Edistyneen oppijansuomen avainrakenteita

Korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin

Ilmari Ivaska

1 Johdanto

Tarkasteltaessa tyypillisiä ensikielisen ja edistyneen ei-ensikielisen kielenkäytön välisiä eroja nostetaan erottavaksi tekijäksi usein sanastolliset ja rakenteelliset frekvenssierot (ks. esim. Ringbom 1998). Aineistoihin perustuvassa tutkimuksessa frekvenssit yhdessä muun esitetyn evidenssin kanssa siis joko tukevat tai eivät tue esitettyä hypoteesia eroista tai yhtäläisyyksistä. Frekvenssilähtöisyyden voi kuitenkin valjastaa toisen kielen tutkimuksen tarpeisiin myös soveltamalla niin sanottua tilastollisen avaimuuden (engl. *keyness*) käsitettä eli tutkimalla sellaisia vertailtavien aineistojen välisiä frekvenssieroja, jotka ovat tilastollisesti merkitseviä (Scott & Tribble 2006: 58–59). Tutkin tässä artikkelissa[1] jotain muuta kuin suomea ensikielenään käyttävien mutta suomen taidoissaan hyvin edistyneiden humanistisen alan yliopisto-opiskelijoiden (S2) tenttivastauksia ja vertaan niitä ensikielenään suomea käyttävien (S1) vastaaviin teksteihin. Keskityn tarkastelemaan piirteitä, joissa tarkastellut kielimuodot eroavat toisistaan määrällisesti eniten. Tutkimuskysymykseni ovat: 1) Missä morfologisissa muodoissa tai morfologisten muotojen yhdistelmissä on suurimmat frekvenssierot S1- ja S2-aineistojen välillä? 2) Mikä näiden morfologisten muotojen tyypillisessä käytössä voi selittää havaittuja eroja?

Tutkimus vastaa osaltaan siihen, millaiset kielelliset seikat erottavat vielä edistyneitäkin suomenoppijoita ensikielisistä suomenpuhujista. Tämä voi puolestaan tuoda uutta tietoa suomen oppimisen perustavanlaatuisista erityispiirteistä. Samaan aikaan tarkastelen sitä, miten eri aineistojen välisiä tyypillisiä eroja voidaan lähestyä aineistovetoisesti – ilman ennalta esitettyä hypoteesia mahdollisista eroista. Lisäksi tutkimus voi valottaa tarkastellun tekstilajin – akateemisen tenttivastauksen – erityispiirteitä.

1. Kiitän Kirsti Siitosta, Jarmo Jantusta, Maarit Muttaa sekä Virittäjän anonyymeja arvioijia heidän esittämistään hyödyllisistä kommenteista.

Tulokset voivatkin tarjota uusia näkökulmia keskusteluun akateemisen maahanmuuton kielitaitokysymyksistä.

Tutkimusmenetelmänä on niin sanottu avainrakenneanalyysi, jonka avulla samankielisistä mutta jonkin taustamuuttujan perusteella toisistaan erotettavista aineistoista etsitään niitä toisistaan erottavia kielen rakenteita ja tulkitaan erojen mahdollisia syitä. Avainrakenneanalyysi on luonteeltaan monimenetelmäinen siten, että määrällinen analyysi ohjaa laadullista ja laadullinen analyysi tarkentaa määrällisesti saatuja tuloksia (Ivaska & Siitonen 2011; Ivaska 2012; monimenetelmäisyydestä ks. Creswell, Plano Clark & Garret 2008: 67–70). Lähtökohtanani ovat sanojen morfologiset muodot, joiden frekvensseissä havaittavia eroja tarkastelen. Tämän jälkeen paneudun niihin morfologisiin muotoihin, joissa on suurin frekvenssiero aineistojen välillä. Tarkastelen morfologisten muotojen tyypillistä käyttöä ja siinä ilmeneviä eroja muodoissa esiintyvän leksikaalis-funktionaalisen vaihtelun ja tyypillisen esiintymisympäristön avulla. Aineistossa toistuvat määrälliset erot auttavat paikantamaan mahdollisia kielimuodolle tyypillisiä avainrakenteita, ja löytyvien erojen tarkempi laadullinen analyysi pyrkii tavoittamaan sen kielen ilmiön, joka synnyttää eron. Tutkin sekä yksittäisten sanamuotojen että sanamuotojen tyypillisten yhteisesiintymien frekvensseissä ilmeneviä eroja. Huolimatta morfologiseen muotoon perustuvasta lähtöasetelmasta analyysin lopputulokseksi nousee keskenään hyvin erilaisia konstruktioita leksikaalisesti spesifeistä (esim. sanan *esimerkiksi* käyttö) abstraktien konstruktioiden distribuutiossa ilmeneviin eroihin (esim. aikamuotojen käyttö). Kaikkia tarkasteltavia eroja yhdistää se, että ne ovat luonteeltaan toistuvia. Esitettävät havainnot kuvaavat tyypillisiä kirjoitetun S1-suomen ja edistyneiden S2-oppijoiden kirjoitetun suomen välisiä eroja, mutta menetelmää voitaneen soveltaa minkä tahansa kieliaineistojen tyypillisten erojen analyysiin pyrkivässä tutkimuksessa.

Tämä artikkeli rakentuu seuraavasti: Luvussa 2 esittelen aineistolähtöisen määrällisen tutkimuksen sidoksia kielen teoriaan ja määrittelen tutkimukseni keskeistä käsitteistöä, luvussa 3 esittelen käyttämäni aineiston ja soveltamani metodit, luvussa 4 raportoin suurimmat morfologisissa muodoissa havaitut frekvenssierot, joita käytän lähtökohtana luvussa 5 analysoidessani havaittujen erojen taustalla olevia ilmiöitä. Luvussa 6 teen kokoavaa tarkastelua ja arvioin käytetyn metodin soveltuvuutta.

2 Aineistolähtöinen näkökulma rakenteeseen

2.1 Käyttöpohjaisuus ja konstruktiokielioppi

Kielen käyttöpohjaisella mallilla tarkoitetaan heterogeenistä teoreettista viitekehystä, jonka keskeisimpänä yhdistävänä tekijänä on käytetyn kielen ja todellisten kielellisten ilmausten ensisijaisuuden ja frekvenssin merkityksen korostaminen puhuttaessa kielen rakenteellisista yksiköistä. Kieli voidaan nähdä käyttöpohjaisen mallin valossa todennäköisinä kielellisinä tapoina, jotka kehkeytyvät todellisen kielenkäytön ilmauksista. Kehkeytymistä ohjaavat käytettyjen ilmausten samankaltaisuudet ja erilaisuudet sekä niiden varaan rakentuvat yleistykset. (Hopper 1987; Bybee & Thompson 1997.) Kielen

omaksumisen kannalta näin on onnistuttu verrattain hyvin kuvaamaan omaksumista edeltävän kielellisen syötöksen ja tämän syötöksen laukaiseman omaksumisen välistä suhdetta (esim. Bates & MacWhinney 1987; Tomasello 1992; Lieven, Pine & Baldwin 1997). Keskeistä on se, että kompetenssia ja performanssia ei nähdä toisistaan erillään ja että yksittäiset kielen ilmaukset eivät ole vain abstraktin kielijärjestelmän reaalistumia, vaan ne muodostavat osan joka hetki muuttuvasta dynaamisesta mekanismien joukosta (Hunston & Francis 2000: 17).

Mekanismit, joiden varassa käyttöpohjaisen kieliopin katsotaan operoivan, rakentuvat tyypillisesti rakenteellis-funktionaalisten samankaltaisuuksien varaan. Toisin sanoen ne perustuvat kielen analogiseen perusluonteeseen (Skousen 1989; Itkonen 2005). Tähän perusluonteeseen nojaavat vähintään implisiittisesti myös useimmat niin sanotun konstruktiokieliopin edustajat (esim. Fillmore 1985; Goldberg 1995, 2006; Croft 2001), samoin kuin käyttöfrekvenssejä korostavat fraseologisen tutkimuksen edustajat (esim. Gries 2008). Konstruktiokieliopin parissa on runsaasti erilaisia painotuksia, mutta keskeistä on se, että kielelliset ainekset muodostavat pareja tiettyjen merkitysten kanssa ja että kielioppi voidaan nähdä yksittäisistä lekseemeistä skemaattisempiin kokonaisuuksiin yltävinä hierarkkisina konstruktioina (Kotilainen 2007: 19–22). Nojaudun tässä tutkimuksessa Goldbergin määritelmään konstruktiosta:

> All levels of grammatical analysis involve constructions: learned pairings of form with semantic or discourse function, including morphemes or words, idioms, partially lexically filled and fully general phrasal patterns. – – Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency. (Goldberg 2006: 5.)

Goldbergin määritelmän (vrt. Goldberg 1995: 4) mukaan konstruktiomerkitykseltä ei siis välttämättä vaadita sen osien merkityksestä poikkeavaa kokonaismerkitystä, mutta muodon ja merkityksen muodostaman parin on tällöin oltava riittävän frekventti. Toisaalta kaikki usein yhdessä esiintyvät muotopiirteet eivät suinkaan muodosta konstruktioita, sillä muodon on yhdistyttävä johonkin tiettyyn merkitykseen ollakseen konstruktio. Käsillä olevan kaltaisessa määrällisessä tutkimuksessa toistuvuutta korostava näkemys on luonteva, sillä aineistojen väliset frekvenssierot ovat nimenomaan jonkin kielenpiirteen toistuvuudessa havaittuja eroja. Konstruktioiden hierarkkisuus ja eritasoisten konstruktioiden toimiminen toistensa rakenneosina on niin ikään tämän tutkimuksen kannalta perusteltu näkökulma – tutkimuksen lähtökohtana ovat frekvenssierot luonnollisessa kielenkäytössä esiintyvissä rakenneyksiköissä. Tarkastelussa esiin nousevat konstruktiot itsessään ovat niin sisäisten osiensa puolesta kuin konstruktioidenvälisestikin hyvin erilaisia keskenään sekä abstraktisuudeltaan että kompleksisuudeltaan.

2.2 Korpusvetoinen kontrastiivinen tutkimus ja tilastollinen avaimuus oppijankielen tutkimuksessa

Korpustutkimuksessa erotetaan usein toisistaan korpuspohjainen ja korpusvetoinen tutkimus. Siinä missä korpuspohjaisessa tutkimuksessa selvitetään jonkin ennalta esitetyn hypoteesin paikkansapitävyyttä korpusaineiston avulla, korpusvetoisessa tutkimuksessa tutkimuskysymykset muotoutuvat aineistolähtöisesti esimerkiksi aineistolle tyypillisten ilmiöiden ympärille. (Tognini-Bonelli 2001: 85–100.) Lähestyminen vastaa siis informaatioteknologian alalta tuttua tiedonlouhintaa (engl. *data mining*), eli tavoitteena on löytää aineistolle tyypillisiä toistuvia piirteitä siten, että piirteitä ei ole määritelty ennalta (ks. esim. Fayyad, Piatetsky-Shapiro & Smyth 1996). Korpusvetoinen asetelma rakentuu usein kontrastiivisesti siten, että aineiston tyypillisyydet ja epätyypillisyydet kuvaavat suhdetta johonkin toiseen aineistoon. Tähän ajatukseen nojaa tilastollisen avaimuuden käsite, jolla tarkoitetaan kahden aineiston välisten frekvenssien vertailua ja sellaisten seikkojen (tutkimuskysymyksen mukaan esimerkiksi sanojen) etsimistä, joiden frekvensseissä on tilastollisesti merkitsevä ero aineistojen välillä (Scott & Tribble 2006: 58; Scott 2010: 48). Termillä avaimuus halutaan korostaa sitä, että tulokset voivat paljastaa sellaista tietoa, johon tutkija ei muuten pääse käsiksi (Scott 2010: 44). Ensikielisten ja ei-ensikielisten kielenkäyttäjien kirjoitetun kielen vertaileminen tällä tavoin voi nostaa esiin eroja, jotka laadullinen arviointi jättäisi huomiotta. Lisäksi laskennallisen aineistolähtöisyyden etuna voi joskus olla se, että tietokoneohjelma näkee tekstin ilman kokonaismerkityksiä, mikä on ihmiselle haastavaa (mas. 45). Kun aineistosta kyetään erottamaan tarvittavat frekvenssit ja eroja paikannetaan tilastollisten testien avulla, voidaan tutkimus keskittää niihin kielenpiirteisiin, jotka erottavat vertailtavia aineistoja toisistaan.

Korpustutkimuksen viitekehyksessä S1- ja S2-aineistojen välisten määrällisten erojen voi nähdä kuvastavan sitä, mitä Sinclair (1991: 109–115) kutsuu vapaan valinnan periaatteeksi ja idiomiperiaatteeksi. Vapaata valintaa edustavissa kielellisissä tilanteissa voidaan soveltaa yhtä todennäköisesti mitä tahansa tilanteeseen kieliopillisesti soveltuvaa elementtiä, kun taas idiomiperiaatteen mukaisesti tietyt valinnat ovat toisia todennäköisempiä eli idiomaattisempia. Ohjaavana tekijänä ovat aiemmin koetut kielelliset ilmaukset ja niiden yleisyydet. Tuntuu luontevalta, että toisen kielen oppijoiden kokemuspohja poikkeaa ensikielisistä kielenkäyttäjistä. On mahdollista, että oppijat toimivat ensikielisiä enemmän vapaan valinnan periaatteen varassa (Granger 1998) tai että oppijat soveltavat ensikielensä idiomiperiaatteita myös toiseen kieleen (Paquot 2008). Lisäksi toisen kielen oppijat voivat toimia idiomiperiaatteen mukaisesti mutta periaatteen ohjaamat ratkaisut eroavat ensikielisisten tekemistä poikkeavan omaksumisprosessin, erilaisen kielellisen syötöksen ja erilaisten käyttöyhteyksien vuoksi (Vetchinnikova 2012).

Kontrastiivinen tutkimusasetelma lienee perinteisimpiä toisen kielen tutkimuksessa käytettyjä, onpa vertailtavina sitten oppijan lähtö- ja kohdekielet, oppijoiden ja ensikielisten tuottamat kielet tai vaikkapa kahden erilaisen oppijankielen kielimuodot. Asetelma on tutkimuksen kannalta haastava: yhtäältä siksi, että oppijankieltä ei voi eikä tule nähdä kohdekielen epätäydellisenä muotona (Bley-Vroman 1983), ja

toisaalta siksi, että edes prototyyppisen yksikieliset kielenkäyttäjät eivät välttämättä selviydy tehtävistä, joita edistyneiltä toisen kielen oppijoilta vaaditaan (Hulstijn 2011). Laajojen aineistojen avulla vertailevat tutkimusasetelmat pystyvät kuitenkin ottamaan huomioon sisäisen variaation ja kääntämään sen metodologiseksi vahvuudeksi (Osborne 2013), ja tarkastelun keskiöön nousevat aineistojen tyypillisyydet ja niissä ilmenevät systemaattiset erot – myös sellaiset, joita tutkija ei itse tule ajatelleeksi (Jarvis 2008).

Oppijankieltä käsittelevässä korpustutkimuksessa tyypillisenä asetelmana on niin sanottu kontrastiivinen välikielianalyysi (*contrastive interlanguage analysis*, CIA, Granger 1996, 2013). Tavoitteena on jonkin taustamuuttujan perusteella jaoteltujen, samankielisten aineistojen vertaaminen korpuksen avulla. Huolellisesti kerätyt korpusaineistot takaavat tulosten vertailtavuuden ja toistettavuuden, ja aineistojen koko mahdollistaa yli- ja aliedustumisten kaltaisten määrällisten erojen tutkimisen (Granger 1996: 45). Vertaan tässä tutkimuksessa kontrastiivisen välikielianalyysin hengessä kahden suomen kielimuodon morfologis-syntaktisia eroja ja yhtäläisyyksiä. Suomi toisena kielenä -alalla morfologis-syntaktista tutkimusta on runsaasti mutta kontrastiiviseen välikielianalyysiin perustuvat tutkimusasetelmat ovat harvinaisia. Poikkeuksia ovat kuitenkin muun muassa Jantunen (2011), joka tutkii oppijansuomelle tyypillisiä sanastollisia ja kieliopillisia piirteitä ja niiden suhdetta saadun kielenopetuksen määrään, sekä Spoelman (2013), joka tutkii kieltenvälisiä vaikutuksia ja kielitaitotasojen suhdetta oppijansuomen partitiivin käyttöön. Käyttämäni aineistovetoisen menetelmän vuoksi en ole etukäteen kohdistanut tutkimusta mihinkään tiettyihin kielen ilmiöihin, ja olenkin päätynyt esittelemään aiempia tutkimustuloksia luvussa 5 sen mukaan, mitä aineistosta kulloinkin nousee tarkempaan tarkasteluun.

2.3 Rakenteet ja grammit korpustutkimuksessa

Termillä rakenne voidaan viitata kielitieteessä hyvin erikaltaisiin kielen ilmiöihin foneemien yhdistelmistä laajojen tekstijaksojen keskinäisiin suhteisiin. Korpustutkimuksessa sana- ja lausetason rakennetta lähestytään usein sanojen ja niiden yhdistymisen kannalta siten, että esiintyessään yhdessä sanat muodostavat n-grammeja – monisanaisia yksiköitä (eli sanakimppuja tai -klustereita), joiden pituus on n. Granger ja Paquot (2008) suosittelevat tällaisten grammien ja niitä koskevan tarkastelun jakamista edelleen kahtia laadullisesti tulkittaviksi fraseologisiksi yksiköiksi ja frekvenssiperustaisesti analysoitaviksi n-grammeiksi, joita kuvataan käyttödistribuutioiden sekä osallistuvien sanojen kokonais- ja yhteisesiintymisten suhteen avulla. Korpuksilla tehtävässä grammitutkimuksessa grammit ovat usein lekseemejä, mutta tarkastelu paljastaa samalla abstraktimpia toistuvia rakenteellisia malleja (vrt. Hunston & Francis 2000; Conrad & Biber 2004). Toinen vaihtoehto on ottaa jokin kielen rakennetta kuvaava abstraktio tarkastelun lähtökohdaksi. Pazos ja Pamies (2008) kannattavat lingvistisen metatiedon huomioonottavaa lähestymistapaa tutkiessaan fraseologisten yksiköiden automaattista tilastollista tunnistamista. Heidän (ma.) mukaansa tunnistaminen toimii merkittävästi paremmin silloin, kun leksikaali-

sen tiedon rinnalla käytetään myös sanaluokka-annotaatiota. Kielenpiirteiden yhteisesiintyminen yhdessä toistuvan kokonaismerkityksen kanssa on tyypillistä myös konstruktiokieliopilliselle kuvaukselle, ja Gries (2008: 12–15) huomauttaakin fraseologisen tutkimuksen ja konstruktiokieliopin olevan terminologisista eroista huolimatta monilta osin toistensa kaltaisia.

Oppijankieltä käsittelevässä korpustutkimuksessa kielen fraseologista luonnetta ja sanojen välisten yhteisesiintymien vaikutusta kielenoppimiseen on tutkittu runsaasti koko alan olemassaolon ajan 1990-luvulta alkaen (ks. esim. Granger 1998; Nesselhauf 2004). Kuten Jantunen (2009: 361) kuitenkin huomauttaa, tarkastelu on pääosin jättänyt huomiotta leksikaalista yhteisesiintymistä abstraktimmat rakenteelliset tekijät. Tämä puolestaan juontunee lähinnä aiemman tutkimuksen englantikeskeisyydestä ja kielten typologisista eroista: suomen kaltaisissa, morfologisesti mutkikkaissa kielissä tutkimusasetelmat on syytä muotoilla kielen erityispiirteet huomioiden (vrt. Martin 2007; Suni 2012: 415–420), mikä puolestaan tekee korpusten soveltamisen haasteelliseksi ja teknisesti monimutkaisemmaksi.

Sanaluokka-annotointiin nojaavaa tutkimusta on sovellettu jonkin verran myös oppijankieleen, esimerkiksi tutkittaessa sellaisia 3-grammeja, jotka tyypillisesti erottavat englantia toisena kielenä opiskelevien oppijoiden kieltä ensikielisten kirjoittamasta kielestä tai jotka erottavat ensikieleltään erilaisia oppijaryhmiä toisistaan (Aarts & Granger 1996; Wiersma, Nerbonne & Lauttamus 2011). Tarkasteltavien kielenpiirteiden valinta riippuu tutkijan intresseistä ja tutkittavasta kielestä sekä käytössä olevasta aineistosta. Jotta määrällisten erojen tarkasteleminen on mielekästä, on aineiston oltava riittävän suuri. Näin ollen vertailussa tarvittavien frekvenssitietojen keräämisen on onnistuttava ainakin osin automatisoidusti, eli käytännössä vaatimuksena on jonkinlainen tietokoneen avulla luettava ja käsiteltävä korpusaineisto. Mikäli tutkimus keskittyy leksikaalisen pintatason sijaan esimerkiksi sanaluokkiin, sanojen morfologisiin muotoihin tai syntaktisiin funktioihin, vaaditaan aineistolta lisäksi soveltuvaa kieliopillisen metatiedon merkintätapaa eli annotointia.

Monisanaisten yksiköiden kuvaukset (ks. esim. Aarts & Granger 1996; Wiersma, Nerbonne & Lauttamus 2011) sisältävät useimmiten vähintään implisiittisen oletuksen siitä, että sanat esiintyvät peräkkäin samassa järjestyksessä, mikä voi puolestaan jättää tarkasteltavalle rakenteelle tyypillisen variaation tai moniosaisuuden vaille tutkijan huomiota (Sinclair 2001: 353). Monisanaisia yksiköitä voidaan kuitenkin tarkastella myös niin sanottuina skipgrammeina (*skipgram*) – jolloin sanojen ei tarvitse olla peräkkäin, kunhan ne ovat lähekkäin ja samassa järjestyksessä (Guthrie ym. 2006) – tai konkgrammeina (*concgram*) – jolloin ainoana kriteerinä on sanojen lähekkäisyys (Cheng, Greaves & Warren 2006). Näin tavoitetaan paremmin grammeissa esiintyvä rakenteellinen variaatio, minkä lisäksi kielelle tyypilliset toistuvat mallit nousevat esiin jo verrattain pienestä aineistosta (Guthrie ym. 2006: 1225).

Aiemmissa tutkimuksissa tulkinta on pääosin jätetty tilastollisesti havaittujen frekvenssierojen toteamiseen eikä erojen syiden taustalla olevia kielellisiä ilmiöitä ole juuri pyritty selittämään. Tilastollista avaimuutta hyväkseen käyttävä tutkimus ei kuitenkaan pääty erojen raportointiin. Kuten Scott (2010: 56–57) toteaa, vertailu nostaa esiin sellaisia seikkoja, joissa aineistot poikkeavat määrällisesti toisistaan. Tut-

kijan tehtävänä on analysoida tämän poikkeaman luonne ja selvittää sen mahdollisia syitä. Eroja voi tutkia aineistolähtöisesti vertailemalla grammien sisäistä vaihtelua sekä niiden tyypillistä esiintymisympäristöä (Francis 1993). Stefanowitch ja Gries (2003) keskittyvät sisäiseen vaihteluun kollostruktionaaliseksi (*collostructional*) analyysiksi kutsumassaan metodissa, jossa he pyrkivät yhdistämään korpuslähestymisen ja konstruktiokieliopin vahvuuksia tutkimalla valitsemiensa konstruktioiden tyypillistä leksikaalista vaihtelua ja konstruktioiden ja niissä esiintyvien sanojen välistä suhdetta. Tyypillistä esiintymisympäristöä puolestaan voi lähestyä tarkastelemalla analysoitavan grammin kotekstuaalisia ominaisuuksia, kuten kollokaatioita – leksikaalisia myötäesiintymiä – ja kolligaatioita – kieliopillisia myötäesiintymiä (Sinclair 1991: 115–117; Hunston 2001; termeistä ks. Firth 1968 [1957]). Kuten todettua, toistuvakaan yhteisesiintyminen ei vielä itsessään takaa kyseessä olevan konstruktion. Sisäisen ja kotekstuaalisen vaihtelun tarkastelu on ensiarvoisen tärkeää, jotta voidaan arvioida, mikä kielen ilmiö nousee aineistoja erottavaksi avainrakenteeksi.

3 Tutkimusaineisto ja metodit

3.1 Edistyneiden suomenoppijoiden korpus LAS2

Tutkimuksen aineisto on osa Turun yliopistossa koottua Edistyneiden suomenoppijoiden korpusta (jatkossa LAS2). Korpus sisältää edistyneiden S2-oppijoiden tekstejä, jotka on kirjoitettu osana suomenkielistä akateemista diskurssia Suomessa (Ivaska, tulossa 2014). Korpus jakautuu ajallisesti rajoitettujen tekstien (mm. tenttivastaukset), julkaistavaksi tarkoitettujen tekstien (mm. tutkielmien käsikirjoitukset) sekä yksityisten tekstien (mm. kurssipäiväkirjat) osakorpuksiin. Kaikkiin osakorpuksiin kuuluu myös S1-vertailuaineisto. Tekstejä yhdistää se, että niiden ensisijaisena funktiona ei ole ollut kielitaidon arvioiminen. Käytän tässä tutkimuksessa yksinomaan spontaanien tekstien osakorpukseen kuuluvia tekstejä, jotka ovat kaikki tenttivastauksia.

S2-tekstejä on aineistossani yhteensä 275 tekstiyksikköä (120 965 sanetta) ja S1-tekstejä on 56 tekstiyksikköä (30 399 sanetta). S2-aineistoa on 31 informantilta, S1-aineiston kukin teksti on eri kirjoittajalta. S2-aineiston informantit ovat kielitaustaltaan heterogeeninen joukko, sillä edustettuja ensikieliä on yhteensä 12: englanti, islanti, japani, komi, liettua, puola, saksa, slovakki, tšekki, unkari, venäjä ja viro. Tekstilaji on kuitenkin melko homogeeninen, sillä kaikki tekstit ovat kielen ja kulttuurin alan yliopisto-opintoihin liittyviä tenttivastauksia. Näin ollen on perusteltua olettaa, että ilmiöt, jotka erottavat S1- ja S2-aineistoa toisistaan, liittyvät nimenomaan informanttien ensikielisyyteen. Edelleen voidaan olettaa, että havaitut erot kuvastavat kielitaitoerojen lisäksi tekstilajin hallitsemiseen liittyviä eroja. Korpuksen kultakin informantilta on toistaiseksi arvioitu yleiseurooppalaisen viitekehyksen (CEFR) mukaisesti 2–4 ajallisesti rajoitettua tekstiä. Arvioiden perusteella tekstit asettuvat yleensä välille B2–C1 (tilanne 9.12.2013: B1 4 %, B2 32 %, C1 62 %, C2 3 %).

Aineisto on lemmattu ja annotoitu sanakohtaisesti sanaluokkien, sanojen morfologisten muotojen ja niiden syntaktisten funktioiden osalta. Aineistoa ei ole annotoitu erikseen lausetyyppien tai nollapersoonailmausten kaltaisten lausetason ilmiöiden osalta, mutta tekstit on jaoteltu kappaleisiin, virkkeisiin ja lauseisiin. Aineisto on tallennettu XML-muodossa, ja se seuraa pääpiirteissään Turun yliopiston lauseopin X-arkistossa (LaX) ja Mikael Agricolan morfosyntaktisessa tietokannassa (Inaba 2007) käytettyä TEI-standardin sovellusta. (Käytetystä annotoinnista ja annotointiprosessista tarkemmin Ivaska, tulossa 2014.) Olen tehnyt kaikki korpushaut LAS2:n omilla hakutyökaluilla (ma.), tilastollisissa analyyseissa olen käyttänyt R-ohjelmointiympäristöä (R 2013). Frekvenssihavainnot on normalisoitu siten, että kustakin tekstistä lasketut frekvenssit kuvaavat esiintymien määrää tuhatta sanaa kohti. Näin kunkin tekstin painoarvo on keskiarvolaskelmissa ja tilastollisissa analyyseissa sama tekstien pituudesta riippumatta.

3.2 Avainrakenneanalyysi: frekvenssieroista funktionaalis-distributionaalisiin eroihin

Sovellan tässä tutkimuksessa avainrakenneanalyysia (Ivaska & Siitonen 2011; tarkemmin Ivaska 2012), jossa yhdistyvät vertailtavien kieliaineistojen tilastollisesti merkitsevien frekvenssierojen paikallistaminen aineistovetoisesti ja havaittujen erojen kvantitatiivis-kvalitatiivinen analyysi niiden tyypillisen käytön kannalta. Menetelmä on perusluonteeltaan kontrastiivinen, ja toisen kielen tutkimuksen kontekstissa se soveltaakin kontrastiivista välikielianalyysia. Keskeistä on siis se, että sekä tutkittavien yksiköiden valinta että niiden tyypillisen käytön tarkastelu ovat aineistovetoisia.

Keräsin aineistosta kaikkien sanojen morfologisten muotojen 1-, 2- ja 3-grammien kokonaisfrekvenssit. Käyttämäni korpuksen morfologinen annotointi kohdistuu kokonaisiin sanoihin eikä yksittäisiin morfeemeihin, ja tarkoitankin tässä tutkimuksessa grammilla yhtä sanaa, joka voi sisältää yhden tai useamman morfeemin. Näin grammeja voi myös tarkastella siltä kannalta, mitä lemmoja, sanaluokkia ja syntaktisia funktioita niissä esiintyy. 1-grammi koostuu yhdestä tällaisesta grammista, 2-grammi kahdesta ja 3-grammi kolmesta grammista. Esiintymien kynnysarvona on 80 kokonaisesiintymää. Kuten Biber, Conrad ja Cortes (2004: 376) huomauttavat, kynnysarvojen määrittäminen on aina osin sattumanvaraista. Tässä tutkimuksessa on verrattain matala kynnysarvo, jotta myös harvinaiset tai oppijankielestä jopa kokonaan puuttuvat ilmiöt voisivat seuloutua mukaan tarkasteluun. Käytettävät tilastolliset menetelmät ovat kuitenkin sellaiset, että ne paikantavat toistuvia aineistojen välisiä eroja, mikä vähentää kynnysarvon vaikutusta tuloksiin. Monisanaiset grammit on määritelty skipgrammilähestymistä soveltaen. Samaan n-grammiin laskettavilla grammeilla ei ole maksimietäisyyttä, mutta niiden on oltava samassa lauseessa. Esimerkki 1 ja luetelma 1 kuvastavat frekvenssien laskutapaa:

(1) Lapset haluavat nähdä rahaa.
<pl nom><fin ind pres pl3><inf1><sg part>

Skipgrammit
1-grammit:
<pl nom>
<fin ind pres pl3>
<inf1>
<sg part>
2-grammit:
<pl nom><fin ind pres pl3>
<pl nom><inf1>
<pl nom><sg part>
<fin ind pres pl3><inf1>
<fin ind pres pl3><sg part>
<inf1><sg part>
3-grammit:
<pl nom><fin ind pres pl3><inf1>
<pl nom><fin ind pres pl3><sg part>
<pl nom><inf1><sg part>
<fin ind pres pl3><inf1><sg part>

Luetelma 1.
Esimerkissä 1 esiintyvät skipgrammit.

Esimerkin 1 lause koostuu neljästä sanasta, eli morfologisia 1-grammeja on neljä, 2-grammeja on kuusi ja 3-grammeja on neljä. Korpuksen frekvenssilistatyökalu laskee tällä tavoin aineiston kustakin tekstiyksiköstä erikseen kaikki morfologisten muotojen grammien frekvenssit, normalisoi frekvenssit 1 000:ta sanaa kohti ja tuottaa tuloksista csv-tiedoston.

Frekvenssien laskemisen jälkeen järjestin kaikki grammit sen mukaan, miten hyvin ne erottavat aineistoja toisistaan. Tilastollisena menetelmänä on automaattiseen kategoriointiin suunniteltu satunnaismetsä-menetelmä (*random forest*, algoritmista tarkemmin ks. Breiman 2001). Menetelmän avulla voidaan etsiä muuttujia, joiden arvot ennustavat parhaiten aineiston eri alaryhmiä – kuten eri kielimuotoja – tai jotka korreloivat parhaiten jonkin muuttujan kanssa (ks. esim. Tagliamonte & Baayen 2012). Tein vertailun R:n party-paketin (Hothorn ym. 2006; Strobl ym. 2007; Strobl ym. 2008) cforest-metodin avulla. Karkeistaen menetelmä toimii seuraavasti: Aineistosta rajataan ensin pois noin kolmasosa tulevan luokittelun arviointia varten (*out-of-bag*). Tämän jälkeen jäljelle jäävästä aineistosta valitaan sattumanvaraisesti tietty määrä muuttujia ja katsotaan, miten hyvin ne erottavat selitettävät muuttujat toisistaan. Tämä satunnaisvalinta toistetaan useita kertoja parhaiten eroa selittävien muuttujien löytämiseksi, minkä jälkeen muuttujat arvioidaan sen mukaan, miten hyvin niiden avulla pystytään ennustamaan ulkopuolelle rajatusta aineistonosasta kirjoittajan ensikielisyyttä. Lopuksi muuttujat järjestetään alenevaan järjes-

tykseen party-paketin varimp-metodin avulla sen perusteella, miten hyvin ne ennustavat selitettävän muuttujan arviointiin rajatusta aineistosta.[2]

Satunnaismetsät ovat tutkijan kannalta käytännöllinen menetelmä niihin sisäsyntyisesti kuuluvan ristikkäisvalidoinnin ansioista. Tästä syystä malli ei ylisovitu (*overfit*), ja havaitut erot (ja näin ollen myös erojen taustalla vaikuttavat ilmiöt) ovat luotettavammin yleistettävissä muihin samankaltaisiin aineistoihin. Tulos perustuu nimensä mukaisesti satunnaisotantaan, ja vaihtelu otannassa saattaa hieman muuttaa muuttujien järjestystä. Aiempien havaintojen mukaan sen vaikutus on kuitenkin vähäinen (Liaw & Wiener 2002: 21). Lisäksi käsittelen tässä tutkimuksessa 50:tä eroa parhaiten selittävää muuttujaa tasaveroisesti, joten tarkalla järjestyksellä on vähän käytännön merkitystä tutkimustuloksille.

Grammien järjestämisen jälkeen erotin saadusta grammilistauksesta 50 ensimmäistä tapausta eli 50 n-grammia, joiden frekvenssit edellä kuvatun tilastollisen testin mukaan ennustavat parhaiten S1:n ja S2:n välistä eroa (ks. lukua 4). Olen pyrkinyt kohdistamaan analyysin niihin kielenkäytön seikkoihin, jotka todella ovat merkityksellisiä aineistoja erottavia tekijöitä. Tästä syystä poimin listasta kaikki 1-grammit sekä ne 2- ja 3-grammit, joissa kyseiset 1-grammit esiintyvät. Käsittelen näistä 1-grammeja, joskin tekstiyhteyden analyysi saattaa osoittaa 2- tai 3-grammin kuvaavan paremmin aineistoja erottavaa konstruktiota. Olen tällä tavoin pyrkinyt välttämään sen, että grammeja, jotka kuvaavat toisistaan erillisiä aineistoja erottavia ilmiöitä ja vain sattuvat esiintymään samoissa lauseissa, analysoitaisiin toisiinsa vaikuttavina kielenpiirteinä. Toimin samalla tavoin myös 2-grammien kanssa, ja jäljelle jäävät 3-grammit analysoin 3-grammeina.[3]

Analysoin tämän jälkeen kaikki valikoituneet grammit niiden sisäisen ja kotekstuaalisen vaihtelun osalta. Keräsin aineistosta kaikki kunkin morfologisten muotojen perusteella määritellyn grammin esiintymät. LAS2-korpuksesta voidaan kerätä ja tallettaa csv-tiedostoksi grammin kunkin sanan sisäinen ja kotekstuaalinen vaihtelu annotoinnin eri tasoilla[4]. Vertailemalla vaihtelua aineistojen välillä voidaan tehdä luotettavia, määrällisiin käyttötendensseihin perustuvia arvioita tarkasteltavan ilmiön konstruktionaalisesta luonteesta ja tulkintoja analysoitavia kielimuotoja tyypillisesti erottavista kielenpiirteistä. Olen lisäksi tarvittaessa tehnyt tarkentavia hakuja ilmiön laajuuden selvittämiseksi. Verratessani S1- ja S2-aineiston yksittäisten piirteiden välisiä frekvenssieroja sovellan Mann-Whitneyn U-testiä, joka on luonteeltaan epäparametrinen eikä näin ollen kohdista tarkasteltavien arvojen jakaumalle rajoitteita (ks. Hollander & Wolfe 1973: 68–75).[5]

---

2. Satunnaisotantojen jyvityksenä oli koko ajan 7 531. Selitettävänä muuttujana oli kirjoittajan ensikielisyys ja testattavina muuttujina kaikki kynnysarvon ylittäneet morfologiset n-grammit. Käytin kullakin kerralla satunnaisesti valittavien muuttujien määränä metodin oletusasetusta eli viittä muuttujaa. Satunnaisvalinta toistettiin muuttujien suurehkon määrän vuoksi oletusasetuksista poiketen 4 000 kertaa.

3. Muuttujien merkitsevyyden laskemisessa käytetty varimp-metodi mahdollistaa myös muuttujien välisen kollineaarisuuden huomioonottamisen. Tällöin samoja elementtejä sisältävien grammien poistaminen olisi toisarvoista. Se osoittautui kuitenkin laskennallisesti kestämättömän raskaaksi.

4. Olen aineistoanalyysissa tarkastellut sisäistä ja kotekstuaalista vaihtelua annotoinnin kaikilla tasoilla jokaisen tarkasteltavan grammin jokaisen jäsenen osalta. Nostan analyysissani kuitenkin esiin vain ne seikat, jotka parhaiten selittävät havaittua eroa.

5. Kaikki artikkelissa mainittavat erojen p-arvot on saatu soveltamalla Mann-Whitneyn U-testiä. Tilastollisen merkitsevyyden arviointi on tehty tasolla 0.01. Satunnaismetsä-menetelmällä tavoitettujen

4 Potentiaalisia avainrakenteita: suurimmat frekvenssierot ensikielisen suomen ja edistyneen oppijansuomen välillä

Aineistossa on yhteensä 3 816 erilaista morfologisten muotojen n-grammia, jotka täyttävät aineistokriteerit. Näistä 80 on 1-grammeja, 826 on 2-grammeja ja 2 910 on 3-grammeja. Grammit on järjestetty satunnaismetsä-menetelmän avulla laskevaan järjestykseen sen mukaan, miten hyvin niiden frekvenssit selittävät S1- ja S2-aineiston välistä eroa. Taulukossa 1 on listauksen 50 ensimmäistä grammia, joista kahdeksan grammin mikään osa ei ole osana mitään toista 50:n ensimmäisen grammin joukossa olevaa grammia. Näin aineistosta valikoituu edellä kuvatuista syistä tarkempaan analyysiin viisi 1-grammia, kaksi 2-grammia ja yksi 3-grammi.

Taulukko 1.
S1- ja S2-aineistot toisistaan parhaiten erottavat morfologisten muotojen n-grammit.

Järjestys	n-grammi	Grammi sisältää jonkin toisen top 50 -grammin
1	<sg tra><inf1>	kyllä
2	<fin pass ind pres><inf1><adv>	kyllä
3	<fin ind pres sg3><sg gen><sg gen>	kyllä
4	<sg tra><sg nom><inf1>	kyllä
5	<fin cond pres sg3><sg part>	kyllä
6	<sg tra><sg nom><sg nom>	kyllä
7	<pcp1 sg gen>	ei
8	<sg tra><cnj><cnj>	kyllä
9	<inf1><cnj>	kyllä
10	<sg tra><cnj><sg nom>	kyllä
11	<inf1><cnj><inf1>	kyllä
12	<sg ill><inf1>	kyllä
13	<inf1><pl ill><pl ill>	kyllä
14	<fin cond pres sg3><inf1><sg nom>	kyllä
15	<fin ind pres sg3><sg gen><cnj>	kyllä
16	<sg tra><cnj>	kyllä
17	<inf1><sg gen>	kyllä
18	<pl part><inf1>	kyllä
19	<inf1><fin ind pres sg3><cnj>	kyllä
20	<sg nom><sg gen><inf1>	kyllä
21	<sg tra><fin ind pres sg3>	kyllä

erojen p-arvot ovat kuitenkin huomattavasti pienempiä. Eron kattavuutta kuvaava r-arvo on saatu jakamalla Mann-Whitneyn testistä saatava Z-arvo havaintojen kokonaismäärän neliöjuurella.

22	<sg tra><infı><cnj>	kyllä
23	<fin ind pres sg3><sg gen>	ei
24	<fin ind pret sg3><sg nom>	kyllä
25	<fin ind pres sg3><sg gen><sg nom>	kyllä
26	<pl gen><infı><pl part>	kyllä
27	<pl nom><infı><pl ill>	kyllä
28	<fin pass ind pres><sg ill>	ei
29	<sg gen><fin cond pres sg3><infı>	kyllä
30	<fin ind pret sg3>	ei
31	<pcp1 sg gen><sg nom>	kyllä
32	<fin ind pres sg3><infı><sg tra>	kyllä
33	<fin pass ind pres><infı><cnj>	kyllä
34	<infı><adv>	kyllä
35	<infı><pl ill><cnj>	kyllä
36	<sg tra><sg gen><sg nom>	kyllä
37	<sg tra><infı><sg nom>	kyllä
38	<sg ill><fin ind pres sg3><sg gen>	kyllä
39	<infı><sg gen><sg nom>	kyllä
40	<fin cond pres sg3>	ei
41	<infı><sg ill><sg ill>	kyllä
42	<fin ind pres sg3><infı><sg gen>	kyllä
43	<sg tra>	ei
44	<pl part><fin ind pres pl3><pl nom>	ei
45	<sg tra><sg gen>	kyllä
46	<infı>	ei
47	<infı><cnj><pl ill>	kyllä
48	<sg gen><sg gen><infı>	kyllä
49	<infı><sg nom>	kyllä
50	<infı><adv><sg gen>	kyllä

Käsiteltäväksi valikoituvat seuraavat 1-grammit: VA-partisiipin yksikön genetiivi (7: <pcp1 sg gen>), yksikön kolmannen persoonan indikatiivin imperfekti (30: <fin ind pret sg3>), yksikön kolmannen persoonan konditionaalin preesens (40: <fin cond pres sg3>), yksikön translatiivi (43: <sg tra>) ja A-infinitiivin lyhyt muoto (46: <infı>). Taulukko 2 havainnollistaa kunkin käsiteltävän 1-grammin frekvenssiä aineistoissa sekä aineistojen välisen eron merkitsevyyttä.

Taulukko 2.
Käsiteltävät 1-grammit, niiden frekvenssit S1- ja S2-aineistossa sekä eron tilastollista merkitsevyyttä kuvaavan Mann-Whitneyn U-testin tulokset.

n-grammi	ka. / 1 000 sanetta		med. / 1 000 sanetta		U-testi
	S1	S2	S1	S2	
<pcp1 sg gen>	2.913879	0.7222875	2.598525	0	U = 11584.5 Z = 7.2766 p = 2.432e-12 r = .40
<fin ind pret sg3>	4.474611	21.47804	0	13.62398	U = 3454.5 Z = -6.5408 p = 8.703e-12 r = .36
<fin cond pres sg3>	5.551942	1.891491	4.774799	0	U = 11408 Z = 6.2635 p = 3.487e-10 r = .34
<sg tra>	11.99828	6.542514	11.44745	5.624297	U = 11157 Z = 5.3236 p = 4.678e-08 r = .29
<inf1>	37.42165	22.92515	37.39421	21.2766	U = 11630 Z = 6.0214 p = 4.55e-10 r = .33

Käsiteltäväksi valikoituvat seuraavat 2-grammit: yksikön kolmannen persoonan indikatiivin preesens ja sitä seuraava yksikön genetiivi (23: <fin ind pres sg3><sg gen>) sekä passiivin indikatiivin preesens ja sitä seuraava monikon illatiivi (28: <fin pass ind pres><sg ill>). Taulukko 3 havainnollistaa näiden 2-grammien frekvenssiä aineistoissa sekä aineistojen välisen eron merkitsevyyttä.

Taulukko 3.
Käsiteltävät 2-grammit, niiden frekvenssit S1- ja S2-aineistossa sekä eron tilastollista merkitsevyyttä kuvaavan Mann-Whitneyn U-testin tulokset.

n-grammi	ka. / 1 000 sanetta		med. / 1 000 sanetta		U-testi
	S1	S2	S1	S2	
<fin ind pres sg3> <sg gen>	35.78595	22.11402	32.69515	19.04762	U = 11474 Z = 5,7824 p = 2,418e-09 r = .32
<fin pass ind pres> <sg ill>	2.096298	0.8069145	1.693685	0	U = 10434 Z = 5.3889 p = 1.934e-07 r = .30

3-grammeista käsiteltäväksi puolestaan valikoituu monikon partitiivista, sitä seuraavasta monikon kolmannen persoonan indikatiivin preesensistä ja sitä seuraavasta monikon nominatiivista koostuva grammi (44: <pl part><fin ind pres pl3><pl nom>). Taulukko 4 havainnollistaa tämän 3-grammin frekvenssiä aineistoissa sekä aineistojen välisen eron merkitsevyyttä.

Taulukko 4.
Käsiteltävä 3-grammi, sen frekvenssi S1- ja S2-aineistossa sekä eron tilastollista merkitsevyyttä kuvaavan Mann-Whitneyn U-testin tulokset.

n-grammi	ka. / 1 000 sanetta		med. / 1 000 sanetta		U-testi
	S1	S2	S1	S2	
<pl part> <fin ind pres pl3> <pl nom>	1.704859	0.2235148	0	0	U = 9407.5 Z = 5.3343 p = 1.044e-06 r = .29

Kohdistan siis luvussa 5 esitettävän tarkemman analyysin edellä esitettyihin kahdeksaan n-grammiin. On kuitenkin syytä huomata, että analyysin perusteella myös jotkin muut taulukossa 1 esiintyvät n-grammit kuvannevat samaa kielen ilmiötä (ks. esim. 2-grammin <fin ind pres sg3><sg gen> analyysi luvussa 5.1).

5 Tyypillisiä ensikielisen suomen ja edistyneen oppijansuomen eroja

5.1 Morfosyntaktinen kompleksisuus

Katson kolmen eri n-grammin frekvenssieron perustuvan S1-kielen suurempaan morfosyntaktiseen kompleksisuuteen ja runsaampaan variaatioon. Nämä ovat 2-grammi <fin ind pres sg3><sg gen>, 1-grammi <pcp1 sg gen> ja 3-grammi <pl part><fin ind pres pl3><pl nom>.

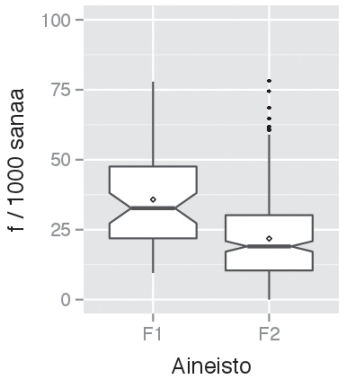
Indikatiivin kolmannen persoonan preesensmuodon ja sitä seuraavan yksikön genetiivin sisältämä 2-grammi esiintyy S1-aineistossa huomattavasti S2-aineistoa useammin, ja ero on tilastollisesti merkitsevä (ks. taulukko 3). Kuvion 1 laatikkojanat[6] kuvaavat frekvenssien jakaumaa aineistoissa (ks. viereinen sivu).

Funktioiden sisäiset jakaumat osoittavat, että ensimmäinen grammi on lauseen pääverbi, kun toinen grammi puolestaan on useimmiten substantiivin määrite (S1: 76 %; S2: 78 %, ks. esim. 2), joskus myös nominaalinen objekti (S1: 9 %; S2: 8 %). Erot

6. Laatikkojana-kuviossa laatikko kuvaa havaintojen keskimmäistä 50:tä prosenttia ja laatikoista alkava viiva ylimmäistä ja alimmaista 25:tä prosenttia havainnoista lukuun ottamatta keskiarvosta voimakkaasti poikkeavia havaintoja, jotka on merkitty erikseen pisteellä. Kuvion F1 tarkoittaa S1-aineistoa ja F2 puolestaan S2-aineistoa. Laatikon keskellä oleva viiva osoittaa mediaanin ja pallo keskiarvon sijainnin. Merkinnät ovat samat kaikissa tämän artikkelin laatikkojana-kuvioissa.

siinä, mitä funktiota grammit edustavat, eivät siis selitä aineistojen välistä eroa, mistä syystä paneudun tarkemmin yleisintä funktiota kuvaaviin tapauksiin.

(2) Se *on sanan* perusosa ilman taivutuspäätteitä. (S2: las2-7tt01te02[7])



Kuvio 1.
2-grammin <fin ind pres sg3><sg gen> jakauma laatikkojanalla kuvattuna S1- ja S2-aineistoissa.

Kumpaakin grammia käytetään aineistoissa likimain yhtä paljon yksittäin, eikä finiittiverbin yksikön kolmannen persoonan preesensin (S1: 72,1 / 1 000 sanetta vs. S2: 74,0 / 1 000 sanetta; U = 8704, Z = –0.622, p = .5352, r = .034) ja genetiivimuotoisen substantiivin määritteen (S1: 47,1 / 1 000 sanetta vs. S2: 44,0 / 1 000 sanetta; U = 8704, Z = 1.5381, p = .1243, r = .085) yleisyydessä ole juuri eroa.

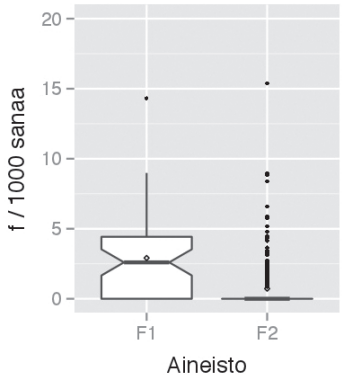
Taulukko 5.
2-grammin <fin ind pres sg3><sg gen> jälkimmäisen jäsenen tyypillinen seuraava eli oikeanpuoleinen (R1 ja R2) koteksti lausefunktioiden osalta.

	R1-koteksti		R2-koteksti	
<sg gen>	S1	S2	S1	S2
	<nmod> 33 %	<advl> 30 %	<cl> 23 %	<cl> 38 %
	<advl> 21 %	<nmod> 25 %	<nmod > 23 %	<advl> 14 %
	<npobj> 12 %	<npobj> 11 %	<advl> 15 %	<nmod> 14 %
	<compl:s> 9 %	<compl:s> 8 %	<lkeyhd> 9 %	<lkeyhd> 6 %
	<npsubj> 6 %	<npsubj> 8 %	<npobj> 8 %	<npobj> 5 %

7. Merkinnän alkuosa (tässä: S2) kertoo aineiston, josta esimerkki on ja loppuosa (tässä: las2-7tt-01te02) sen sijainnin korpuksessa. Käytän samaa merkintätapaa läpi tämän artikkelin.

Taulukko 5 kuvaa 2-grammin <fin ind pres sg3><sg gen> jälkimmäistä jäsentä seuraavien lausefunktioiden frekvenssejä. Havaittu ero näyttää kietoutuvan substantiivin määritteiden käyttöön, sillä tekstiyhteyden tarkastelu paljastaa, että S1-aineistossa on verbinjälkeisessä asemassa S2-aineistoa enemmän valinnaisia nominaalisia jäseniä. S1-aineistossa 2-grammin jälkimmäistä jäsentä seuraa useammin toinen substantiivin määrite (<nmod>) ja S2-aineistossa 2-grammin jälkimmäinen jäsen on useammin lähellä lauseen loppua (<cl>). Tulkintaa tukee myös se, että vastaavasta verbinmuodosta ja kahdesta yksikön genetiivistä koostuva 3-grammi on taulukon 1 grammilistauksessa sijalla 3. Määritteiden käyttö S2-oppijoita erottavana tekijänä on linjassa aiemman tutkimuksen kanssa: esimerkiksi Kannisto (2012) osoittaa samaa korpusta käyttäen, että S1-aineistossa on S2-aineistoa enemmän lausekkeita, joissa on vähintään kaksi määritettä, vaikka lauseiden kokonaispituudet ovat S1- ja S2-aineistoissa samankaltaiset (Salmi 2010: 62–63). Ukkola (2009) puolestaan toteaa substantiivilausekkeiden määritteiden ylipäätään lisääntyvän kielitaidon kehittyessä. Tulosten valossa vaikuttaa selvältä, että vielä edistyneetkin S2-oppijat käyttävät substantiivin määritteitä ensikielisiä vähemmän nimenomaan verbinjälkeisessä asemassa. Ero näyttää perustuvan useiden määritteiden käyttöön verbinjälkeisessä asemassa ylipäätään, ei niiden suurempaan määrään jonkin tietyn lauseenjäsenen yhteydessä.

VA-partisiipin yksikön genetiivistä koostuva 1-grammi on kummassakin aineistossa melko harvinainen, mutta se esiintyy S1-aineistossa S2-aineistoa enemmän, ja ero on tilastollisesti merkitsevä (ks. taulukko 2 s. 173). Kuvion 2 laatikkojanat kuvaavat frekvenssien jakaumaa aineistoissa.



Kuvio 2.
1-grammin <pcp1 sg gen> jakauma laatikkojanalla kuvattuna S1- ja S2-aineistoissa.

Funktioiden sisäiset jakaumat osoittavat, että 1-grammi edustaa tyypillisesti joko substantiivin määritettä (46 %; esim. 3) tai referatiivirakenteen predikaattia (39 %, esim. 4).

(3) – – jossa idästä *tulevan* ilman tähden lampotila on jopa 2 yli nolla – – (S2: las2-23tt01te16)

(4) Suomen kielen taito on osoittettu *heikkentyvän* (S2: las2-30tt01te06)

Funktioiden sisäinen jakauma osoittaa, että referatiivirakenne (S1: 49 %; S2: 22 %) kattaa huomattavasti suuremman osan S1-aineiston esiintymistä kuin S2-aineiston esiintymistä. Rakenne on melko harvinainen molemmissa aineistossa, mutta grammin kokonaisfrekvenssin ero selittyy juuri tällä käyttöerolla – kyseinen referatiivirakenne esiintyy S2-aineistossa vain 15:ssä aineiston 275 tekstistä – ja ero on tilastollisesti merkitsevä (S1: 1,3 / 1 000 sanetta vs. S2: 0,1 / 1 000 sanetta; U = 10607, Z = 7.9545, p = $2.955e^{-12}$, r = .44).

Monikon partitiivista, monikollisesta indikatiivin preesensin verbistä ja monikon nominatiivista koostuva 3-grammi on niin ikään harvinainen molemmissa aineistoissa. S1-aineistossa se esiintyy kuitenkin S2-aineistoa useammin, ja ero on tilastollisesti merkitsevä (ks. taulukko 4 s. 174). 3-grammin ensimmäinen sana on tyypillisesti joko predikatiivi (<compl:s> 66 %) tai objekti (<npobj> 22 %), toinen sana on predikaatti (<pred>) ja kolmas sana on joko subjekti (<npsubj> 62 %) tai substantiivin määrite (<nmod> 31 %). 3-grammi kuvaa siis tyypillisesti joko predikatiivi- tai transitiivilausetta, jossa on käänteinen sanajärjestys. Esimerkki 5 kuvaa transitiivista lausetta ja esimerkki 6 predikatiivilausetta.

(5) – – mutta *määritteitä voivat* saada myös *substantiivit* adjektiivit ja adverbit (S2: las2-20tt01te09)

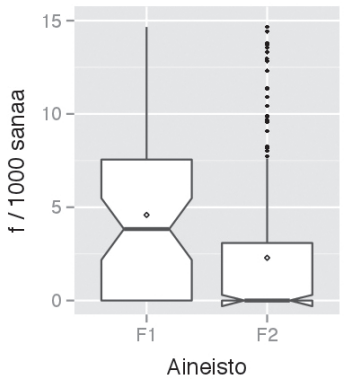
(6) Niistä *vanhimpia* ovat indoeurooppalaiset ja arjalaiset *lainasanat*. (S2: las2-7tt01te02)

Taulukko 6.
Aineiston sellaisten transitiivi- ja predikatiivilauseiden frekvenssit ja ero, joiden sanajärjestys on OVS tai ComplVS.

n-grammi	ka. / 1 000 sanetta		med. / 1 000 sanetta		U-testi
	S1	S2	S1	S2	
<npobj> <pred> <npsubj>	2,6	2,0	0	0	U = 8631 Z = 1.6335 p = .1025 r = .09
<compl:s> <pred> <npsubj>	5,9	3,6	4,7	0	U = 9840 Z = 3.4981 p = .00047 r = .19

Aineistojen väliseen frekvenssieroon vaikuttaa olevan useita päällekkäin limittyviä syitä. Kuten taulukko 6 osoittaa, ero näyttää liittyvän erityisesti käänteisen sanajärjestyksen predikatiivilauseisiin, sillä sanajärjestykseltään käänteisten transitiivilauseiden frekvenssissä ei ole tilastollisesti merkitsevää eroa aineistojen välillä. Lisäksi

predikatiivisten lauseiden kokonaismäärässä ei ole merkitsevää eroa aineistojen välilllä (S1: 46,0 / 1 000 sanetta vs. S2: 40,7 / 1 000 sanetta; U = 8972.5, Z = , p = .05111, r = .11). Kuvion 3 laatikkojanat kuvaavat niiden predikatiivilauseiden frekvenssejä, joissa on käänteinen sanajärjestys.



Kuvio 3.
3-grammin <compls><pred><npsubj> jakauma laatikkojanalla kuvattuna S1- ja S2-aineistoissa.

Tyypillisiä inversiotapauksia aineistossa ovat sellaiset, joissa viitataan edeltävään kontekstiin joko implisiittisesti (esim. 7, aiemmin tekstiyhteydessä käsitelty *verbit*) tai eksplisiittisesti (esim. 8, tekstiyhteydessä käsiteltäviin matkoihin viittaava *näiden*). Tapaukset kuvaavat usein retorista elaboraatiosuhdetta, jossa inversiotapaus tarkentaa aiemmassa kontekstissa sanottua (vrt. Komppa 2012: 46–48). Kompan mukaan sidosteisuus voi kuvata S2-kirjoittajan taitoja siten, että heikommat kirjoittajat käyttävät enemmän konjunktioiden kaltaisia eksplisiittisiä keinoja, kun taas taitavammilla kirjoittajilla on käytössään erilaisia keinoja (mts. 186–187), jona edellä kuvattua inversiotakin voinee pitää.

(7) Yksi tapa on luokitella ne niiden paikkaisuuden perusteella. *Nollapaikkaisia ovat verbit* – – (S1: las2-vtt01vert001)

(8) Hänen matkojensa laajuus on n. 20000 km. *Merkittäviä ovat* näiden matkojen *tulokset*: (S2: las2-9tt01te02)

Predikatiivin monikollisuus on ylipäätään yleisempää käänteisen sanajärjestyksen predikatiivilauseissa (S1: 23 % kaikista predikatiiveista monikollisia, 44 % sanajärjestykseltään käänteisten lauseiden predikatiiveista monikollisia; S2: 16 % kaikista predikatiiveista monikollisia, 32 % sanajärjestykseltään käänteisten lauseiden predikatiiveista monikollisia). Lisäksi S1-aineistossa näiden sanajärjestykseltään käänteisten lauseiden monikolliset predikatiivit ovat aina partitiivissa, kun taas S2-aineistossa monikollisista predikatiiveista 53 % on nominatiivissa. Suoran sanajärjestyksen predikatiivilauseissa vastaavaa eroa ei näy, sillä S1-aineistossa monikollisista predika-

tiiveista 90 % on partitiivissa ja S2-aineistossa näin on 85 %:ssa lauseita. Esimerkki 9 havainnollistaa sellaista predikatiivilausetta, jossa on käänteinen sanajärjestys ja jossa monikollinen predikatiivi on nominatiivissa.

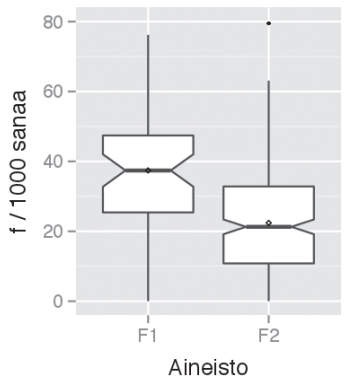
(9) Suomalais-ugrilaisen kantakielen sijasysteemi muodostui kahdesta sijatyypistä: kieliopillisista sijoista ja paikallissijoista.**Paikallissijat olivat erosijat* tulosijat ja olosijat (S2: las2-10tt01te03)

Esimerkin 9 sanajärjestykseltään käänteisen predikatiivilauseen predikatiivina on substantiivi, jonka sijanmerkintä on kohdekielen kannalta yllättävä vaikkakaan ei yksiselitteisen norminvastainen, ja subjektin ja predikatiivin looginen merkityssuhde on yksi yhteen (vrt. Sadeniemi 1950). Tällaisia ovat kahta adjektiivista tapausta lukuun ottamatta kaikki sanajärjestykseltään käänteisissä lauseissa olevat nominatiiviset predikatiivit. Käänteinen sanajärjestys on siis jo itsessään yleisempi S1-aineistossa, minkä lisäksi predikatiivin sijanmerkintään liittyvät erot korostuvat näissä tapauksissa. Nominatiivisten monikon predikatiivien on havaittu yliyleistyvän S2-kielessä partitiivin kustannuksella (Spoelman 2013: 307–310), ja inversion kaltaisen rakenteen epäprototyyppisyyden on havaittu lisäävän muun epäprototyyppisyyden mahdollisuutta (Ivaska 2011). Esimerkissä 9 saattaakin olla kyse konstruktion osien roolien hämärtymisestä.

5.2 Modaalisuuden ilmaiseminen tai ilmaisematta jättäminen

Kahdessa tilastollisessa analyysissa löytyneessä aineistojen välisessä frekvenssierossa syynä näyttää olevan modaalisuuden ilmaisemiseen tai ilmaisematta jättämiseen liittyvät erot. Näin on 1-grammeissa <inf1> ja <fin cond pres sg3>.

A-infinitiivi (<inf1>) esiintyy S1-aineistossa S2-aineistoa useammin, ja ero on tilastollisesti merkitsevä (ks. taulukko 2 s. 173). Kuvio 4 kuvaa A-infinitiivien jakautumista aineistoissa.



Kuvio 4.
1-grammin <inf1> jakauma laatikkojanalla kuvattuna S1- ja S2-aineistoissa.

Infinitiivin tarkastelu osoittaa, että se esiintyy tyypillisesti joko osana lauseen predikaattia (64 %, esim. 10), infinitiivisubjektina (esim. 11) tai e-NP:n määritteenä (esim. 12).

(10) Opettaja voi *puhua* helpotettua kieltä – – (S1: las2-vtt01vert050)

(11) – – siksi on helpompi *rekonstruoida* (S2: las2-26tt01te05)

(12) – – koska opettajalla ei ole mahdollisuutta *selittää* oppilaille kasvokkain (S2: las2-1tt01te12)

Aineistojen välinen ero liittyy verbiketjujen määrään, ja niissä erityisesti modaalisuutta ilmaiseviin konstruktioihin. Kuten taulukko 7 osoittaa, S1-aineistossa verbiketjun finiittiverbeistä viisi yleisintä verbiä ovat kaikki luonteeltaan modaalisia, ja nämä kattavat kaikista verbiketjuista yli 90 %. S2-aineistossa yleisimmät verbit sen sijaan vaihtelevat niin merkitykseltään kuin jakaumaltaankin enemmän ja mukana on myös laajemmin modaalisiksi käsitettäviä tahdon ja aikeen ilmauksia, kuten *haluta* ja *yrittää*.

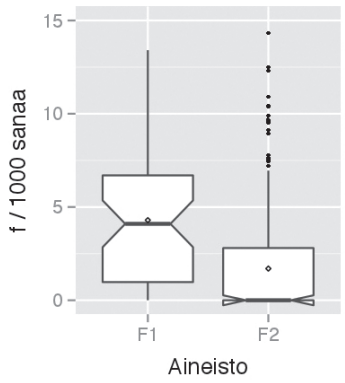
Taulukko 7.
Tyypillisimpien finiittiverbien leksikaalinen jakauma finiittisestä verbistä ja A-infinitiivistä koostuvissa verbiketjuissa.

Finiittiverbi		<inf1>
S1	S2	
voida 75 %	voida 59 %	
saattaa 10 %	pitää 8 %	
pitää 2 %	haluta 6 %	
tulla 2 %	alkaa 6 %	
täytyä 2 %	yrittää 3 %	
[…]	[…]	

Yksikön kolmannen persoonan konditionaalinen preesensmuotoinen verbi esiintyy S1-aineistossa S2-aineistoa useammin, ja ero on tilastollisesti merkitsevä (ks. taulukko 2 s. 173). Kuvio 5 (ks. viereinen sivu) kuvaa näiden konditionaalien jakautumista aineistoissa.

Lekseemien sisäinen jakauma osoittaa, että konditionaali esiintyy tyypillisesti modaalisissa ilmauksissa (ks. taulukko 8). Erityisesti *tulla*-verbi on yleisempi S1-aineistossa. S1-aineistossa konditionaalia seuraavassa kotekstissa on useammin A-infinitiivi (S1: 54 %; S2: 44 %), minkä lisäksi S1-aineistossa konditionaalia edeltää S2-aineistoa useammin genetiivinen subjekti (S1: 20 %; S2: 11 %). Frekvenssiero näyttääkin nivoutuvan erityisesti välttämättömyyttä ilmaisevan ’X:n *tulisi* <inf1>’ -konstruktion käyttöön (ks. esim. 13).

(13) Kielen oppimisen *tulisi* tähdätä käytännönläheisissä kommunikaatiotilanteissa pärjäämiseen (S1: las2-vtt01vert058)



Kuvio 5.
1-grammin <fin cond pres sg3> jakauma laatikkojanalla kuvattuna S1- ja S2-aineistoissa.

Konstruktiotulkintaa tukee myös se, että pelkän konditionaalin lisäksi yksikön genetiivistä, konditionaalista ja A-infinitiivistä koostuva 3-grammi on aineistoja toisistaan erottavien grammien listalla sijalla 29 (ks. taulukko 1 s. 171–172).

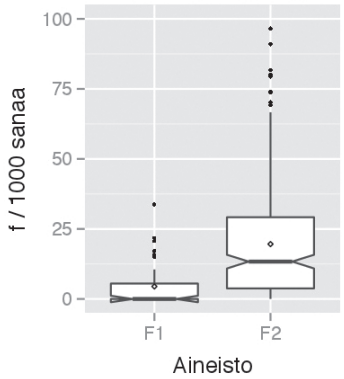
Taulukko 8.
Tyypillisimpien yksikön kolmannen persoonan konditionaalisten preesensin verbien leksikaalinen jakauma.

<fin cond pres sg3>	
S1	**S2**
olla 32 %	olla 37 %
tulla 25 %	pitää 15 %
voida 15 %	voida 15 %
pitää 11 %	tulla 4 %
saada 2 %	osata 2 %
[…]	[…]

Sekä A-infinitiivin että konditionaalin käyttöerot aineistojen välillä heijastavat aineistojen välistä eroa modaalisuuden ilmaisemisessa, ja modaalikonstruktiot vaikuttavat kaikkiaan yleisemmiltä S1-aineistossa. Tulokset ovat samansuuntaisia aiemman tutkimuksen kanssa. Seilonen (2013) toteaa sekä passiivisten modaaliverbien (mts. 64–69) että *voi*-konstruktioiden (mts. 103–107) yleistyvän kielitaidon karttuessa. Niiranen (2008) puolestaan sivuaa modaalisia verbikonstruktioita tutkiessaan oppimiskontekstin vaikutusta suomen oppimiseen. Hänen mukaansa kaksikielisessä ympäristössä suomea oppineet informantit käyttävät modaalisesta verbistä ja A-infinitiivistä koostuvaa verbikonstruktiota selvästi luokkahuoneympäristössä suomea opiskelleita oppijoita enemmän (mts. 300–302).

5.3 Imperfektin käyttö

Tilastollisen analyysin perusteella yksikön kolmannen persoonan indikatiivin imperfekti on yleisempi S2-aineistossa, ja ero on tilastollisesti merkitsevä (ks. taulukko 2 s. 173). Kuvio 6 havainnollistaa grammin jakaumia aineistoissa.



Kuvio 6.
1-grammin <fin ind pret sg3> jakauma laatikkojanalla kuvattuna S1- ja S2-aineistoissa.

Kaikkien aikamuotojen tarkastelu rinnakkain osoittaa, että ero keskittyy nimenomaan imperfektiin, kun taas preesens on yleisempi aikamuoto S1-aineistossa. Taulukko 9 esittää eri aikamuotojen frekvenssin aineistoissa sekä aineistojen välisten erojen tilastollisen merkitsevyyden.[8]

Taulukko 9.
Aikamuotojen frekvenssit ja niiden välisen eron tilastollinen merkitsevyys osa-aineistojen välillä.

Aikamuoto	ka. / 1 000 sanetta		med. / 1 000 sanetta		U-testi
	S1	S2	S1	S2	
preesens	106,2	94,4	106,3	98,6	U = 9111 Z = 2.1617 p = .03041 r = .12
imperfekti	7,2	29,8	2,0	18,9	U = 3551.5 Z = -6.3736 p = 3.25e-11 r =.35

8. Tässä jaottelussa preesens-predikaateiksi on laskettu ne, jotka ovat preesensmuodossa ja joiden kanssa samassa lauseessa ei esiinny NUT/TU-partisiippia osana predikaattia, imperfekti-predikaateiksi taas ne, jotka ovat imperfektimuodossa ja joiden kanssa samassa lauseessa ei esiinny NUT/TU-partisiippia osana predikaattia. Perfekti- ja pluskvamperfekti-predikaateiksi on laskettu ne, joissa on NUT/TU-partisiippi ja tarvittavassa muodossa oleva finiittinen *olla*-verbi.

perfekti	7,2	9,3	5,2	6,0	U = 7459 Z = -0.3708 p = .7118 r = .02
pluskvamperfekti	0,2	1,2	0	0	U = 6591 Z = -2.3667 p = .018 r = .13

Imperfektin morfologisten ilmiasujen jakauma on osa-aineistoissa samankaltainen. Indikatiivin yksikön ja monikon kolmas persoona sekä passiivin indikatiivi kattavat valtaosan kaikista esiintymistä, eivätkä niiden jakaumat juuri eroa toisistaan (ks. taulukko 10).

Taulukko 10.
Imperfektissä olevien verbien tyypillinen morfologinen jakauma.

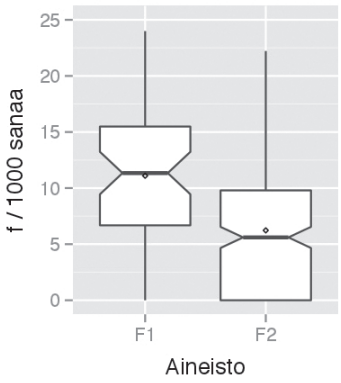
<pret>	
S1	**S2**
<fin ind pret sg3> 59 % <fin pass ind pret> 21 % <fin ind pret pl3> 17 % […]	<fin ind pret sg3> 61 % <fin ind pret pl3> 19 % <fin pass ind pret> 16 % […]

Ero näyttää siis perustuvan siihen, että S2-aineistossa käytetään enemmän imperfektiä kuin S1-aineistossa ja yliedustuminen syntyy ennen kaikkea preesensin kustannuksella. Samasta korpuksesta tehdyssä aiemmassa tutkimuksessa Siitonen ja Niemelä (2011) toteavat tempuksen liittyvän ennen kaikkea tekstien aiheeseen, mutta he vertaavat vain preesensin ja perfektin suhdetta, ja aineisto on tähän tutkimukseen verrattuna pieni. Räisänen (2005) toteaa tempusten aspektuaalisten piirteiden tuottavan oppijoille ongelmia, mikä voi johtaa aiempien opittujen kielten järjestelmien soveltamiseen tai prototyyppisten valintojen yliyleistymiseen. Näin ollen imperfekti voi yliyleistyä niin perfektin (mts. 58) kuin pluskvamperfektinkin (mts. 65) kustannuksella. Nyt esitetyt tulokset eivät kuitenkaan tue aspektitulkintaa, sillä ainoa selvä korrelaatio eri aikamuotojen distribuutiossa on preesensin ja imperfektin välillä.

5.4 Leksikaalisesti rajatut konstruktiot

Tilastollisessa analyysissa nousi esiin kaksi sellaista grammia, joissa ero näyttää selittyvän jonkin leksikaalisesti rajatun konstruktion käyttöerolla aineistojen välillä. Nämä grammit ovat 1-grammi <sg tra> ja 2-grammi <fin pass ind pres><pl ill>.

Yksikön translatiivi esiintyy S1-aineistossa S2-aineistoa useammin, ja ero on tilastollisesti merkitsevä (ks. taulukko 2 s. 173). Kuvio 7 havainnollistaa frekvenssien jakaumaa aineistojen välillä.



Kuvio 7.
1-grammin <sg tra> jakauma laatikkojanalla kuvattuna S1- ja S2-aineistoissa.

Sisäinen vaihtelu osoittaa, että aineistojen välinen ero selittyy lähinnä *esimerkiksi*-sanan adverbiaalisen käytön suuremmalla määrällä. Nämä tapaukset kattavat S1-aineiston <sg tra>-grammeista yli puolet, kun S2-aineistossa näin on vain kolmasosassa tapauksista, ja ero on niin ikään tilastollisesti merkitsevä (S1: 6,1 / 1 000 sanetta vs. S2: 1,9 / 1 000 sanetta; U = 11462.5, Z = 6.2629, p = $2.752e^{-10}$, r = .34). Esimerkki 14 kuvaa tällaista käyttöä.

(14) Arvioimista voi vaikeuttaa *esimerkiksi* se, – – (S1: las2-vtt01verto65)

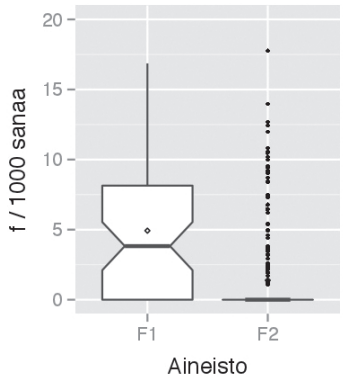
Toista leksikaalisesti rajatun konstruktion käyttöeroa kuvastaa indikatiivisesta passiiviverbistä ja yksikön illatiivista koostuva 2-grammi. Frekvenssien tarkastelu osoittaa, että tendenssi on samansuuntainen sekä yksiköllisillä (ks. taulukko 3 s. 173) että monikollisilla illatiiveilla (S1: 2,1 / 1 000 sanetta vs. S2: 0,8 / 1 000 sanetta; U = 9986.5, Z = 5.9369, p = $5.935e^{-08}$, r = .33), joten otan seuraavassa huomioon sekä yksiköllisen että monikollisen illatiivin sisältävät 2-grammit (S1: 4,9 / 1 000 sanetta vs. S2: 1,6 / 1 000 sanetta; U = 11566, Z = 7.1514, p = $4.58e^{-12}$, r = .39). Kuvio 8 havainnollistaa frekvenssien jakaumaa aineistojen välillä (ks. viereinen sivu).

Havainto vastaa aiemman tutkimuksen tuloksia, sillä passiivisten predikaattien frekvenssin ja niissä esiintyvän leksikaalisen vaihtelun on havaittu korreloivan kielitaidon kanssa nimenomaan B2- ja C1-tasojen välillä, johon valtaosa tämänkin tutkimuksen aineistosta sijoittuu (Seilonen 2013: 58–59). Sisäinen ja kotekstuaalinen vaihtelu osoittaa, että grammi edustaa usein joko transitiivista luokittelukonstruktiota ’X jaetaan/luokitellaan/järjestetään/jne. Y:ihin / N:ään Y:hyn / Y:hyn ja Z:aan’ (esim. 15) tai prosessia kuvaavaa intransitiivista konstruktiota ’X:ssä viitataan/keskitytään/pureudutaan/jne. Y:hyn’ (esim. 16).

(15) Lausesemantiikkaa tarkasteltaessa *sanat jaetaan* sisältö- ja *muotosanoihin* (S1: las2-vtt01verto25)

(16) *Demonstratiivideiksiksessa viitataan* kieliaktissa havaittavaan tai tekstissä ja kontekstissä mainittuun *tarkoitteeseen* (S2: las2-6tt01te08)

S1-aineistossa luokittelukonstruktio kattaa noin 34 % kaikista grammin esiintymistä, S2-aineistossa näitä on 26 %. Prosessikonstruktioita puolestaan on S1-aineistossa likimain 27 % kaikista grammeista ja S2-aineistossa 13 % kaikista grammeista. Molemmat siis kattavat suuremman osan S1-aineiston grammeista. Ero ei ole järin suuri, mutta se voi kuvata passiivin skemaattistumista lekseemeittäin ja merkityspiireittäin.



Kuvio 8.
2-grammin <fin pass ind pres><ill> jakauma laatikkojanalla kuvattuna S1- ja S2-aineistoissa.

6 Tulkintoja ja pohdintaa

Olen edellä tarkastellut avainrakenneanalyysin avulla edistyneen oppijansuomen ja ensikielisen suomen välisiä tyypillisiä eroja. Vastauksena ensimmäiseen tutkimuskysymykseen aineistosta valikoitui tilastollisen analyysin perusteella tarkasteltavaksi kahdeksan sellaista sanojen morfologisten muotojen perusteella määriteltyä n-grammia, joiden frekvenssit eroavat aineistojen välillä. Analysoin näiden grammien tyypillisen käytön perusteella, mihin konstruktioihin havaitut erot todennäköisimmin niveltyvät eli mitkä ovat aineistoja toisistaan erottavia avainrakenteita. Vastauksena toiseen tutkimuskysymykseen jaottelin erot analyysin perusteella morfosyntaktiseen kompleksisuuteen, modaalisuuden ilmaisemiseen, aikamuotojen käyttöeroihin sekä leksikaalisesti rajattujen konstruktioiden käyttöön liittyviin tapauksiin. Verbinjälkeiset substantiivin määritteet, VA-partisiipista muodostetut referatiivirakenteet ja sanajärjestykseltään käänteiset predikatiivilauseet ovat S1-aineistossa S2-aineistoa yleisempiä. Nämä voivat osoittaa S1-aineiston suurempaa rakenteellista kompleksisuutta. Modaalisesta verbistä ja A-infinitiivistä koostuvat predikaatit sekä konditionaaliset predikaatit ovat S1-aineistossa S2-aineistoa yleisempiä, ja voidaankin olettaa, että modaalisuuden ilmaiseminen erottaa aineistoja toisistaan. Imperfekti on sen sijaan huomatta-

vasti yleisempi S2-aineistossa kuin S1-aineistossa, mikä kuvastaa aikamuotojen käyttöeroja aineistojen välillä. Adverbiaalinen *esimerkiksi*-sana on puolestaan yleisempi S1-aineistossa, samoin jaottelukonstruktio 'X jaetaan Y:hyn ja Z:aan' ja prosessikonstruktio 'X:ssä viitataan Y:hyn'. Tarkastelun lähtökohtana olleet grammit eivät usein itsessään muodosta konstruktiokieliopin mukaisia toistuvia muoto–merkitys-pareja, vaan kotekstuaalisen analyysin perusteella erot ovat selitettävissä nimenomaan eroina tiettyjen konstruktioiden käytössä.

Nähdäkseni konstruktiokielioppi tarjoaa luontevan teoreettisen viitekehyksen tutkimukselle. Aloin tarkastelun yksiulotteisesti sanojen morfologisista muodoista, mutta lopulliset tulkinnat soveltavat konstruktiokieliopin ajatusta moniulotteisista ja keskenään hyvin erilaisista konstruktioista toisten konstruktioiden osana. Avainrakenneanalyysi tarjoaa määrällisiin eroihin perustuvan aineistolähtöisen metodisen kehyksen, jonka avulla paikannetaan aineistoja erottavia määrällisiä eroja ilman ennakko-oletuksia ja keskitetään tämän jälkeen kvantitatiivis-kvalitatiivinen analyysi ja tulkinta kulloinkin kyseessä olevan ilmiön tyypilliseen käyttöön ja siinä havaittaviin eroihin aineistojen välillä. Metodin vahvuutena on se, että tilastollisesti havaittavat toistuvat erot yhdistyvät kielen konstruktionaaliseen luonteeseen konkreettisesti eikä tutkittavien kielenpiirteiden valinta ole tutkijan intuition varassa. Kielenpiirteen aineistojen välisissä frekvensseissä havaittavan, tilastollisesti merkitsevän eron lisäksi eron takana olevan ilmiön tavoittaminen on ensiarvoisen tärkeää – vain näin voidaan paikantaa aineistoja toisistaan erottavia konstruktioita eli avainrakenteita. Keskenään korreloivien rakenteellisten piirteiden katsotaan kuvaavan luontevasti aineistojen välisiä eroja (vrt. Biber, Gray & Poonpon 2011), ja tyypillisen sisäisen ja kotekstuaalisen vaihtelun vertailu on yksi mahdollinen tapa havainnoida näitä. Samalla metodi ohjaa tarkastelua toistuvuutta ja variaatiota korostavien näkemysten väliselle keskitielle (vrt. Jantunen 2009: 375). Kielen eri tasojen rinnakkainen tarkastelu antaa erojen todellisista syistä holistisemman kuvan ja helpottaa erojen konstruktionaalisen luonteen arviointia.

Osaa tarkastelluista ilmiöistä on käsitelty aiemmissa S2-alan tutkimuksissa. Substantiivilausekkeiden käytön on havaittu erottavan S1- ja S2-oppijoita (Kannisto 2012), modaalisuuden ilmausten ja passiivin käytön erottavan eritasoisia (Seilonen 2013) ja modaalisuuden myös oppimistaustaltaan erilaisia oppijoita (Niiranen 2008), kun taas imperfektin on huomattu voivan yliyleistyä muiden aikamuotojen kustannuksella (Räisänen 2005). Osa konstruktioista puolestaan on sellaisia, että niitä ei tietääkseni ole tarkasteltu ainakaan suomi toisena kielenä -tutkimuksessa. Tässä tutkimuksessa kaikki käsitellyt ilmiöt kuitenkin nousivat esiin aineistosta.

Jäljelle jää silti kysymys siitä, mistä erot lopulta johtuvat. Syiden kattava selvittäminen vaatii kunkin ilmiön tarkempaa diskurssi- ja tekstilajikohtaista analysointia, mutta joitakin havaintoihin perustuvia pohdintoja voi tehdä. Monien erojen taustalla vaikuttanevat tarkastellun tekstilajin – akateemiseen diskurssiin kuuluvan tenttivastauksen – ominaispiirteet. Informantit ovat tekstejä tuottaessaan opiskelleet suomalaisessa yliopistossa, mutta valtaosa heistä on aiemmin opiskellut muualla, pääosin jollain muulla kielellä. Kielenkäyttökonventiot aina kirjallisista tenteistä alkaen saattavat olla heille uusia kielitaitotasosta riippumatta. Leksikaalisesti rajatuissa konstruktioissa yhteys lie-

nee selvä: kyse on tekstilajille ominaisista merkityksistä. Erot passiivikonstruktioissa voivat myös kuvata sitä, että konstruktioiden käyttö laajenee merkityskonteksti kerrallaan – passiivimuoto yksinään ei noussut aineistoa erottelevien morfologisten muotojen listalle. Tekstilaji voinee vaikuttaa lisäksi myös aikamuotojen käyttöön: monissa kohdin imperfektin käyttö korostaa tehtyä tutkimusta toimintana, kun akateemisessa suomessa tutkimuksen tuottaman tiedon kuvauksessa suositaan preesensiä. Modaalisuuden käyttö saattaa sekin liittyä tenttivastauksen luonteeseen, sillä aineistoja erottavat modaalikonstruktiot ovat luonteeltaan lieventäviä ja muiden vaihtoehtojen mahdollisuutta osoittavia. Rakenteelliseen kompleksisuuteen tässä tutkimuksessa liittämäni seikat ovat niin ikään varmasti tekstilajikohtaisia, ja konventioiden tuttuus vaikuttanee myös käyttöfrekvensseihin.

Referatiivirakenteet, predikatiivilauseiden käänteinen sanajärjestys ja määritteiden runsaus lienevät esimerkkejä rakenteellisesta kompleksisuudesta. Huomionarvoista on mielestäni se, että aineistoja erottaviksi piirteiksi osoittautuivat ennen kaikkea tapaukset, joissa on samanaikaisesti useita prototyyppirakenteesta poikkeavia piirteitä. Referatiivirakenteet poikkeavat morfosyntaktisesti monin tavoin finiittisestä lauseesta. Erot käänteisessä sanajärjestyksessä puolestaan näkyvät lähinnä monikollisten predikatiivien yhteydessä, joiden sijanmerkintä poikkeaa joiltain osin yksiköllisestä. Verbinjälkeisten substantiivin määritteiden ero johtuu ennen kaikkea tapauksista, joissa määritteitä on samassa lauseessa useita. Kompleksiset tekijät yksinään siis eivät erota S1- ja S2-aineistoja toisistaan, vaan sen tekevät tapaukset, joissa esiintyy samanaikaisesti useita rakenteellisesti kompleksisia ja prototyyppisimmästä tapauksesta poikkeavia piirteitä (vrt. Ivaska 2011). Kaikki havaitut erot ovat hienovaraisia ja vivahteikkaita kielellisiä keinoja, joiden hallitseminen kuvannee erittäin edistynyttä kielitaitoa. Tämä korostaa kielitaidon tekstilajikohtaisuutta ja kielen käyttöpohjaista luonnetta: kielenkäyttäjät toimivat heille tutuin kielellisin tavoin, ja erot käyttökontekstin tuttuudessa näkyvät myös eroina todennäköisissä kielellisissä tavoissa.

Tässä artikkelissa olen vertaillut S1- ja S2-aineistoja, mutta metodia voinee soveltaa minkä tahansa kielimuotojen välisten erojen tarkasteluun. Jatkotutkimuksen on paneuduttava siihen, miten keskittää analyysi pelkän eron sijaan taustalla olevaan ilmiöön. Esimerkiksi tässä tutkimuksessa havaittu yksikön translatiivin frekvenssiero ei liity mitenkään kyseiseen morfologiseen muotoon vaan *esimerkiksi*-sanan käyttöeroon. Tämän artikkelin havainnot perustuvatkin homogeeniseen aineistoon, ja ainakin osa havaituista eroista liittyy todennäköisesti nimenomaan eroihin tutkitussa tekstilajissa – akateemisissa tenttivastauksissa. Tästä syystä huolellinen sisäisen ja kotekstuaalisen vaihtelun analyysi on ensiarvoisen tärkeää. Suuremmat aineistot mahdollistavat hienosyisempien tilastollisten menetelmien käytön myös näiden erojen selvittämisessä.

Lähteet

Aarts, Jaan – Granger, Sylviane 1998: Tag sequences in learner corpora. A key to interlanguage grammar and discourse. – Sylviane Granger (toim.), *Learner English on computer*

s. 132–141. London: Longman.
Bates, Elizabeth – MacWhinney, Brian 1987: Competition, variation and language learning. – Brian MacWhinney (toim.), *Mechanisms of language acquisition* s. 157–193. Hillsdale: Lawrence Erlbaum Associates.
Biber, Douglas – Conrad, Susan – Cortes, Viviana 2004: If you look at… Lexical bundles in university teaching and textbooks. – *Applied Linguistics* 25 s. 371–405.
Biber, Douglas – Gray, Bethany – Poonpon, Kornwipa 2011: Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? – *TESOL Quarterly* 45 s. 5–35.
Bley-Vroman, Robert 1983: The comparative fallacy in interlanguage studies. The case of systematicity. – *Language Learning* 33 s. 1–17.
Breiman, Leo 2001: Random forests. – *Machine Learning* 45 s. 5–32.
Bybee, Joan – Thompson, Sandra 1997: Three frequency effects in syntax. – *Berkeley Linguistics Society* 23 s. 65–85.
CEFR = *Common European framework for languages. Learning, teaching, assessment* 2006. Cambridge: Cambridge University Press.
Cheng, Winnie – Greaves, Chris – Warren, Martin 2006: From n-gram to skipgram to concgram. – *International Journal of Corpus Linguistics* 11 s. 411–433.
Conrad, Susan – Biber, Douglas 2004: The frequency and use of lexical bundles in conversation and academic prose. – *Lexicographica* 20 s. 56–71.
Creswell, John. W. – Plano Clark, Vicki. L. – Garrett, Amanda 2008: Methodological issues in conducting mixed methods research designs. – Manfred Max Bergman (toim.), *Advances in mixed methods research* s. 66–83. London: Sage.
Croft, William 2001: *Radical Construction Grammar. Syntactic theory in typological perspective*. Oxford: Oxford University Press.
Fayyad, Usama – Piatetsky-Shapiro, Gregory –Smyth, Padhraic 1996: From data mining to knowledge discovery in databases. – *AI Magazine* 17 s. 37–54. http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf (10.10.2013).
Fillmore, Charles J. 1985: Syntactic intrusions and the notion of grammatical construction. – *Berkeley Linguistics Society* 11 s. 73–86.
Firth, John Rupert 1968 [1957]: *Selected papers of J. R. Firth*. Toimittanut F. R. Palmer. London: Longman.
Francis, Gill 1993: A corpus-driven approach to grammar. Principles, methods and examples. – Mona Baker, Gill Francis & Elena Tognini-Bonelli (toim.), *Text and technology. In honour of John Sinclair* s. 137–156. Amsterdam: John Benjamins.
Goldberg, Adele 1995. *Constructions. A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.
Goldberg, Adele 2006: *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.
Granger, Sylviane 1996: From CA to CIA and back. An integrated approach to computerized bilingual and learner corpora. – Karin Aijmer, Bengt Altenberg & Mats Johansson (toim.), *Languages in contrast* s. 37–51. Lund: Lund University Press.
—— 1998: Prefabricated patterns in advanced EFL writing. Collocations and formulae. – Anthony Paul Cowie (toim.), *Phraseology. Theory, analysis, and applications* s. 145–160. Oxford: Clarendon Press.
—— 2013: Contrastive interlanguage analysis. A reappraisal. Esitelmä konferenssissa Learner Corpus Research 2013. Bergen 27.–29.9.2013.

Granger, Sylviane – Paquot, Magali 2008: Disentangling the phraseological web. – Sylviane Granger & Fanny Meunier (toim.), *Phraseology. An interdisciplinary perspective* s. 27–49. Amsterdam: John Benjamins.

Gries, Stefan Th. 2008: Phraseology and linguistic theory. – Sylviane Granger & Fanny Meunier (toim.), *Phraseology. An interdisciplinary perspective* s. 3–25. Amsterdam: John Benjamins.

Guthrie, David – Allison, Ben – Liu, Wei – Guthrie, Louise – Wilks, Yorick 2006: A closer look at skipgram modelling. – *Proceedings of fifth international conference on language resources and evaluation (LREC), Genoa, Italy*. http://www.lrec-conf.org/proceedings/lrec2006/ (24.07.2013).

Hollander, Myles – Wolfe, Douglas A. 1973: *Nonparametric statistical methods*. New York: John Wiley & Sons.

Hopper, Paul 1987: Emergent grammar. – *Berkeley Linguistics Society* 13 s. 139–157.

Hothorn, Torsten – Buehlmann, Peter – Dudoit, Sandrine – Molinaro, Annette – Van Der Laan, Mark 2006: Survival ensembles. – *Biostatistics* 7 s. 355–373.

Hulstijn, Jan 2011: Language proficiency in native and nonnative speakers. An agenda for research and suggestions for second-language assessment. – *Language Assessment Quarterly* 8 s. 229–249.

Hunston, Susan 2001: Colligaction, lexis pattern, and text. – Mike Scott & Geoff Thompson (toim.), *Patterns of text. In honour of Michael Hoey* s. 13–34. Amsterdam: John Benjamins.

Hunston, Susan – Francis, Gill 2000: *Pattern grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Inaba, Nobufumi 2007: Mikael Agricolan teokset tietokannan muodossa. – Kaisa Häkkinen & Tanja Vaittinen (toim.), *Agricolan aika* s. 147–161. Helsinki: BTJ.

Itkonen, Esa 2005: *Analogy as structure and process*. Amsterdam: John Benjamins.

Ivaska, Ilmari 2011: Lausetyyppien sekoittuminen edistyneessä oppijansuomessa. Näkökulmana eksistentiaalilause. – *Lähivõrdlusi. Lähivertailuja* 21 s. 65–85.

—— 2012: Keystructure analysis of formally defined structures of learner Finnish. Esitelmä konferenssissa Learner Language, Learner Corpora. Oulu 5.–6.10.2012.

—— (tulossa 2014): The corpus of advanced learner Finnish (LAS2). Database and toolkit to study academic learner Finnish. – *Apples: Journal of Applied Language Studies* 8(3). http://apples.jyu.fi.

Ivaska, Ilmari – Siitonen, Kirsti 2011: Avainrakenneanalyysi. Tapa tutkia oppijankielen lauserakennetta korpusvetoisesti. – *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 3 s. 35–47. http://ojs.tsv.fi/index.php/afinla/issue/view/694.

Jantunen, Jarmo Harri 2009: ”Minulla on aivan paljon rahaa”. Fraseologiset yksiköt suomen kielen opetuksessa. – *Virittäjä* 113 s. 356–381.

—— 2011: Avainsana-analyysi annotoidun oppijankieliaineiston tutkimisessa. Alustavia havaintoja. – *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 3 s. 48–61. http://ojs.tsv.fi/index.php/afinla/issue/view/694.

Jarvis, Scott 2008: The detection-based approach. An overview. – Scott Jarvis & Scott A. Crossley (toim.), *Approaching language transfer through text classification. Explorations in the detection-based approach* s. 1–33. Bristol: Multilingual Matters.

Kannisto, Sara 2012: *Substantiivilausekkeiden määritteet edistyneiden suomenoppijoiden kirjoituksessa*. Pro gradu -tutkielma. Turun yliopiston suomen kielen oppiaine.

Komppa, Johanna 2012: *Retorisen rakenteen teoria suomi toisena kielenä -ylioppilaskokeen kirjoitelman kokonaisrakenteen ja kappalejaon tarkastelussa*. Helsinki: Helsingin yliopiston

suomen kielen, suomalais-ugrilaisten ja pohjoismaisten kielten ja kirjallisuuksien laitos.
Kotilainen, Lari 2007: *Konstruktioiden dynamiikkaa*. Helsinki: Helsingin yliopiston suomen kielen, suomalais-ugrilaisten ja pohjoismaisten kielten ja kirjallisuuksien laitos.
LaX = Lauseopin X-arkisto. Kotimaisten kielten keskus & Turun yliopiston kieli- ja käännöstieteiden laitos, Lauseopin arkisto. Turku.
Liaw, Andy – Wiener, Matthew 2002: Classification and regression by randomForest. – *R News* 2 (3). http://www.r-project.org/doc/Rnews/ (18.12.2013).
Lieven, Elena – Pine, Julian – Baldwin, Gillian 1997: Lexically-based learning and early grammatical development. – *Journal of Child Language* 24 s. 187–219.
Martin, Maisa 2007: A square peg into a round hole? Fifteen years of research into Finnish as a second language. – *Nordand. Nordisk tidsskrift for andrespråksforskning* 2 s. 63–85.
Nesselhauf, Nadja 2004: *Collocations in a learner corpus*. Studies in Corpus Linguistics 14. Philadelphia: John Benjamins.
Niiranen, Leena 2008: *Effects of learning contexts on knowledge of verbs. Lexical and inflectional knowledge of verbs among pupils learning Finnish in Northern Norway*. Julkaisematon väitöskirja. University of Tromsø. http://hdl.handle.net/10037/2109 (20.5.2014).
Osborne, John 2013: Comparisons are odorous. Native-speaker data in learner corpus research. Esitelmä konferenssissa Learner Corpus Research 2013. Bergen 27.–29.9.2013.
Paquot, Magali 2008: Exemplification in learner writing. A cross-linguistic perspective. – Fanny Meunier & Sylviane Granger (toim.), *Phraseology in language learning and teaching* s. 101–119. Amsterdam: John Benjamins.
Pazos Bretaña, Jose-Manuel – Pamies Bertrán, Antonio 2008: Combined statistical and grammatical criteria for the retrieval of phraseological units in an electronic corpus. – Sylviane Granger & Fanny Meunier (toim.), *Phraseology. An interdisciplinary perspective* s. 391–406. Amsterdam: John Benjamins.
R 2013 = R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Ringbom, Håkan 1998: Vocabulary frequencies in advanced learner English. A cross-linguistic approach. – Sylviane Granger (toim.), *Learner English on computer* s. 41–52. London: Longman.
Räisänen, Johanna 2005: *Suomen tempusten semantiikka tšekin- ja venäjänkielisten suomenoppijoiden välikielissä*. Pro gradu -tutkielma. Turun yliopiston suomen kielen oppiaine.
Sadeniemi, Matti 1950: Totaalisesta ja partiaalisesta predikatiivista. – *Virittäjä* 54 s. 46–53.
Salmi, Niina 2010: *Virkkeiden ja lauseiden piirteitä edistyneiden suomenoppijoiden teksteissä*. Pro gradu -tutkielma. Turun yliopiston suomen kielen oppiaine.
Scott, Mike 2010: Problems in investigating keyness, or clearing the undergrowth and marking out trails... – Marina Bondi & Mike Scott (toim.), *Keyness in texts* s. 43–57. Studies in Corpus Linguistics 41. Amsterdam: John Benjamins.
Scott, Mike – Tribble, Christopher 2006: *Textual patterns. Key words and corpus analysis in language education*. Studies in Corpus Linguistics 22. Amsterdam: John Benjamins.
Seilonen, Marja 2013: *Epäsuora henkilöön viittaaminen oppijansuomessa*. Jyväskylä Studies in Humanities 197. Jyväskylä: Jyväskylä University Printing House.
Siitonen, Kirsti – Niemelä, Jenny 2011: Mitä pitkittäistutkimus voi paljastaa edistyneiden suomenoppijoiden kielitaidosta? – *Lähivõrdlusi. Lähivertailuja* 21 s. 242–279.
Sinclair, John 1991: *Corpus, concordance, collocation*. Oxford: Oxford University Press.

—— 2001: Review. – *International Journal of Corpus Linguistics* 6 s. 339–359.

Skousen, Royal 1989: *Analogical modeling of language*. Dodrecht: Kluwer Academic Publishers.

Spoelman, Marianne 2013: *Prior linguistic knowledge matters. The use of partitive case in Finnish learner language*. Acta Universitatis Ouluensis B Humaniora 111. Oulu: Oulun yliopisto.

Stefanowitsch, Anatol – Gries, Stefan Th. 2003: Collostructions. Investigating the interaction of words and constructions. – *International Journal of Corpus Linguistics* 8 s. 209–243.

Strobl, Carolin – Boulesteix, Anne-Laure – Zeileis, Achim – Hothorn, Torsten 2007: Bias in random forest variable importance measures. Illustrations, sources and a solution. – *BMCBioinformatics* 8 (25). http://www.biomedcentral.com/1471-2105/8/25 (15.11.2012).

Strobl, Carolin – Boulesteix, Anne-Laure – Kneib, Thomas – Augustin, Thomas – Zeileis, Achim 2008: Conditional variable importance for random forests. – *BMCBioinformatics* 9 (307). http://www.biomedcentral.com/1471-2105/9/307 (18.12.2013).

Suni, Minna 2012: The impact of Finno-Ugric languages in second language research. Looking back and setting goals. – *Lähivõrdlusi. Lähivertailuja* 22 s. 407–438.

Tagliamonte, Sali – Baayen, R. Harald 2011: Models, forests and trees of York English. Was/were variation as a case study for statistical practice. – *Language Variation and Change* 24 s. 135–178.

Tognini-Bonelli, Elena 2001: *Corpus linguistics at work*. Studies in Corpus Linguistics 6. Amsterdam: John Benjamins.

Tomasello, Michael 1992: *First verbs. A case study of early grammatical development.* Cambridge: Cambridge University Press.

Ukkola, Annette 2009: *Taas yksi nollatutkimus. Substantiivilausekkeiden määritteet S2-oppijoiden teksteissä*. Pro gradu -tutkielma. Jyväskylän yliopiston suomen kielen oppiaine.

Vetchinnikova, Svetlana 2012: Idiom principle in second language use. Esitelmä konferenssissa Learner language, learner corpora. Oulu 5.–6.10.2012.

Wiersma, Wybo – Nerbonne, John – Lauttamus, Timo 2011: Automatically extracting typical syntactic differences from corpora. – *Literary and Linguistic Computing* 26 s. 107–124.

SUMMARY

Key structures in advanced learner Finnish: Corpus approach towards structural differences between two language forms

When the language of advanced learners is compared with that of native speakers, frequency observations often serve as a starting point for more detailed analysis. Frequency observation makes it possible to observe when these two language forms function similarly and when they do not. This article focuses on the differences between written advanced learner Finnish (F2) and writing by native speakers (F1). It does so in a corpus-driven manner by analysing those morphological forms and their combinations (n-grams) that differ most in terms of their respective frequencies. The research method applied – key-structure analysis – detects statistically the most extensive frequentative differences between texts, then focuses on the resulting n-grams by analysing them in terms of their inner and co-textual variation. By doing so, the methodology can identify those constructions that actually differ between the two language forms.

The results show that even advanced F2 texts contain features that are used differently to those in F1 texts. Such differences are likely to depict both features used differently in the two language forms in general and features specific to the text genre under investigation – the written, academic exam essay. The constructions identified here are divided into four groups: constructions of modality, morphosyntactically complex constructions, tense constructions, and lexically specific constructions. These constructions are of a very different nature, but they were all detected in a data-driven manner without drawing any hypotheses regarding the potential differences. Key-structure analysis can thus be considered a useful methodological procedure, when the aim is to extrapolate from quantitative differences to the actual linguistic phenomena that constitute the difference itself. This method can well be applied to any research design of a contrastive nature.

Edistyneen oppijansuomen avainrakenteita: Korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin

Kun edistyneiden kielenoppijoiden kieltä verrataan ensikielisiin kielenpuhujiin, tarkastellaan usein eri kielenpiirteiden käyttöfrekvenssejä. Näin voidaan tarkastella sitä, milloin nämä kielimuodot toimivat keskenään samalla lailla ja milloin ne eroavat toisistaan. Tässä artikkelissa edistyneen oppijansuomen ja ensikielisen suomen välisiä eroja tutkitaan korpusvetoisesti tarkastelemalla sellaisia morfologisia muotoja ja morfologisten muotojen yhdistelmiä (n-grammeja), joiden välillä aineistoissa on suurimmat frekvenssierot. Tutkimusmetodina on avainrakenneanalyysi, jossa vertailtavista aineistoista etsitään ensin tilastollisesti suurimmat erot, minkä jälkeen löydetyt n-grammit analysoidaan niiden sisäisen ja kotekstuaalisen vaihtelun kannalta. Näin päästää käsiksi niihin konstruktioihin, jotka todella erottavat kielimuotoja toisistaan.

Tulokset osoittavat, että vielä edistyneidenkin suomenoppijoiden kirjoituksessa on piirteitä, joiden käyttö eroaa ensikielisten kirjoituksesta. Erot kuvannevat sekä ensikielisiä ja edistyneitä suomenoppijoita erottavia ilmiöitä ylipäätään että tutkittuun tekstilajiin – akateemiseen tenttivastaukseen – liittyviä erityispiirteitä. Löydetyt konstruktiot on tässä artikkelissa jaettu neljään ryhmään: modaalisuus-konstruktioihin, morfosyntaktisesti kompleksisiin konstruktioihin, aikamuotoihin ja leksikaalisesti spesifeihin konstruktioihin. Konstruktiot ovat luonteeltaan hyvin erilaisia keskenään, mutta ne kaikki tavoitettiin aineistolähtöisesti ilman ennakko-oletuksia aineistojen välisistä eroista. Avainrakenneanalyysia voidaankin pitää luontevana metodologisena prosessina, kun halutaan edetä määrällisesti havaituista eroista niihin kielenpiirteisiin, joissa verratut aineistot eroavat toisistaan. Metodia voitaneen soveltaa mihin tahansa kahta tai useampaa aineistoa vertaavaan tutkimusasetelmaan.

Kirjoittajan yhteystiedot (address):
etunimi.sukunimi@utu.fi