kompleksisuuden ja tarkkuuden kehitystä suomea toisena kielenä ja vieraana kielenä opiskelevien teksteissä tutkinut Tilma (2014) havaitsi kompleksisuuden lisääntyvän molemmissa ryhmissä, kun mittareina käytettiin virketyyppejä sekä virkkeiden ja lauseiden sisältämien morfeemien määrää. Oppijansuomen kompleksisuutta ja tarkkuutta dynaamisten systeemien teorian (DST) näkökulmasta tutkineet Spoelman ja Verspoor (2010) puolestaan havaitsivat, että yhdeltä oppijalta kolmen vuoden aikana kerätyssä aineistossa kompleksisuus lisääntyi epälineaarisesti, kun sitä tarkasteltiin sanojen, substantiivilausekkeiden ja virkkeiden kompleksisuutta kuvaavien suhdelukujen avulla. Tämän artikkelin aineistossa tehdyssä määrällisessä tutkimuksessa (Mylläri, 2020a) S2-tekstien syntaktista kompleksisuutta ja sen kehitystä tarkasteltiin seitsemän kansainvälisessä toisen kielen oppimisen tutkimuksessa yleisesti käytetyn määrällisen mittarin avulla. Tutkimuksen tulokset osoittivat, että ryhmätasolla sanojen määrä T-yksikössä, lauseiden määrä T-yksikössä ja sivulauseiden osuus kaikista lauseista kasvoivat ja rinnasteisten päälauseiden määrä virkkeessä pieneni kielitaidon kehittyessä, kun aineiston tehtävätyyppejä tarkasteltiin rinnakkain. Mittareiden tavoittama kehitys

2 Alisaari käytti T-yksikön sanamäärää sujuvuuden mittarina (vrt. Wolfe-Quintero ym., 1998).

ei kuitenkaan ollut lineaarista ja varianssi taitotasojen sisällä oli suurta, mikä osittain selittää tutkimuksen määrällisten mittareiden heikkoa kykyä erotella taitotasoja.

Myös konjunktioiden käyttöä oppijansuomessa on tutkittu. Alkeistason suomenoppijoiden sanastoa tutkinut Määttä (2012) totesi, että jo alkuvaiheessa teksteissä käytettiin samoja konjunktioita, jotka esiintyivät oppikirjassa: *ja*, *kun* ja *mutta*. Alakouluikäisten S1- ja S2-oppilaiden leksikaalista osaamista tutkinut Honko (2013) puolestaan havaitsi, että ylemmillä luokilla konjunktioiden käyttö monipuolistui ja niiden määrä yleisimpien sanojen joukossa lisääntyi, minkä lisäksi hän havaitsi virkerakenteiden kompleksistuvan ja monipuolistuvan alakoulun aikana. Pajunen & Vainio (2021b) ovat tarkastelleet S1-tekstien alisteisia elementtejä suhteessa kaikkiin lauseita yhdistäviin elementteihin ja todenneet, että kun tekstit jaettiin kahteen ryhmään niiden T-yksiköiden keskimääräisen sanamäärän (MLTU) tai kompleksisuusindeksin (CI) mediaanien ylä- ja alapuolelle, alisteisten elementtien osuus oli suurempi teksteissä, joiden MLTU tai CI asettuivat ryhmän mediaaniarvon yläpuolelle. Lisäksi he totesivat, että kehityksellisesti relatiivilauseet lisääntyvät ja adverbiaalilauseet vähenevät.

Tässä tutkimuksessa syntaktista kompleksisuutta lähestytään konjunktioiden syntaktisen käytön avulla, ja tarkasteluun otetaan mukaan sekä rinnasteisuus että alisteisuus. Ilmiötä tarkastellaan ryhmätasolla suhteessa kielitaidon taitotasoon. Tarkastelu rajataan vain konjunktioihin ja niillä ilmaistuihin kytköksiin. Käsiteltävät konjunktiot ja niiden jako rinnastus- ja alistuskonjunktioihin esitellään seuraavassa luvussa.

3 AINEISTO JA MENETELMÄT

Tutkimusaineisto sisältää 667 tekstiä (48 876 sanaa) aikuisilta S2-kirjoittajilta ja 411 tekstiä (16 590 sanaa) yläkouluikäisiltä S2-kirjoittajilta sekä 453 tekstin (19 826 sanan) vertailuaineiston S1-nuorilta (taulukko 1). Se on koostettu Jyväskylän yliopiston Ceflinghankkeen[3] S2-aineistosta ja S1-vertailuaineistosta, joka on Cefling-hankkeessa kerätty yläkouluikäisiltä S1-kirjoittajilta samoilla tehtävillä kuin S2-nuorten aineisto. Tehtävätyyppeinä ovat epämuodollinen viesti (esim. sähköposti ystävälle tai opettajalle), muodollinen viesti (esim. reklamaatio verkkokauppaan) ja mielipideteksti annetusta aiheesta (esim. kännykän käytöstä koulussa). Alkuperäiset tekstit on kirjoitettu käsin rajoitetussa ajassa ja ilman apuvälineitä. S2-tekstit on Cefling-hankkeessa sijoitettu Eurooppalaisen viitekehyksen mukaisille taitotasoille A1–C2 tavalla, jota voidaan pitää varsin luotettavana sekä tilastollisin että laadullisin menetelmin tarkasteltuna (Huhta, Alanen, Tarnanen, Martin & Hirvelä, 2014). Aikuisten aineistossa on tekstejä kaikilta taitotasoilta, nuorten aineistossa tasoilta A1–B2, tosin nuorten aineiston taitotasolla B2 on vain pieni määrä tekstejä. S1-tekstejä ei ole arvioitu taitotasoille. (Jantunen & Pirkola, 2015; Mustonen, 2015.) Aineisto on aiemman syntaktisen kompleksisuuden tutkimuksen yhteydessä koodattu xml-muotoon ja siihen on merkitty virke- ja lauserajat sekä lauseiden keskinäiset suhteet. Teksteistä on poistettu sanasta sanaan tehtävänannoista kopioidut otsikot sekä erityisesti viesteille tyypilliset tervehdykset ja lähettäjän

3 Cefling = Linguistic Basis of the Common European Framework for L2 English and L2 Finnish. (http://www.jyu.fi/hytk/fi/laitokset/kivi/tutkimus/hankkeet/paattyneet-tutkimushankkeet/cefling/suom)

tai vastaanottajan yhteystiedot silloin, kun ne ovat sisältäneet vain verbittömiä rakenteita. (Mylläri, 2020a; aineiston segmentoinnista tarkemmin Mylläri, 2020b.)

Tässä tutkimuksessa eri tehtävätyypit on yhdistetty. Konjunktioita ja niiden käyttöä tarkastellaan suhteessa kielitaidon taitotasoon ja erilaiset kirjoittajakohtaiset taustatekijät, kuten ensikieli tai Suomessa vietetty aika[4], sekä niiden mahdollinen vaikutus tekstien syntaktisiin piirteisiin on rajattu tarkastelun ulkopuolelle. Aikuisten ja nuorten tekstejä tarkastellaan kuitenkin erikseen, sillä aiemmissa Cefling-aineistoa käyttävissä tutkimuksissa (Kajander, 2013; Reiman, 2011; Seilonen, 2013) on havaittu, että joissakin kielenpiirteissä ja niiden muuttumisessa taitotasolta toiselle on eroja aikuisten ja nuorten välillä. S1-vertailuaineiston tekstit on tässä tutkimuksessa yhdistetty yhdeksi ryhmäksi (ensikielisen vertailuaineiston tarpeellisuudesta ks. esim. Pallotti, 2009).

TAULUKKO 1. Aineiston teksti- ja sanamäärät kirjoittajaryhmittäin ja tehtävätyypeittäin (EM = epämuodolliset viestit, M = muodolliset viestit, MP = mielipidetekstit).

Kirjoittaja-ryhmä / S2-taitotaso	**Tekstejä**				**Sanoja**			
	EM	**M**	**MP**	**Yhteensä**	**EM**	**M**	**MP**	**Yhteensä**
Aikuiset								
A1	39	22	50	**111**	1 582	752	2 261	**4 595**
A2	39	27	37	**103**	1 533	1 494	2 272	**5 299**
B1	41	42	43	**126**	2 453	2 290	5 142	**9 885**
B2	39	34	35	**108**	2 206	1 983	4 166	**8 355**
C1	26	45	46	**117**	1 528	3 337	5 876	**10 741**
C2	14	58	30	**102**	970	5 152	3 879	**10 001**
S2-nuoret								
A1	25	33	32	**90**	677	860	775	**2 312**
A2	79	40	39	**158**	2 879	1 489	1 589	**5 957**
B1	64	40	40	**144**	2 971	1 980	2 232	**7 183**
B2	12	7	·	**19**	677	461		**1 138**
Vertailuryhmä								
S1-nuoret	163	162	128	**453**	6 186	6 931	6 709	**19 826**

4 Aineiston S2-kirjoittajat ovat äidinkieleltään heterogeeninen ryhmä, eri ensikieliä on yli 20 (Jantunen & Pirkola, 2015). Kirjoittajat ovat voineet oppia suomea toisena tai vieraana kielenä tai molempina.

Tarkasteltaviksi konjunktioiksi valittiin Ison suomen kieliopin (VISK, 2004) pykälissä 813–816 käsitellyt konjunktiot, jotka jaoteltiin syntaktisen käytön mukaan rinnastus- ja alistuskonjunktioihin (taulukko 2). Vertailukonjunktio *kuin* laskettiin alistuskonjunktioksi, vaikka sillä on myös rinnastuskonjunktion piirteitä (VISK § 819). Konjunktioiden esiintymät etsittiin siten, että aineistosta kerättiin ensin koneellisesti kaikkien lauseiden ensimmäiset sanat, joiden joukosta tarkasteltavaksi valitut konjunktiot poimittiin käsin. Näin mukaan saatiin myös kirjoitusasultaan odotuksenmukaisesta poikkeavat lauseenalkuiset esiintymät. Esimerkiksi sanat *etta, ette* ja *silla* sekä puhekieliset muodot *et*, *ko*, *ku*, *kuha(n)* ja *jossei* otettiin mukaan tarkasteluun, jos ne kontekstin perusteella olivat tulkittavissa konjunktioiksi. Näin löytyneiden eri kirjoitusasujen kaikki esiintymät kerättiin aineistosta koneellisesti virkekonteksteineen, jolloin mukaan saatiin myös muut kuin lauserajalla olevat esiintymät.

Tarkastelun ulkopuolelle rajattiin tapaukset, joissa sanaa käytettiin muuten kuin konjunktiona. Löydettyjen esiintymien joukosta poistettiin käsin kysymyssanana käytetty *koska*, adverbina käytetty *vaan* (esim. *missä vaan*), sekä sanat *et* (< *ei*), *mut* (< *minut*) ja *sillä* (< *se*), kun niitä käytettiin kieltoverbin tai pronominin taivutusmuotoina. Lisäksi pois rajattiin *vaikka* ja *vaikkapa*, kun niitä käytettiin muuten kuin konjunktiona (esim. *tavataan vaikka keskiviikkona*). Konjunktiot koodattiin aineistoon, ja esiintymien määrät laskettiin koneellisesti. Parikonjunktio *sekä – että* koodattiin ja laskettiin yhdeksi rinnastuskonjunktioksi. Lopuksi tarkistettiin koneellisesti, ettei aineistossa ole sellaisia konjunktioita, jotka on mainittu VISK:n pykälissä 813–816 mutta joita ei esiintynyt lauserajoilla. Esiintymien koneellisessa poiminnassa ja laskemisessa hyödynnettiin tätä tutkimusta varten kirjoitettuja Python-skriptejä.

TAULUKKO 2. Tutkimuksessa tarkasteltavat konjunktiot.

Aineistossa esiintyvät konjunktiot		**Aineistossa esiintymättömät konjunktiot**
Rinnastuskonjunktiot	**Alistuskonjunktiot**	
-kä, eli, elikä, ja, mutta (mut), muttei, sekä, sillä, tai, vaan, vai; sekä – että, joko – tai, joko – taikka, niin – kuin	*ellei, että (et), ettei (etteikö), jos (jossei), jotta, jottei, koska, kuin, kun (ku, ko), kunhan, kunnes, mikäli, vaikka, vaikkapa; ennen kuin, niin kuin, kun taas*	*jahka, jollei, joskin, josko, kunpa, saati, sun, taikka, ynnä, vaikkakaan, vaikkakin*

Konjunktioiden esiintymistä tarkastellaan ensisijaisesti ryhmätasolla ja aineiston tilastollisessa kuvailussa kunkin taitotason tekstejä käsitellään tekstimassana. Koska tekstien lukumäärä ja sanamäärä vaihtelevat taitotasolta toiselle, taitotasojen ja kirjoittajaryhmien rinnakkaisessa tarkastelussa käytetään konjunktioiden normalisoituja esiintymistaajuuksia suhteessa tuhanteen sanaan. Eri konjunktioiden absoluuttiset ja normalisoidut esiintymistaajuudet on esitetty liitteen 1 taulukoissa 1A (aikuisten aineisto) ja 1B (nuorten aineisto

sekä S1-vertailuaineisto). Aineistossa esiintyvien konjunktioiden valikoiman tarkastelussa käytetään myös konjunktioiden suhteellisia frekvenssejä (f%), jotka on laskettu jakamalla yksittäisen konjunktion esiintymien määrä kullakin taitotasolla kaikkien saman taitotason konjunktioesiintymien kokonaismäärällä.

Taitotasojen tilastollista vertailua varten laskettiin myös tekstikohtaiset konjunktioiden esiintymistaajuudet, jotka normalisoitiin ryhmätason tarkastelusta poiketen suhteessa sataan sanaan, koska kaikilla taitotasoilla samoin kuin S1-aineistossa tekstien keskimääräinen pituus on alle sata sanaa. Vertailussa käytettiin ei-parametrisiä testejä, sillä aineisto ei Shapiro-Wilkin ja Levenen testien mukaan täyttänyt parametristen tilastollisten testien taustaoletuksia jakauman normaaliuden ja varianssin homogeenisuuden osalta. Tilastollisina testeinä käytettiin Kruskal-Wallisin testiä ja parittaisissa vertailuissa Mann-Whitneyn U-testiä vastaavaa Wilcoxonin järjestyslukujen summan testiä, jonka p-arvoihin tehtiin Bonferroni-korjaus. Tilastollisen merkitsevyyden rajana käytettiin arvoa $p < 0{,}05$. Tilastolliset analyysit tehtiin R-ohjelmiston versiolla 3.4.4 RStudion version 1.1.456 avulla.

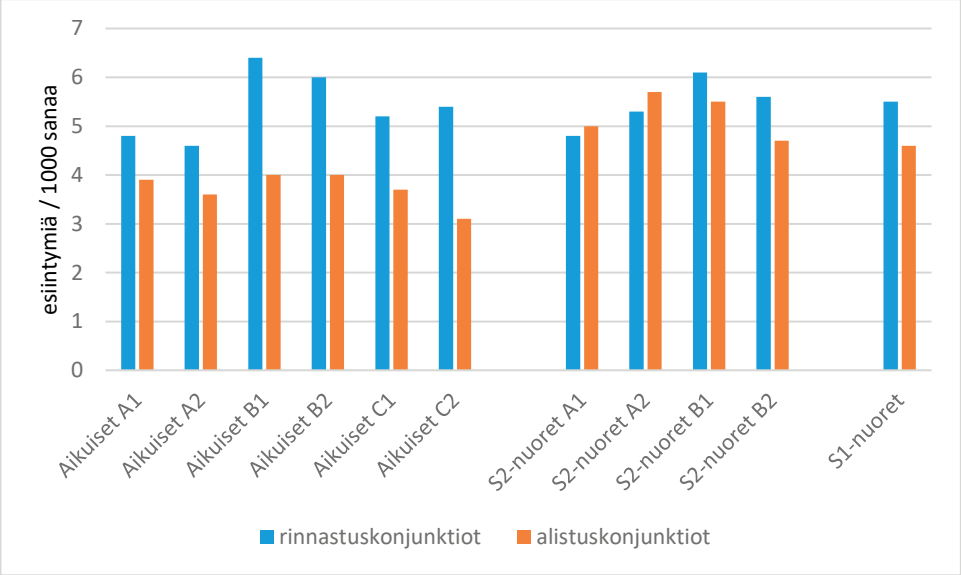
Tarkastelu aloitetaan rinnastus- ja alistuskonjunktioiden esiintymistä eri taitotasoilla. Tässä yhteydessä testataan, onko taitotasojen välillä tilastollisesti merkitseviä eroja kaikkien konjunktioiden sekä erikseen rinnastus- ja alistuskonjunktioiden normalisoiduissa esiintymistaajuuksissa. Sen jälkeen selvitetään, mitä kielenyksiköitä rinnastuskonjunktioilla yhdistetään ja miten suuri osa aineiston sivulauseista sisältää alistuskonjunktion. Lopuksi kartoitetaan kullekin taitotasolle tyypillistä konjunktioiden varantoa selvittämällä, mitä konjunktioita eri taitotasoilla esiintyy, sekä laskemalla yksittäisten konjunktioiden suhteelliset frekvenssit ja se, miten monessa eri tekstissä ne esiintyvät (vrt. Michel, 2013).

4 TULOKSET

Sekä rinnastuskonjunktioita että alistuskonjunktioita käytetään aineiston kaikilla taitotasoilla. Jo taitotasolla A1 aikuisten aineistossa 89 % ja nuorten aineistossa 86 % teksteistä sisältää vähintään yhden konjunktion ja noin puolessa teksteistä (aikuisten aineistossa 51 prosentissa ja nuorten aineistossa 46 prosentissa) on käytetty sekä alistus- että rinnastuskonjunktioita molempia vähintään kerran. Aineiston alimmilla taitotasoilla on myös täysin konjunktiottomia tekstejä: taitotasolla A1 aikuisten aineistossa 12 (11 %) ja nuorten aineistossa 13 (14 %), taitotasolla A2 aikuisten aineistossa 7 (7 %) ja nuorten aineistossa 3 (2 %). Lisäksi aikuisten aineiston taitotasolla B1 on yksi teksti, jossa ei käytetä konjunktioita. Vaikka ilmiö näyttää olevan tyypillinen S2-aineiston alimmille taitotasoille, myös S1-nuorten vertailuaineistossa on 20 täysin konjunktiotonta tekstiä (4 % kaikista S1-teksteistä). Tyypillisesti kirjoittajat käyttävät sekä rinnastus- että alistuskonjunktioita.

Aikuisten aineistossa sekä rinnastus- että alistuskonjunktioiden normalisoitu esiintymistaajuus on suurimmillaan taitotasoilla B1 ja B2, nuorten aineistossa puolestaan alistuskonjunktioiden taitotasoilla A2 ja B1 ja rinnastuskonjunktioiden taitotasolla B1 (kuvio 1). Rinnastuskonjunktioiden normalisoitu esiintymistaajuus on lähes kaikilla taitotasoilla suurempi kuin alistuskonjunktioiden eli rinnastuskonjunktiot ovat yleisempiä kuin alistuskonjunktiot. Ainoa poikkeus tästä ovat S2-nuorten aineiston taitotasot A1 ja

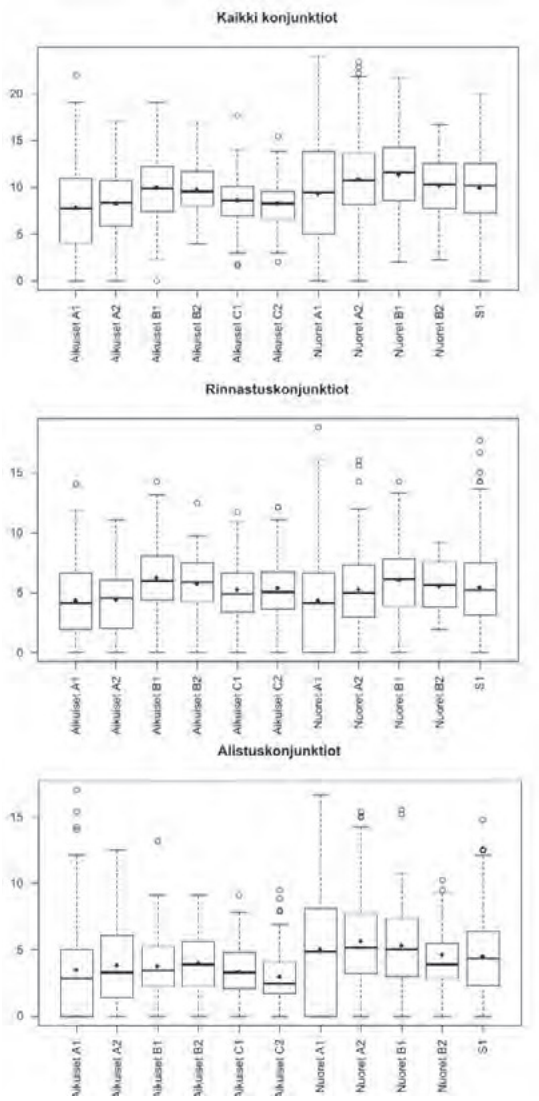
A2, joilla alistuskonjunktioiden normalisoitu esiintymistaajuus on hieman suurempi kuin rinnastuskonjunktioiden. Tältä osin aikuisten ja nuorten aineistot poikkeavat toisistaan. S1-vertailuaineistossa rinnastuskonjunktiot ovat yleisempiä kuin alistuskonjunktiot, mikä vastaa yleissuomesta aiemmin saatuja tuloksia (Saukkonen, Haipus, Niemikorpi & Sulkala, 1979, s. 10).



KUVIO 1. Rinnastus- ja alistuskonjunktioiden normalisoidut esiintymistaajuudet.

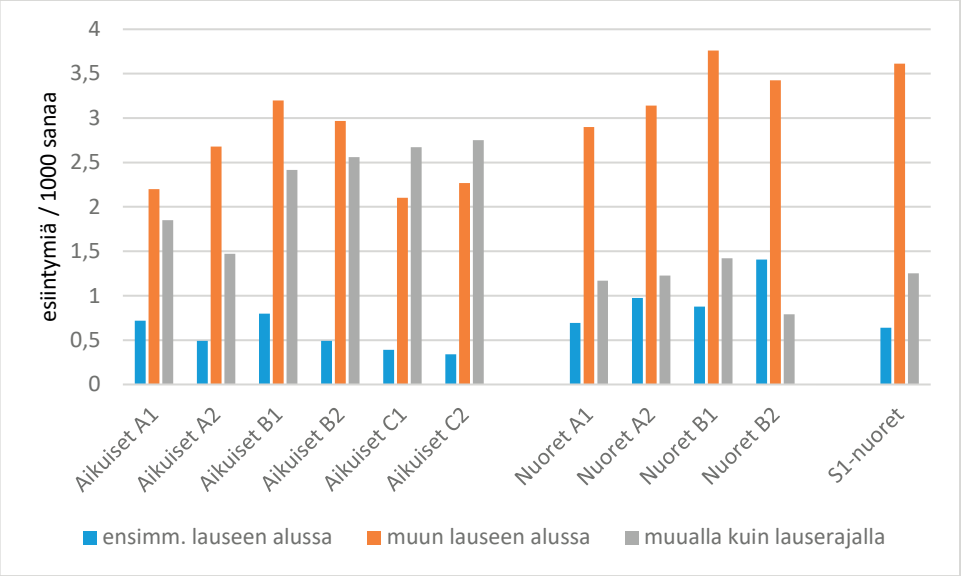
Kaikkien konjunktioiden, rinnastuskonjunktioiden ja alistuskonjunktioiden normalisoituja esiintymistaajuuksia (kuvio 2) verrattiin tilastollisin menetelmin. Aikuisten aineistossa taitotasojen välillä on Kruskal-Wallisin testin mukaan tilastollisesti merkitseviä eroja kaikkien konjunktioiden (*Khiin neliö* = 35,29; *df* = 5; $p < 0{,}001$), rinnastuskonjunktioiden (*Khiin neliö* = 36,17; *df* = 5; $p < 0{,}001$) ja alistuskonjunktioiden (*Khiin neliö* = 16,15; *df* = 5; $p = 0{,}006$) normalisoidussa esiintymistaajuudessa. Kaikkien konjunktioiden esiintymistaajuus on suurimmillaan B-tasolla, ja erot sekä alimpiin että ylimpiin taitotasoihin ovat Wilcoxonin järjestyslukujen summan testin mukaan tilastollisesti merkitseviä. Taitotaso B1 eroaa tilastollisesti merkitsevästi taitotasoista A1 ($W = 4933$; $p = 0{,}001$), A2 ($W = 4983$; $p = 0{,}038$), C1 ($W = 5657$; $p = 0{,}026$) ja C2 ($W = 4521$; $p = 0{,}002$), taitotaso B2 puolestaan taitotasoista A1 ($W = 4194{,}5$; $p = 0{,}002$), C1 ($W = 5657$; $p = 0{,}025$) ja C2 ($W = 3752$; $p = 0{,}001$). Rinnastuskonjunktioiden normalisoidussa esiintymistaajuudessa on aikuisten aineistossa tilastollisesti merkitseviä eroja tasojen A ja B välillä. Taitotaso A1 eroaa tilastollisesti merkitsevästi taitotasoista B1 ($W = 4653$; $p < 0{,}001$) ja B2 ($W = 4290$; $p = 0{,}004$), samoin taitotaso A2 taitotasoista B1 ($W = 4332$; $p < 0{,}001$) ja B2 ($W = 3963$; $p = 0{,}005$). Aikuisten aineistossa alistuskonjunktioiden normalisoidussa esiintymistaa-

juudessa ainoa tilastollisesti merkitsevä ero on taitotasojen B2 ja C2 välillä ($W = 7071$; $p = 0{,}006$). Nuorten aineistossa sekä kaikkien konjunktioiden että rinnastuskonjunktioiden normalisoitu esiintymistaajuus kasvavat taitotasolta A1 taitotasolle B1 ja alistuskonjunktioiden taitotasolta A1 taitotasolle A2. Kruskal-Wallisin testin mukaan nuorten aineistossa on tilastollisesti merkitseviä eroja taitotasojen välillä kaikkien konjunktioiden (*Khiin neliö* = 10,14; $df = 3$; $p = 0{,}017$) ja rinnastuskonjunktioiden (*Khiin neliö* = 16,98; $df = 3$; $p < 0{,}001$) normalisoidussa esiintymistaajuudessa. Parittaisten vertailujen perusteella taitotasojen A1 ja B1 välillä on tilastollisesti merkitsevä ero sekä kaikkien konjunktioiden ($W = 5023$; $p = 0{,}023$) että rinnastuskonjunktioiden ($W = 4474{,}5$; $p < 0{,}001$) esiintymistaajuudessa. Alistuskonjunktioiden esiintymistaajuudessa ei nuorten aineistossa ole Kruskal-Wallisin testin perusteella tilastollisesti merkitseviä eroja taitotasojen välillä (*Khiin neliö* = 2,86; $df = 3$; $p = 0{,}414$).



KUVIO 2. Tekstikohtaiset normalisoidut esiintymistaajuudet (esiintymiä / 100 sanaa).

Koska rinnastuskonjunktioita voidaan käyttää erilaisten syntaktisten elementtien yhdistämiseen, rinnastuskonjunktioiden esiintymiä tarkasteltiin myös jakamalla ne kolmeen ryhmään (kuvio 3): virkkeen ensimmäisen (tai ainoan) lauseen aloittaviin, virkkeen sisällä kahta lausetta yhdistäviin sekä muualla kuin lause- tai virkerajalla esiintyviin (vrt. Vyatkina, 2013). Yleisimpiä ovat virkkeen sisällä lauseita yhdistävät rinnastuskonjunktiot, joita on aikuisten aineistossa taitotasoilla A1–B2 noin 50 %, S2-nuorten aineistossa noin 60 % ja vertailuaineistossa 66 % kaikista rinnastuskonjunktioista. Aikuisten aineiston kaksi ylintä taitotasoa poikkeavat muusta aineistosta. Niillä rinnastuskonjunktioista noin 40 % yhdistää saman virkkeen lauseita ja noin puolet (C1-tasolla 52 % ja C2-tasolla 51 %) rinnastuskonjunktioista sijoittuu muualle kuin lause- tai virkerajalle. Virkkeenalkuisten rinnastuskonjunktioiden osuus vaihtelee aikuisten aineistossa A1-tason 15 prosentista C2-tason 6 prosenttiin, S2-nuorten aineistossa 14 ja 25 prosentin välillä. S1-nuorten teksteissä rinnastuskonjunktioista 12 % aloittaa virkkeen, mikä on hyvin lähellä suomenkielisen asiaproosan aineistosta aiemmin raportoituja tuloksia, joiden mukaan 13 % rinnastuskonjunktiolla alkavista päälauseista sijaitsi virkkeen alussa (Hakulinen, Karlsson & Vilkuna, 1980).



KUVIO 3. Rinnastuskonjunktioiden normalisoidut esiintymistaajuudet esiintymispaikan mukaan ryhmiteltyinä.

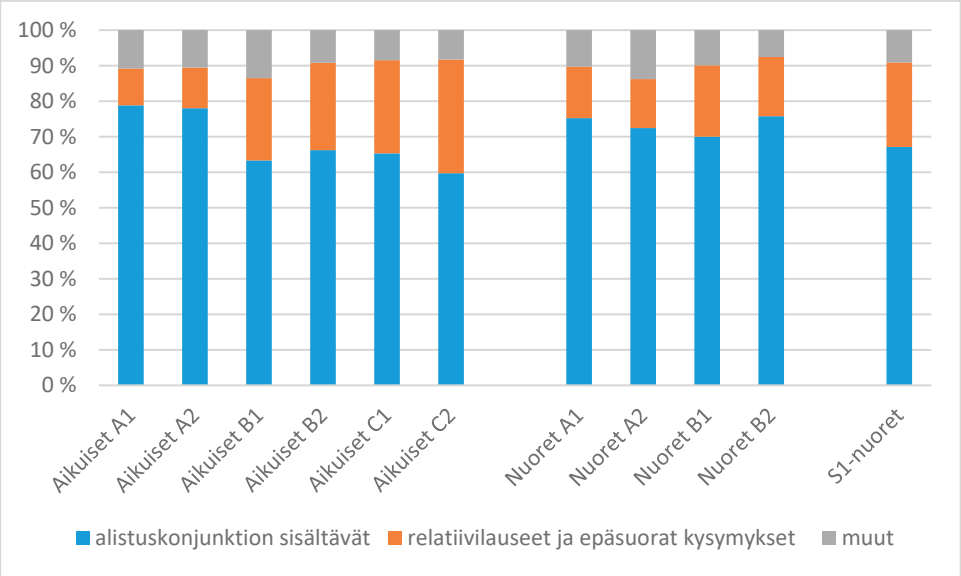
Kaikki sivulauseet eivät välttämättä sisällä alistuskonjunktiota. Siksi alistuskonjunktioiden esiintymiä tarkasteltiin myös suhteessa sivulauseiden määrään. Tarkasteltua varten aineiston sivulauseet luokiteltiin kolmeen ryhmään: 1) alistuskonjunktion sisältävät sivulauseet, 2) relatiivilauseet ja epäsuorat kysymykset sekä 3) muut sivulauseet (vrt. Vilkuna, 2003, s. 68). Sekä rinnastuskonjunktion että alistuskonjunktion sisältävät sivulauseet (esimerkki1) ja esimerkin 2 kaltaiset sivulauseet luokiteltiin alistuskonjunktion sisältäviksi. Muihin sivulauseisiin laskettiin sivulauseet, joissa lauseraja tai lauseen alisteisuus on tulkinnanvarainen, sekä edelliselle sivulauseelle rinnasteiset sivulauseet, joiden pintarakenteesta alistuskonjunktio on jätetty pois (esimerkki 3). Tällaisissa sivulauseissa ellipsi ei välttämättä rajoitu alistuskonjunktioon (esimerkki 4).

Esimerkki 1. *Toivon että otatte minuun yhteyttä pikaisesti ja että saan uuden kahvinkeittimen tilalle.* (F-704 aikuiset B2)

Esimerkki 2. *Halusin kysyä, että milloin meillä on äidinkielen koe ja mitkä kappaleet pitää lukea?* (F-041 nuoret B1)

Esimerkki 3. *Huomasin heti, että paketti oli huonossa kunnossa ja levyn kansi oli rikki.* (F-653 aikuiset B1)

Esimerkki 4. *Maailma olisi paljon kauniimpi paikka, jos kaikki kävelisivät tai ottaisivat bussia*. (F-689 aikuiset B2)



KUVIO 4. Alistuskonjunktion sisältävien sivulauseiden, relatiivilauseiden ja epäsuorien kysymyslauseiden sekä muiden sivulauseiden osuus kaikista sivulauseista.

Suurin osa aineiston sivulauseista sisältää alistuskonjunktion: aikuisten aineistossa 60–79 %, S2-nuorten aineistossa 70–76 % ja S1-nuorten aineistossa 67 % (kuvio 4). Relatiivilauseiden ja epäsuorien kysymysten osuus kasvaa aikuisten aineistossa A1-tason 10 prosentista C2-tason 32 prosenttiin, nuorten aineistossa A-tason noin 14 prosentista B1-tason 20 prosenttiin. S1-nuorten vertailuaineistossa relatiivilauseiden ja epäsuorien kysymysten osuus on 24 %. Muiden sivulauseiden osuus vaihtelee S2-aineistossa 8 ja 14 prosentin välillä ja on S1-vertailuaineistossa 9 % kaikista sivulauseista. Tämän aineiston perusteella alistuskonjunktiot voivat toimia sivulauseiden likiarvona alimmilla taitotasoilla, mutta jo niillä esiintyy myös muita kuin alistuskonjunktion sisältäviä sivulauseita.

Kun konjunktioiden valikoimaa tarkastellaan ryhmätasolla, havaitaan, että koko aineistossa yleisimmät konjunktiot *ja*, *koska*, *että*, *mutta*, *kun, jos* ja *tai* ovat käytössä jo taitotasolla A1. Muita heti taitotasolla A1 ilmaantuvia konjunktioita ovat *kuin* ja aikuisten aineistossa *eli*, nuorten aineistossa lisäksi *vai*, *eikä*, *vaikka*, *sillä* ja *ellei*. Osa näistä konjunktioista tosin on taitotasolla A1 harvinaisia ja niistä on vain yksittäisiä esiintymiä. Aikuisten aineistossa eri konjunktioiden määrä kasvaa selvimmin taitotasojen A2 ja B1 välillä, jolloin se yli kaksinkertaistuu 11 konjunktiosta 25 eri konjunktioon. Osaa tästä muutoksesta selittävät vain kerran esiintyvät konjunktiot, joita on taitotasolla B1 kahdeksan ja taitotasolla B2 seitsemän, kun niitä tasoilla A2 ja C1 on vain kaksi molemmilla. Nuorten aineistossa valikoiman laajeneminen on tasaisempaa ja vain kerran esiintyvien konjunktioiden määrä vaihtelee neljän (taitotasot A1 ja B2) ja viiden (taitotasot A2 ja B1) välillä. Valikoiman laajeneminen ei ryhmätasolla ole selkeästi kumulatiivista, sillä kaikki jollakin taitotasolla esiintyvät konjunktiot eivät välttämättä esiinny seuraavalla tai muilla ylemmillä taitotasoilla (taulukot 3 ja 4). Käytettyjen konjunktioiden valikoiman laajuutta tarkasteltaessa on otettava huomioon, ettei konjunktioiden käyttö jakaudu tasaisesti tekstien välillä. Mikään konjunktioista ei esiinny kaikissa teksteissä millään taitotasolla eikä S1-aineistossa, ei edes aineistossa kaikkein yleisin konjunktio *ja*.

TAULUKKO 3. Konjunktiot yleisyysjärjestyksessä ja useammin kuin kerran esiintyvien konjunktioiden suhteelliset frekvenssit (f%) aikuisten aineistossa taitotasoittain.

Aikuiset A1	**Aikuiset A2**	**Aikuiset B1**	**Aikuiset B2**	**Aikuiset C1**	**Aikuiset C2**
ja (43) *koska* (20) *että* (10) *mutta* (10) *kun* (8) *jos* (5) *kuin* (2) *tai* (1) *eli* (1)	*ja* (42) *että* (16) *koska* (15) *mutta* (10) *jos* (6) *kun* (5) *tai* (3) *kuin* (2) *vai* (0) *eikä*, *eli*	*ja* (41) *että* (16) *mutta* (11) *jos* (6) *koska* (6) *kun* (6) *tai* (6) *kuin* (3) *eli*, *vaikka* (1) *eikä* (1) *sekä* (1) *sekä – että* (0) *vai* (0) *niin kuin* (0) *ettei*, *vaan* (0) *elikä*, *joko – tai*, *jotta*, *kunhan*, *kunnes*, *muttei*, *niin – kuin*, *sillä*	*ja* (40) *että* (15) *mutta* (10) *jos* (7) *kun* (6) *koska* (4) *tai* (4) *kuin* (3) *eikä* (2) *sekä* (1) *eli* (1) *sekä – että*, *vaikka* (1) *jotta*, *vaan* (1) *ettei* (1) *sillä*, *vai* (0) *elikä*, *joko – tai*, *kunnes*, *mikäli*, *muttei*, *niin kuin*, *niin – kuin*	*ja* (42) *että* (18) *jos* (7) *mutta* (6) *kun* (5) *tai* (3) *kuin* (3) *koska* (3) *eikä*, *vaikka* (2) *ettei* (2) *sekä* (1) *sillä* (1) *vai* (1) *eli*, *ennen kuin*, *jotta*, *vaan* (1) *sekä – että* (1) *ellei*, *niin kuin* (0) *joko – tai*, *muttei* (0) *kunhan*, *vaikkapa*	*ja* (45) *että* (17) *mutta* (6) *jos* (6) *kun* (4) *koska* (3) *tai* (3) *kuin* (3) *eikä* (2) *sekä* (2) *vaan* (1) *eli* (1) *ettei*, *vai* (1) *sillä* (1) *jotta* (1) *ennen kuin*, *vaikka* (1) *sekä – että* (1) *kunhan* (0) *kun taas*, *kunnes*, *mikäli*

TAULUKKO 4. Konjunktiot yleisyysjärjestyksessä ja useammin kuin kerran esiintyvien konjunktioiden suhteelliset frekvenssit (f%) S2-nuorten aineistossa taitotasoittain ja S1-nuorten aineistossa.

S2-nuoret A1	S2-nuoret A2	S2-nuoret B1	S2-nuoret B2	S1-nuoret
ja (31) *koska* (21) *jos* (12) *kun* (8) *mutta* (8) *että* (7) *tai* (7) *kuin*, *vai*, *vaikka* (1) *eikä* (1) *eli*, *ellei*, *sillä*	*ja* (31) *koska* (17) *että* (14) *jos* (10) *mutta* (8) *kun* (7) *tai* (7) *ettei* (1) *eikä*, *kuin* (1) *eli*, *sillä*, *vaikka* (1) *jotta*, *kunhan*, *niin kuin*, *sekä – että*, *vai*	*ja* (32) *että* (17) *koska* (11) *jos* (10) *mutta* (10) *tai* (7) *kun* (7) *eikä* (1) *eli* (1) *ettei* (1) *kuin*, *vaikka* (1) *jotta* (0) *vaan* (0) *kunhan*, *vai* (0) *ennen kuin*, *joko – tai*, *sekä*, *sekä – että*, *sillä*	*ja* (34) *että* (19) *kun* (13) *mutta* (11) *jos*, *koska*, *tai* (6) *eikä* (3) *eli*, *kuin*, *vaikka*	*ja* (31) *että* (14) *jos* (12) *mutta* (10) *kun* (9) *tai* (7) *koska* (6) *eikä* (3) *sillä* (2) *kuin* (1) *ettei* (1) *eli*, *vaikka* (1) *vaan* (1) *sekä* (0) *vai* (0) *jotta*, *niin kuin* (0) *joko – tai* (0) *ellei*, *ennen kuin*, *kunhan* (0) *mikäli* (0) *etteikö*, *joko – taikka*, *jottei*, *kunnes*, *niin – kuin*, *sekä – että*

S2-aineiston kaikilla taitotasoilla ja S1-nuorten vertailuaineistossa yleisin konjunktio on rinnastuskonjunktio *ja*. Sen esiintymät kattavat aikuisten aineistossa 40–45 %, S2-nuorten aineistossa 31–34 % ja S1-nuorten vertailuaineistossa 31 % kaikista konjunktioiden esiintymistä. Aineiston toiseksi yleisin konjunktio on aikuisten aineiston taitotasolla A1 ja nuorten aineiston taitotasoilla A1 ja A2 alistuskonjunktio *koska*, jonka osuus on taitotasolla A1 noin 20 % kaikista konjunktioista. Seuraavilla taitotasoilla toiseksi yleisimmäksi konjunktioksi nousee *että*, joka taitotason A1 jälkeen kattaa aikuisten aineistossa 15–18 % ja nuorten aineistossa 14–19 % kaikkien konjunktioiden esiintymistä. Se on toiseksi yleisin konjunktio myös S1-vertailuaineistossa, jossa sen osuus kaikista konjunktioista on 14 %. Toiseksi yleisin rinnastuskonjunktio on kaikilla taitotasoilla ja myös S1-aineistossa *mutta*, joka kattaa 6–11 % kaikista konjunktioiden esiintymistä ja on aikuisten aineiston eri taitotasoilla kolmanneksi tai neljänneksi yleisin konjunktio, nuorten aineistossa nel-

jänneksi tai viidenneksi yleisin.

Suurin osa aineiston konjunktioista on yksiosaisia. Pari- ja liittokonjunktiot ovat harvinaisia niin S2- kuin S1-aineistossa. Yleisin on *sekä – että*, joka ilmaantuu aikuisten teksteihin taitotasolla B1 mutta esiintyy S2-nuorten teksteissä vain kahdesti ja S1-vertailuaineistossa vain kerran. Aikuisten aineistossa yleisin liittokonjunktio on vain C-tasolla esiintyvä *ennen kuin*, S1-aineistossa puolestaan *niin kuin*, joka esiintyy joitakin kertoja aikuisten aineiston kolmella ylimmällä taitotasolla. Kumpikin esiintyy S2-nuorten aineistossa vain kerran. Myös osa yksiosaisista konjunktioista on aineistossa harvinaisia. Tällainen on esimerkiksi *kunnes*, joka esiintyy S2-aineistossa vain aikuisten teksteissä muutaman kerran ja S1-vertailuaineistossa kerran. S1-vertailuaineistossa on kolme vain kerran esiintyvää konjunktiota, jotka eivät esiinny S2-aineistossa lainkaan, S2-aikuisten aineistosta puolestaan löytyy B- ja C-tasolta neljä konjunktiota, joita ei S2- ja S1-nuorten aineistossa ole käytetty lainkaan. Näiden esiintymistaajuus jää kuitenkin pieneksi, ja osa esiintyy koko aineistossa vain kerran. Aikuisten ja nuorten aineisto eroavat toisistaan myös konjunktioiden puhekielisten kirjoitusasujen osalta. Puhekielisiä kirjoitusasuja esiintyy lähes yksinomaan nuorten aineistossa[5], ja ne ovat yleisempiä S1-nuorten kuin S2-nuorten teksteissä.

Aineistossa yleisimmät konjunktiot ilmaantuvat S2-oppijoiden teksteihin heti taitotasolla A1, tosin joidenkin konjunktioiden kohdalla kyse on yksittäisistä esiintymistä. Kerran tai kaksi esiintyviä konjunktioita ei voi pitää taitotasolle tyypillisinä, mutta ne havainnollistavat, ettei harvinaisen tai ylemmille taitotasoille tyypillisen konjunktion käytön ja kielitaidon tason välillä välttämättä ole yhteyttä. Tähän samaan viittaa se, ettei käyttävissä olevien konjunktioiden varannon laajeneminen näy ryhmätasolla systemaattisena teksteissä käytettyjen konjunktioiden valikoiman laajentumisena eivätkä kaikki jollakin taitotasolla käytetyt konjunktiot välttämättä esiinny seuraavalla tai muilla ylemmillä taitotasoilla. Esimerkiksi aikuisten aineistossa harvinainen ja vain taitotasolla C1 muutaman kerran esiintyvä *ellei* esiintyy nuorten aineistossa vain kerran, taitotasolla A1.

5 POHDINTA

Tämän tutkimuksen aineistossa S2-kirjoittajat käyttävät sekä rinnastus- että alistuskonjunktioita kaikilla taitotasolla. Konjunktioiden ilmaantuminen S2-teksteihin tukee aiempia tuloksia konjunktioiden käytöstä jo kielitaidon alkuvaiheessa (Määttä, 2012; oppijansaksassa Vyatkina, 2012, 2013). Tulokset eivät tältä osin tue EVK:n taitotasokuvauksista välittyvää näkemystä, että syntaktisten kytkösten merkitseminen konjunktioilla alkaisi yleistyä vasta taitotasolla A2, sillä tämän tutkimuksen aineistossa erot konjunktioiden normalisoidussa esiintymistaajuudessa eivät näytä olevan tilastollisesti merkitseviä taitotasojen A1 ja A2 välillä. Toisaalta tulokset myös osittain tukevat EVK:n taitotasokuvauksia, sillä täysin konjunktiottomien tekstien osuus kaikista teksteistä on suurin juuri taitotasolla A1 ja taitotasolla A2 niiden osuus on aikuisten aineistossa puolet ja nuorten aineistossa kolme neljäsosaa pienempi kuin taitotasolla A1. Jo taitotasolla A1 yli 80 % teksteistä kuitenkin sisältää vähintään yhden konjunktion.

Konjunktioita käytetään normalisoitu-

5 Puhekielisistä kirjoitusasuista vain *et* esiintyy yhden kerran aikuisten aineistossa (taitotasolla B1).

jen esiintymistaajuuksien perusteella eniten suhteessa kokonaissanamäärään aikuisten aineistossa taitotasoilla B1 ja B2, nuorten aineistossa taitotasoilla A2 ja B1. Normalisoiduissa esiintymistaajuuksissa on tilastollisesti merkitseviä eroja taitotasojen välillä aikuisten aineistossa lähinnä tasojen A ja B välillä, nuorten aineistossa taitotasojen A1 ja B1 välillä. Vaikka tasojen A ja C välillä ei ole tilastollisesti merkitseviä eroja konjunktioiden normalisoiduissa esiintymistaajuuksissa, tulosta ei voi tulkita niin, että konjunktioiden käyttö tai syntaktinen kompleksisuus olisivat samanlaista aineiston alimmilla ja ylimmillä taitotasoilla. Valittu tarkastelutapa kuvaa konjunktioiden suhteellista yleisyyttä eikä ota huomioon laadullisia eroja konjunktioiden käytössä tai syntaktisessa kompleksisuudessa. Lisäksi lauseiden keskimääräinen sanamäärä kasvaa aineistossa taitotason noustessa (Mylläri, 2020a), millä voi olla vaikutusta normalisoituihin esiintymistaajuuksiin.

Tilastollisessa analyysissa näkyviin taitotasojen välisiin eroihin vaikuttaa myös taitotasojen sisäinen varianssi, joka kertoo siitä, että samalle taitotasolle sijoitetut tekstit eroavat toisistaan siinä, miten paljon niissä on käytetty konjunktioita. Sekä aikuisten että nuorten aineistossa varianssi on suurinta taitotasolla A1, minkä jälkeen se vähenee. Taitotasojen sisäistä varianssia ja taitotasojen välisten tilastollisesti merkitsevien erojen vähyyttä voi osin selittää se, että aineiston eri taitotasoilla eri tehtävätyyppien osuus teksteistä vaihtelee, sillä tehtävätyyppien tai tekstilajien välillä voi olla eroja syntaktisessa kompleksisuudessa (esim. Michel, 2017; Pallotti, 2009). Lisäksi Cefling-aineistossa on havaittu tehtävätyyppien välisiä eroja joidenkin kielenpiirteiden kehityksessä taitotasolta toiselle (Kajander, 2013; Reiman, 2014; Seilonen, 2013). Taitotasojen sisäiseen varianssiin ja taitotasojen välisten tilastollisesti merkitsevien erojen vähyyteen voivat vaikuttaa myös eri taustatekijöistä johtuvat yksilölliset erot, joiden vaikutus konjunktioiden käyttöön ja sen avulla tarkasteltavaan syntaktiseen kompleksisuuteen on rajattu tämän tutkimuksen ulkopuolelle.

Tutkimuksen aineistossa rinnastuskonjunktiot ovat alistuskonjunktioita yleisempiä kaikilla muilla taitotasoilla kuin nuorten aineiston taitotasoilla A1 ja A2. Alistuskonjunktioiden osuus kaikista konjunktioista ei ylemmillä taitotasoilla ole suurempi kuin alemmilla. Tulos poikkeaa Vyatkinan (2012) tuloksista, joiden mukaan rinnastuskonjunktioiden ja alistuskonjunktioiden esiintymistaajuuden välillä vallitsi ryhmätasolla negatiivinen korrelaatio niin, että alistuskonjunktioiden määrän lisääntyessä rinnastuskonjunktioiden määrä väheni. Rinnastuskonjunktioiden käytön tarkempi tarkastelu osoitti, että aikuisten aineiston C-tasolla yli puolet rinnastuskonjunktioiden esiintymistä sijaitsee muualla kuin lauserajalla. Tämä tulos tukee näkemystä, että kielitaidon ylemmillä taitotasoilla syntaktinen kompleksisuus voi olla erilaista kuin perus- tai keskitasolla ja että ylimmillä taitotasoilla kompleksisuuden kehitys voi näkyä lauseiden yhdistämisen sijasta lausekkeiden yhdistämisessä (esim. Norris & Ortega, 2009).

Sivulauseiden tarkempi tarkastelu puolestaan osoitti, että alistuskonjunktion sisältävien sivulauseiden osuus kaikista sivulauseista on ylemmillä taitotasoilla pienempi kuin alemmilla taitotasoilla, mikä viittaa siihen, että kirjoittajien käyttämät alisteiset rakenteet monipuolistuvat kielitaidon ylemmillä taitotasoilla. Jos syntaktista kompleksisuutta

tarkastellaan kielenpiirteiden monipuolisuutena (esim. Ortega, 2003), tulosta voidaan tulkita myös syntaktisen kompleksisuuden lisääntymisenä ryhmätasolla. Alistuskonjunktion sisältävien sivulauseiden vähenevä osuus kaikista sivulauseista on myös linjassa sen kanssa, että relatiivilauseiden on todettu lisääntyvän kehityksellisesti niin toisessa kielessä (esim. Ortega, 2009) kuin ensikielessä (esim. Pajunen & Vainio, 2021b). Relatiivilauseiden ja epäsuorien kysymysten osuuden kasvu voi vaikuttaa myös siihen, että alistuskonjunktioiden normalisoitu esiintymistaajuus on C-tasolla pienempi kuin B-tasolla, sillä sivulauseiden osuus kaikista lauseista ei tässä aineistossa kasva tilastollisesti merkitsevästi näiden tasojen välillä (Mylläri, 2020a).

Yleisimpien konjunktioiden ilmaantuminen ja konjunktioiden suhteellisen suuri yhteenlaskettu esiintymistaajuus jo taitotasolla A1 viittaavat siihen, että aikuiset ja yläkouluikäiset S2-kirjoittajat hallitsevat erilaisia syntaktisia yhdistämiskeinoja jo kielitaidon alimmilla taitotasoilla. Tämä ei sinänsä ole yllättävää, sillä sekä rinnasteisten että alisteisten suhteiden ilmaisemisen voi olettaa olevan aikuisille ja myös yläkouluikäisille suomenoppijoille tuttua aiemmin opituista kielistä, jolloin kyse on enemmän olemassa olevan taidon siirtämisestä uuteen kieleen kuin uusista syntaktisten kytkösten luomisen keinoista. Tulokset tukevat näkemyksiä siitä, että toisen kielen taidon kehittymistä ei välttämättä ole perusteltua tarkastella sivulauseiden määrään tai T-yksikön pituuteen pohjautuvilla mittareilla, jotka on kehitetty ensikieltä oppivien lasten kielitaidon kehittymisen seuraamiseen (esim. Bardovi-Harlig, 1992). Samalla ne tukevat näkemyksiä (esim. Martin ym., 2010), että tämä pätee oppijansuomen kompleksisuuden mittaamiseen.

Viestintä tai kielitaito eivät itsessään edellytä kompleksisuutta tai monimutkaisten rakenteiden käyttöä (esim. De Clercq & Housen, 2017). Tämä näkyy myös konjunktioiden käytössä. Vaikka konjunktioiden monipuolinen käyttö tai jonkin alemmilla taitotasoilla harvinaisen konjunktion esiintyminen tekstissä voi olla merkki edistyneestä kielitaidosta, minkään yksittäisen konjunktion puuttuminen tai konjunktioiden vähäinen määrä ei välttämättä tarkoita heikkoa kielitaitoa. Esimerkiksi seuraava epämuodollinen viesti sijoittuu taitotasolle C1, vaikka se ei ole syntaktisesti kompleksinen, jos kompleksisuutta tarkastellaan sivulauseiden tai konjunktioiden avulla:

> *Hyvät asukkaat! Kevät on tullut!*
> *Pidetään pihan*
> *SIIVOUSTALKOOT*
> *lauantaina 25.4.2005 klo 11-14*
> *Päälle mukavat vaatteet. Mukaan reipas ja iloinen mieli. Tarjolla limua ja makkaraa!*
> (F-1026 aikuiset C1)

Yksinkertaisempi syntaksi ei välttämättä tarkoita huonoa kielitaitoa, vaan voi olla merkki siitä, että kirjoittaja osaa sopeuttaa ilmaisuaan tilanteeseen (Kuiken & Vedder, 2019).

6 LOPUKSI

Pelkkä konjunktioiden tarkastelu antaa rajatun kuvan tekstien sisältämistä syntaktisista kytköksistä, sillä se jättää tarkastelun ulkopuolelle esimerkiksi asyndeettiset rinnastukset ja muut kuin konjunktion sisältävät sivulauseet. Konjunktioiden esiintymillä voi kuitenkin olla merkitystä oppijankielen

syntaktisen kompleksisuuden mittaamisen näkökulmasta, sillä ne ovat tekstien pintarakenteessa näkyviä merkkejä kirjoittajan käyttämistä syntaktisista kytköksistä ja niiden poimiminen aineistosta on helpommin automatisoitavissa kuin esimerkiksi lauseiden tai T-yksiköiden laskeminen. Niiden avulla voidaan myös tehdä havaintoja syntaktisesta kompleksisuudesta kielellisten keinojen monipuolisuutena, kun tarkastellaan eri taitotasoille sijoittuvissa teksteissä esiintyvien konjunktioiden valikoimaa tai alistuskonjunktion sisältävien lauseiden osuutta kaikista sivulauseista. Konjunktioiden esiintymien tarkastelu voi näin täydentää muilla mittareilla saatua tietoa oppijankielen syntaktisesta kompleksisuudesta.

Konjunktioiden käyttöä ja sen kautta syntyvää kuvaa syntaktisesta kompleksisuudesta on tässä tutkimuksessa tarkasteltu suhteessa kielitaidon taitotasoon. Tavoitteena oli kartoittaa ilmiötä ryhmätasolla, ja tutkimuksen keskiössä olivat eri taitotasoilla mahdolliset ja tyypilliset piirteet sekä konjunktioiden käytössä näkyvät erot taitotasojen välillä. Tällainen tarkastelu ei tavoita kielitaidon tai syntaktisen kompleksisuuden yksilöllistä kehitystä, joka voi olla erilaista kuin ryhmätason kehitys (esim. Larsen-Freeman, 2006). Pitkittäisaineiston avulla olisi mahdollista tarkastella myös yksilötason kehitystä ja verrata sitä nyt havaittuihin ryhmätason kehityskulkuihin.

Tässä artikkelissa konjunktioita ja niiden avulla luotuja rinnastus- ja alistussuhteita on tarkasteltu syntaksin alaan kuuluvana ilmiönä lause- ja virketasolla määrällisin menetelmin. Kokonaisten tekstien tarkastelu voisi tarjota uudenlaisen näkökulman syntaktiseen kompleksisuuteen (esim. Rimmer, 2009). Laadullisempi tutkimusote voisi myös tehdä näkyväksi sellaisia konjunktioiden käytön puolia, joita tämän tutkimuksen määrällinen tarkastelu ei tavoita, ja näin täydentää nyt saatua tietoa oppijansuomen syntaktisesta kompleksisuudesta.

LÄHTEET

Alanen, R., Huhta, A. & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. Teoksessa I. Bartning, M. Martin & I.Vedder (toim.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research,* (s. 21–56). Edisegno srl.

Alisaari, J. (2016). *Songs and poems in the second language classroom. The hidden potential of singing for developing writing fluency.* Väitöskirja. Turun yliopisto.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2), 390–395.

Benevento, C. & Storch, N. (2011). Investigating writing development in secondary school learners of French. *Assessing Writing, 16*(2), 97–110.

Bernardini, P. & Granfeldt, J. (2019). On cross-linguistic variation and measures of linguistic complexity in learner texts: Italian, French and English. *International Journal of Applied Linguistics, Special issue*, 1–22.

Biber, D., Gray, B. & Poonpon, K. (2011) Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35.

Bulté, B. & Housen, A. (2012). Defining and operationalising L2 complexity. Teoksessa A. Housen, I. Vedder & F. Kuiken, (toim.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA,* (s. 21–46). John Benjamins Publishing Company.

Cristofaro, S. (2014). Is there really a syntactic category of subordination? Teoksessa L. Visapää, J. Kalliokoski & H. Sorva (toim.), *Contexts of subordination. Cognitive, typological and discourse perspectives*, (s. 1–17). John Benjamins Publishing Company.

Crossley, S. & McNamara, D. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66–79.

De Clercq, B. & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal, 101*(2), 315–334.

Ellis, R. & Barkhuizen, G. (2005). *Analysing learner language*. Oxford University Press.

EVK 2003 = Eurooppalainen viitekehys (2003). *Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys*. WSOY.

Grant, L. & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123–145.

Gyllstad, H., Granfeldt, J., Bernardini, P. & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. Teoksessa L. Roberts, I. Vedder & J.H. Hulstijn (toim.), *Eurosla Yearbook 14*, (s. 1–30). John Benjamins Publishing Company.

Hakulinen, A., Karlsson, F. & Vilkuna, M. (1980). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus.* Helsingin yliopisto.

Herlin, I., Visapää, L. & Kalliokoski, J. (2014). Introduction. Teoksessa L. Visapää, J. Kalliokoski & H. Sorva (toim.), *Contexts of subordination. Cognitive, typological and discourse perspectives*, (s. 1–17). John Benjamins Publishing Company.

Honko, M. (2013). *Alakouluikäisen leksikaalinen tieto ja taito: Toisen sukupolven suomi ja S1-verrokit*. Väitöskirja. Tampereen yliopisto.

Housen, A., De Clercq, B., Kuiken, F. & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second language Research, 35*(1), 3–21.

Housen, A., Vedder, I. & Kuiken, F. (toim.) (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins Publishing Company.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M. & Hirvelä, T. (2014) Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3), 307–328.

Jagaiah, T., Olinghouse, N. G. & Kearns, D. M. (2020). Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and writing quality. *Reading and Writing, 33*, 2577–2638.

Jantunen, J. & Pirkola, S. (2015). Oppijansuomen sähköiset tutkimusaineistot. Nykytilanne. *Virittäjä, 119*(1), 88–103.

Kajander, M. (2013). *Suomen eksistentiaalilause toisen kielen oppimisen polulla*. Väitöskirja. Jyväskylän yliopisto.

Khushik, G. & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across Common European framework of reference levels A1, A2, and B1. *Applied Linguistics, 41*(4), 506–532.

Kuiken, F. & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics, 29*, 192–210.

Lambert, C. & Kormos, J. (2014). Complexity, accuracy and fluency in task-based research: Toward more developmentally-based measures of second language acquisition. *Applied Linguistics, 35*, 607–614.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590–619.

Martin, M. (2013). The complex simple – a problematic adjective in the CEFR writing scales. *Nordand, 8*(2), 63–85

Martin, M., Mustonen, S., Reiman, N. & Seilonen, M. (2010). On becoming an independent user. Teoksessa I. Bartning, M. Martin & I.Vedder (toim.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, (s. 57–80). Edisegno srl.

Michel, M. (2013). The use of conjunctions in cognitively simple versus complex oral L2 tasks. *The Modern Language Journal, 97*(1), 178–195.

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. Teoksessa S. Loewen & M. Sato (toim.), *The Routledge handbook of instructed second language acquisition*, (s. 50–68). Routledge.

Mustonen, S., (2015). *Käytössä kehittyvä kieli. Paikat ja tilat suomi toisena kielenä -oppijoiden teksteissä.* Väitöskirja. Jyväskylän yliopisto.

Mylläri, T. (2020a). Measuring syntactic complexity in learner Finnish. *Apples – Journal of Applied Language Studies 14*(2), 67–92.

Mylläri, T. (2020b). Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Assosiation, 4*(1), 13–23.

Määttä, T. (2012). Oppikirjan sanaston vaikutuksesta ruotsinkielisten alkeistason suomenoppijoiden kirjallisiin tuotoksiin. *Lähivertailuja, 22*, 188–218.

Norris, J. & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics, 30*(4), 555–578.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518.

Ortega, L. (2009). *Understanding second language acquisition*. Hodder Education.

Pajunen, A. & Vainio, S. (2021a). Kielen rakenteen hallinta peruskoululaisilla ja nuorilla aikuisilla. Teoksessa A. Pajunen & M. Honko (toim.), *Suomen kielen hallinta ja sen kehitys: peruskoululaiset ja nuoret aikuiset*. Suomalaisen Kirjallisuuden Seuran toimituksia, 1472, (s. 367–420). Suomalaisen Kirjallisuuden Seura.

Pajunen, A. & Vainio, S. (2021b). Syntaktista rakennetta tiivistävät ja kompleksistavat tekijät peruskoululaisilla ja nuorilla aikuisilla. Teoksessa A. Pajunen & M. Honko (toim.), *Suomen kielen hallinta ja sen kehitys: peruskoululaiset ja nuoret aikuiset*. Suomalaisen Kirjallisuuden Seuran toimituksia, 1472, (s. 421–479). Suomalaisen Kirjallisuuden Seura.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590–601.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117–134.

Reiman, N. (2011). Two faces of complexity: Structural measures and diversity of constructions. *Nordand, 6*(2), 9–23.

Reiman, N. (2014). Yläkoulun S2-oppilaiden transitiivi-ilmausten käyttö Eurooppalaisen viitekehyksen taitotasoilla. *Lähivõrdlusi. Lähivertailuja, 24*, 183–220.

Rimmer, W. (2009). Can what counts in complexity be counted? *University of Reading: Language studies working papers, 1*, 25–34.

Saukkonen, P., Haipus, M., Niemikorpi, A. & Sulkala, H. (1979). *Suomen kielen taajuussanasto.* WSOY.

Seilonen, M. (2013). *Epäsuora henkilöön viittaaminen oppijansuomessa.* Väitöskirja. Jyväskylän yliopisto.

Spoelman, M. & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics, 31*(4), 532–553.

Taguchi, N., Crawford, W. & Wetzel, D. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly, 47*(2), 420–430.

Tilma, C. (2014). *The dynamics of foreign versus second language development in Finnish writing.* Väitöskirja. Jyväskylän yliopisto.

Vilkuna, M. (2003). *Suomen lauseopin perusteet*. Edita.

VISK = Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. & Alho, I. (2004). *Iso suomen kielioppi*. Suomalaisen Kirjallisuuden Seura. Verkkoversio, viitattu 15.2.2021. Saatavissa: http://scripta.kotus.fi/visk.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal, 96*(4), 576–598.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal 97 (S1)*, 11–30.

Wolfe-Quintero, K., Inagaki, S. & Kim, H. (1998). *Second language development in writing: measures of fluency, accuracy, and complexity.* University of Hawai'i, Second Language Teaching and Curriculum Center.

CONJUNCTIONS AND SYNTACTIC COMPLEXITY: DEVELOPMENT OF THE USE OF CONJUNCTIONS IN WRITTEN LEARNER FINNISH

Taina Mylläri, Department of Language and Communication Studies, University of Jyväskylä

Learner language development is often analysed by measuring complexity, accuracy and fluency. Complexity can be defined as the range and sophistication of the structures available to the learner, yet syntactic complexity is typically analysed using quantitative measures tapping subordination. This article focuses on the development of syntactic complexity in written learner Finnish across the CEFR proficiency levels by exploring changes in the use of conjunctions. The data are drawn from the corpus of the Jyväskylä University CEFLING project. The development of syntactic complexity in terms of coordination and subordination is explored by tracking changes in the use of coordinate and subordinate conjunctions. Syntactic complexity as the range of forms used by the learner is explored by investigating the emergence and frequency of individual conjunctions. The results question the role of subordination in the development of syntactic complexity in learner Finnish, as they show that both coordinate and subordinate conjunctions are already used by learners on the lowest CEFR levels. The results also indicate that development in the use of conjunctions may be different for adult learners and adolescent learners.

Keywords: conjunctions, learner Finnish, syntactic complexity

LIITE 1

TAULUKKO 1A. Konjunktioiden esiintymät (n) ja esiintymistaajuus tuhatta sanaa kohden (n/1000) aikuisten aineistossa.

	YKI A1		YKI A2		YKI B1		YKI B2		YKI C1		YKI C2		Yhteensä	
	n	n/1000	n	n/1000	n	n/1000	n	n/1000	n	n/1000	n	n/1000	n	n/1000
ja	172	37	186	35	421	43	330	40	402	37	380	38	1 891	39
mutta	39	8	42	8	113	11	81	10	60	6	52	5	387	8
tai	5	1	14	3	60	6	36	4	28	3	25	3	168	3
eli	3	1	1	0	10	1	9	1	6	1	11	1	40	1
vai			2	0	4	0	3	0	7	1	10	1	26	1
eikä			1	0	8	1	14	2	18	2	18	2	59	1
sekä					6	1	10	1	11	1	15	2	42	1
sekä – että					5	1	7	1	5	0	5	1	22	0
vaan					2	0	6	1	6	1	12	1	26	1
sillä					1	0	3	0	8	1	8	1	20	0
muttei					1	0	1	0	2	0			4	0
joko – tai					1	0	1	0	2	0			4	0
elikä					1	0	1	0					2	0
niin – kuin					1	0	1	0					2	0
Rinnastus-konjunktiot yhteensä	219	48	246	46	634	64	503	60	555	52	536	54	2 693	55
koska	79	17	64	12	62	6	37	4	25	2	27	3	294	6
että	41	9	71	13	162	16	129	15	174	16	145	15	722	16
kun	32	7	24	5	61	6	54	6	50	5	32	3	253	5
jos	21	5	25	5	63	6	62	7	66	6	48	5	285	6
kuin	8	2	8	2	26	3	28	3	27	3	21	2	118	2
vaikka					10	1	7	1	18	2	6	1	41	1
niin kuin					3	0	1	0	3	0			7	0
ettei					2	0	5	1	17	2	10	1	34	1
jotta					1	0	6	1	6	1	7	1	20	0
kunnes					1	0	1	0			1	0	3	0
kunhan					1	0			1	0	2	0	4	0
mikäli							1	0			1	0	2	0
ennen kuin									6	1	6	1	12	0
ellei									3	0			3	0
vaikkapa									1	0			1	0
kun taas											1	0	1	0
Alistus-konjunktiot yhteensä	181	39	192	36	392	40	331	40	397	37	307	31	1 800	37
Kaikki konjunktiot yhteensä	**400**	**87**	**438**	**83**	**1 026**	**104**	**834**	**100**	**952**	**89**	**843**	**84**	**4 493**	**92**

TAULUKKO 1B. Konjunktioiden esiintymät (n) ja esiintymistaajuus tuhatta sanaa kohden (n/1000) nuorten aineistossa.

	S2-nuoret A1		**S2-nuoret A2**		**S2-nuoret B1**		**S2-nuoret B2**		**Yhteensä**		**S1-nuoret**	
	n	**n/1000**	**n**	**n/1000**	**n**	**n/1000**	**n**	**n/1000**	**n**	**n/1000**	**n**	**n/1000**
ja	71	31	203	34	262	36	40	35	576	35	613	31
mutta	17	7	54	9	83	12	13	11	167	10	199	10
tai	15	6	46	8	61	8	7	6	129	8	131	7
vai	3	1	1	0	2	0			6	0	9	0
eikä	2	1	5	1	11	2	3	3	21	1	53	3
eli	1	0	4	1	9	1	1	1	15	1	18	1
sillä	1	0	4	1	1	0			6	0	38	2
sekä – että			1	0	1	0			2	0	1	0
vaan					3	0			3	0	12	1
sekä					1	0			1	0	10	1
joko – tai					1	0			1	0	5	0
niin – kuin											1	0
joko – taikka											1	0
Rinnastus-konjunktiot yhteensä	110	48	318	53	435	61	64	56	927	56	1 091	55
koska	48	21	114	19	89	12	7	6	258	16	122	6
jos	27	12	68	11	85	12	7	6	187	11	231	12
kun	18	8	47	8	58	8	15	13	138	8	186	9
että	16	7	95	16	137	19	23	20	271	16	287	14
kuin	3	1	5	1	5	1	1	1	14	1	25	1
vaikka	3	1	4	1	5	1	1	1	13	1	18	1
ellei	1	0							1	0	3	0
ettei			6	1	8	1			14	1	23	1
jotta			1	0	4	1			5	0	6	0
kunhan			1	0	2	0			3	0	3	0
niin kuin			1	0					1	0	6	0
ennen kuin					1	0			1	0	3	0
mikäli											2	0
etteikö											1	0
jottei											1	0
kunnes											1	0
Alistus-konjunktiot yhteensä	116	50	342	57	394	55	54	47	906	55	918	46
Kaikki konjunktiot yhteensä	**226**	**98**	**660**	**111**	**829**	**115**	**118**	**104**	**1 833**	**110**	**2 009**	**101**

IV

MONIVERBISET KONSTRUKTIOT JA OPPIJANSUOMEN KOMPLEKSISUUS KIELITAIDON ERI TASOILLA

by

Mylläri, Taina 2023

Lähivõrdlusi. Lähivertailuja 33, 152–180.

https://doi.org/10.5128/LV33.05

doi:10.5128/LV33.05

Moniverbiset konstruktiot ja oppijansuomen kompleksisuus kielitaidon eri tasoilla

TAINA MYLLÄRI
Vilnan yliopisto, Jyväskylän yliopisto

Tiivistelmä. Kielitaitoa ja sen kehitystä voidaan tarkastella oppijankielen tarkkuuden, sujuvuuden ja kompleksisuuden avulla. Yksinkertaisimmillaan tarkkuudella tarkoitetaan kielen virheettömyyttä ja sujuvuudella kielen tuottamisen helppoutta. Kompleksisuutta puolestaan voidaan ajatella monipuolisuutena: mitä enemmän kielisysteemissä on osasia ja mitä useammalla tavalla niitä voidaan yhdistää, sitä kompleksisempi kielisysteemi on. Syntaktista kompleksisuutta kuitenkin mitataan usein määrällisillä mittareilla, jotka eivät välttämättä tavoita syntaktisten rakenteiden monipuolistumista. Tässä tutkimuksessa keskitytään finiittiverbin ja infiniittisen verbinmuodon sisältäviin konstruktioihin ja niiden käyttöön oppijansuomessa kielitaidon eri taitotasoilla. Tavoitteena on selvittää, tapahtuuko konstruktioissa sellaisia muutoksia, jotka eivät tule näkyviin perinteisillä syntaktisen kompleksisuuden mittareilla. Aineisto koostuu Jyväskylän yliopiston Cefling-korpuksen aikuisten suomenoppijoiden mielipideteksteistä. Tulokset osoittavat, että tarkasteltavat konstruktiot eivät ylemmillä taitotasoilla ole välttämättä pitempiä kuin alemmilla taitotasoilla mutta niissä on enemmän variaatiota. Samalla tulokset tukevat näkemyksiä, joiden mukaan useimmin käytetyt syntaktisen kompleksisuuden mittarit eivät tavoita oppijankielen kompleksisuuden kaikkia ulottuvuuksia ja että ainakaan kaikissa kielissä syntaktista, morfologista ja leksikaalista kompleksisuutta ei välttämättä voi rajata erillisiksi ilmiöiksi.

Avainsanat: toisella kielellä kirjoittaminen; Eurooppalainen viitekehys; syntaktinen kompleksisuus; verbikonstruktiot; suomi

1. Johdanto

Kielitaito on muutakin kuin sujuvuutta ja tarkkuutta. Toisen kielen oppimisen tutkimuksessa näitä kahta täydentämässä käytetään usein kompleksisuutta, ja yhdessä nämä kolme tunnetaan lyhenteellä CAF (engl. *complexity*, *accuracy*, *fluency*). Vaikka myös sujuvuus ja tarkkuus voidaan määritellä monin tavoin, vaikeimmin määriteltäväksi on osoittautunut kompleksisuus, jota on tutkittu sekä itsenäisenä kielitaidon osa-alueena että osana kompleksisuuden, tarkkuuden ja sujuvuuden muodostamaa kolmikkoa (Ellis & Barkhuizen 2005; Michel 2017; Housen ym. 2019).

Oppijankielen kompleksisuus on moniulotteinen ilmiö (Norris & Ortega 2009). Se voidaan jakaa leksikaaliseen ja kieliopilliseen kompleksisuuteen, joista jälkimmäinen voidaan vielä jakaa syntaktiseen ja morfologiseen kompleksisuuteen (ks. Bulté & Housen 2012). Näistä eniten tutkittu on syntaktinen kompleksisuus (De Clercq & Housen 2017). Sitä mitataan tavallisesti lauseisiin ja niistä muodostuviin laajempiin yksiköihin perustuvilla määrällisillä mittareilla (esim. Bulté & Housen 2012; Ellis & Barkhuizen 2005; Wolfe-Quintero ym. 1998). Nämä mittarit tavoittavat kuitenkin vain osan mahdollisesta oppijankielen rakenteiden kehityksestä, sillä esimerkiksi keskenään hyvin erilaiset T-yksiköt eli päälauseen ja sille alisteisten rakenteiden muodostamat kokonaisuudet voivat sisältää täysin saman määrän sanoja (Biber ym. 2011; Kyle & Crossley 2018; Rimmer 2006), ja esimerkiksi lauseet *se näyttää hyvältä* ja *se näyttää toimivan* ovat näiden mittareiden mukaan yhtä kompleksisia. Oppijankielen kompleksisuuden määrällisen mittaamisen rinnalle tarvitaankin myös laadullista tutkimusta (Larsen-Freeman 2009; Martin 2013; Reiman 2011b). Esimerkiksi oppijankielen konstruktioiden tarkastelu voi nostaa näkyväksi sellaisia kompleksisuuden kehityskulkuja, joita määrälliset mittarit eivät tavoita (Reiman 2011a, 2011b, 2014).

Tässä artikkelissa oppijansuomen kompleksisuutta tarkastellaan finiittiverbistä ja infiniittisestä verbinmuodosta koostuvien moniverbisten konstruktioiden avulla. Näissä infiniittisenä verbinmuotona voi olla A-infinitiivi (*alkaa tehdä*), MA-infinitiivi (*rupeaa tekemään*) tai partisiippimuoto (*näkyy tekevän*). Moniverbisten konstruktioiden ilmaantuminen oppijankieleen voi näkyä syntaktisen kompleksisuuden määrällisissä mittareissa: esimerkiksi *voin tehdä* on rakenteena pitempi kuin *teen* ja kasvattaa siksi lauseen tai T-yksikön sanamäärää. Samoin erilaisten infiniittisten rakenteiden yhdistäminen voi tuottaa sanamäärältään pitempiä ilmauksia: kun konstruktio *opin lukemaan* yhdistetään esimerkiksi modaaliseen *voida*-verbiin, ilmaus pitenee kolmisanaiseksi (*voin oppia lukemaan*). Sen sijaan konstruktion leksikaalinen ja morfologinen laajeneminen jäävät määrällisten mittareiden tavoittamattomiin silloin, kun ne eivät vaikuta konstruktion sanamäärään (Martin ym. 2010).

Moniverbiset konstruktiot ilmantuvat oppijansuomeen heti taitotasolla A1, ja kielitaidon tason noustessa morfologinen ja leksikaalinen variaatio lisääntyy: A-infinitiivin lisäksi moniverbisissä rakenteissa aletaan käyttää MA-infinitiiviä ja partisiippimuotoja samalla kun rakenteissa käytettyjen verbien kirjo laajenee (esim. Haapala 2008; Kynsijärvi 2007; Paavola 2008). Oppijansuomen moniverbisissä konstruktioissa ja niiden kompleksisuudessa tapahtuu siis myös sellaisia muutoksia, joita ei voi havaita lauseiden tai T-yksiköiden sanamääriin perustuvilla määrällisillä mittareilla. Tässä artikkelissa tavoitteena on selvittää, mitä nämä muutokset voivat kertoa oppijansuomen kompleksisuudesta kielitaidon eri tasoilla, ja näin monipuolistaa aiemmassa tutkimuksessa (esim. Mylläri 2020) määrällisillä mittareilla saatuja tuloksia oppijansuomen kompleksisuudesta. Artikkelin tutkimuskysymykset ovat:

1. Millaisia moniverbisiä konstruktiota oppijansuomen taitotasoilla A1–C2 käytetään?
2. Millaista oppijansuomen kompleksisuus on, kun sitä tarkastellaan moniverbisten konstruktioiden avulla?

Tutkimusaineistona on 241 suomi toisena kielenä (S2) -mielipidetekstiä, jotka sijoittuvat Eurooppalaisen viitekehyksen (EVK 2003) taitotasoille A1–C2. Aineistoa tarkastellaan tekstimassana taitotasoittain. Konstruktioissa tapahtuvia muutoksia kuvataan määrällisesti ja aineistoesimerkkien avulla.

2. Taustaa

2.1. Oppijankielen kompleksisuudesta ja sen mittaamisesta

Oppijankielen kompleksisuutta voidaan tarkastella käytettyjen kielenpiirteiden monipuolisuutena (esim. Ortega 2003: 492) tai kielisysteemin sisältämien osien ja niiden välisten kytkösten määränä ja monipuolisuutena (esim. Bulté & Housen 2012). Syntaktista kompleksisuutta on tyypillisesti mitattu sellaisilla määrällisillä mittareilla kuin T-yksikön keskipituus sanoina (Mean Length of T-Unit eli MLT tai MLTU) tai lauseen keskipituus sanoina (Mean Length of Clause, MLC) (esim. Ortega 2003). Näistä ensimmäisen katsotaan mittaavan syntaktista kompleksisuutta yleisellä tasolla ja jälkimmäisen lausetasolla (Bulté & Housen 2012; Norris & Ortega 2009). Etenkin T-yksikön pituuteen perustuvia mittareita on kuitenkin arvosteltu siitä, että ne korostavat alisteisten rakenteiden osuutta kompleksisuuden muiden ulottuvuuksien kustannuksella (esim. Bulté & Housen 2012; Kyle & Crossley 2018) ja jättävät huomiotta esimerkiksi rinnastamisen (Bardovi-Harlig 1992). Mittareissa keskitytään tavallisesti kielenyksiköiden keskimääräisiin pituuksiin, ja syntaktisten rakenteiden monipuolisuutta mitataan vain harvoin (De Clercq & Housen 2017).

Oppijansuomen kompleksisuutta rakenteiden variaation näkökulmasta samasta aineistosta kuin tässä tutkimuksessa ovat tarkastelleet Reiman (2011a) ja Seilonen (2013). Reiman (2011a) on tutkinut transitiivikonstruktioiden käyttöympäristöjä ja havainnut, että taitotason noustessa käyttöyhteydet laajenevat ja transitiivikonstruktioita aletaan esimerkiksi yhdistää erilaisiin infinitiivirakenteisiin. Oppijansuomen nollapersoonaisia ilmauksia tutkinut Seilonen (2013) puolestaan on

havainnut muun muassa, että nesessiivisten ilmausten rakenteellinen ja leksikaalinen kirjo kasvaa kielitaidon tason noustessa: nesessiivinen *pitää*-konstruktio ilmaantuu oppijoiden teksteihin heti taitotasolla A1, nesessiivi-ilmaukset *tulee tehdä* ja *on tehtävä* taitotasolla B2, ja taitotasolla C2 nesessiivi-ilmauksissa on käytössä yhdeksän eri verbiä. Kielenyksiköiden pituuteen perustuvat syntaktisen kompleksisuuden määrälliset mittarit tavoittavat näistä muutoksista vain ne, jotka lisäävät lauseen tai T-yksikön sanamäärää, mutta eivät morfologisen ja leksikaalisen variaation lisääntymisen kaltaisia muutoksia, jotka eivät näy konstruktion sanamäärässä (Martin ym. 2010). Tämä saattaa osaltaan selittää sitä, miksi perinteiset syntaktisen kompleksisuuden määrälliset mittarit eivät näytä erottelevan oppijansuomen taitotasoja kovin hyvin (Mylläri 2020).

Oppijansuomen syntaktista kompleksisuutta ovat tutkineet määrällisten mittareiden avulla myös Tilma (2014) sekä Spoelman ja Verspoor (2010). Oppijansuomen syntaktista kehitystä yhdeksän kuukautta kattavassa pitkittäisaineistossa tarkastellut Tilma (2014) havaitsi, että aineiston teksteissä sekä lauseet että virkkeet pitenivät oppimisen edetessä, kun niiden pituutta mitattiin morfeemeina. Käyttämistään syntaktisen kompleksisuuden mittareista Tilma (2014) totesi oppijansuomeen parhaiten soveltuvaksi lauseen keskipituuden morfeemeina. Yhdeltä suomenoppijalta kolmen vuoden aikana kerättyä pitkittäisaineistoa tutkineet Spoelman ja Verspoor (2010) puolestaan totesivat, että syntaktisen kompleksisuuden kehitys ei tutkimuksessa käytettyjen mittareiden mukaan ollut lineaarista.

2.2. Suomen kielen moniverbisistä konstruktioista

Myös oppijansuomen verbiketjuja ja muita moniverbisiä rakenteita on tarkasteltu suhteessa taitotasoon. Tämän tutkimuksen aineiston kanssa osittain päällekkäisillä aineistoilla tehdyt tutkimukset ovat osoittaneet, että moniverbisten rakenteiden esiintymistaajuus ja niissä käytettyjen verbien kirjo kasvavat ja että konstruktioiden rakenteellinen variaatio

lisääntyy taitotasolta toiselle (Haapala 2008; Kynsijärvi 2008; Paavola 2008; Puhakka 2010; Seilonen 2013). Ensimmäisenä finiittiverbien yhteyteen ilmaantuvat A-infinitiivi, sitten MA-infinitiivi ja partisiippimuodot (Haapala 2008; Paavola 2008). Modaaliset ja nesessiiviset verbiketjut ilmaantuvat suomenoppijoiden teksteihin jo alimmilla taitotasoilla, merkitykseltään erikoistuneet konstruktiot viimeisinä (esim. Seilonen 2013). Lisäksi konstruktioiden leksikaalinen variaatio kasvaa, mikä pätee sekä finiittiverbeihin että infiniittisiin verbinmuotoihin (Haapala 2008; Kynsijärvi 2008; Paavola 2008; Puhakka 2010; Seilonen 2013).

Tässä tutkimuksessa keskitytään oppijansuomen syntaktiseen kompleksisuuteen ja sen kehittymiseen tarkastelemalla finiittiverbin ja vähintään yhden infiniittisen verbinmuodon sisältävien konstruktioiden käyttöä ja sen muuttumista kielitaidon tasolta toiselle. Nämä verbikonstruktiot ovat syntaktisen kompleksisuuden näkökulmasta mielenkiintoisia siksi, että niissä kielitaidon kehittyminen voi näkyä sekä ilmausten pitenemisenä että niiden monipuolistumisena. Lisäksi rakenteiden tarkastelussa konstruktioina yhdistyvät syntaktinen, morfologinen ja leksikaalinen kompleksisuus. Tarkasteltaviksi moniverbisiksi konstruktioiksi valittiin modaaliverbistä tai muusta abstraktista verbistä ja infiniittisestä verbinmuodosta koostuvat verbiketjut (esim. *voin tehdä*, *pyrin tekemään*, *näytät tekevän*) (VISK: § 496) sekä sellaiset vakiintuneet apu- ja pääverbistä koostuvat rakenteet, joita voidaan pitää verbiliittoina (esim. *olen tekemättä*, *tulen tekemään*, *on tehtävä*, *tulin tehneeksi*) (VISK: § 451). Lisäksi tarkasteluun otettiin mukaan rakenteet, joissa A-infinitiivi toimii finiittiverbin objektina (esim. *haluan tehdä*, *yritän tehdä*) (VISK: § 493), ja rakenteet, joissa MA-infinitiivi toimii finiittiverbin täydennyksenä (esim. *menen tekemään*, *rentoudun lukemalla*) (VISK: § 494).

Tässä artikkelissa tarkasteltavissa moniverbisissä konstruktioissa on käytössä kaksi eri infinitiiviä. Niistä selvästi yleisempi on A-infinitiivi, joka on Lauseopin arkiston koodatussa yleiskielisessä aineistossa kaksi kertaa niin yleinen kuin MA-infinitiivi (VISK: § 1228). Infinitiivien yleisyydessä on kuitenkin eroja murteiden ja kirjakielen välillä (Herlin & Visapää 2005: 17). ”Iso suomen kielioppi” luettelee 16 verbiä, jotka

voivat muodostaa verbiketjun yhdessä A-infinitiivin kanssa (VISK: § 496), sekä 24 verbiä, jotka voivat saada objektiksi infinitiivilausekkeen, ja toiset 24 verbiä, joiden objektina voi olla NP, infinitiivilauseke tai lause (VISK: § 469). Näiden lisäksi A-infinitiiviä voidaan käyttää *olla*-verbin kanssa verbiliitossa *olla (vähällä) tehdä* (VISK: § 541) sekä verbien *tulla* ja *kuulua* kanssa nesessiivisissä ilmauksessa (*kuuluu tehdä*, *tulee tehdä*) (VISK: § 1577). MA-infinitiivin kanssa verbiketjun muodostavia verbejä on puolestaan lueteltu 7 (VISK: § 496). Lisäksi MA-infinitiiviä voidaan käyttää verbin täydennyksenä (VISK: § 470). ”Iso suomen kielioppi” luettelee noin 80 eri verbiä, joiden muottitäydennyksenä MA-infinitiivi voi toimia (VISK: § 479). Lisäksi MA-infinitiivi voi ”Ison suomen kieliopin” mukaan muodostaa verbiliiton verbien *olla* (*on tekemässä*, *on tekemäisillään*, *on tekemättä*), *jäädä* (*jäädä tekemättä*), *jättää* (*jättää tekemättä*), *pitää* (*pitää tekemän*) ja *tulla* (futuurinen *tulla tekemään*) kanssa (VISK: § 541). MA-infinitiivin yleisin sijamuoto on illatiivi (Herlin & Visapää 2005: 17; VISK: § 1228). Lauseopin arkiston koodatussa yleiskielisessä aineistossa muut sijamuodot ovat yleisyysjärjestyksessä abessiivi, inessiivi, adessiivi ja elatiivi (VISK: § 1228).

Myös tämän tutkimuksen aineiston kanssa osittain päällekkäisissä oppijansuomen aineistoissa A-infinitiivi on yleisin infiniittinen verbinmuoto (Haapala 2008; Kynsijärvi 2007; Paavola 2008). Edistyneen oppijansuomen piirteitä toisessa aineistossa tutkinut Ivaska (2014) on havainnut, että A-infinitiivien käyttö S2-teksteissä poikkeaa sen käytöstä äidinkielisessä vertailuaineistossa. Ivaska (2014: 179–180) totesi, että S2-teksteissä A-infinitiiviä esiintyy vähemmän ja että ero liittyy verbiketjujen määrään ja siihen, mitä finiittiverbejä niissä käytetään. Ketjuuntuvia verbirakenteita tutkineen Seppälän (2013) mukaan ICLFI-korpuksen teksteissä yleisimmät A-infinitiivin vaativat verbit ovat *voida*, *haluta*, *saada*, *yrittää*, *alkaa* ja yleisimmät MA-infinitiivin vaativat *lähteä*, *mennä*, *oppia*, *pystyä*, *auttaa*, *käydä*, *istua*, *olla* sekä *tulla.* Nesessiiviset rakenteet eivät olleet mukana Seppälän tutkimuksessa.

Infinitiivien lisäksi moniverbisissä konstruktioissa voidaan käyttää verbin partisiippimuotoja. Vaikutelmaverbit *kuulua*, *näkyä*, *näyttää*,

osoittautua, *tuntua* ja *vaikuttaa* voivat muodostaa VA-partisiipin kanssa verbiketjun (VISK: § 496). Verbit *olla*, *tulla* ja *saada* puolestaan voivat muodostaa verbiliiton VA-partisiipin (*on tehtävä*, *on tehtävissä*, *on tekevä*, *on tekevinään*, *tulee tehtäväksi*, *saa tehtäväkseen*) tai NUT-partisiipin (*on tehneenä*, *tulee tehneeksi*, *tulee tehtyä ~ tehdyksi*, *saa tehtyä ~ tehdyksi*) kanssa, verbi *ottaa* VA-partisiipin kanssa (*ottaa tehtäväkseen*) ja verbi *joutua* NUT-partisiipin kanssa *joutua* (*joutuu pidätetyksi*) (VISK: § 451). Oppijansuomeen partisiipin sisältävät moniverbiset konstruktiot näyttävät ilmaantuvan myöhemmin kuin infinitiivin sisältävät (esim. Seilonen 2013) eivätkä kaikki erilaiset finiittiverbistä ja partisiipista koostuvat konstruktiot välttämättä ole käytössä ylimmillä taitotasoillakaan (Haapala 2008).

Moniverbisistä konstruktioista tekee kompleksisuuden kannalta mielenkiintoisia lisäksi se, että niitä voidaan laajentaa myös infiniittisiä muotoja rinnastamalla (1) tai ketjuttamalla (2).

(1) Minun mielessäni on aina semmonen halu että **haluaisin katsoa ja kokea** jotain ihan erilaista ja ihmeellistä. (B2, F-698)
(2) **Voisitko käydä** kaupassa **keskustelemassa** asiasta, **tai** vaikkapa **lähettää** kauppaan vastauksen. (C1, F-1075)

Syntaktisen kompleksisuuden tutkimuksessa lausetason rinnastamista pidetään usein tyypillisenä kielitaidon alimmille tasoille, lausekkeiden pitenemistä ja monipuolistumista puolestaan kielitaidon ylemmille tasoille (esim. Kuiken & Vedder 2019; Norris & Ortega 2009). Rinnastaminen ei kuitenkaan välttämättä ole pelkkää yksinkertaisten ilmausten liittämistä toisiinsa. Esimerkiksi Haapala (2008: 71) on kiinnittänyt huomiota siihen, että oppijansuomen verbiketjuissa esiintyy myös rinnastamista ja ellipsiä. Tekstuaalista ellipsiä voidaankin pitää sidosteisuuden keinona (Shore 2020: 574). Moniverbisen konstruktion osat voivat myös hajaantua kuten esimerkissä (2). Ketjuuntuvien verbirakenteiden osien väliin voivat adverbien lisäksi sijoittua esimerkiksi subjekti tai objekti (Vilkuna 2003: 277). Sekä elliptiset että hajaantuvat rakenteet ovat kompleksisuuden kannalta mielenkiintoisia, sillä niissä näkyy taito muunnella ja varioida konstruktioita.

3. Aineisto ja menetelmät

Tutkimusaineistona on käytetty Jyväskylän yliopiston Cefling-hankkeen (Cefling 2007–2009) S2-aineiston aikuisten mielipidetekstejä. Aineisto kattaa Eurooppalaisen viitekehyksen taitotasot A1–C2. Alkuperäiset tekstit on kirjoitettu testitilanteessa käsin ilman apuvälineitä rajoitetussa ajassa (Jantunen & Pirkola 2015). Alimmilla taitotasoilla tehtävänä on ollut kirjoittaa palautetta tehtävänannossa nimetystä palvelusta, keskitasolla ja ylimmillä taitotasoilla tehtävänä on ollut kirjoittaa mielipidekirjoitus annetusta aiheesta esimerkiksi lehden mielipidepalstalle lähetettäväksi. Tekstien arviointi taitotasoille on tehty Cefling-hankkeessa, ja sitä voidaan pitää luotettavana sekä tilastollisin että laadullisin menetelmin tarkasteltuna (Huhta ym. 2014). Tässä tutkimuksessa käytetty aineisto (241 tekstiä, 23 596 sanaa) on aiemman syntaktisen kompleksisuuden tutkimuksen yhteydessä koodattu xml-muotoon ja siitä on poistettu sanasta sanaan tehtävänannoista kopioidut otsikot sekä verbittömistä rakenteista koostuvat aloitukset ja lopetukset, esimerkiksi erilaiset tervehdykset ja yhteystiedot (Mylläri 2020). Aineiston tekstien jakautuminen taitotasoille on esitetty taulukossa 1.

Taulukko 1. *Aineiston teksti- ja sanamäärät taitotasoittain*

Aineisto	**Taitotaso**						**Yhteensä**
	A1	**A2**	**B1**	**B2**	**C1**	**C2**	
Tekstejä	50	37	43	35	46	30	241
Sanoja	2 261	2 272	5 142	4 166	5 876	3 879	23 596

Tätä tutkimusta varten aineistosta poimittiin käsin kaikki finiittiverbien esiintymät. Esiintymät luokiteltiin moniverbisiin ja muihin finiittiverbikonstruktioihin. Liittomuodot tulkittiin verbin taivutusparadigmaan kuuluviksi (esim. VISK: § 450) eikä niitä luokiteltu moniverbisiksi. Myös *on hauskaa* -lauseet rajattiin tutkimuksen ulkopuolelle, sillä samoja esiintymiä on käsitellyt Seilonen (2013) analysoidessaan nollapersoonaisen *on kiva* -konstruktion kehitystä koko Cefling-aineiston tasolla.

Mikäli löytynyt moniverbinen konstruktio oli osa laajempaa moniverbistä kokonaisuutta (esim. 3), koko ilmaus koodattiin aineistoon samaan konstruktioon kuuluvaksi siitä riippumatta, saivatko infiniittiset verbinmuodot omia määritteitään. Konstruktiot luokiteltiin finiittiverbin ja A-infinitiivin, finiittiverbin ja MA-infinitiivin tai finiittiverbin ja partisiippimuodon sisältäväksi ensimmäisen finiittiverbiä seuraavan infiniittisen verbinmuodon mukaan. Esimerkin (3) virkkeeseen koodattiin yksi finiittiverbin ja A-infinitiivin sisältävä moniverbinen konstruktio, jossa finiittiverbiin *täytyy* on yhdistetty kaksi keskenään rinnasteista A-infinitiiviä (*opettaa* ja *selittää*) ja jossa A-infinitiiveistä ensimmäiseen on lisäksi yhdistetty kaksi keskenään rinnasteista MA-infinitiiviä (*kirjoittamaan* ja *lukemaan*). Tällainen koodaus ei ota huomioon lauserajoja, mutta se tekee näkyväksi erilaiset tavat yhdistellä verbi + verbi -rakenteita.

(3) Koulun t**äutuu opettaa** oppilaan **kirjoittamaan ja lukemaan** hyvin **sekä selittää** mitkä ovat seuramukset jos he eivät opiskele hyvin. (B2, F-676)

Erikseen koodattiin esiintymät, joissa finiittiverbin kanssa käytetty verbinmuoto ei ollut tulkittavissa miksikään tarkasteluun valituista infiniittisistä muodoista (4) (oppijansuomen odotuksenmukaisesta poikkeavien muotojen tulkinnasta ks. esim. Brunni ym. 2015). Tällaisten esiintymien määrä aineistossa on varsin pieni, mutta niistä muodostettiin oma ryhmänsä, jota kutsutaan nimellä finiittiverbi + muu verbinmuoto. Näin aineiston määrälliseen tarkasteluun saatiin mukaan myös ne moniverbiset konstruktiot, joissa verbin infiniittinen muoto oli odotuksenvastainen (ks. myös Niiranen 2010). Yhden tai kahden kirjaimen poikkeamat vakiintuneesta kirjoitusasusta tulkittiin infinitiivimuodoiksi (5). Finiittiverbin infiniittisen laajennuksen kohdekielisyydestä poikkeavia muotoja ovat tämän tutkimuksen aineiston kanssa osittain päällekkäisessä aineistossa käsitelleet tarkemmin esimerkiksi Haapala (2008) ja Paavola (2008).

(4) ja **ei saa häritse** Sinun ystäväsi luokaan (B1, F-638)
(5) Siihen menessa **pitä mietia** kuinka **voi säästä** turhia matkoja. (B1, F-663)

Sekä finiittiset että infiniittiset verbinmuodot lemmattiin eri verbien esiintymien määrän laskemiseksi. Lemmaksi koodattiin verbin A-infinitiivi. Kirjoitusvirheiden vaikutusta variaatioon kontrolloitiin tulkitsemalla vakiintuneesta kirjoitusasusta poikkeavat muodot oletetun tavoitemuodon tavoin. Esimerkissä (5) finiittiverbeiksi koodattiin siis *pitää* ja *voida*, infiniittisten muotojen puolestaan *miettiä* ja *säästää*.

Aineiston määrällisessä tarkastelussa käytetään sekä absoluuttisia että suhteellisia frekvenssejä. Absoluuttisten frekvenssien avulla saadaan näkyviin esiintymien lukumäärän vaikutus variaatioon etenkin silloin, kun esiintymiä on vähän: esimerkiksi taitotasoilla A1 ja A2 on molemmilla kaksi finiittiverbin ja MA-infinitiivin sisältävän konstruktion esiintymää, joten näillä taitotasoilla voi kummallakin olla enintään kaksi eri verbiä finiittiverbinä. Suhteelliset frekvenssit on tässä tutkimuksessa laskettu aiemmista Cefling-aineistosta tehdyistä tutkimuksista poiketen suhteessa kaikkien finiittiverbien määrään, ei suhteessa tuhanteen sanaan. Tällä pyritään pienentämään aiemmassa tutkimuksessa (Mylläri 2020) havaitun lauseiden pitenemisen vaikutusta suhteellisiin frekvensseihin.

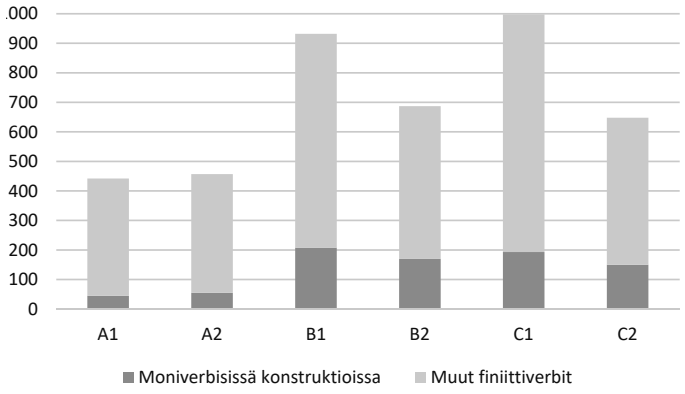
Aineiston tekstejä tarkastellaan taitotasoittain tekstimassana, ja aineistosta etsitään eri taitotasoilla esiintyviä ja niille tyypillisiä piirteitä. Tavoitteena on löytää ryhmätasolla näkyviä trendejä kompleksisuuden kehityksessä. Yksilöllinen kielenkehitys, joka voi olla hyvin erilaista kuin ryhmätason kehitys (esim. Larsen-Freeman 2006; 2009), on rajattu tämän tutkimuksen ulkopuolelle.

Tulosten esittely aloitetaan aineiston määrällisellä kuvauksella. Aluksi esitellään tutkittavien moniverbisten konstruktioiden absoluuttiset frekvenssit ja moniverbisten konstruktioiden osuus kaikista finiittiverbikonstruktioista taitotasoittain ja tarkastellaan konstruktioiden leksikaalista variaatiota niissä esiintyvien verbien kirjon avulla. Sen jälkeen moniverbisten konstruktioiden käyttöä eri taitotasoilla

tarkastellaan aineistoesimerkkien avulla. Ensimmäisenä kartoitetaan aineistossa yleisimmän moniverbisen konstruktion eli *voida*-verbistä ja A-infinitiivistä koostuvan konstruktion käyttöä, sitten finiittiverbin ja MA-infinitiivin sisältävien konstruktioiden ja lopuksi finiittiverbin ja partisiippimuodon sisältävien konstruktioiden käyttöä eri taitotasoilla.

4. Tulokset

Finiittiverbien absoluuttiset frekvenssit eri taitotasoilla sekä moniverbisten konstruktioiden osuus kaikista finiittiverbien esiintymistä on esitetty kuviossa 1.

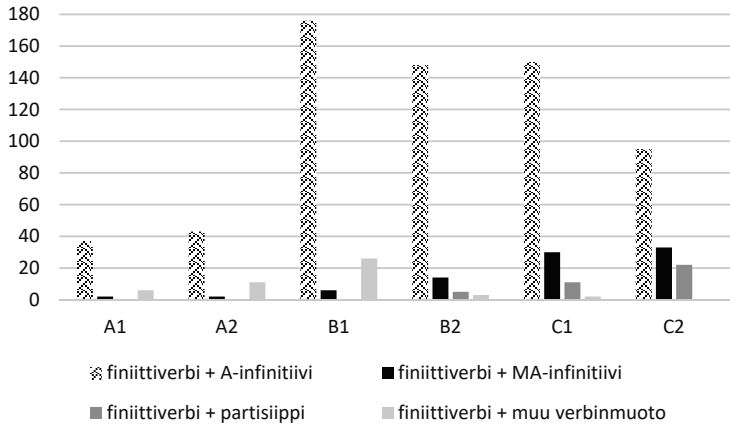


Kuvio 1. *Finiittiverbien absoluuttiset frekvenssit ja jakautuminen moniverbisissä konstruktiossa esiintyviin ja muihin finiittiverbeihin*

Ensimmäiset finiittiverbin ja infiniittisen verbinmuodon sisältävät moniverbiset konstruktiot ilmaantuvat aineiston teksteihin heti taitotasolla A1. Saman havainnon ovat tehneet Haapala (2008), Kynsijärvi (2008), Paavola (2008), Puhakka (2010) ja Seilonen (2013). Nyt tarkasteltavien moniverbisten konstruktioiden absoluuttiset frekvenssit ovat suurimmillaan taitotasoilla B1 ja C1. Tätä voi selittää se, että näillä taitotasoilla aineiston kokonaissanamäärä on suurin. Myös taitotasoilla

B2 ja C2 tekstien kokonaissanamäärät ovat lähes kaksinkertaisia taitotasoihin A1 ja A2 verrattuna, mikä voi osaltaan selittää sitä, että konstruktioiden esiintymiä on ylemmillä taitotasoilla alempia enemmän. Tulokset ovat melko samansuuntaisia kuin aiemmat tulokset. Haapalan (2008) mukaan erilaisia infiniittirakenteita on eniten keskitasolla. Paavolan (2008) aineistossa moniverbisten rakenteiden suhteellinen määrä kasvoi taitotasolle B1 saakka ja väheni sen jälkeen hieman. Kynsijärven tarkastelemien *olla*-verbin sisältävien moniverbisten rakenteiden määrä puolestaan oli kielitaidon ylimmällä tasolla kolminkertainen keskitasoon verrattuna. Tässä tutkimuksessa tarkasteltavien moniverbisten konstruktioiden osuus kaikista finiittiverbien esiintymistä on suurin taitotasoilla B2 ja C2, ja taitotasoilla B1–C2 niiden osuus kaikista finiittiverbiesiintymistä on noin kaksinkertainen taitotasoihin A1 ja A2 verrattuna.

Myös tämän tutkimuksen aineistossa yleisin infiniittinen verbinmuoto on kaikilla taitotasoilla A-infinitiivi, joka ilmaantuu aineistoon heti taitotasolla A1 (kuvio 2).



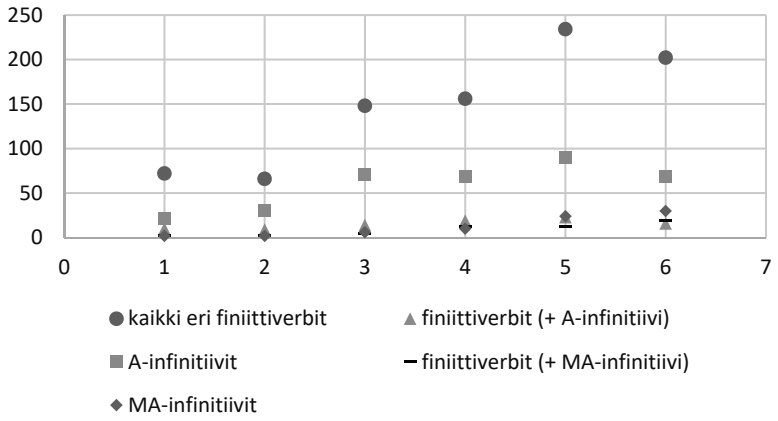
Kuvio 2. *Eri rakennekonstruktioiden absoluuttiset frekvenssit taitotasoittain*

Taitotasoilla A1 ja A2 noin neljä viidesosaa tarkasteltavista moniverbisistä konstruktiosta sisältää A-infinitiivin, ja C2-tasolla tällaisten esiintymien osuus on yli 60 % moniverbisistä konstruktioista. Myös MA-infinitiivit ilmaantuvat finiittiverbin laajennuksiksi jo taitotasolla A1, tosin taitotasoilla A1–B1 ne ovat vielä harvinaisia ja esiintymiä on vain joitakin. Taitotasolla B2 MA-infinitiivin sisältävien konstruktioiden absoluuttinen esiintymistaajuus on kolminkertainen taitotasoon B1 verrattuna, ja niiden osuus kasvaa aina taitotasolle C2 saakka mutta jää silläkin vain noin viidesosaan kaikista tarkasteltavista moniverbistä konstruktiosta. Tulos on samansuuntainen kuin Paavolan (2008) havainto, että A-infinitiivitäydennykset ilmaantuvat suomenoppijoiden teksteihin aikaisemmin kuin MA-infinitiivitäydennykSet. Moniverbisiin *olla*-verbin sisältäviin rakenteisiin MA-infinitiivi ilmaantuu Kynsijärven (2010) mukaan vasta C-tasolla. Partisiipit ilmaantuvat tarkasteltaviin konstruktioihin viimeisinä, taitotasolla B2, ja partisiipin sisältävä rakenne on tarkasteltavissa konstruktioissa selvästi harvinaisin.

Kuvion 2 ryhmä finiittiverbi + muu verbinmuoto sisältää esiintymät, joissa finiittiverbiä seuraava verbinmuoto on muu kuin infinitiivi- tai partisiippimuoto, esimerkiksi verbin persoonamuoto (6) tai verbin vartalo (7). Joidenkin esiintymien kohdalla ryhmittely on tulkinnanvaraista: esimerkin 7 konstruktiossa *voit kuntele* finiittiverbiä seuraava muoto poikkeaa selvästi *kuunnella*-verbin A-infinitiivistä, sen sijaan konstruktiossa *voit ota* kyse voi olla myös oikeinkirjoituksen horjunnasta (ks. myös Haapala 2008; Paavola 2008).

(6) Minä ajatellen että luokan **pitaisi on** lämpimämpi. (A2, F-629)
(7) Oman auton **voit ota** esimerkiksi vaunut jos lapset tärvitsevät, ruoka, lelut, vatet ja **voit kuntele** mitä sinä haluat. (B1, F-654)

Moniverbisissä konstruktioissa esiintyvien verbien leksikaalista variaatiota tarkasteltiin laskemalla aineistosta erikseen kaikki finiittisessä verbinmuodossa, A-infinitiivissä tai MA-infinitiivissä esiintyneet eri verbit sekä vielä erikseen kaikki A-infinitiivin tai MA-infinitiivin kanssa esiintyneet eri finiittiverbit (kuvio 3).



Kuvio 3. *Verbien leksikaalinen variaatio taitotasoittain*

Kun moniverbisten konstruktioiden absoluuttinen frekvenssi aineistossa kasvaa, niiden leksikaalinen variaatio lisääntyy. Käytettyjen verbien kirjo kasvaa sekä moniverbisten konstruktioiden finiittiverbeissä että infinitiiveissä, minkä lisäksi myös kaikkien finiittiverbien kirjo kasvaa. Nyt tarkasteltavassa aineistossa A-infinitiivin kanssa moniverbisen konstruktion muodostavien eri verbien lukumäärä kasvaa taitotasolle C1 saakka, MA-infinitiivin laajennukseksi saavien finiittiverbien puolestaan taitotasolle C2 saakka. Myös infinitiiveissä verbien kirjo kasvaa taitotasolle C1 saakka. Konstruktioiden leksikaalisen variaation lisääntyminen näkyy myös taitotasolla toistuvien verbien määrässä: vain kerran esiintyvien infinitiivien osuus kaikista taitotason A-infinitiiveistä on suurin taitotasolla C2. Myös Paavolan (2008) mukaan verbien kirjo kasvaa selvästi B-tasolla ja ylimmillä taitotasoilla *voida*-ketjujen A-infinitiiveissä on eniten eri verbejä ja eniten sellaisia verbejä, jotka eivät toistu.

Aineistossa A-infinitiivin kanssa esiintyvien verbien kirjoa verrattiin myös ”Isossa suomen kieliopissa” lueteltuihin A-infinitiivin kanssa esiintyviin verbeihin (§ 469). Pykälässä on verbiketjun muodostavia verbejä 16 ja infinitiivilausekkeen mahdollisesti objektiksi saavia 48. Näistä

tämän tutkimuksen aineistossa yleisimpiä ovat verbiketjun muodostavat verbit, joista aineiston moniverbisissä konstruktioissa on käytetty yhdeksää (*voida*, *pitää*, *täytyä*, *tarvita*, *saada*, *kannattaa*, *saattaa*, *alkaa*, *ehtiä*), muualla aineistossa finiittiverbinä neljää (*sopia*, *tahtoa*, *tavata*, *mahtaa*) ja A-infinitiivinä yhtä (*uhata*) ja joista kahta (*meinata* ja *taitaa*) ei esiinny aineistossa lainkaan. A-infinitiivin mahdollisesti objektiksi saavista verbeistä (VISK: §469) aineistossa on käytössä noin puolet. A-infinitiivin kanssa näistä verbeistä esiintyy 12 (*haluta*, *osata*, *yrittää*, *jaksaa*, *aikoa*, *uskaltaa*, *tahtoa*, *ajatella*, *muistaa*, *päättää*, *ra(a)skia*, *viitsiä*), muualla aineistossa 11 (*tietää*, *ymmärtää*, *nähdä*, *huomata*, *unohtaa*, *luvata*, *välittää*, *keksiä*, *pelätä*, *arvata*, *kehdata*). Seppälän (2013) tutkimuksessa yleisimmät A-infinitiivin kanssa esiintyvät verbit ovat *voida*, *haluta*, *saada*, *yrittää* ja *alkaa*, joista neljä ensimmäistä ovat myös tämän tutkimuksen aineistossa yhdeksän yleisimmän A-infinitiivin kanssa esiintyvän verbin joukossa. Eroa selittää se, että Seppälän tarkastelussa ei ollut mukana nesessiiviverbejä.

Lisäksi aineistossa esiintyy A-infinitiivin kanssa 13 muuta finiittiverbiä, jotka ovat yleisyysjärjestyksessä *pystyä*, *tulla*, *aloittaa*, *antaa*, *joutua*, *toivoa*, *auttaa*, *jäädä*, *neuvoa*, *opettaa*, *oppia*, *pyytää*, *valita*. Näistä vain kuusi ensimmäistä esiintyy aineistossa A-infinitiivin kanssa useammin kuin kerran. *Pystyä* ja *joutua* esiintyvät sekä A-infinitiivin että MA-infinitiivin kanssa (ks. myös Paavola 2008: 60–61; Seppälä 2013: 325–327). Verbin *aloittaa* yhdistäminen A-infinitiiviin ei ole oppijansuomessa harvinaista, Seppälän (2013: 331) mukaan se esiintyy useammin A-infinitiivin kuin teonnimen kanssa.

Moniverbisten konstruktioiden yleisin finiittiverbi on *voida* kaikilla muilla taitotasoilla kuin B1, jolla hieman sitä yleisempi on nesessiivinen *pitää*. Koko aineistossa yleisin konstruktio onkin *voida* + A-infinitiivi. Tätä selittävät sekä A-infinitiivin että *voida*-verbin yleisyys aineiston kaikilla taitotasoilla. *Voida* on koko aineiston toiseksi yleisin finiittiverbi heti *olla*-verbin jälkeen ja sen esiintymistä 90 % on A-infinitiivikonstruktiossa. Myös Seppälän (2013) mukaan *voida* on selvästi yleisin A-infinitiivin kanssa esiintyvä finiittiverbi. Seuraavaksi konstruktion

voida + A-infinitiivi käyttöä eri taitotasoilla tarkastellaan aineistoesimerkkien avulla.

Tyypillisesti verbin *voida* kanssa käytetään A-infinitiiviä. Aineistosta löytyy vain yksi esiintymä, jossa kirjoittaja käyttää sen sijaan MA-infinitiiviä (8) (ks. myös Seilonen 2013: 106–107). Lisäksi etenkin taitotasolla B1 on esiintymiä, joissa finiittiverbiä seuraa muu kuin infinitiivi (9).

(8) minä olen mielestani siksi sopivää lähikaupassa on hyvää käyda heiti kuin tarvitsen jotaikin **voihakemaan** (A1, FF-277)
(9) Oman auton hyvää puoli **voit mene** ihan van. (B1, F-669)

Taitotasolla A1 *voida*-konstruktiota käytetään sekä nollapersoonaisissa lauseissa (10) että persoonamuotoisena subjektin kanssa. Finiittiverbi voi jo tällä taitotasolla olla negatiivinen (11).

(10) **voi olla** että niillä ei ole mitä sinä just halut. (A1, F-273)
(11) Rakastan juustoa, mutta **en voi ostaa** monta erilaista merkijä. (A1, F-270)

Taitotasolla A2 konstruktioon ilmaantuvat ensimmäiset rakenteellisen variaation merkit. Konstruktion osien väliin ilmaantuvat ensimmäiset adverbit (12). Jo taitotasolla A2 verbi + verbi -rakenteet voivat alkaa ketjuuntua, vaikka infinitiiviin liitetyt infiniittiset täydennykset eivät vielä olisikaan täysin kohdekielen mukaisia (13).

(12) **Ei voi nopesti löytää** tavaran hinnan. (A2, F-280)
(13) Elokuvat tai laulut (suomeksi) **voivat auttaa oppia** kieltä. (A2, F-622)

Taitotasolla B1 konstruktiossa aletaan käyttää konjunktioilla *ja* sekä *tai* rinnastettuja infinitiivejä (14). Lisäksi ilmaukset pitenevät, kun erilaisia verbi + verbi -rakenteita aletaan yhdistellä (15).

(14) Jos perhe matkustaa autolle, **voi pysähtyä** missä vain **ja katsoa** mitä haluat. (B1, F-640).
(15) Kaupungissa on aina kiire, **voi olla vaikea löydä** ravintola, **voi ola** jono museossa. (B1, F-661)

Myös taitotasolla B2 infinitiivien rinnastaminen on melko yleistä. Lisäksi ilmaukset pitenevät, kun *voida*-konstruktioon yhdistetään muita verbi + verbi -rakenteita (16, 17). Näissä esimerkeissä näkyvät myös konstruktion osien hajaantuminen, morfologinen variaatio niin finiittiverbissä kuin infiniittisessä verbinmuodossa sekä erilaiset idiomaattiset verbi-ilmaukset. Lisäksi tältä taitotasolta konstruktiosta löytyy ensimmäinen esiintymä, jossa A-infinitiiviin on lisätty liitepartikkeli (*voi oppiakin*, F-679).

(16) Myös koulussa on tarkeää että opettaja **voisi huomata** jos jollakulla opiskelijalla on ongelmia, opettaja **voi auttaa häntä ratkaisemaan ongelmia tai etsimään apua** jostakin henkilöstä tai yhteisöltä. (B2, F-679)

(17) Kun on kymmenentuhansia autoja, se **voi käydä hermoille aina etsiä** parkkipaikkaa. (B2, F-709)

C-tasolla *voida* + A-infinitiivi -konstruktion laajeneminen infinitiivejä rinnastamalla vähenee ja infiniittisten muotojen yhdisteleminen muilla tavoin lisääntyy (18). Myös muut kuin finiittiverbin laajennuksena toimivat infiniittiset muodot yleistyvät ja niitä käytetään samoissa virkkeissä moniverbisten konstruktioiden kanssa (19, 20).

(18) ”Järkeä verorahojen käytön” puolesta **kirjoittanut** lukija väittää, että **graffitia voisi käyttää** laajemmin **piristämään** harmaata betonia. (C1, F-1024)

(19) **Ei voi olla salittua** että jotkut **häiritsevät** jatkuvasti ihmisten kotirauhaa **soittamalla myydäkseen** turhia juttuja. (C2, F-1088)

(20) **Ei tarvitse odottaa** kylminä sateisina päivinä pienellä pysäkillä ilman katosta **palellen**, vaan **voi asettua istumaan** autoon **ja lähteä** valittuun paikkaan. (C1, F-687)

Esimerkissä 20 leksikaalinen variaatio näkyy sekä finiittiverbissä että A-infinitiivissä. Ensimmäinen moniverbisistä konstruktioista sisältää kieltoverbin *ei*, ja *voida*-verbin sisältävässä konstruktiossa infiniittisiä verbinmuotoja yhdistetään sekä ketjuuntumalla (*voi asettua istumaan* eikä esim. *voi istua*) että rinnastamalla (*voi asettua ja lähteä*). Lisäksi kirjoittaja käyttää virkkeessä E-infinitiiviä (*palellen*).

Seuraavaksi tarkastellaan finiittiverbin ja MA-infinitiivin sisältävien konstruktioiden käyttöä eri taitotasoilla. Taitotasoilla A1 ja A2 MA-infinitiivi on harvinainen ja sijamuodon valinnassa on selvästi horjuntaa. MA-infinitiivin oikean sijamuodon valinnan ongelmallisuuden oppijansuomessa on havainnut myös Seppälä (2013). Moniverbinen konstruktio voi kuitenkin olla myös täysin kohdekielen mukainen, vaikka muissa verbirakenteissa olisi vielä ongelmia (21) tai jäädä muodon kohdekielisyydestä huolimatta merkitykseltään epäselväksi (22): kurssipalautteessa käyttämällään *kieli oppii puhumaan* kirjoittaja voi tarkoittaa, että kieltä pitäisi oppia myös puhumaan tai mahdollisesti että kieltä oppii puhumalla.

(21) Minä **olin syömässä** ravintolassa Helsingissa, minä nähnyt 3 huonoa asiaa ja 1 hyvä asia (A1, F-1012)
(22) Me puhumme vähän kursilla, vain opettaja. | **Kieli oppii puhumaan, ei vain kuuntelemaan**. Kun meidän täytyy sanoa tai kysyä jotakin, se on oikein vaikea. (A2, F-622)

Myös taitotasolla B1 MA-infinitiivin sisältäviä moniverbisiä konstruktioita käytetään vain vähän. Konstruktiot kuitenkin monipuolistuvat, vaikka esiintymät ovat vielä yksittäisiä. Tällä taitotasoilla aineiston teksteihin ilmaantuvat merkitykseltään futuurinen *tulla tekemään* (23) sekä konstruktio *jäädä tekemättä* (24). Näistä *tulla tekemään* ilmaantuu Puhakan (2010) mukaan aikuisten aineiston teksteihin jo taitotasolla A2 ja nuorten aineiston teksteihin heti taitotasolla A1. Taitotason B1 teksteissä MA-infinitiivi voidaan myös yhdistää finiittiverbin ja A-infinitiivin sisältävään konstruktioon (25).

(23) Mutta ilman auto **tule olemaan** vaikea että ainoa mitä voidaan tehdä tämän asian eteen on se että keksitään keinoja millä voidaan vältä näitä vahingoja Mikä auto jätä. (B1, F-648)
(24) Kansa unohtuu ja kaikki luvattut asiat **jäävät tekemättä**. (B1, F-403)
(25) Vanhemat **pitää opettaa oman lapset kunnioittammaan** opettajat, henkilökunta ja muuita oppilaaita. (B1, F-641)

Taitotasolla B2 finiittiverbin ja MA-infinitiivin sisältävien konstruktioiden absoluuttinen frekvenssi on noin kaksinkertainen taitotasoon B1 verrattuna, samoin konstruktioissa käytettyjen finiittiverbien kirjo. Miltei kaikissa esiintymissä MA-infinitiivi on illatiivissa. Illatiivi on myös Seppälän (2013) mukaan MA-infinitiivin yleisin sijamuoto. Taitotasolla B2 konstruktiot voivat hajaantua (26), MA-infinitiivejä voidaan rinnastaa (27) ja niitä voidaan yhdistää finiittiverbin ja A-infinitiivin sisältäviin konstruktioihin (28).

(26) Tämä kaikki analysointi **auttaa** heitä paremmin **auttamaan** heidän lastensa. (B2, F-676)
(27) Koulu on paikka, jossa lapsi **oppii kirjoittamaan ja lukemaan.** (B2, F-708)
(28) Minusta meidän **täytyy pyrkkia käyttämään** autoja vähemmän näissä olosuhteissa. (B2, F-700)

Taitotasoilla C1 ja C2 finiittiverbin ja MA-infinitiivin sisältävien konstruktioiden frekvenssi on kaksinkertainen taitotasoon B2 verrattuna. Myös morfologinen ja leksikaalinen variaatio kasvavat. Eri verbejä on taitotasolla C2 eniten sekä finiittiverbeinä että MA-infinitiiveinä. Finiittiverbien kirjo kasvaa lähes puolitoistakertaiseksi taitotasojen C1 ja C2 välillä, infinitiivinä käytettyjen verbien kirjo puolestaan on yli kaksinkertainen taitotasolla C1 taitotasoon B2 verrattuna ja kasvaa vielä hieman taitotasolle C2.

Vaikka eri sijamuotoja on eniten käytössä taitotasoilla C1 ja C2, myös niillä on vain yksittäisiä esiintymiä MA-infinitiivin adessiivista ja abessiivista. Yleisin sijamuoto on illatiivi, joka kattaa kolme neljäsosaa kaikista esiintymistä. Toiseksi yleisin on inessiivi (29, 30), jota käytetään etenkin leksikaalistuneessa ilmauksessa *olla olemassa*. Finiittiverbin ja MA-infinitiivin muodostamista verbiliitoista C-tasolla esiintyvät *olla tekemättä*, *jäädä tekemättä* ja *jättää tekemättä*. Taitotasolla C2 *jättää tekemättä* esiintyy myös osana moniverbistä konstruktiota, jonka finiittiverbinä on *yrittää* (31). Lisäksi taitotasolla C2 on aineiston ainoa *ruveta*-verbin esiintymä (32).

(29) Itse **olisin** luontokohdetta **ihailemassa**. (C1, F-702)
(30) Ehdokkaat **istuvat** studiossa **kertomassa** kuka perheestä vie roskat ulos ja vaihtelevat ruokaohjeita. (C2, F-1006)
(31) Kun minä olen lähellä he todella **yrittävät jättää** rumat sanat **sanomatta**. (C1, F-1039)
(32) Kun olin syönyt suklaata, niin hampaani **rupesi särkemään**. (C2, F-245)

Muista MA-infinitiivin kanssa verbiketjun muodostavista seitsemästä verbistä (VISK: § 496) on aineiston moniverbisissä konstruktioissa käytössä vain yksi, *pyrkiä*, joka esiintyy aineistossa finiittiverbinä seitsemän kertaa mutta vain kahdesti moniverbisessä konstruktiossa (taitotasoilla B2 ja C2). Muista mahdollisista verbeistä *sattua* esiintyy aineiston teksteissä vain muissa kuin moniverbisissä konstruktioissa, verbit *lakata*, *pakata* ja *tupata* puolestaan eivät esiinny aineistossa lainkaan. Kaikki aineiston moniverbisten konstruktioiden finiittiverbinä käytetyt *alkaa*-verbit esiintyvät A-infinitiivin kanssa.

Kun aineiston moniverbisissä konstruktioissa esiintyviä finiittiverbejä verrataan ”Ison suomen kieliopin” (VISK: § 479) muotteihin, joissa MA-infinitiivilauseketta voidaan käyttää, havaitaan, että eniten on suuntautumis- ja vaikuttamismuoteissa käytettäviä verbejä (18 eri verbiä) ja että estämis- ja estymismuoteissa käytettävistä verbeistä aineiston moniverbisissä konstruktioissa esiintyy vain yksi, *estää* (33).

(33) **Eikä** mitkään valvontakamerat ja poistamiset **estä** nuoria **tekemästä** niitä. (C2, F-1025)

Finiittiverbin ja partisiippimuodon sisältävät konstruktiot ilmaantuvat aineiston teksteihin tarkasteltavista rakenteista viimeisenä, taitotasolla B2. Puhakan (2010) aineistossa, joka sisältää nyt tarkasteltujen mielipidetekstien lisäksi viestejä, rakenne *tulee tehtyä* ilmaantuu aikuisten aineiston teksteihin jo yhtä taitotasoa aikaisemmin. Nominilausekkeisiin ensimmäiset partisiipit ilmaantuvat jo taitotasolla A2, tosin niissäkin ne alkavat yleistyä vasta taitotasolla B2. Oppijansuomen melko vähäiseen partisiippien käyttöön on kiinnittänyt huomiota myös Ivaska (2014: 176–177), joka on todennut, että VA-partisiippi esiintyy edistyneessä

oppijansuomessa etenkin referatiivirakenteen predikaattina harvemmin kuin äidinkielisessä vertailuaineistossa.

Suurin osa partisiipin sisältävistä moniverbisistä konstruktioista on erilaisia verbiliittoja. ”Ison suomen kieliopin” (VISK: § 496) luettelemista havaintoverbin ja partisiipin muodostamista verbiketjuista aineistossa on vain yksi esiintymä taitotasolla C2 (34). Aineistossa yleisin finiittiverbin ja partisiipin sisältävä konstruktio on taitotasolla B2 ilmaantuva nesessiivinen *on tehtävä*, joka taitotasolla C1 kattaa kaksi kolmasosaa ja taitotasolla C2 yli puolet konstruktion esiintymistä (rakenteen käytöstä ks. myös Seilonen 2013). Muista mahdollisista verbiliitoista on aineistossa lähinnä yksittäisiä esiintymiä: *tulla tehtyä* esiintyy taitotasoilla B2–C2 kerran kullakin, *olla tehtävissä* taitotasolla C2 kahdesti ja *tulla tehneeksi* taitotasolla C2 kerran.

(34) Ainona ratkaisuna **näyttää olevan muuttaa** muualle **kertomatta** uutta osoitetta! (C2, F-1088)

Myös partisiipin sisältäviä moniverbisiä rakenteita voidaan rinnastaa sekä yhdistää muihin infiniittisiin muotoihin, kuten esimerkissä (34).

5. Lopuksi

Tässä artikkelissa esiteltyjen tutkimustulosten mukaan moniverbiset konstruktiot ilmaantuvat aineiston teksteihin heti taitotasolla A1 ja alkavat yleistyä taitotasolla B1. Moniverbisissä konstruktioissa esiintyvien finiittiverbien osuus kaikista finiittiverbeistä lähes kaksinkertaistuu taitotasojen A2 ja B1 välillä. Tulos on linjassa aiempien tulosten (Haapala 2008; Kynsijärvi 2008; Paavola 2008; Puhakka 2010; Seilonen 2013) kanssa. Esiintymien määrän muutoksen voi olettaa näkyvän myös syntaktisen kompleksisuuden määrällisissä mittareissa. Samasta aineistosta aiemmin tehdyn määrällisen tutkimuksen mukaan sekä lauseet että T-yksiköt ovat taitotasoilla A1 ja A2 lyhyempiä kuin muilla taitotasoilla ja erot ovat tilastollisesti merkitseviä (Mylläri 2020). Moniverbisten konstruktioiden yleistyminen ei yksin selitä näitä eroja, mutta voi osaltaan vaikuttaa niihin.

Keskitasolla ja ylimmillä taitotasoilla moniverbisten konstruktioiden suhteellisessa frekvenssissä ei tapahdu suuria muutoksia, mutta konstruktioiden variaatio lisääntyy: niissä käytettyjen verbien kirjo laajenee ja A-infinitiivin rinnalle käyttöön tulevat MA-infinitiivin eri sijamuodot sekä partisiippimuodot. Ylimmillä taitotasoilla on käytössä eniten erilaisia konstruktioita. Myös tämä tulos on linjassa aiempien tutkimustulosten kanssa (Haapala 2008; Kynsijärvi 2008; Paavola 2008; Puhakka 2010; Seilonen 2013). Muutokset moniverbisissä konstruktioissa voivat näkyä syntaktisen kompleksisuuden määrällisissä mittareissa. Kun infiniittisiä muotoja rinnastetaan ja infiniittiset muodot saavat omia infiniittisiä laajennuksia, moniverbisten konstruktioiden piteneminen voi näkyä myös lauseiden ja T-yksiköiden pitenemisenä. Toisaalta rinnastuksissa käytetyt elliptiset rakenteet voivat tiivistää ilmaisua ja näin pienentää lauseiden ja T-yksiköiden sanamäärää.

Leksikaalinen ja morfologinen variaatio sen sijaan eivät näy syntaktisen kompleksisuuden määrällisissä mittareissa. Mittareiden ulottumattomiin jää myös se, että ylemmillä taitotasoilla käyttöön tulevat merkitykseltään eriytyneet konstruktiot ja että ilmausten idiomaattisuus lisääntyy. Teksteissä käytettyjen verbien ja infiniittisten verbinmuotojen kirjon laajeneminen kuitenkin kasvattavat kielisysteemin osien ja erilaisten yhdistämistapojen määrää ja monipuolistavat näin oppijan käytössä olevia kielenpiirteitä, mikä voitaisiin tulkita kompleksisuuden lisääntymiseksi (vrt. Bulté & Housen 2012; Ortega 2003). Tulokset moniverbisten konstruktioiden sanastollisesta ja rakenteellisesta laajenemisesta ilman muutoksia konstruktion sanamäärässä tukevat osaltaan näkemyksiä, joiden mukaan rakenteiden monipuolisuus pitäisi ottaa aiempaa paremmin huomioon syntaktisen kompleksisuuden mittaamisessa (De Clercq & Housen 2017) ja ettei kompleksisuuden eri puolia välttämättä tavoiteta pelkillä määrällisillä mittareilla (Larsen-Freeman 2009; Martin 2013; Reiman 2011b).

Tässä tutkimuksessa oppijansuomen kompleksisuutta on tarkasteltu ryhmätasolla poikittaisaineistossa, joka on melko pieni ja koostuu vain yhdestä tehtävätyypistä. Vaikka ryhmätason tarkastelulla ei tavoiteta

yksilöllistä vaihtelua, se antaa yleiskuvan kompleksisuuden kehittymisestä. Lisäksi variaatio voi tulla yksittäisiä tekstejä selvemmin esiin ryhmätason tarkastelussa. Kun konstruktioiden varanto laajenee, yksittäisen konstruktion puuttuminen tuotoksesta ei välttämättä tarkoita, että se puuttuisi kielivarannosta. Esimerkiksi nesessiivisyyden ilmausten kirjo on laajin C-tasolla (Seilonen 2014), jolloin on mahdollista ja myös todennäköistä, että etenkin lyhyessä tekstissä kirjoittaja käyttää vain joitakin hallussaan olevista keinoista. Poikittaisaineisto ei myöskään tavoita yksilön kielen kehitystä vaan kertoo tarkasteltavan kielenpiirteen esiintymisestä eri taitotasoilla. Lisäksi tulosten yleistämisessä on otettava huomioon, että syntaktisessa kompleksisuudessa voi olla eroja tehtävätyyppien ja tekstilajien välillä (esim. Michel 2017; Pallotti 2009).

Vaikka tämän tutkimuksen tuloksia ei sellaisenaan voi käyttää erottelemaan taitotasoja toisistaan tai sijoittamaan yksittäisiä tekstejä taitotasoille, ne antavat yleiskuvan oppijansuomen kompleksisuuden moniulotteisuudesta. Lisäksi ne osoittavat, että suomen kaltaisessa morfologisesti rikkaassa kielessä syntaktinen, morfologinen ja leksikaalinen kompleksisuus voivat kietoutua yhteen monella eri tavalla, myös sellaisilla, jotka eivät tule näkyviin syntaktisen kompleksisuuden usein käytetyillä lauseen tai T-yksikön sanamääriin perustuvilla määrällisillä mittareilla. Samoin sanavaraston kasvun tai abstraktien rakenteiden ilmaantumisen tarkastelu näyttäisivät tuovan esiin vain osan kompleksisuuden monista ulottuvuuksista. Tässä tutkimuksessa käytetty konstruktiolähtöinen tarkastelutapa yhdistää eri näkökulmia ja voisi siten esimerkiksi kielen tilanteiseen vaihteluun kohdistuvaan tutkimukseen yhdistettynä syventää kuvaa oppijankielen kompleksisuudesta ja sen kehittymisestä edelleen.

Lähteet

Alanen, Riikka, Ari Huhta, Mirja Tarnanen 2010. Designing and assessing L2 writing tasks across CEFR proficiency levels. – Inge Bartning, Maisa Martin, Ineke Vedder (eds.). Communicative Proficiency and Linguistic Development: Intersections Between SLA and Language Testing Research. Rome: Edisegno srl., 21–56.

Bardovi-Harlig, Kathleen 1992. A second look at T-Unit analysis: Reconsidering the sentence. – Tesol Quarterly 26 (2), 390–395. https://doi.org/10.2307/3587016

Biber, Douglas, Bethany Gray, Kornwipa Poonpon 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? – TESOL Quarterly 45 (1), 5–35. https://doi.org/10.5054/tq.2011.244483

Brunni, Sisko, Marja-Liisa Lehto, Jarmo H. Jantunen, Valtteri Airaksinen 2015. How to annotate morphologically rich learner language: Principles, problems and solutions. – Bergen Language and Linguistics Studies 6, 133–152. https://doi.org/10.15845/bells.v6i0.812

Bulté, Bram, Alex Housen 2012. Defining and operationalising L2 complexity. – Alex Housen, Ineke Vedder, Folkert Kuiken (eds.). Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. Language Learning & Language Teaching 32. Amsterdam: John Benjamins Publishing Company, 21–46. https://doi.org/10.1075/lllt.32.02bul

De Clercq, Bastien, Alex Housen 2017. A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. – The Modern Language Journal 101 (2), 315–334. https://doi.org/10.1111/modl.12396

Ellis, Rod, Gary Barkhuizen 2005. Analysing Learner Language. Oxford: Oxford University Press.

EVK 2003 = Eurooppalainen viitekehys 2003. Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys. Helsinki: WSOY.

Haapala, Terhi 2008. Finiittiverbeistä verbiketjuihin: verbiytimien kompleksistuminen S2-oppijoiden kielessä. Pro gradu -tutkielma. Tampereen yliopisto.

Herlin, Ilona, Laura Visapää 2005. Elävä kielioppi. Suomen infiniittisten rakenteiden dynamiikkaa. Suomalaisen Kirjallisuuden Seuran toimituksia 1021. Helsinki: Suomalaisen Kirjallisuuden Seura.

Housen, Alex, Bastien De Clercq, Folkert Kuiken, Ineke Vedder 2019. Multiple approaches to complexity in second language research. – Second language Research 35 (1), 3–21. https://doi.org/10.1177/0267658318809765

Huhta, Ari, Riikka Alanen, Mirja Tarnanen, Maisa Martin, Tuija Hirvelä 2014. Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. – Language Testing 31 (3), 307–328. https://doi.org/10.1177/0265532214526176

Ivaska, Ilmari 2014. Edistyneen oppijansuomen avainrakenteita. Korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin. – Virittäjä 118 (2), 161–193.

Jantunen, Jarmo H., Silja Pirkola 2015. Oppijansuomen sähköiset tutkimusaineistot: Nykytilanne. – Virittäjä 119 (1), 88–103.

Kuiken, Folkert, Ineke Vedder 2019. Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. – International Journal of Applied Linguistics 29 (2), 192–210. https://doi.org/10.1111/ijal.12256

Kyle, Kristopher, Scott A. Crossley 2018. Measuring syntactic complexity in L2 writing using fine grained clausal and phrasal indices. – The Modern Language Journal 102 (2), 333–349. https://doi.org/10.1111/modl.12468

Kynsijärvi, Taru 2007. *Se johtuu siitä, että minulla oli muistinmenetys: Olla*-verbirakenteiden kehkeytyminen oppijankielessä. Pro gradu -tutkielma. Jyväskylän yliopisto.

Larsen-Freeman, Diane 2006. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. – Applied Linguistics 27 (4), 590–619. https://doi.org/10.1093/applin/aml029

Larsen-Freeman, Diane 2009. Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. – Applied Linguistics 30 (4), 579–589. https://doi.org/10.1093/applin/amp043

Martin, Maisa 2013. The complex simple: A problematic adjective in the CEFR writing scales. – Nordand 8 (2), 63–85.

Martin, Maisa, Sanna Mustonen, Nina Reiman, Marja Seilonen 2010. On becoming an independent user. – Inge Bartning, Maisa Martin, Ineke Vedder (eds.). Communicative Proficiency and Linguistic Development: Intersections Between SLA and Language Testing Research. Rome: Edisegno srl., 57–80.

Michel, Marije 2017. Complexity, accuracy, and fluency in L2 production. – Shawn Loewen, Masatoshi Sato (eds.). The Routledge Handbook of Instructed Second Language Acquisition. Florence: Routledge, 50–68. https://doi.org/10.4324/9781315676968-4

Mylläri, Taina 2020. Measuring syntactic complexity in learner Finnish. – Journal of Applied Language Studies 14 (2), 67–92. https://doi.org/10.47862/apples.99134

Niiranen, Leena 2010. Tapaustutkimus kolmen suomenoppijan kompleksisista verbikonstruktioista. – Lähivõrdlusi. Lähivertailuja 20, 155–190. https://doi.org/10.5128/LV20.05

Norris, John M., Lourdes Ortega 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. – Applied Linguistics 30 (4), 555–578. https://doi.org/10.1093/applin/amp044

Ortega, Lourdes 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. – Applied Linguistics 24 (4), 492–518. https://doi.org/10.1093/applin/24.4.492

Paavola, Vilja 2008. *Haluaisitko menna muunkansa kalastaman?* Verbiketjujen kehkeytyminen suomi toisena kielenä -oppijoiden kielessä. Pro gradu -tutkielma. Jyväskylän yliopisto.

Pallotti, Gabriele 2009. CAF: Defining, refining and differentiating constructs. – Applied Linguistics 30 (4), 590–601. https://doi.org/10.1093/applin/amp045

Puhakka, Martta 2010. *Sit se meni ja tuli hetken päästä takas*: Verbit *mennä* ja *tulla* suomi toisena kielenä -oppijoiden teksteissä. Pro gradu -tutkielma. Jyväskylän yliopisto.

Reiman, Nina 2011a. Transitiivikonstruktio ikkunana syntaksin kehitykseen: Infiniittiset rakenteet ja passiivi taidon indikaattoreina S2-oppijoiden teksteissä. – AFinLAe Soveltavan kielitieteen tutkimuksia 3, 142–157.

Reiman, Nina 2011b. Two faces of complexity: structural measures and diversity of constructions. – Nordand 6 (2), 9–23.

Reiman, Nina 2014. Yläkoulun S2-oppilaiden transitiivi-ilmausten käyttö Eurooppalaisen viitekehyksen taitotasoilla. – Lähivõrdlusi. Lähivertailuja 24, 183–220. https://doi.org/10.5128/LV24.07

Rimmer, Wayne 2006. Measuring grammatical complexity: The Gordian knot. – Language Testing 23 (4), 497–519. https://doi.org/10.1191/0265532206lt339oa

Seilonen, Marja 2013. Epäsuora henkilöön viittaaminen oppijansuomessa. Jyväskylä Studies in Humanities 197. Jyväskylä: Jyväskylän yliopisto.

Seppälä, Tanja 2013. Oppijansuomen kolligaatit ketjuuntuvissa verbirakenteissa. – Lähivertailuja 23, 315–340. https://doi.org/10.5128/LV23.13

Shore, Susanna 2020. Lauseita ja vesinokkaeläimiä: Perinteisestä funktionaaliseen lauseoppiin. Suomalaisen Kirjallisuuden Seuran toimituksia 1460. Helsinki: Suomalaisen Kirjallisuuden Seura.

Spoelman, Marianne, Marjolijn Verspoor 2010. Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. – Applied Linguistics 31 (4), 532–553. https://doi.org/10.1093/applin/amq001

Tilma, Corinne 2014. The dynamics of foreign versus second language development in Finnish writing. Jyväskylä Studies in Humanities 233. Jyväskylä: Jyväskylän yliopisto.

Vilkuna, Maria 2003. Suomen lauseopin perusteet. Helsinki: Edita.

VISK = Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen, Irja Alho 2004. Iso suomen kielioppi. Helsinki: Suomalaisen Kirjallisuuden Seura. Verkkoversio: http://scripta.kotus.fi/visk

Wolfe-Quintero, Kate, Shunji Inagaki, Hae-Young Kim 1998. Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity. Technical Report No. 1. Honolulu: Second Language Teaching and Curriculum Center.

Taina Mylläri
Vilnius University, Centre for Scandinavian Studies
Universiteto gatve 5, LT-01131 Vilnius, Lithuania
taina.myllari@flf.vu.lt
Kieli- ja viestintätieteiden laitos
PL 35, FI-40014 Jyväskylän yliopisto, Finland
taina.myllari@jyu.fi

Verb constructions and complexity across proficiency levels in learner Finnish

TAINA MYLLÄRI
Vilnius University, University of Jyväskylä

Learner language development can be analysed by measuring complexity, accuracy and fluency.

Complexity, our focus here, can be defined as the range and sophistication of learner language. Syntactic complexity is typically analysed by quantitatively measuring the length of production units or the amount of subordination rather than by exploring syntactic variation and diversity in learner language. In this article, the development of syntactic complexity in written learner Finnish across the CEFR proficiency levels is studied by exploring changes in the use of non-finite verb constructions. The aim of the study is to bring to light differences in complexity that are not captured by the traditional quantitative measures of syntactic complexity.

The data in the study comprise 241 written learner Finnish texts (23,596 words) from the University of Jyväskylä Cefling project corpus and they cover all CEFR levels, from A1 to C2. The data are explored both quantitatively and qualitatively. The focus of the study is on the changes in the use of verb constructions containing a finite verb and at least one non-finitive verb form, and on how those changes may reflect the development of syntactic complexity.

The results show that the constructions studied do not necessarily grow in length but instead become more varied both lexically and structurally as proficiency increases. Such changes are not revealed by quantitative measures of syntactic complexity focusing on the length of production units. Hence, the results support calls to adopt a more qualitative approach to investigating syntactic complexity. They also suggest that, in some languages, syntactic, morphological and lexical complexity cannot always be separated.

Keywords: learner language; CEFR proficiency levels; syntactic complexity; infinitive constructions; Finnish