Oppijansuomen n-grammit: korpusvetoinen tutkimus B1-kielitaitotason toistuvista monisanaisista rakenteista

Maisterintutkielma
Juha-Matti Pekkala
Suomen kieli
Kieli- ja viestintätieteiden laitos
Jyväskylän yliopisto
2020

JYVÄSKYLÄN YLIOPISTO

Tiedekunta – Faculty Humanistis-yhteiskuntatieteellinen tiedekunta	Laitos – Department Kieli- ja viestintätieteiden laitos
Tekijä – Author Juha-Matti Pekkala	
Työn nimi – Title Oppijansuomen n-grammit: korpusvetoinen tutkimus B1-kielitaitotason toistuvista monisanaisista rakenteista.	
Oppiaine – Subject Suomen kieli	Työn laji – Level Maisterintutkielma
Aika – Month and year Lokakuu 2020	Sivumäärä – Number of pages 117 sivua + liitteet

Tiivistelmä – Abstract

Tutkimuksessa kartoitetaan oppijansuomen B1-kielitaitotason kirjoitetuissa teksteissä ilmeneviä n-grammeja. Kielentutkimuksessa *n-grammeilla* tarkoitetaan tutkittavassa kieliaineistossa usein toistuvia, *n*-määrästä sanoja koostuvia sanaketjuja. Niiden ei tarvitse olla esimerkiksi kieliopillisesti täydellisiä rakenteita tai idiomeja. (Biber, Johansson, Leech, Conrad & Finegan 1999: 989–990; Granger & Paquot 2008: 38–39.) Tutkimus sijoittuu teoreettiselta viitekehykseltään tiiveimmin fraseologiaan, joka on kielentutkimuksen ala, jossa tutkitaan ennen kaikkea kielen käyttöä ja sen erilaisia valmisrakenteisia yksiköitä. N-grammit lukeutuvat sanojen syntagmaattisia myötäesiintymiä ilmaiseviin fraseologisiin yksikköihin.

Tutkimuksen merkittävimpänä tavoitteena on selvittää, millaisia n-grammeja *Eurooppalaisella viitekehyksellä* (EVK 2003) B1-kielitaitotasolle arvioidut suomenoppijat kirjoitetuissa teksteissään tuottavat ja mitä n-grammit kykenevät B1-tasoisen oppijansuomen leksikaalisista ja rakenteellisista piirteistä kertomaan. Samalla tutkimuksella halutaan laajentaa fraseologista oppijansuomen tutkimusta n-grammien osalta sekä kokeilla erilaisten metodien toimivuutta n-grammien tarkastelussa. Tutkimuksen aineistona toimii *Kansainvälinen oppijansuomen korpus* (ICLFI), jota lähestytään korpusvetoisella tutkimusmetodilla. Tutkimuksen aluksi korpuksen B1-taitotasoarvioinnin saaneista teksteistä haetaan korpusohjelmalla kaikki niissä esiintyvät, ennalta määrätyt raja-arvot ylittävät 3-, 4-, 5- ja 6-grammit, jotka kootaan yhdeksi listaksi frekvenssiensä mukaan järjestäen. Näitä n-grammeja analysoidaan tämän jälkeen leksikon osalta siten, että n-grammien sanoista laaditaan omat sananmuoto- ja lemmalistansa, ja rakenteiden puolesta niin, että huomio kiinnitetään ensi sijassa sellaisiin n-grammeihin, joihin sisältyy finiittiverbi, verbiliitto tai osa verbiliitosta. Verbillisistä n-grammeista selvitetään, mitä tempuksia ja syntaktisia lausetyyppejä (VISK § 891) niissä ilmenee ja kuinka paljon.

Tutkimuksen tulokset osoittavat, että leksikkonsa puolesta B1-tasoiset suomenoppijat käyttävät paljon natiivisuomestakin tuttua sanastoa, mutta mukana on myös joitain selkeitä leksikaalisia yliedustumia. Ne selittyvät pitkälti tekstien tehtävänannoilla. Sananmuotojen perusteella oppijat suosivat verbeissä runsaasti yksikön ensimmäisen persoonan muotoja. N-grammien lemmoissa edustuvat sanaluokkien osalta eniten nominit (45 %) ja verbit (32 %). Rakenteiden puolesta oppijansuomessa käytetään tempuksien osalta ylivoimaisesti eniten preesensiä (88,6 % verbillisten n-grammien esiintymistä). Syntaktisista lausetyypeistä käytetyin on kopulalause, jota ilmentää 36,4 prosenttia verbillisten n-grammien esiintymistä. Muina yleisluontoisina huomioina todetaan muun muassa, että n-grammeista yli 80 prosenttia on 3-grammeja ja että n-grammeissa esiintyy verrattain vähän kielenvastaisia muotoja. Moduksista indikatiivi on yhtä konditionaalimuotoista n-grammia lukuun ottamatta ainoa verbillisissä n-grammeissa käytetty tapaluokka. Verbittömät n-grammit toimivat tulosten perusteella pääosin asioiden rinnastajina sekä suhteuttavat tapahtumia aikaan. Tutkimustulokset lisäävät tietoa oppijankielen fraseologisista piirteistä ja viitoittavat osaltaan tietä mahdolliselle tulevalle tutkimukselle aiheesta.

Asiasanat – Keywords fraseologia, klusterit, korpustutkimus, n-grammit, suomi vieraana kielenä

Säilytyspaikka – Depository

Muita tietoja – Additional information

SISÄLLYS

KONKORDANSSIT

KUVIOT

TAULUKOT

1 JOHDANTO

Vanhakantaisen käsityksen mukaan kielessä sanasto ja kielioppi toimivat toisistaan irrallisina yksikköinä, jotka myös opitaan omina osa-alueinaan sana sanalta ja rakenne kerrallaan. Sittemmin esiin on noussut teorioita, jotka esittävät, että kielenoppiminen tapahtuisikin sujuvammin niin, että kielenoppija omaksuu yksittäisten sanojen ja niistä irrallisten konstruktioiden sijaan useammasta sanasta koostuvia, enemmän tai vähemmän valmisrakenteisia sanakimppuja tai -könttiä, joita aletaan varioida ja joiden abstraktiivisuustaso kasvaa oppimisen ja kielenkäytön myötä (ks. esim. N. Ellis 1997; Wray 2000; N. Ellis, O'Donnell & Römer 2013; McCauley & Christiansen 2015). Tällaista näkemystä kielenoppimisesta korostaa muun muassa fraseologinen näkökulma kielestä ja sen omaksumisesta. Sanakönttiin (engl. *chunks*) voivat lukeutua niin fraasit, rutiini-ilmaukset kuin idiomitkin – esimerkiksi *Hyvää huomenta*, *Kaunis päivä tänään*, *En osaa sanoa*, *selvä pyy* ja *heittää helmiä sioille* – mutta myös muut usein kielessä ilmenevät, rakenteeltaan enemmän tai vähemmän keskeneräisetkin sanayhtymät, kuten vaikkapa *hän sanoi minulle että*, *siihen asti kunnes* ja *Mitä sinä siinä teet*. Tällaisten könttärakenteiden hallitseminen on vieraan ja toisen kielen oppimisen yhteydessä avain idiomaattisen, natiivinkaltaisen kielen omaksumiseen. Samalla ne auttavat kielenoppijaa sekä kielen käyttämisessä että sen sujuvoittamisessa, kun jokaista opittavaa kokonaisuutta ei tarvitse purkaa yksittäisinä sanoina ja kieliopillisina kategorioina analysoitaviksi osasiksi.

Käsillä olevan tutkimuksen ytimessä ovat sellaiset kolmesta tai useammasta peräkkäisestä sanasta koostuvat, verrattain usein tutkittavassa aineistossa toistuvat sanaköntät tai -kimput, joita suomenoppijat[1] tuottavat. Tällaisia voivat olla vaikkapa ilmaukset *minulla on yksi*, *nousen aamulla kello*, *mutta se ei ole* ja *olen sitä mieltä että*[2]. Kielentutkimuksessa tällaisia toistuvasti tavattavia sanakönttiä nimitetään (muun muassa) *n-grammeiksi*. Tavoitteenani on selvittää korpusaineistoon nojautuen, mitkä ovat suomenoppijoiden tyypillisimmin tuottamia n-grammeja ja kuinka paljon he näitä käyttävät, millaisista kielellisistä rakenteista nämä n-grammit koostuvat ja mitä ne omalta osaltaan kykenevät kertomaan oppijansuomesta kielimuotona. Samalla tutkimuksessa sivutaan myös sitä, miten n-grammeja kyetään ylipäänsä lähestymään tutkimusmenetelmällisellä tasolla. Tutkimus kohdistuu *Eurooppalaisella viitekehyksellä* (EVK 2003) kielitaitotasolle B1 arvioituihin suomenoppijoiden teksteihin, jolloin sen avulla

[1] *Suomenoppija* on tässä tutkimuksessa termi, jota käytetään kuvaamaan muuta kuin suomea ensikielenään puhuvaa, suomen kieltä opiskelevaa henkilöä. Tällaisen henkilön oppimisen kohteena olevan kielen kielimuotoa nimitetään tutkimuksessa *oppijankieleksi*, tarkemmin rajattuna *oppijansuomeksi*.
[2] Esimerkit tutkimusaineistosta.

kerrytetään tietoa etenkin siitä, millaisia useampisanaisia yksikköjä suomen kielessä jo perustaitotason saavuttaneet, niin kutsutulla kynnystasolla operoivat, suomenoppijat käyttävät hyväkseen.

Tutkimus linkittyy teoriataustansa puolesta kontekstuaalisen semantiikan, tarkemmin ottaen fraseologian, sekä oppijankielen tutkimuksen viitekehyksiin. Kontekstuaalinen semantiikka tutkii merkitysten muodostumista kontekstien kautta siinä missä fraseologia taas on kielitieteen haara, joka kartoittaa ennen kaikkea kielen käyttöä, sen valmisrakenteisuutta ja erilaisia kieleen vakiintuneita useammasta sanasta koostuvia yksikköjä, kuten n-grammeja. Tutkimuksen tavoitteena on ennen kaikkea laajentaa käsitystä ja tietoutta oppijansuomen yhdestä kielitaitotasosta ja sen erilaisista ominaispiirteistä. Sillä halutaan tuoda esiin myös niitä mahdollisuuksia, joita n-grammianalyysi voi fraseologiseen oppijansuomen tutkimukseen tuoda, sillä tähän mennessä n-grammit ovat jääneet siinä melko vähäiselle huomiolle.

Kielentutkimuksessa *n-grammeilla* (*n-grams*), toisilta nimiltään muun muassa *klustereilla* (*clusters*), *sanakimpuilla* (*lexical bundles*) ja *ryppäillä*[3], tarkoitetaan kielessä taajaan toistuvia, vähintään kaksi sanaa sisältäviä sanaketjuja. Näiden sanaketjujen ei tarvitse olla esimerkiksi kieliopillisesti täydellisiä rakenteita tai mitenkään idiomaattisia ilmauksia, eivätkä ne usein olekaan yhdellä sanalla korvattavissa kuten varsinaiset idiomit. (Biber, Johansson, Leech, Conrad & Finegan 1999: 989–990; Granger & Paquot 2008: 38–39; Jantunen 2009b: 359.) Nimityksessä *n-grammi* kirjain *n* osoittaa, monestako sanasta rakenne koostuu; esimerkiksi ilmaus *ei ole kovin* on siis 3-grammi[4] (Jantunen 2009b: 359).

Tutkimuksen tutkimusaineistona toimii Oulun yliopistossa koottu *Kansainvälinen oppijansuomen korpus* (Jantunen, Brunni & Oulun yliopisto 2013). Korpus[5] koostuu ulkomaisista yliopistoista kerätyistä suomen kielen opiskelijoiden tuottamista teksteistä, eli se on suomi vieraana kielenä[6] -korpus. Jokainen korpukseen lukeutuvista teksteistä on arvioitu *Eurooppalaisen viitekehyksen* (EVK 2003) kielitaitotasojen (A1–C2) mukaisesti kahden arvioijan toimesta, ja mikäli heidän arviointinsa ovat poikenneet toisistaan, on mukana ollut vielä kolmas arvioija (Jantunen & Pirkola 2015: 97). Kuten edellä mainittiin, kohdistetaan tämä tutkimus ainoastaan

[3] Tässä tutkimuksessa tukeudutaan jatkossa lukuisista erilaisista nimityksistä systemaattisesti termiin *n-grammi* kaikenlaisiin vakiintuneisiin useamman sanan yhtymiin viittaajana riippumatta siitä, mitä nimitystä niistä on lähdeteoksissa käytetty (ks. luku 2.2.2).

[4] 3-grammille vaihtoehtoinen nimitys on tri-grammi. 2-grammille vastaava nimitys on bi-grammi. (Esim. D. Guthrie, Allison, Liu, L. Guthrie & Wilks 2006; Shibuya & Jensen 2015: 26.)

[5] Termillä *korpus* tarkoitetaan järjestelmällisesti kerättyä, suhteellisen laajaa näytettä luonnollisesta puhutusta tai kirjoitetusta kielestä (Tieteen termipankki s.v. *korpus*; ks. tämän tutkimuksen luku 3.1).

[6] *Suomella vieraana kielenä* viitataan suomen kielen opiskeluun lähtökohtaisesti muualla kuin Suomessa erotuksena *suomi toisena kielenä* (S2) -käsitteestä, jolla tarkoitetaan yleensä Suomessa tapahtuvaa suomen opiskelua (ks. käsitteistä esim. Latomaa & Tuomela 1993; *suomi toisena ja vieraana kielenä* -alasta Martin 1999).

B1-kielitaitotasoarvionnin saaneisiin suomenoppijoiden teksteihin. Niistä koostetaan tutkimusta varten niin sanottu osakorpus, josta etsitään korpusohjelmalla n-grammeja erilaisin raja-arvoin. Syy, miksi tutkimus on rajattu vain yhteen kielitaitotasoon, eikä siinä tutkita koko *Kansainvälisestä oppijansuomen korpuksesta* löytyviä n-grammeja, on lähinnä se, että koko korpusta tutkimalla analysoitavien n-grammien määrä nousisi ainakin löyhempiä raja-arvoja käyttämällä erittäin suureksi ja näin ollen liian haastavaksi käsitellä. Juuri B1-kielitaitotaso on valikoitu tarkastelun kohteeksi osittain siksi, koska B1-taitotasolle arvioidut tekstit muodostavat jo yksinään kuitenkin lähes puolet koko *Kansainvälisestä oppijansuomen korpuksesta*, mutta toisaalta myös siksi, koska B1-taitotasoa voidaan pitää eräänlaisena rajana aloittelevan ja jo itsenäisen peruskielitaidon saavuttaneen kielenoppijan välillä (rajauksesta tarkemmin luvussa 2.3.3). Korpusaineistoa lähestytään tutkimuksessa korpusvetoisesti, mikä tarkoittaa, että tarkemman tutkimuksen ja analyysin kohteeksi päätyvät kielenainekset nousevat esiin korpuksesta itsestään ilman, että tutkija valikoi niitä ennakolta (Tognini-Bonelli 2001: 84).

Kansainvälisen oppijansuomen korpuksen avulla pyrin siis tutkimuksellani selvittämään ennen kaikkea, mitkä ovat *Eurooppalaisen viitekehyksen* mukaan B1-kielitaitotasolle arvioitujen suomenoppijoiden teksteissä frekvensseiltään eli esiintymistaajuuksiltaan kaikkien tyypillisimmin tavattavat n-grammit ja millaisia ne ovat kielellisiltä elementeiltään. Käytän tutkimuksen toteutuksessa apunani internetistä vapaasti saatavaa *AntConc*-nimistä konkordanssiohjelmaa (Anthony 2019), jonka avulla pystytään luomaan listauksia eripituisista ja frekvensseiltään vaihtelevista n-grammeista käyttäjän asettamien hakuehtojen mukaisesti. Tutkimuskysymykseni ovat seuraavat:

1. Mitkä ovat *Kansainvälisen oppijansuomen korpuksen* B1-kielitaitotason frekventeimmät 3-, 4-, 5- ja 6-grammit?
    a. Millaisesta leksikosta ja kielellisistä rakenteista n-grammit muodostuvat?
    b. Mitä n-grammien perusteella voidaan päätellä B1-oppijansuomesta kielimuotona?
2. Miten erilaiset metodologiset valinnat ohjaavat n-grammeista saatavaa dataa ja sen tulkintaa?

Kuten ensimmäisestä tutkimuskysymyksestä voidaan lukea, rajasin n-grammeiksi tässä tutkimuksessa kolmesta, neljästä, viidestä ja kuudesta sanasta koostuvat sanaketjut (rajauksista tarkemmin luvussa 3). Ensimmäisen tutkimuskysymyksen kautta on tarkoitus kerätä aluksi etenkin kvantitatiivista eli määrällistä tietoa oppijansuomessa tavattavista n-grammeista. N-grammeja analysoidaan myös kvalitatiiviselta eli laadulliselta kantilta ensimmäisen tutkimuskysymyksen alakysymyksiin a. ja b. vastaamiseksi. Toinen tutkimuskysymys taas korostaa sitä, että aivan vastaava oppijansuomen n-grammitutkimusta ei ole vielä tehty, jolloin tässä tutkimuksessa samalla kokeillaan yhdenlaisia mahdollisia tulokulmia suomenoppijoiden n-

grammeihin. Tutkimuksen luvut 5 ja 6 ovat siis osaltaan vastaamassa tutkimuksen yhteen vetävän luvun 7 kanssa tähän kysymykseen.

Hypoteesinani tutkimukselleni on, että suomenoppijoiden taajimmin tuottamat n-grammit ovat fraasinomaisia rakenteita, jotka ovat etenkin niissä käytetyn sanaston puolesta paljolti motivoituneita siitä, minkälaisten tekstilajien ja tehtävänantojen yhteyksissä ne esiintyvät. Tutkittavien tekstien rajautuminen B1-kielitaitotasolle sekä n-grammien vaatimus taajasta toisteisuudesta saavat myös uskomaan, että n-grammeissa tuskin esiintyy juurikaan järin monimutkaista sanastoa tai kielellisiä ilmiöitä. Esimerkiksi verbien tempuksista eli aikamuodoista tyypillisimmin käytetyksi voisi suomenoppijoiden kielessä arvella preesensin. Verbien moduksista eli tapaluokista n-grammiaineistossa esiintyy todennäköisesti suurimmissa määrin indikatiivia sekä jonkin verran konditionaalia. N-grammien leksikon voisi ajatella rakentuvan pitkälti sellaisista sanoista, jotka kuuluvat natiivisuomessakin kaikkein taajimmin käytettyihin.

Tutkimus jatkaa oppijankielen sanakimppurakenteisiin keskittyvää tutkimusta, jota on tehty kansainvälisesti fraseologisessa tutkimusperinteessä jonkin verran (esim. Biber & Conrad 1999; Cortes 2004; 2008; 2012; Paquot 2013; 2014; Salazar 2014), mutta suomessa huomattavasti vähemmän. Tutkimus linkittyy kuitenkin yleisesti myös aiempaan oppijasuomen fraseologiseen tutkimukseen, jota nyt siis edelleen laajennetaan yhden aiemmin vähemmälle huomiolle jääneen fraseologisen yksikön osalta. Suomen oppiminen tarvitsee nähdäkseni lisää tutkimusta etenkin fraseologisesta näkökulmasta käsin, jotta tietoisuus fraseologian merkityksestä kielen oppimisen ja sen opetuksen piirissä kasvaisi. Tutkimustulosten myötä oppijansuomi kyetään myös mahdollisesti näkemään ainakin hitusen tuoreemmassa valossa.

Tutkimuksessa kartoitetaan ensin tutkimuksen teoreettista viitekehystä luvussa 2, jossa kuvataan kontekstuaalista semantiikkaa ja fraseologiaa kielentutkimuksen osa-alueina, määritellään spesifimmin n-grammin käsite ja siihen läheisesti liittyvät ilmiöt sekä esitellään aiempaa tutkimusta n-grammeista. Siinä pureudutaan myös oppijankielen määritelmään, sen merkittävimpiin piirteisiin ja sitä koskevaan tutkimukseen. Luvussa 3 luodaan katsaus sekä ylipäänsä korpusaineistoihin kielentutkimuksessa että tarkemmin *Kansainväliseen oppijansuomen korpukseen* ja käydään läpi tutkimuksen käyntiin sysäävä tutkimusmenetelmä eli korpusvetoinen tutkimus. Siinä myös rajataan tutkimusaineisto lopulliseen muotoonsa. Neljännessä luvussa vastataan tutkimuksen ensimmäiseen tutkimuskysymykseen esittämällä listaus lopullisen tutkimusaineiston eli ICLFI:n B1-kielitaitotason tekstien toistuvimmista n-grammeista.

Luvuissa 5 ja 6 kuvaillaan, kuinka n-grammilistauksesta tehdään laskelmia erilaisten kielenpiirteiden ja leksikaalisten elementtien edustumisten suhteen. Näistä saatavaa kvantitatiivista tietoa pohditaan luvuissa myös kvalitatiivisesta näkökulmasta sekä verrataan sitä

aiempaan tutkimustietoon erilaisista oppijansuomen aspekteista. Tässä yhteydessä vastataan alustavasti ensimmäisen tutkimuskysymyksen alakysymyksiin. Tutkimuksen viimeisessä luvussa vedetään yhteen tulokset, jotka tutkimuksesta saadaan, ja pohditaan, miten ne loppujen lopuksi valaisevat oppijansuomea kielimuotona. Aivan lopuksi esitetään vielä arviointia tutkimuksen onnistumisesta sekä muutamia potentiaalisia aihioita mahdollisille jatkotutkimuksille.

2 FRASEOLOGIA, N-GRAMMIT JA OPPIJANKIELEN TUTKIMUS

2.1 Kontekstuaalinen semantiikka ja fraseologia kielentutkimuksessa

2.1.1 Lähtökohdat ja tutkimuskohteet

Semantiikka eli merkitysoppi on kielitieteen alalaji, joka tutkii erilaisten merkkijärjestelmien, usein ihmiskielien, välittämiä ilmauksia ja niiden merkityksiä (esim. Kangasniemi 1997: 22; Kuiri 2012: 7). Semantiikka voidaan jaotella muun muassa *sanasemantiikkaan* eli *leksikaaliseen semantiikkaan*, joka tutkii nimensä mukaisesti lekseemien merkityksiä, ja *lausesemantiikkaan* eli *syntaktiseen semantiikkaan*, jossa kiinnostuksen kohteena on vastaavasti se, mitä lauseilla tarkoitetaan (Kangasniemi 1997: 22). *Kontekstuaalisella semantiikalla* taas viitataan analyysiin, jossa merkityksiä etsitään ennen kaikkea sanojen esiintymiskontekstien kautta. *Kontekstilla* voidaan tässä yhteydessä tarkoittaa joko ilmausta lähimmin ympäröivää tekstikontekstia tai laajempaa, koko sosiaalisen ympäristön huomioon ottavaa kontekstia. (Jantunen 2004: 7.) Kontekstin käsitettä lähelle osuu termi *koteksti*, joka on Sinclairin (1991: 172) mukaan kuitenkin kontekstia tarkkarajaisempi ja jolla viitataan nimenomaisesti sanaa tai fraasia likimmin ympäröiviin sanoihin. Korpustutkimuksessa on usein olennaista määritellä jo tutkimuksen alkuvaiheessa tutkittavan ilmauksen tarkastelukoteksti eli se tekstialue, jonka kautta ilmauksen kotekstuaalisia ominaisuuksia selvitetään (Jantunen 2004: 12). N-grammitutkimus on siinä määrin poikkeuksellista, että siinä sanayhtymät voidaan usein irrottaa kokonaan koteksteistaan ja konteksteistaan ja tutkia niitä pelkästään itsenäisinä yksikköinä. Kuitenkin esimerkiksi nyt käsillä olevassa tutkimuksessa etenkin rakenteellinen analyysi vaatii onnistuakseen paikoitellen myös n-grammien lähimmän tekstiympäristön tarkastelua ja tuntemista.

Kontekstuaalisessa semantiikassa korostetaan merkityksen ja muodon sekä sanaston ja kieliopin yhteyttä (Jantunen 2004: 7). Konventionaalisissa kieliopeissa nämä kielen osa-alueet on tupattu näkemään erillisinä (ks. esim. Kuiri 2012: 8–9), mutta esimerkiksi Sinclairin (1991: 102–104, 108) mukaan on mieletöntä pyrkiä erottamaan leksikkoa, syntaksia ja semantiikkaa toisistaan. Kontekstuaalisessa semantiikassa pyritäänkin abstraktien käsitteiden sijaan tutkimaan pikemminkin todellisia toistuvia kielenkäyttöyhteyksiä ja täten kuvaamaan niiden

usuaalista[7] merkitystä. Kontekstuaalisen semantiikan mukaan kielenkäytössä kunkin käytettävän ilmauksen valintaa ohjaavat ennen kaikkea ilmausten keskinäiset syntagmaattiset ja paradigmaattiset suhteet. (Jantunen 2004: 8–9.) Syntagmaattiset ja paradigmaattiset suhteet esitteli alun alkaen Saussure (2014) 1900-luvun alussa. *Syntagmaattiset suhteet* ilmenevät lineaarisella jatkumolla, *syntagmalla*, johon erilaiset kielen elementit, kuten foneemit ja sanat, järjestyvät jonoksi. Syntagma vaatii siis vähintään kahta peräkkäistä kielen yksikköä osakseen ja konkretisoituu sanatasolla etenkin kollokaatioina (ks. luku 2.2.1), kuten *hevoset hirnuivat*. (Karlsson 2008: 18, 232; Saussure 2014: 226.) Syntagmaattisiin suhteisiin lukeutuu siten myös tämän tutkimuksen ydinkäsite n-grammi.

Syntagmassa olevat elementit, esimerkiksi lekseemit, ovat taas *paradigmaattisessa suhteessa*[8] sellaisiin elementteihin, joiden kanssa ne ovat vaihdettavissa (Karlsson 2008: 18; Saussure 2014: 229–236). *Hevoset hirnuivat* -esimerkissä *hevosten* paikalle ei sovi semantiikan puolesta mikä tahansa muu olio, vaan kyseisen sanan kanssa paradigmaattisessa suhteessa ovat lähinnä muut hevossukuiset eläimet, kuten *poni*, tai vaikkapa *hevosen* synonyymit, kuten *hepo* ja *polle*. Synonymiaa voidaankin pitää yhtenä merkittävimmistä paradigmaattisista suhteista (Karlsson 2008: 219). Elementit, jotka ovat keskenään vaihtosuhteessa, muodostavat *paradigman* (mts. 18). Edellisessä esimerkissä *hirnua*-verbi olisi myös muutettavissa eri tempukseen, koska taivutusmuotojen sarjatkin edustavat omia paradigmojaan (ks. VISK § 53). Syntagmaattiset ja paradigmaattiset suhteet voidaan hahmottaa horisontaalisena ja vertikaalisena janana, jossa syntagmaattiset suhteet järjestyvät siis (länsimaisten) kirjoitettujen tekstien tapaan horisontaaliseen jatkumoon siinä missä paradigmaattiset suhteet taas voidaan mieltää vertikaalisiksi vaihtoehtojen sarjoiksi. Kirjoitettua tai puhuttua kieltä tutkittaessa tutkitaan nimenomaan syntagmaattisia suhteita, sillä paradigmat ilmaisevat ainoastaan niitä vaihtoehtoja, joita kunkin elementin tilalle olisi ollut kieltä tuotettaessa periaatteessa mahdollista valita. (Sinclair 2004: 168.)

Kontekstuaalisessa semantiikassa ollaan siis kiinnostuneita ennen kaikkea siitä, mitkä säännöt ja preferenssit ohjaavat yhden ilmauksen valintaa sen paradigmasta muiden ilmausten kanssa syntagmassa käytettäväksi (Jantunen 2004: 8). Kielioppiteorioiden suosiossa on etenkin aiemmin ollut hyvin pitkälti ilmausten paradigmaattinen ulottuvuus, jolloin on nähty, että teksti muodostuu itsenäisistä leksikaalisista yksiköistä, joita on yksikertaisesti aseteltu peräkkäin toistensa kanssa välittämättä suuremmin siitä, käytetäänkö kieltä todellisuudessa tähän tapaan

[7] *Usuaalinen merkitys* tarkoittaa sanan (tai ilmauksen) yleistä, tilanteesta riippumatonta merkitystä, jonka ihmiset käsittävät yhteisesti samalla tavalla. Tästä erotetaan *okkasionaalinen merkitys*, joka viittaa abstraktimpaan, puhujan tilanteessa sanalle antamaan merkitykseen, jonka tämä olettaa myös kuulijan sille antavan. (Kuiri 2012: 27.)
[8] Paradigmaattisista suhteista Saussure (2014) käyttää nimitystä *assosiatiiviset suhteet*.

vaiko ei (vapaan valinnan periaate, ks. luku 2.1.2). Tällöin kiinteämmät ja kielen käytön kannalta merkityksellisemmät useamman sanan yhdistelmät ovat jääneet vähemmälle tarkastelulle. Korpusaineistojen yleistymisen myötä myös eri pituisiin, toistuviin useampisanaisiin yksikköihin on kuitenkin alettu kiinnittää enemmän huomiota. (Sinclair 2004: 140.)

Juuriltaan kontekstuaalinen semantiikka juontaa etenkin J.R. Firthin ja John Sinclairin tunnetuksi tekemän niin sanotun brittiläisen koulukunnan tutkimusperinteeseen (Stubbs 1996: 22). Seuraavaksi esiteltävä fraseologia taas lähtee liikkeelle pitkälti kontekstuaalisen semantiikan näkemyksistä kielenkäytöstä ja sen suhteesta konteksteihin (Lehto 2018: 79).

Fraseologia on kielitieteen ala, joka tutkii kielen erilaisia vakiintuneita rakenteita ja toistuvasti keskenään yhteisesiintyviä kielenyksiköitä (esim. Cowie 2006; Gries 2008: 4, Jantunen 2009b: 360–361). Fraseologia esitetään usein leksikologian eli sanastontutkimuksen osa-alana, joka yksittäisten lekseemien sijaan tutkii kuitenkin ennemmin kielen *monisanaisia yksiköitä* (*multi-word units*) (Granger 2005: 165). Fraseologian keskiössä on ajatus siitä, että suurin osa ihmisten kielenkäytöstä pohjautuu eräänlaiseen kielen elementti- ja valmisrakenteisuuteen ja on täten fraseologista. Tämän vuoksi fraseologisessa tutkimuksessa painotetaan kontekstuaalisen semantiikan tapaan nimenomaan kielen käytön tutkimista. (Jantunen 2009b: 360.) Aiemmin fraseologian ymmärrettiin viittaavan lähinnä jähmettyneiden ilmausten, kuten idiomien, tutkimukseen, mutta sittemmin katsantokanto asiaan on laajentunut (Gries 2008: 3). Tähän ovat vaikuttaneet muun muassa huomiot, joita korpusaineistojen perusteella on tehty siitä, että suurella osalla monisanaisista yksiköistä ei ole olemassa mitään spesifiä jähmettynyttä tai kanonista muotoaan, vaan ne ovat enemmän tai vähemmän varioivia (Cowie 2006: 582). Fraseologialla voidaankin nykyään tarkoittaa pelkkien kaikkein kiteytyneimpien idiomien, fraasien ja rutiiniilmausten tutkimisen sijaan muidenkin vakiintuneiden leksikaalisten elementtirakenteiden tarkastelua (Granger 1998; Jantunen 2009b). Kielellisiin rakenteisiin, joita fraseologia tutkii, lukeutuvat jähmettyneimpien konstruktioiden osalta esimerkiksi juuri idiomit ja sanonnat ja vähemmän kiinteiden syntagmaattisten suhteiden puolelta taas esimerkiksi kollokaatiot, kolligaatiot[9] ja n-grammit (Granger & Paquot 2008: 27–28; Jantunen 2009b: 358–359). Fraseologian alalla terminologian puolesta osaltaan ongelmallista on ollut tällaisia rakenteita kuvaavien käsitteiden sekoittuminen keskenään, jolloin samanlaisista kielenyksiköistä on voitu eri yhteyksissä puhua eri termein ja päinvastoin toisistaan selkeästi poikkeavista suhteista yhdellä ja samalla käsitteellä (Granger & Paquot 2008: 28).

[9] *Kolligaatio* tarkoittaa sanan yhteyttä sen kanssa toistuvasti esiintyvän kieliopillisen kategorian kanssa; tähän lukeutuvat suomen kielessä muun muassa verbien rektiot (Stubbs 2007b: 178; Jantunen 2009b: 359).

Erilaisten termien määritelmien ohella samaten koko fraseologian ala on ajan saatossa saanut osakseen useampia vaihtoehtoisia määritelmiä (ks. Granger & Paquot 2008). Kaksi merkittävintä lähestymistapaa fraseologiaan ovat Itä-Euroopasta ja Venäjältä lähtöisin oleva vanhempi perinne sekä Sinclairin aloittama modernimpi lähestymistapa. Ensin mainittu traditio näkee juuri idiomaattisimmat kielenyksiköt keskeisimpinä fraseologisen tutkimuksen kannalta. Kyseinen traditio vakiinnutti fraseologian omaksi tieteenhaarakseen, loi alan pääkäsitteistön ja tarjosi kriteerejä fraseologisten yksikköjen erittelyyn. Jälkimmäinen perinne taas pohjautuu pitkälti korpusvetoisesti analysoitavaan dataan, jonka kautta on onnistuttu pääsemään käsiksi muunkinlaisiin kuin kaikkein jähmeimpiin, joka tapauksessa huomattavan yleisiin sanayhtymiin. Tässä perinteessä tärkeässä roolissa on Sinclairin (1991) esittämä idiomiperiaate (ks. luku 2.1.2) merkityksen muodostuksesta. (Granger 2005; Granger & Paquot 2008: 28–29.) Siinä missä aiemmassa fraseologisessa tutkimuksessa ilmauksia pyrittiin siis jäsentämään erilaisten kielellisten kriteerien mukaan, nykyään tavoitteena on pikemminkin korpusmenetelmin nostaa aineistosta esiin sellaisia leksikaalisia yhteisesiintymiä, jotka eivät välttämättä suoranaisesti istu mihinkään ennalta annettuihin kategorioihin (Granger 2005: 166). Tässä tutkielmassa tukeudutaan jälkimmäisen perinteen näkemykseen fraseologiasta. Tähän lähestymistapaan viitataan usein tilastollisena tai frekvenssiperustaisena lähestymistapana johtuen siinä hyödynnettävästä korpuksin kerrytettävästä kvantitatiivisesta datasta.

Fraseologia on avainroolissa merkitysten muodostumisessa. Lähes aina kieltä puhuttaessa tai kirjoitettaessa valitaan yksittäisten sanojen sijaan käytettäviksi pikemminkin erilaisia sanojen yhdistelmiä. (Warren 2011: 161–162.) Sanat saavatkin useimmiten merkityksensä ennen kaikkea niistä kielenkäyttötilanteista ja muista sanoista, joiden konteksteissa ne esiintyvät (N. Ellis 2008: 1), mikä on myös yksi kontekstuaalisen semantiikan ydinnäkemyksiä (ks. Firth 1968 [1957]). Sinclair (1991: 108) väittää jopa, että yksittäisillä sanoilla on harvoin varsinaisesti mitään suurempaa merkitystä sellaisinaan vaan ainoastaan palasina useammasta sanasta koostuvia komponentteja, jotka yhdessä muodostavat tekstin. Suurin osa teksteistä taas koostuu yleisistä sanoista, jotka esiintyvät yleisissä malleissa tai näiden mallien variaatioissa. Kielen fraseologiseksi hahmottavan ajattelutavan keskiössä onkin ajatus eräänlaisesta *kaavamaisesta kielestä* tai *kielenkäytöstä* (*formulaic language*) (ks. Wray 2002a; 2008). Biberin ym. (1999: 996) mukaan etenkin keskustelu on erityisen kaavamaista ja koostuu hyvin pitkälti vakiintuneista sanakimppurakenteista. Korpustutkimusten yhteydessä onkin esitetty arvioita siitä, että jopa noin 80 prosenttia tavanomaisesta kielenkäytöstä olisi jollain tapaa kaavamaista (ks. Altenberg 1998).

Kaavamaista kieltä ja sen lähi-ilmiöitä on fraseologian keskeisimpien käsitteiden tapaan kuvattu lukuisin erilaisin termein, jotka asettuvat enemmän tai vähemmän päällekkäin toistensa kanssa. Wray (2002a: 8–9) listaa näihin monien muiden muassa termit *jähmettyneet fraasit* (*frozen phrases*), *rutiiniformulat* (*routine formulae*) ja *valmiit ilmaukset* (*ready-made expressions*). Käsitteen laaja-alaisuuden vuoksi Wray (mts. 9) suosii ilmiöstä selkeämmin määriteltävissä olevaa, kaavamaista kieltä tarkempirajaisempaa nimitystä *kaavamaiset jaksot*[10] (*formulaic sequences*). Tällaisten jaksojen hän määrittelee tarkoittavan seuraavaa:

> a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (Wray 2002a: 9).

Määritelmä tiivistää ajatuksen siitä, että kielellisiä ilmauksia varastoidaan muistiin ja otetaan sieltä käytettäviksi kokonaisina sanaketjuina, ei yksittäisinä lekseemeinä. Useimmat natiivit kielenpuhujat pystyvätkin tunnistamaan keskenään samoja sanaketjuja, mikä johtuu yksinkertaisesti siitä, että he jakavat keskenään yhteisen kielen (Wray 2008: 11). Kaavaimaiset jaksot ovat täten myös eräänlainen yläkäsite lukuisille erilaisille sanayhtymille ja kielellisille ilmaisukeinoille, ja niiden hallitsemista kaikkinensa voi kielenoppimisen kontekstissa pitää yhtenä merkittävimmistä astinlaudoista natiivinkaltaista kieltä tavoiteltaessa (ks. esim. Wray 2000; 2002a; Paquot & Granger 2012). Luvussa 2.3.2 pureudutaan vielä hieman tarkemmin kielen kaavamaisuuden ja kielen fraseologisuuden oppimisen yhteyteen.

Kiinnostus fraseologiaa kohtaan on kasvanut viimeisten vuosikymmenten saatossa (Gries 2008: 3), ja fraseologinen tutkimus onkin lisääntynyt huomattavissa määrin aina 1980-luvulta lähtien (N. Ellis 2008: 4). Tähän lienee vaikuttanut etenkin se, että siinä missä fraseologiaa pidettiin varsinkin aiemmin lähinnä yhtenä sanastontutkimuksen haarana, on se tällä vuosituhannella saanut jalansijaa myös omana kielitieteen osa-alueenaan (Granger & Meunier 2008: ix). Korpusaineistojen huomattava yleistyminen on samaten osaltaan vauhdittanut fraseologisen tutkimuksen kasvua kielitieteen parissa (Jantunen 2009b: 360–361). Fraseologian alalla tutkittavista sanojen syntagmaattisista suhteista etenkin kollokaatiot ovat olleet suosittuja tutkimuksen kohteita varsinkin kansainvälisessä (ks. esim. Stubbs 1995; Sinclair, Jones & Daley 2004; Williams 2008; Mollin 2009) mutta myös suomalaisessa tutkimuksessa (ks. esim. Jantunen 2001; Pirkola 2016; Kangas 2018).

[10] Ainakin Mustonen (2015: 55) on käyttänyt ilmiöstä myös suomennosta *idiomaattiset ilmaukset*.

2.1.2 Sanojen kontekstuaalisen valinnan periaatteet ja leksikaalisen primingin teoria

Kontekstuaalisen semantiikan kuten osaltaan myös fraseologian ytimessä on siis se, millaiset preferenssit ohjaavat sanojen tai ilmausten valintaa niiden koko paradigmasta käytettäväksi toistensa kanssa syntagmassa. Sinclair (1991: 109–110) on esittänyt kaksi tunnettua tulkintaperiaatetta tälle niin sanotulle sanojen kontekstuaaliselle valinnalle. Ensimmäistä tulkintaa sanojen valintapreferenssistä hän nimittään *vapaan valinnan periaatteeksi* (*open choice principle*), joka tarkoittaa sanojen rajoittamatonta yhdistelyä toistensa kanssa, kunhan vain lopputulemana oleva ilmaus on kieliopillisesti pätevä. Vapaan valinnan periaatteeseen nojaamalla sanojen välinen syntagma muodostetaan siis täyttämällä kielellisiä aukkoja millä tahansa sanoilla ja sananmuodoilla, jotka aukkoihin syntaksinsa puolesta sattuvat sopimaankaan (ns. *slot filler* -tyyppinen ratkaisu). Liki kaikkien kieliopeista voisi nähdä perustuvan vähintäänkin implisiittisesti tälle periaatteelle, sillä niissä asetetaan useimmiten rakennekuvaus etusijalle merkitykseen nähden ja muutoinkin siitä erilleen (ks. Sinclair 2004: 166–167). Vapaan valinnan periaate on verrattavissa Chomskyn (esim. 2006: 88) näkemykseen kielen luovasta puolesta, jossa natiivin kielenkäyttäjän ajatellaan kykenevän sekä ymmärtämään että käyttämään rajatonta määrää kielellisiä ilmauksia (Singleton, Leśniewska & Witalisz 2007: 210). Kieltä ei kuitenkaan käytännössä katsoen koskaan tosielämässä toteuteta pelkkään vapaan valinnan periaatteeseen nojaten, eikä kielenkäyttö välttämättä ole aivan niin luovaa kuin Chomsky sen hahmottaa olevan.

Toisena vaihtoehtona vapaan valinnan periaatteelle Sinclair (1991: 110) esittääkin *idiomiperiaatteen* (*idiom principle*). Idiomiperiaate on hänen mukaansa se valintaperiaate, jonka pohjalta kieltä todellisuudessa rakennetaan. Idiomiperiaatteessa olennaisinta on ajatus siitä, että kielenkäyttö ei perustu pelkkiin satunnaisuuksiin, sillä kielen jaottelu on usein yhteydessä maailman luonteeseen, jolloin esimerkiksi tiettyjä rekistereitä käytettään vain tietyissä tilanteissa. Tämä jo osaltaan useimmiten rajoittaa huomattavasti kussakin tilanteessa käytettävissä olevia sanoja ja niiden paradigmaattisia vaihtoehtoja. Koska tietyt sanat tuppaavat vetämään herkästi muun muassa tiettyjä kollokaatteja puoleensa ja esiintymään alttiimmin tietyissä semanttisissa ympäristöissä (Sinclair 1991: 112), suositaan näiden sanojen valintaa niiden paradigmasta muiden – kuitenkin näennäisesti mahdollisten – vaihtoehtojen sijaan. Olemme toisin sanoen taipuvaisempia puhumaan esimerkiksi *vahvasta* tai *laihasta* kuin *voimakkaasta* tai *heikosta kahvista*. Kielenkäyttäjällä onkin hallussaan ja hyödynnettävissään lukemattomia useampisanaisia rakenteita ja fraaseja, jotka perustuvat ainoastaan yksittäisille kontekstuaaliset suhteet huomioon ottaville valinnoille, vaikkakin ne rakentuvat monesta sanasta. Tällaiset rakenteet ohjaavat vahvasti kielenkäyttöä. (Mts. 109–110.) Idiomiperiaatetta hyödyntämällä

kyetäänkin vastaamaan muun muassa puhetilanteissa vaadittavaan ripeään kielenkäyttöön, kun jokaista sanavalintaa ei tarvitse punnita omana erillisenä yksikkönään (Singleton ym. 2007: 210). Idiomiperiaate ei nimestään huolimatta siis liity ainoastaan idiomeihin ja niiden käyttöön kielessä vaan myös yksittäisten sanojen kotekstuaalisiin suhteisiin, joihin lukeutuvat kollokaatioiden lisäksi muun muassa n-grammit (Jantunen 2009b: 358). Huomionarvoista on myös, että sanojen kontekstuaalinen valinta ei lopulta jakaudu mustavalkoisesti joko kokonaan vapaaseen valintaan tai täysin rajoittuneisiin myötäesiintymisiin, vaan ääripäiden väliin jää laaja alue, jossa lievemmätkin sanojen yhdistymistä ohjaavat niin kutsutut *myötäesiintymäpreferenssit* ovat mahdollisia (Jantunen 2004: 15).

Idiomiperiaatteen ilmiönä voi nähdä kuuluvan hyvin vahvasti yhteen edellä esiteltyjen kielen kaavamaisten jaksojen kanssa (ks. Wray 2002a: 12–15; 2002b: 119), sillä molemmissa painotetaan ilmausten poimimista muistista käyttöön yhtä sanaa laajempina, analysoimattomina yksikköinä. Kielenkäyttäjällä onkin tuhansittain tällaisia valmiita rakenteita varastoituneena käytettäväkseen hänen niin sanotussa *mentaalisessa leksikossaan* (*mental lexicon*) (Gries 2008: 17–18), joka voidaan määritellä löyhästi ”kognitiiviseksi järjestelmäksi, joka mahdollistaa kyvyn tietoiseen ja tiedostamattomaan leksikaaliseen toimintaan” (Jarema & Libben 2007: 2). Tuon mentaalisen leksikon käyttöä voi nähdä osaltaan kielenkäyttötilanteissa ohjaavan – idiomiperiaatteen rinnalla – muun muassa sen, miten vastaanottaja ja tämän tarpeet ja odotukset pyritään ottamaan kulloisessakin kontekstissa huomioon (ks. Wray 2008: 20–21). Yhdessä merkittävimmistä rooleista mentaalisen leksikon valjastamisessa tarkoituksenmukaiseen käyttöön toimii myös niin kutsuttu leksikaalinen priming.

Michael Hoeyn (2005) kehittämä, psykolingvistiikkaan nojautuva *leksikaalisen primingin*[11] (*lexical priming*) käsite ja teoria limittyvät tiiviisti yhteen idiomiperiaatteen kanssa (ks. Hoey 2009). Hoeyn (2005: 1) teoria pyrkii muun muassa Sinclairilta (2004: 164–176) tutun ajattelutavan tapaan kääntämään keskenään ympäri leksikon ja kieliopin perinteiset roolit argumentoimalla, että kielioppi on tulosta kompleksisesta ja systemaattisesti rakennetusta leksikosta eikä päinvastoin, kuten asian on tyypillisesti nähty olevan. Hoeyn (2005: 8) mukaan opittaessa sanoja niiden käyttöyhteyksissään puheen ja kirjoituksen kautta tullaan samalla oppineeksi myös lukuisia muita sellaisia sanoja, jotka kuuluvat toistensa kanssa jaettuihin

[11] *Priming* on ennen kaikkea psykologian käsite, ja se voidaan suomentaa ainakin *virittäytymiseksi* tai *aktivoitumiseksi*. Psykolingvistiikan alalla primingilla viitataan ennakointiin, virittäytymiseen tai reagointiin, joka perustuu johonkin ärsykkeeseen tai ennakkotietoon ja jota testataan usein esimerkiksi sanaparitestein. Leksikaalinen priming on lähellä tätä, mutta pidettävä psykolingvistiikan käsitteestä kuitenkin erillään leksikaalisen primingin kuvatessa nimenomaan kielen fraseologista luonnetta eikä niinkään reaktiota koetilanteessa. (Ks. Hoey 2005: 7–8; Jantunen & Brunni 2012: 73–74; Pace-Sigge 2013: 152–153.)

käyttökonteksteihin. Tällöin sanoista kukin saa oman ’priminginsa’, eli ikään kuin virittyy toimimaan odotuksenmukaisesti tietyissä käyttöyhteyksissä. Hoeyn (mts. 13) mukaan kaikki sanat ovat virittäytyneet tämänkaltaisesti, jolloin jokaisella sanalla on olemassa muun muassa omat kollokaationsa, semanttiset assosiaationsa, pragmaattiset funktionsa ja kieliopilliset toimintonsa. Sanojen ja niihin liittyvän primingin toistuvan kohtaamisen myötä olemme myös itse taipuvaisempia toistamaan niitä samoissa konteksteissa ja genreissä samoin kieliopillisin mallein sekä pragmaattisin ja tekstuaalisin keinoin, kuin mihin olemme ne omien kokemustemme perusteella linkittäneet (Hoey 2009: 36). Tähän lukeutuu myös tieto siitä, että tietyntyyppisiä sanojen yhdistelmiä käytetään vain tietyntyyppisissä teksteissä (Hoey 2004: 23).

Leksikaalisessa primingissa sanat eivät lataudu ainoastaan niiden käyttökonteksteilla ja niiden yhteydessä usein tavattavilla muilla sanoilla vaan myös näiden muiden sanojen vastaavilla konteksteilla ja koteksteilla (Hoey 2005: 8). Pelkän kollokaation lisäksi primingiin liittyvätkin vahvasti myös muut niin sanotut fraseologiset yksiköt, kuten semanttinen preferenssi ja semanttinen prosodia[12] (mts. 22–24). Leksikaalinen priming onkin kaikkinensa olennainen osa fraseologiaa. Jantunen ja Brunni (2012) tiivistävät leksikaalisen primingin ja fraseologian yhteyden seuraavasti:

> Priming liittyy siis laajasti fraseologiaan: ilmauksilla on odotuksenmukaiset käyttöyhteytensä alkaen tyypillisistä kerasanoista eli kollokaateista aina tyylillisiin ja tekstuaalisiin preferensseihin asti. Käyttöyhteyksien omaksumisessa kertautumisella on erittäin suuri merkitys. Samalla tavalla kuin sanojenkin omaksuminen perustuu toistumiseen, myös kontekstuaalisten suhteiden hallintaan tarvitaan prosessi, jossa tietyt rakenteet toistuvat taajaan. (Jantunen & Brunni 2012: 74–75.)

Siinä missä korpustutkimusten kautta kyetään selvittämään sanojen keskinäistä esiintymistä, on leksikaalinen priming tekijä, joka selittää, *miksi* juuri tietyt sanat ovat taipuvaisia esiintymään keskenään (Pace-Sigge 2013: 151). Toisaalta korpusdata ei sekään kykene osoittamaan ihmisten yksilöllisiä primingeja vaan ainoastaan datan tuottaneen joukon todennäköisiä sanojen virittymissuhteita, sillä korpukset eivät edusta kenenkään yksilön henkilökohtaista kokemusta kielestä (Hoey 2004: 24; 2005: 14–15). Hoey (2007: 9) painottaakin, että priming on kaikesta huolimatta ensisijaisesti ihmisen eikä niinkään sanan ominaisuus. Vaikka voidaankin ajatella, että tietty sana suosii priminginsa osalta esimerkiksi tiettyjä kollokaatteja, on ilmiössä kuitenkin lopulta kyse pikemminkin siitä, että suurin osa kielenpuhujista jakaa keskenään saman kollokationaalisen primingin kyseiselle sanalle. Kaikkien ihmisten primingit ovat silti

[12] *Semanttisesta preferenssistä* ja *semanttisesta prosodiasta* ensin mainitulla tarkoitetaan sanan ja sen kotekstissa esiintyvän merkityspiirteen yhteisesiintymää; esimerkiksi OIKEIN-astemäärite preferoi ’hyvyyttä’ merkitseviä sanoja. Jälkimmäinen taas viittaa abstraktimpaan sanojen myötäesiintymään, diskurssifunktioon, joka ilmentää ennen kaikkea puhujan asenteita. (Sinclair 2004: 142; Stubbs 2007b: 178; Jantunen 2009b: 359–360.)

jollain tavalla erilaisia, sillä jokaisella kielenkäyttäjällä on taustallaan omat henkilökohtaiset kielelliset kokemuksensa, jotka ovat osaltaan muovanneet tämän primingeja (vrt. Wray 2008: 11). Priming ei myöskään ole muuttumatonta: joka kerta kun sanaa tai sanayhtymää käytetään tai se kohdataan uudestaan, siihen sitoutunut priming joko vahvistuu tai heikkenee riippuen siitä, tavataanko sana esimerkiksi tutussa tai vieraassa kontekstissa tai päätetäänkö sitä käyttää tietoisesti tavalla, joka syrjäyttää sen senhetkisen primingin (Hoey 2004: 23–24; 2005: 9). Hoey (2005: 11) väittääkin, että mieliimme – toisin sanoen mentaalisiin leksikkoihimme – on koodattu korpuslingvistiikasta tuttujen konkordanssien (ks. luku 3.2.2) tapaan jokaiselle tuntemallemme sanalle oma niin sanottu *mentaalinen konkordanssinsa*, johon kaikki edellä mainittu informaatio sanojen kanssa ajan saatossa kohdatuista käyttöyhteyksistä on sisällytetty. Tämän tutkimuksen yhteydessä leksikaalisella primingilla on merkitystä etenkin kielenoppimisen yhteydessä. Tätä sivutaan tarkemmin luvussa 2.3.2.

2.2 Toistuvat useampisanaiset jaksot eli n-grammit

2.2.1 Fraseologiset yksiköt ja sanojen syntagmaattinen myötäesiintyminen

Kuten jo edellä todettiin, fraseologisen näkökulman mukaan kielenkäyttö koostuu mitä suuremmilta osin erilaisista kieleen vakiintuneista rakenteista (ks. esim. Gries 2008; Jantunen 2009b; Granger 2011; Warren 2011). Tällaisista rakenteista voidaan käyttää myös yhteisnimitystä *fraseologiset yksiköt*, joilla viitataan yhtä sanaa laajempiin elementteihin, joiden osat ovat syntagmaattisissa suhteissa toistensa kanssa. Näitä yksiköitä ovat kollokaatio, n-grammi, kolligaatio, semanttinen preferenssi ja semanttinen prosodia. (Jantunen 2009b: 358–360.) N-grammi on näistä tämän tutkimuksen kannalta merkittävin, mutta sitä hyvin likelle osuu myös yksiköistä tunnetuin eli kollokaatio, joka on täten myös syytä esitellä lyhykäisesti tässä yhteydessä. *Kollokaatio* on käsitteenä peräisin Firthiltä (1968 [1957]), jonka paljon lainattu toteamus ”You shall know a word by the company it keeps” (mts. 179) siihen usein liitetään. *Kollokaatio*-termiä käytetään varsin laajasti kielitieteessä, ja sille on olemassa lukuisia erilaisia määritelmiä, joista useimpia yhdistää ainoastaan näkemys siitä, että käsitteellä tarkoitetaan sanojen jonkinlaista syntagmaattista suhdetta toisiinsa (Nesselhauf 2005: 11). Pääsääntöisesti kollokaatioiden ymmärretään kuitenkin olevan kahdesta toistuvasti keskenään yhdessä esiintyvästä sanasta koostuvia, leksikaalisesti ja/tai syntaktisesti kiinteitä sanayhtymiä (Sinclair 1991: 170; Hunston

2002: 12; Nesselhauf 2005: 1; Jantunen 2009b: 358). Tällaisia ovat esimerkiksi ilmaukset *vahva kahvi*, *pieni vauva* ja *kissa naukuu*. Lähtökohtaisesti samaa merkitsevät sanat voidaan useimmiten erottaa toisistaan juuri niiden kollokaatioiden avulla: esimerkiksi englannin *pientä* kuvaavista sanoista *little* ja *small* ensin mainittu tapaa saada kollokaateikseen muun muassa substantiivit *baby* ja *thing*, kun taas jälkimmäinen kollokoi herkemmin esimerkiksi sanojen *print* ja *world* kanssa (Biber & Conrad 1999: 183).

Kollokaatiot voidaan todentaa ja määritellä joko tilastolliselta, assosiatiiviselta tai tekstuaaliselta kantilta[13] (Jantunen 2001: 173; Nesselhauf 2005: 11–13). Tilastollista näkemystä on jalostanut muun muassa Sinclair (1991). Siinä kollokaatioiksi mielletään ainoastaan tilastollisin testein todetut merkitsevät sanojen yhteisesiintymät (mts. 106). Tämänkaltainen kollokaatioiden selvittäminen vaatii useimmiten tukeutumista korpusaineistoihin (Karlsson 2008: 232–233). Kollokaatio on kuitenkin leksikaalisen primingin tavoin jossain määrin myös psykologinen ilmiö: usein yhdessä esiintyvät sanat eivät ole varastoituneet mieliimme erillisinä vaan nimenomaan usean sanan kokonaisuuksina (Hoey 2005: 7). Tästä taas voidaan nähdä yhteys mentaaliseen leksikkoon sekä kaavamaisiin jaksoihin, joita luonnollisesti muodostuu mieliimme etenkin kollokaation myötä. Tällöin kollokaatiot kuvaavat edellä esitettyä lekseemien kontekstuaalisen valinnan idiomiperiaatetta (Sinclair 1991: 115). Näkökulmaa kollokaateista merkityksensä vuoksi mieliimme varastoituneina rakenteina pidetäänkin assosiatiivisena. Tekstuaalisesta näkökulmasta käsin tarkasteltuna kollokaatio taas tarkoittaa mitä tahansa kahta tai useampaa toistensa läheisyydessä esiintyvää sanaa (mts. 170). Toisin sanoen tekstuaalisen näkökannan mukaan jokainen noodista eli korpuksesta haetusta sanasta tietyllä etäisyydellä oleva sana voi olla sen kollokaatti riippumatta siitä, kuinka tiheästi se hakusanan kanssa yhdessä esiintyy (Jantunen 2001: 173). Noodin kanssa kollokoivien sanojen ei siis tarvitse esiintyä aivan välittömässä yhteydessä noodin kanssa, mutta yleensä kollokaatiot suosivat kuitenkin joitakin etäisyyksiä toisia enemmän (Gries 2008: 16).

Kollokaatiolla on siis etenkin assosiatiivisesta näkökulmasta saumaton yhteys edellä esitettyyn Hoeyn (2005) leksikaalisen primingin teoriaan. Hoeyn (2007: 7–8) mukaan sanoja tai sanayhtymiä kohdattaessa niiden esiintymiskonteksteja taltioidaan mieleen alitajuisesti. Tällöin tietoisuuteen kerrytetään samalla koko ajan lisää sanojen kollokaatteja, joita sitten myös päädytään itse toistamaan sanoja käytettäessä. Sanojen kollokationaalisen primingin myötä rakentuu mieliimme myös pitempiä monisanaisia yksiköitä, kun yksi sana kollokoi toisen ja toinen

[13] Tilastollisesta näkökulmasta on käytetty myös ainakin nimitystä *merkittävät myötäesiintymät* (*significant co-occurences*) ja tekstuaalisesta näkökulmasta *satunnaiset myötäesiintymät* (*casual co-occurences*) (ks. Paquot 2007: 127).

taas kolmannen kanssa ja niin edelleen. Tällöin alkuperäinen sana latautuu samalla kunkin rakenteeseen uutena mukaan tulevan lekseemin esiintymiskonteksteilla ja -koteksteilla. Kuten Sinclair (2004: 136) toteaa, jokainen leksikaalinen yksikkö sekä toimii omillaan että on toisaalta myös komponentti muiden yksikköjen valinnalle. Hoey (2005: 8) käyttää ilmiöstä, jossa yhden primingin tuotos päätyy seuraavan primingin kohteeksi, nimitystä *nesting*. Tästä hän antaa kaksi esimerkkiä: ensimmäisessä sana *winter* kollokoi preposition *in* kanssa, kun taas näistä muodostuvan sanayhtymän *in winter* primingiin lukeutuu verbi *BE* (Hoey 2005: 10–11; 2007: 8). Toinen esimerkki koskee sanaa *word*, joka kollokoi verbin *say* kanssa. Ilmaus *say a word* taas kollokoi sanan *against* kanssa, siinä missä *say a word against* on virittynyt kollokoimaan supistuman *won't* kanssa. (Hoey 2005: 11.) Tällä tavoin mieliimme muodostuu alati piteneviä monisanaisia kokonaisuuksia, joita voidaan pitää eräänlaisina potentiaalisina leksikaalisina sanakimppuina, toisin sanoen n-grammeina.

2.2.2 N-grammin määritelmä ja leimallisimmat piirteet

Erilaisia sanojen syntagmaattisista suhteista koostuvia monisanaisia ilmauksia voidaan erotella toisistaan muun muassa niiden idiomaattisuuden mukaan, jolloin kaikkien kiteytyneimpiä rakenteita edustavat varsinaiset idiomit, jotka ovat verrattain vaihtelemattomia ilmauksia. Kollokaatiot sekä tässä luvussa esiteltävät n-grammit taas perustuvat useimmiten pikemminkin tilastollisesti tai frekvenssiperustaisesti todennettaviin sanojen yhteisesiintymiin kuin näkemykseen ilmausten kiinteydestä. (Biber & Conrad 1999: 183; Biber ym. 1999: 988; ks. myös Nesselhauf 2005: 14–15.)

N-grammit (*n-grams*), joista käytetään myös useita muita nimityksiä, kuten *klusterit*[14] (*clusters*) ja *leksikaaliset kimput* tai *köntät*[15] (*lexical bundles*; *lexical chunks*), määritellään kieliaineistossa taajaan toistuviksi peräkkäisten sanojen ketjuiksi, jotka koostuvat *n*-määrästä, yleensä kahdesta tai useammasta, sanoja (Biber ym. 1999: 990; Granger & Paquot 2008: 38–39; Jantunen 2009b: 359; Warren 2011: 154; McEnery & Hardie 2012: 110). Eri kielentutkijat

[14] Vaikka klustereita käytetäänkin tutkimuskirjallisuudessa usein n-grammien synonyymina, ainakin Mahlberg (2013) peräänkuuluttaa käsitteiden erottelua toisistaan. Tätä hän perustelee muun muassa sillä, että siinä missä n-grammien määritelmään kuuluu vaatimus rakenteiden jakaantumisesta useamman tekstien välille, samaa ei ole klustereilla (Mahlberg 2013: 51; ks. myös Biber & Conrad 1999: 184–185). Klusteri soveltuikin ilmiönä paremmin Mahlbergin (2013) tutkimukseen, joka kohdistui Dickensin fiktioon, sillä tutkimuksen tähtäimessä oli kokonaismäärältään suhteellisen pieni otos itsenäisiä tekstejä, tässä tapauksessa Dickensin kirjoittamia romaaneja (ks. mts. 42–61).

[15] Suomenkielisessä tutkimuskirjallisuudessa suomennoksen *leksikaaliset kimput* (myös *sanakimput*) on aiemmin esittänyt ainakin Ivaska (2014a; 2015); *könttä*-termiä *chunksin* suomennoksena on taas käyttänyt muun muassa Reiman (2011).

ovat käyttäneet vuosien saatossa sekalaisesti vaihtelevia termejä viitatessaan pitkälti toisiaan vastaaviin sanakimppurakenteisiin. Stubbsin (2007a: 90) mukaan ilmiölle ei ole olemassakaan yhtä vakiintunutta käsitettä. Tutkijoista *n-grammeihin* viittaavat esimerkiksi Banerjee ja Pedersen (2003), Stubbs (2007a; 2007b; 2009) sekä Ivaska (2014a; 2015), siinä missä *leksikaalisten kimppujen* (*lexical bundles*) käsitettä suosivat muun muassa Biber ym. (1999), Cortes (2004) ja Granger (2014). Termiin *klusteri* taas tukeutuvat esimerkiksi Scott ja Tribble (2006), Mahlberg (2013) sekä Lehto (2018). Myös ainakin *toistuvia sanayhtymiä* (*recurrent word combinations* tai *recurrent word sequences*) ja *monisanaisia rakenteita*/*yksiköitä* (*multi-word constructions*/*units*) on käytetty samaisen ilmiön nimityksenä (ks. esim. Altenberg 1998; Liu 2012; J. Ebeling, S. Ebeling & Hasselgård 2013; Paquot 2008; 2014). *Klusteri*-termi on tutkimuksen kannalta siinä mielessä ongelmallinen, ettei se käyttöalaltaan rajaudu yksinomaan kielitieteeseen, vaan sitä tavataan myös muun muassa kemian (Scottin & Tribblen 2006: 32 mukaan Ball 2005: 30), tietojenkäsittelytieteen (ks. esim. Mueller 2003) ja musiikin (ks. esim. Cowell 1996 [1930]) aloilla. Myös tutkimusmenetelmästä käytetty nimitys *klusterianalyysi* on käsite, joka ulottuu lukuisille tieteenaloille. Sen perusteet ovat ennen kaikkea matematiikan, tilastotieteen ja tietojenkäsittelytieteen konsepteissa, ja se on näillä aloilla usein varsin teknistä (Moisl 2015: 4). Käsillä olevan tutkimuksen puitteissa edellä mainituista käsitteistä termi *n-grammi* vaikuttaa osuvimmalta, minkä vuoksi juuri se on valittu tämän tutkimuksen ydintermiksi. Esimerkiksi termeille *lexical bundles* tai *recurrent word combinations/sequences* ei ole olemassa vakiintuneita suomennoksia, kun taas *klusteri* ei ole käsitteenä yhtä tieteenalaspesifi kuin *n-grammi*, kuten edellä todettiin.

N-grammit ovat kielessä hyvin yleisiä, ja niitä voidaan pitää puhutun ja kirjoitetun diskurssin rakennepalikoina (Biber & Conrad: 1999: 188). N-grammeiksi luokiteltavien sanayhtymien ei tarvitsekaan olla mitenkään idiomaattisia, kieliopillisesti motivoituja tai muodostaa syntaktisesti tai semanttisesti kokonaisia rakenteita (Biber & Conrad 1999: 183; Biber ym. 1999: 990; Jantunen 2009b: 359). Englannin kielessä tyypillisiä keskustelujen yhteydessä ilmeneviä n-grammeja ovat esimerkiksi hyvin tavanomaiset sananketjut *do you know*, *I was going to*, *do you want to go* ja *I don't know what to do*, kun taas akateemisissa teksteissä saattavat sen sijaan toistua muun muassa ilmaukset *in order to*, *as a result of*, *on the other hand* ja *as shown in the figure* (Biber ym. 1999: 994; Cortes 2015: 198). Kieliaineistossa toistuvasti esiintyvän sanaketjun osumien täytyy jakautua useamman eri tekstin välille, jotta sitä voidaan pitää n-grammina. Tällä kriteerillä on tarkoituksena välttää se, etteivät aineiston sanaketjut edustaisi ainoastaan yhden puhujan tai kirjoittajan käyttämän kielen yksilöllisiä piirteitä. (Biber ym. 1999: 992–993.)

Biberin ym. (1999) voidaan katsoa olleen ensimmäisiä, jotka määrittelivät kokonaisvaltaisesti n-grammin käsitteen (nimityksellä *lexical bundles*) kieliopissaan *Longman grammar of spoken and written English*, joka pohjautuu laajaan korpusaineistoon (Salazar 2014: 13; Cortes 2015: 203). Tätä aiemminkin tutkimusta toistuvista sanaketjuista oli kuitenkin jo tehty; Biberin, Conradin ja Cortesin (2004: 373) mukaan ensimmäinen englannin kielestä sanakimppurakenteita hyödyntävää kartoitusta tehnyt oli Altenberg (1998), joka selvitti *London Lund* -korpuksen avulla puhutun englannin fraseologiaa. Toisaalta jonkinlaista kiinteisiin sanayhtymiin keskittyvää tutkimusta on tehty jo vähintäänkin 1950-luvulta asti; erona aiemman ja nykypäivää lähestyvän tutkimuksen välillä on kuitenkin ensin mainitun painottuminen pikemminkin intuitioon kuin empiiriseen tutkimukseen (Cortes 2008: 43–44).

N-grammit eivät useinkaan edusta rakenteeltaan varsinaisesti kokonaisia kielen konstruktioita. Esimerkiksi Biberin ym. (1999: 995) korpushavaintojen perusteella ainoastaan 15 prosenttia keskusteluissa ja alle 5 prosenttia akateemisissa teksteissä ilmenevistä n-grammeista on täydellisiä rakenneyksiköitä. N-grammeja voidaankin ryhmitellä erilaisiin luokkiin muun muassa sen perusteella, minkälaisia rakenteita niihin sisältyy. Esimerkiksi Biber ym. (1999: 1000–1029) luokittelevat englannin kielen keskustelujen ja akateemisten tekstien 4-, 5- ja 6-grammeja erilaisiin kategorioihin sen perusteella, millaisista elementeistä ne rakentuvat. Heidän korpushavaintojensa perusteella akateemisten tekstien n-grammit ovat rakenteeltaan useimmiten nomini- tai prepositiolausekkeita ja funktioiltaan muun muassa fyysisten paikkojen, kokojen ja määrien kuvaajia (esim. *the total number of the*, *the shape of the*), erilaisten prosessien ilmentäjiä (*the development of the*) ja abstraktien ominaisuuksien identifioijia (*the nature of the*). Keskustelun n-grammeihin taas lukeutuu useimmiten persoona- tai demonstratiivipronomini subjektin ominaisuudessa sekä pääverbi, joka on monesti kopula (*it's going to be*, *and there was a*, *I don't know what*, *I tell you what*). Paquot (2014) mainitsee erilaisiksi n-grammien kategorioiksi esimerkiksi kieliopillisesti täydelliset ja epätäydelliset n-grammit, erilaiset lausekkeet ja osat fraaseista. Hän antaa kieliopillisesti täydellisistä n-grammeista esimerkit *by contrast* ja *on the other hand* ja epätäydellisistä taas *the nature of the* ja *is based on the*. Lauseke-esimerkiksi hän antaa ilmauksen *I don't know what* ja fraasin osaksi *the use of*. (Paquot 2014: 216.)

Erilaisiksi n-grammien kategorioiksi voidaan laskea myös kuuluviksi esimerkiksi Nekrasovan (2009: 468) mukaan sananlaskut ja idiomit. *Idiomit* ovat vakiintuneita yhdyssanoja, sanaliittoja ja sanontoja, joiden ilmaisemaa merkityskokonaisuutta ei voida ainakaan järin helposti päätellä pelkästään ilmauksen sisältämien sanojen merkityksistä. Suomen kielen idiomeja ovat esimerkiksi ilmaukset *selkäsauna*, *valkoinen valhe* ja *peukalo keskellä kämmentä*.

(Kangasniemi 1997: 72–73.) Biber ym. (1999: 989) peräänkuuluttavat n-grammien ja idiomien erottelua toisistaan, mutta toisaalta hekin määrittelevät n-grammien olevan ”usein toistuvia ilmauksia riippamatta niiden idiomaattisuudesta ja rakenteellisesta statuksesta” (mts. 990). Määritelmän voi siis nähdä implisiittisesti ilmaisevan, että myös varsinaiset idiomit voivat lukeutua n-grammeihin siinä missä muutkin monisanaiset rakenteet. Idiomit ovat kuitenkin Biberin ym. (1999: 989) mukaan huomattavasti ”tavanomaisia” n-grammeja harvinaisempia: siinä missä n-grammit, kuten englannin *do you want me to* tai *I don't know what*, voivat esiintyä yli 20 kertaa miljoonassa sanassa, hyvin tunnetutkin idiomit, esimerkiksi *kick the bucket* tai *a slap in the face*, ilmenevät esimerkiksi fiktiossa alle viisi kertaa miljoonassa sanassa ja muissa rekistereissä vieläkin harvemmin. Koska taaja esiintymistiheys on yksi n-grammin kriteereistä, ovatkin etenkin monet pitkät idiomit harvoin varsinaisesti n-grammeja johtuen yksinkertaisesti siitä, että niitä käytetään vain harvakseltaan (Biber ym. 2004: 376–377). Idiomit ovat usein myös yleisempiä juuri fiktiivisissä teksteissä kuin esimerkiksi keskusteluissa (Biber ym. 1999: 1025–1026).

Vaikka erilaisia n-grammeja käytetäänkin kielessä taajemmin kuin esimerkiksi idiomeja, saavat n-grammirakenteet silti usein idiomeja vähemmän huomiota osakseen. Tämä johtuu pitkälti siitä, että olemme kielenkäyttäjinä taipuvaisempia havaitsemaan kielestä ennemminkin tiettyjä substantiiveja ja verbejä kuin funktiosanoja, joista n-grammit useimmiten pitkälti koostuvat (Biber & Conrad 1999: 184). N-grammit eivät myöskään monestikaan ole samalla tavalla vakiintuneita ilmauksia kuin idiomit. N-grammit eroavat idiomeista usein myös siinä, että niitä ei välttämättä voida idiomien tapaan korvata yhdellä sanalla. Esimerkiksi englanninkielisen idiomin *kick the bucket* sijaan on mahdollista käyttää yksinkertaisesti sanaa *die*. (Biber ym. 1999: 988–989.) Suomen kielestä vastaava esimerkki on idiomi *heittää veivinsä*, joka voidaan korvata sanalla *kuolla*. Monet n-grammit ovatkin tällaisista tapauksista poiketen merkitykseltään varsin läpinäkyviä ja jokseenkin itsestään selviä, eli niiden merkitys on jäljitettävissä suoraan niistä sanoista, joita n-grammiin lukeutuu[16] (Cortes 2004: 400). Idiomien, kuten *beat about/around the bush*, merkityksen taas voi harvemmin varsinaisesti päätellä pelkästään niistä sanoista, joita se pitää sisällään (Biber ym. 988–989).

Erilaisten rekisterien ja genrejen on todettu vaikuttavan niiden yhteydessä käytettäviin n-grammeihin (esim. Biber 2009; Culpeper & Kytö 2010). N-grammit ovatkin yksi mahdollinen keino päästä kiinni erilaisten tekstilajien ominaispiirteisiin, mitä esimerkiksi yksittäisiin sanoihin keskittyvät sanalistat eivät välttämättä kykene tarjoamaan (Jantunen 2012: 365).

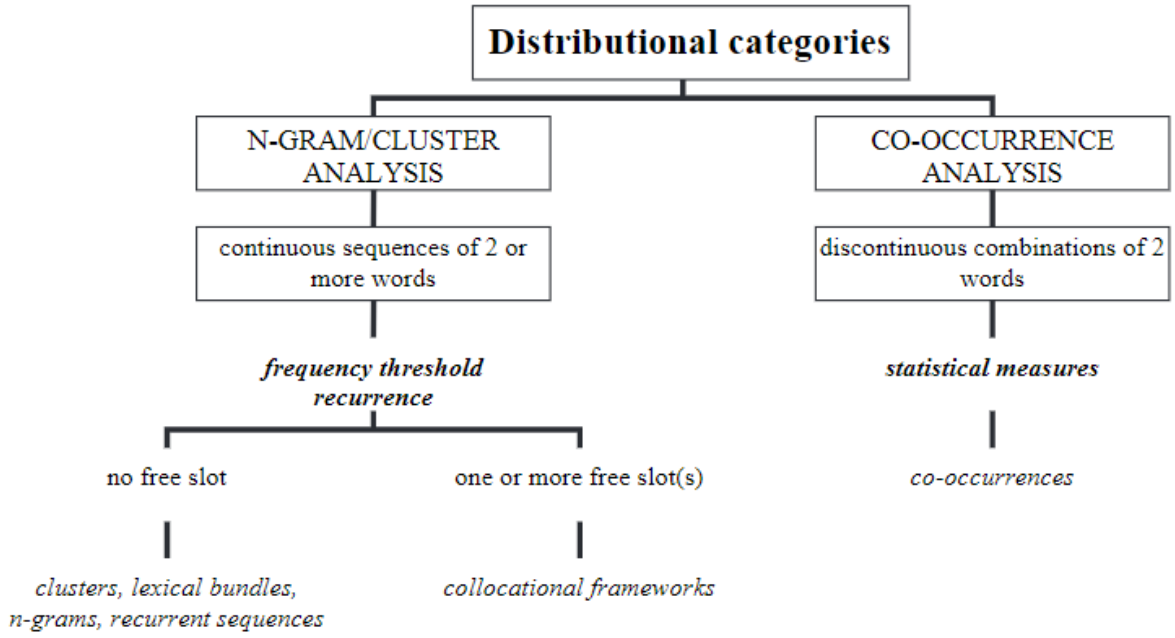
[16] Cortes (2004: 400) antaa tällaisista n-grammeista esimerkit *in the presence of*, *as a result of, I want you to* ja *what do you mean*.

Esimerkiksi keskustelujen kielen ja akateemisen kielen on huomattu eroavan niissä muodostuvien n-grammien osalta huomattavissa määrin toisistaan niin rakenteidensa kuin funktioidensakin osalta (ks. Biber ym. 1999).

2.2.3 N-grammien erosta kollokaatioon ja sanamäärästä

Muun muassa Biber ja Conrad (1999: 183) toteavat, että n-grammit tulee pitää erillään kollokaatioista. N-grammeja erottaa kollokaatioista suurimmilta osin se, että n-grammeissa sanojen esiintyminen ei ole yhtä joustavaa kuin kollokaatioissa[17], vaan sanojen tulee ilmetä ketjuissa välittömästi peräkkäin toistensa kanssa (Gries 2008: 16). Biberin ym. (1999: 989) mukaan n-grammit voidaan kuitenkin jossain määrin nähdä myös eräänlaisina kollokaatiolaajentumina. Etenkin kolmesta sanasta koostuvia n-grammeja, jotka ovat erittäin yleisiä, voidaan pitää jonkinasteisina laajennettuina kollokaatioina. Huomattavasti kolmen sanan n-grammeja harvinaisemmat neljästä, viidestä ja kuudesta sanasta koostuvat n-grammit ovat taas luonteeltaan usein pikemminkin fraasinomaisia. (Mts. 992.) Tarkemmin eroa erilaisten monisanaisten yksikköjen, ennen kaikkea kollokaatioiden ja n-grammien, välillä ovat hahmotelleet Granger ja Paquot (2008) alla olevassa kuviossa, joka pohjaa fraseologian tilastolliseen lähestymistapaan. Malli on vain yksi mahdollinen tapa tulkita monisanaisten yksikköjen hierarkiaa, ja eri tutkijat ovatkin vuosien saatossa esittäneet vaihtelevia luokittelutapoja erilaisille yksiköille (ks. Granger & Paquot 2008: 35–38).

[17] Joustavuudella ja joustamattomuudella Gries (2008: 16) viittaa tarkemmin ottaen siihen, että esimerkiksi sanat *strong* ja *tea* voivat kollokaation osalta ilmentyä esimerkiksi sekä muodoissa *strong tea* tai *the tea is strong*, mutta n-grammeilla ilmentymät ovat usein yhteen muotoon jähmettyneitä. Täten siis *strong tea* ja *the tea is strong* olisivat kumpikin tulkittavissa omiksi n-grammeikseen. (Vrt. kuitenkin luku 2.2.4.)



Kuvio 1. Monisanaisten yksikköjen hierarkiaa Grangerin ja Paquot'n (2008: 39) mukaan.

Kuviossa 1 on esitetty selkeä erottelu n-grammien ja muunlaisten yhteisesiintymien välille. Jälkimmäiset viittaavat Grangerin ja Paquot'n (2008: 39–40) mukaan nimenomaan kollokaatioihin. Kaikkein perustavanlaatuisin n-grammeja ja kollokaatioita toisistaan erottava tekijä on kuvion mukaan se, että n-grammeissa sanat ovat katkeamattomassa yhteydessä toisiinsa ja ne koostuvat kahdesta tai useammasta sanasta, siinä missä kollokaatiot ovat epäyhtenäisempiä ja koostuvat tismalleen kahdesta sanasta. Kollokaatiot todennetaan kuvion jaottelussa luvussa 2.2.1 kuvatun tilastollisen lähestymistavan kautta. Tilastollisuus onkin kuvion mukaan sanojen peräkkäisyyden ohella toinen merkittävimmistä kollokaatiota ja n-grammeja erottavista tekijöistä: kollokaatioiden nostamiseksi aineistosta käytetään tilastollisia lähestymistapoja, kun taas n-grammeja muodostavat ainoastaan sellaiset sanaketjut, jotka toistuvat tutkittavassa aineistossa suhteellisen useasti ylittäen ennalta asetetut frekvenssiraja-arvot. Biberin ym. (1999: 992) mukaan toisteisuus toteutuu tarpeellisissa määrin silloin, kun yksittäinen n-grammi esiintyy esimerkiksi korpusaineistossa suhteutettuna vähintään 10 kertaa miljoonaa sanaa kohden. Muun muassa Wray (2002a: 26) ja Salazar (2014: 13) huomauttavat erilaisten raja-arvoiksi esitettyjen esiintymämäärien olevan kuitenkin lähinnä mielivaltaisia, ja esimerkiksi Cortes (2004) onkin käyttänyt tutkimuksessaan kymmenen esiintymän sijaan 20 osumaa per miljoona sanaa, siinä missä Biberin ym. (2004) tutkimuksessa n-grammin raja-arvoksi asetettiin peräti 40 osumaa miljoonassa sanassa. Erilaiset raja-arvot vaikuttavat luonnollisesti siihen, minkälaisia tilastoja tutkittavan aineiston n-grammeista voidaan muodostaa (Biber ym. 1999: 996).

Biber ym. (2004: 376) painottavat lisäksi, että esiintymämääriin perustuva data ei sinänsä itsessään vielä varsinaisesti kerro paljoakaan; frekvenssidatan kautta voidaan ennemminkin havaita sellaisia kielenkäytön kaavoja, jotka tulee sittemmin pyrkiä selittämään. Nämä kaavat voisivat ilman frekvenssitietoutta jäädä kokonaan huomioimatta. Tässä tutkimuksessa n-grammin riittävän toistuvuuden raja-arvona käytetään 20 esiintymää miljoonaa sanaa kohti (ks. luku 3).

N-grammien hakemista pelkkien esiintymämäärien kautta on kritisoitu siitä, että sillä saavutetaan lähinnä sellaisia n-grammeja, jotka ovat frekventtejä pääosin niihin kuuluvien yksittäisten erittäin frekventtien sanojen vuoksi. Tällaiset sanat ovat yleensä funktiosanoja. (Ks. Salazar 2014: 43.) Salazarin (mts. 19) mukaan frekvenssiperustainen lähestyminen auttaa kuitenkin näkemään muun muassa sen, mitkä leksikaaliset yksiköt ovat tietyille tekstityypeille kaikkein merkittävimpiä. Kollokaatioiden nostamiseksi kieliaineistosta hyödynnetään taas useimmiten niin kutsuttuja MI- ja t-testejä. Samoja testejä on kokeiltu myös enenevissä määrin n-grammien kanssa, ja ne ovat osoittautuneet toimiviksi ainakin 2-grammien tapauksissa (ks. Bestgen & Granger 2014; Garner, Crossley & Kyle 2020). MI-testi (*Mutual Information*) vertaa sanayhtymän frekvenssiä niiden yksittäisten sanojen frekvenssiin, joista yhtymä koostuu ilmentäen täten sanojen todennäköisyyttä esiintyä yhdessä syyn eikä ainoastaan sattuman vuoksi. t-testi taas osoittaa, kuinka varmasti kaksi sanaa muodostavat 2-grammin. MI-testi ei toimi järin hyvin suurifrekventtisten sanojen kanssa, mutta se pystyy osoittamaan matalafrekventtisistä sanoista koostuvia 2-grammeja (esim. *densely populated*), kun taas t-testi korostaa 2-grammeja, jotka rakentuvat hyvin frekventeistä sanoista (esim. *for example*). (Cortes 2015: 206; Garner, Crossley & Kyle 2020: 55.)

Tarpeellisen toistuvuuden ohella myös n-grammeiksi määriteltävien sanayhtymien vähimmäispituuksista on esitetty vaihtelevia näkemyksiä aiemmassa tutkimuksessa: muun muassa Biber ym. (1999) ja Cortes (2004; 2012) rajaavat n-grammeihin kuuluviksi ainoastaan kolmesta tai useammasta sanasta koostuvat sanaketjut, mutta esimerkiksi Granger ja Paquot (2008), Stubbs (2009) sekä Garner ym. (2020) taas mieltävät myös kahden sanan yhtymät n-grammeiksi. Toisaalta n-grammin ilmiönä ei tarvitse koskettaa pelkästään sanatasoa, vaan n-grammeihin lukeutuvat elementit voivat olla myös esimerkiksi kirjaimia tai foneemeja (Mahlberg 2013: 48). Ivaska (2015: 22–23) lisää, että n-grammit voivat periaatteessa koostua myös muun muassa annotointiin eli lingvistiseen metatietoon perustuvista merkinnöistä, jolloin yksittäistä n-grammia voidaan lähestyä esimerkiksi sen suhteen, minkälaisia morfeemeja siihen sisältyy. Ivaska (2014a: 168) onkin hyödyntänyt tutkimuksissaan muiden ohella myös 1-

grammeja, joihin voi tosin sisältyä lukuisia kielellisiä rakenteita.[18] Tässä tutkimuksessa n-grammien nähdään kuitenkin koostuvan muun muassa Biberin ym. (1999: 990) määritelmän tapaan vähintään kolmesta sanasta. 2-grammit ovat aivan liian yleisiä, jotta niitä olisi sekä määränsä että toisaalta ehkä myös laatunsakaan puolesta tarkoituksenmukaista kartoittaa tämän tutkimuksen puitteissa. Toisaalta 2-grammit myös edustavat Salazarin (2014: 42) mukaan useimmiten enemmänkin kollokaatioita.

Grangerin ja Paquot'n (2008: 39) kuviossa (kuvio 1) n-grammit on määritelty omaksi luokakseen vielä lisäksi niiden rakenteesta uupuvien avoimien paikkojen kautta. Kuviossa tarjotaankin toisena vaihtoehtona n-grammeille *collocational frameworks* -nimisiä rakenteita (ks. käsitteestä esim. Renouf & Sinclair 1991). Niillä viitataan fraasikehysten, skipgrammien ja flexgrammien tapaisiin, vapaassa vaihtelussa olevia elementtejä sisältäviin n-grammirakenteisiin, ja niitä sivutaan seuraavassa.

2.2.4 Vaihtoehtoisia lähestymistapoja n-grammeihin

Wray (2008: 16) huomauttaa yleisesti kaavamaisesta kielestä puhuessaan, ettei monikaan kaavamainen jakso ole tallentuneena ihmisen mentaaliseen leksikkoon täysin vaan ainoastaan osittain jähmettyneessä muodossa. Muoto luo ilmauksille eräänlaisen kehyksen, jonka sisällä elementit voivat vaihdella tiettyjen kriteereiden puitteissa. Hän antaa ilmiöstä seuraavan esimerkin:

> For example, the frame underlying 'The elephant was as big as big can be' can be represented as 'NP be-TENSE as ADJ_i as ADJ_i can be'. Completing frames requires insertion rules. By inserting different material into the variable slots, numerous versions can be generated with the same pattern, for example, 'A flea is as small as small can be'; 'That tortoise will be as slow as slow can be'. (Wray 2008: 16.)

Kuvatun kaltaisia, paradigmaattista vaihtelua sisältäviä rakennekehikkoja on tutkimuksissa nimitetty muun muassa *fraasikehyksiksi* (*phrase frames*) (ks. esim. Garner 2016). Fraasikehykset ovat Chengin, Greavesin ja Warrenin (2006) mukaan peräisin Fletcheriltä (2006), ja niitä on hyödyntänyt tutkimuksissaan esimerkiksi Stubbs (2007a & 2007b). Fraasikehys voi olla esimerkiksi rakenne *plays a * part in*, jossa asteriskin paikan voivat ottaa muun muassa sanat *large*, *significant*, *big* tai *major* (Stubbs 2007a: 91). Tässä voidaan nähdä aukko n-grammitutkimuksen kannalta: vaihtoehtoisten täydennysten kautta edellä mainittu rakenne voisi muodostaa useita erilaisia n-grammeja, vaikka kaikkia niistä yhdistääkin täysin sama

[18] Esimerkiksi rakenteen perusteella analysoitu 1-grammi, joka ilmaisee yksikön kolmannen persoonan konditionaalin preesensiä, sisältää yhteensä neljä morfeemia (*fin cond pres sg3*) (ks. Ivaska 2014a: 172).

rakennekehikko. Wrayn (2008) edeltävän esimerkin rakenteenkin voi ajatella potentiaaliseksi n-grammiksi, mutta se ei olisi välttämättä löydettävissä tutkimusaineistosta siitä n-grammeja etsimällä, mikäli paradigmaattisessa suhteessa toistensa kanssa olevat elementit vaihtelisivat aineistossa tiheään.

Joissain yhteyksissä n-grammit sellaisinaan onkin nähty liian rajoittaviksi sekä n-grammien sisäisen sanojen vaihtelemattomuuden että sanojen välittömässä yhteydessä esiintymisen kriteerin vuoksi. Näiden vuoksi huomioimatta ovat voineet esimerkiksi Nesin ja Basturkmenin (2006: 284–285) mukaan jäädä muun muassa n-grammit, joiden osilla on selkeä mutta jostain kohdin katkonainen yhteys, kuten rakenteessa *not only… but also…*. Cheng ym. (2006: 412) taas toteavat, että esimerkiksi ilmaukset *a lot of local people* ja *a lot of different people* ovat molemmat ilmauksen *a lot of people* erilaisia toteumia, jotka vaihtelevuutensa vuoksi jäävät perinteisillä n-grammimenetelmillä tulkittuina mahdollisesti vaille huomiota. Tällaisten puutteellisuuksien eliminoimiseksi n-grammien tueksi onkin kehitetty vielä vaihtoehtoiset *skipgrammit* (*skipgram*) (Warren 2011: 154–155), joita voidaan etsiä korpusaineistoista erilaisin hakutyökaluin. Van Gombel ja van den Bosch (2016) määrittelevät skipgrammit seuraavasti (ks. käsitteestä myös Cheng ym. 2006: 412–413; D. Guthrie ym. 2006; Ivaska 2014a: 23–24; 2015: 166):

> Skipgrams – A fixed-length sequence of p word tokens and q token placeholders/wildcards with total length n ($n = p + q$), the placeholders constitute gaps or skips and a skipgram can contain multiple of these. In turn, a gap can span one or more tokens. For example: “to _ or _ _ be”. (van Gombel & van den Bosch 2016: 2.)

Skipgrammeissa olevien sanojen ei siis tarvitse esiintyä peräkkäin, vaan ainoastaan lähellä toisiaan ja samassa järjestyksessä (D. Guthrie ym. 2006: 1222); ilmiö on täten hyvin lähellä fraasikehyksiä. Kieliaineistosta tehtävät skipgrammihaut eivät myöskään sisällä ainoastaan katkonaisia n-grammeja, vaikka termin nimityksestä niin voisi päätelläkin, vaan myös tavalliset n-grammit kyetään löytämään niiden avulla (Cheng ym. 2006: 412). Van Gombel ja van den Bosch (2016: 2) käyttävät n-grammien ja skipgrammien lisäksi vielä erillistä *flexgrammin* (*flexgram*) käsitettä. Sen he määrittelevät seuraavasti:

> Flexgrams – A sequence with one or more gaps of variable length, which implies the pattern by itself is of undefined length. For example: “to * or * be”. (van Gombel & van den Bosch 2016: 2.)

Flexgrammien ero skipgrammeihin on siis lähinnä se, että sanojen välillä olevien aukkojen pituudet vaihtelevat, eivätkä ole sidottuja tiettyyn sanamäärään skipgrammien tapaan. Skip-

tai flexgrammeja hakemallakaan ei silti poisteta mahdollisuutta, että tutkimuksessa huomiota vaille jäävät *AB*, *BA* -tyyppiset n-grammien sisältämien sanojen sanajärjestyksen vaihtelut. Skipgrammeilla on myös omat rajoitteensa niiden sanapituuden suhteen. (Cheng ym. 2006: 412–413.) Tällaisten vajavuuksien poistamiseksi onkin kehitetty vielä erillinen *konkgrammin* (*concgram*) käsite ja n-grammirakenteiden hakumahdollisuus, jossa ainoana rajoitteena on se, että sanojen tulee esiintyä lähekkäin (Cheng ym. 2006). Käytännön tasolla skip- ja konkgarammit tarkoittavat sitä, että esimerkiksi 4-grammista *Kuka potkaisi pallon pihalle?* on mahdollista skipgrammein haettuna löytää 2-grammit *kuka potkaisi*, *potkaisi pallon*, *pallon pihalle*, *kuka pallon*, *kuka pihalle* ja *potkaisi pihalle*, siinä missä pelkällä perinteisellä n-grammihaulla olisi löydetty näistä ainoastaan kolme ensin mainittua. Konkgrammihaku taas osaisi luokitella vielä esimerkiksi 2-grammit *potkaisi pallon* ja *pallon potkaisi* yhden ja saman n-grammin esiintymiksi. (Ivaska 2015: 22–24.)

Vaikka näkökulmia ja lähestymistapoja toistuviin useampisanaisiin rakenteisiin on olemassa lukuisia, keskitytään tässä tutkimuksessa silti ainoastaan perinteisimpiin n-grammeihin eli sellaisiin sanaketjuihin, jotka koostuvat toistensa kanssa välittömissä kontakteissa olevista sanoista ja joiden sanajärjestys pysyy vakiona. Tällaisiksi n-grammit mieltävät myös esimerkiksi Gries (2008: 16) ja Salazar (2014). Valinta perustuu siihen, ettei vastaavanlaista tutkimusta oppijansuomen osalta ole vielä tehty, jolloin on järkeenkäypää lähteä liikkeelle selvittämällä ensin, millaisia kaikkein tyypillisimpiä toistuvia sanayhtymiä oppijansuomen korpusaineistosta on ylipäänsä löydettävissä. Tutkimuksen aineisto saattaisi myös paisua varsin haastaviin mittasuhteisiin, mikäli siinä haluttaisiin tarkastella sekä perinteisempiä että muun tyylisiä n-grammirakenteita, joten tavallisimmista poikkeavien rakenteiden selvittäminen jää mahdollisten jatkotutkimusten harteille.

2.2.5 Aiempaa n-grammitutkimusta

Korpustutkimuksen yleistymisen myötä myös n-grammit ovat saaneet kielentutkimuksessa lisääntynyttä kiinnostusta osakseen. Esimerkiksi Stubbs (2007b) on kartoittanut tapaustutkimuksessaan n-grammirakenteiden kautta sanan *world* kaavamaisinta käyttöä aineistonaan *British National Corpus*, jonka yksi frekventeimmistä substantiiveista juuri *world* on. Biber ym. (2004) taas selvittivät, minkälaisia n-grammirakenteita yliopiston luokkahuoneopetukseen ja oppikirjoihin keskittyvästä korpuksesta löytyy. He heijastelivat näitä aiemmassa tutkimuksessaan (Biber, Conrad & Cortes 2003) luomaansa taksonomiaan keskustelujen ja akateemisten tekstien n-grammeista. Taksonomian luokat on muodostettu n-grammien diskurssifunktioiden mukaan, ja

ne jakaantuvat fyysiseen ympäristöön tai tekstuaalisiin konteksteihin viittaaviin (esim. *one of the things*), diskurssia organisoiviin (*in this chapter we*), näkökantaa ilmaiseviin (*it is possible to*) ja vuorovaikutteisiin (*I said to him*) n-grammeihin. Biberin ym. (2004) tutkimuksen tuloksista selvisi muun muassa, että opetustilanteiden n-grammit toimivat funktioltaan keskusteluja useammin näkökulmien ja diskurssien organisoijina ja taas akateemisia tekstejä useammin esimerkiksi fyysisiin ja abstrakteihin kokonaisuuksiin viittaajina. Luokkahuoneopetuksen, oppikirjojen, keskustelujen ja akateemisten tekstien konteksteista kaikkein eniten erilaisia n-grammeja esiintyi opetuksessa, mutta keskusteluissa toistettiin kuitenkin yksittäisiä n-grammeja huomattavasti opetustilanteita useammin.

Akateemista kieltä on n-grammein lähestynyt myös muun muassa Cortes (2004), joka selvitti tutkimuksessaan historian ja biologian aloilla julkaistujen tieteellisten tekstien 4-grammeja ja joita hän heijasteli samojen alojen opiskelijoiden tuottamiin teksteihin. Tutkimuksen tuloksista kävi ilmi, että opiskelijat käyttävät teksteissään varsin harvoin ammattikirjoittajilta tuttuja 4-grammeja, joita ovat esimerkiksi ilmaukset *the beginning of the*, *from the perspective of* ja *with the number of*. Tapauksissa, joissa opiskelijat samoja 4-grammeja kuitenkin käyttivät, havaittiin myös eroja ammattikirjoittajien 4-grammien käyttötapoihin nähden: opiskelijat suosivat esimerkiksi 4-grammia *at the same time* ilmaisemaan pikemminkin asioiden lisäystä kuin monen asian samanaikaista tapahtumista. Tuloksista selvisi myös, että opiskelijat toistavat samoja 4-grammeja useita kertoja yksittäisissä teksteissä. Cortes (2012) on samaten kartoittanut akateemisten tekstien johdannoista koostuvassa korpuksessa esiintyviä n-grammeja. Kyseisen tutkimuksen tulosten perusteella tieteellisten tekstien johdannoista on löydettävissä hyvinkin pitkiä n-grammeja, jotka edustavat kieliopillisesti kokonaisia rakenteita, jopa kokonaisia lauseita kuten 9-grammi *the rest of the paper is organized as follows*. Tämä poikkeaa ajatuksesta, jossa n-grammien ei useinkaan mielletä olevan rakenteellisesti kokonaisia yksiköitä. Vastaavia laajoja sanayhtymiä ei aiempien akateemisten tekstien n-grammitutkimusten yhteydessä ollut havaittu ollenkaan n-grammeiksi, mikä osoittaa osaltaan sen, että mitä spesifimpi korpus on kyseessä, sitä frekventeimmäksi yksittäiset n-grammit siinä käyvät.

Fiktionkin kieltä on tutkittu n-grammien avulla: esimerkiksi Mahlbergin (2013) laaja tutkimus pureutuu Charles Dickensin romaaneihin toistuvien sanayhtymien näkökulmasta. Innovatiivisten n-grammimetodien avulla hän onnistui pääsemään käsiksi Dickensin kielessä sellaisiin piirteisiin, joita ei ollut aiemmin kyetty kartoittamaan. Myös ainakin televisiosarjojen kieltä on analysoitu n-grammein: Bednarek (2012) haki amerikanenglantilaisten fiktiivisten tv-sarjojen dialogeista avainsanoja ja 3-grammeja ja vertasi saamiaan tuloksia aiempaan tutkimukseen fiktiivisten sarjojen dialogeista. Näin hän pyrki luomaan yleistyksiä televisiosarjojen dialogien

kielestä. J. Ebeling ym. (2013) ovat sen sijaan tutkineet, miten erilaiset n-grammirakenteet vaihtelevat kahden eri kielen – heidän tapauksessaan norjan ja englannin – välillä fiktiivisissä teksteissä koostuvassa korpuksessa.

Suomessakin n-grammeihin on perehdytty jossain määrin erilaisissa korpustutkimuksissa. Ainakin Kemppanen (2008) on hyödyntänyt n-grammeja väitöskirjassaan, jossa hän tutki *Käännössuomen korpuksesta* (Mauranen 2000) koostamansa *Historiatekstien verrannollisen korpuksen* kautta sekä alun alkaen suomeksi kirjoitettujen että suomeksi käännettyjen historiatekstien ilmentämiä ideologioita muiden muassa n-grammien kautta. Myös ainakin Jantunen (2004) on väitöskirjassaan tutkinut – samaten *Käännössuomen korpukseen*, tosin sen kaunokirjallisiin teksteihin tukeutuen – muiden fraseologisten yksikköjen ohella n-grammien näkökulmasta synonyymiryhmää *hyvin*, *kovin* ja *oikein* (ks. mts. 101–105). Jantunen (2012: 365–366) on sivunnut n-grammeja myös akateemisen kielen osalta listaamalla 20 tieteellisten tekstien frekventeintä 3-grammia. Muutamissa niistä toistuu vastakohtaisuutta ilmentävä adverbi *kuitenkaan*. Jantunen (mts. 366) päätteleekin tällaisten 3-grammien olevan kyseiselle tekstilajille tyypillisiä n-grammeja ja nimittää niitä täten genreklustereiksi tai n-grammeiksi.

Oppijankielestäkin on tehty lukuisia n-grammitutkimuksia etenkin kansainvälisillä tutkimuskentillä. Muutamiin niistä luodaan katsaus luvussa 2.3.4.

2.3 Oppijankieli

2.3.1 Oppijankielen määritelmä ja sen yleismaailmalliset piirteet

Koska tämän tutkimuksen kohteena ovat suomen kielen oppijoiden tekstituotokset eli oppijankieli, on tässä yhteydessä syytä myös määritellä lyhykäisesti oppijankieli käsitteen tasolla sekä kartoittaa joitain piirteitä, jotka on nähty sille ominaisiksi. Yksinkertaisesti ilmaistuna *oppijankielellä* (*learner language*) ymmärretään tarkoitettavan sellaista puhuttua tai kirjoitettua kieltä, jota tuottavat kielenoppijat (R. Ellis & Barkhuizen 2005: 4). Oppijankieli-termin sijaan on usein tukeuduttu myös vaihtoehtoiseen käsitteeseen *välikieli* (*interlanguage*), joka on peräisin Selinkeriltä (1972). Välikielellä on toisaalta saatettu viitata myös eräänlaiseen kielenoppijan tietyssä kielenoppimisensa vaiheessa konstruoimaan mentaaliseen kielioppiin (ks. R. Ellis & Barkhuizen 2005: 54–55; Gass & Selinker 2008: 14). Suomenkielisessä tutkimuksessa *oppijankielen* kanssa rinnan on käytetty muun muassa termejä *vierassuomi* ja *ulkomaalaissuomi* (ks. Nissilä

2011: 39). Tässä tutkimuksessa tukeudutaan kuitenkin systemaattisesti oppijankielen, tarkempirajaisemmin *oppijansuomen*, käsitteeseen, ja tuon kielimuodon käyttäjiin viitataan tutkimuksessa termillä *suomenoppijat*. Esimerkiksi Cook (2016: 16) toisaalta erottelee vielä toisistaan toisen kielen[19] käyttäjät (*L2 user*) ja toisen kielen oppijat (*L2 learner*), joista käyttäjät ovat niitä, jotka hyödyntävät kieltä tosielämässä luokkahuoneopetuksen ulkopuolella ja oppijat taas keitä tahansa, jotka ovat omaksumassa uutta kieltä. Ajatus on siis joiltain osin sama, kuin jaottelussa suomi toisena ja vieraana kielenä -alojen välillä. Cookin (2016) erottelulle merkittävin syy on se, että hän näkee alentavaksi käyttää termin *käyttäjä* sijaan käsitettä *oppija* sellaisestakin henkilöstä, joka on jo vuosia toiminut elinympäristössään toisella kielellään.

Kielenoppijan äidinkielestä voidaan käyttää termejä *lähtökieli*, *lähdekieli* sekä *ensikieli*. Opettelun kohteena olevaa kieltä nimitetään *kohdekieleksi*. (Nissilä 2011: 38.) Toisen – tai yleisemmin vieraan kielen – oppimisesta ja oppimisen tutkimuksen alasta käytetään lyhennettä SLA, joka tulee englannin kielen sanoista *second language acquisition* (R. Ellis & Barkhuizen 2005: 3). Vieraan kielen opettamisen alasta taas käytetään omaa termiään *foreign language teaching* eli FLT (Granger 2002: 5). SLA-tutkimus ei rajaudu ainoastaan kielitieteisiin, vaan siinä yhdistyvät esimerkiksi sosiologian, psykologian ja kasvatustieteiden metodit ja näkökannat (R. Ellis & Barkhuizen 2005: 3).

Jantunen (2008) on esittänyt edeltävien oppijankielen tutkimusten perusteella erilaisia hypoteeseja sille, millaisista yleismaailmallisista piirteistä oppijankieli saattaa koostua riippumatta oppijan ensikielestä tai oppimisen kohdekielestä. Näitä piirteitä hän nimittää *oppijankielen universaaleiksi*. Ensimmäinen hypoteesi oppijankielen universaaleista koskee kieliaineisten epätyypillisiä esiintymistaajuuksia, ennen kaikkea erilaisten piirteiden yliedustumia. Yliedustuvat piirteet voivat olla sanastollisia tai kieliopillisia, ja usein niitä selittävät esimerkiksi vastaopitun rakenteen toistuva käyttäminen pian oppimisen jälkeen, kieliainesten yleistäminen epätyypillisiinkin käyttöyhteyksiin sekä kieltenvälinen siirtovaikutus eli *transferi*. (Mts. 4–6.) Toinen hypoteesi oppijankielen universaaleista piirteistä onkin transferin osuus kielenoppimisessa eli ajatus siitä, että muilla oppijan osaamilla kielillä on rooli tämän tuottamassa oppijankielessä. Etenkin oppijan ensikielen on todettu vaikuttavan oppijankieleen, mutta tämän lisäksi myös muilla oppijan osaamilla kielillä on nähty olevan merkitystä oppijan kielituotoksessa. (Mts. 6–7.) Oppijanenglannin tutkimuksissa oppijan ensikielellä on huomattu olevan vaikutus etenkin oppijoiden tuottamiin kollokaatioihin ja niiden virheellisyyksiin (Paquot & Granger 2012: 140). Transferin voi nähdä jakautuvan sekä positiiviseen että negatiiviseen, joilla

[19] Cookille (2016) toinen kieli (*second language*) tarkoittaa yksinkertaisesti kaikkia ensikieltä seuraavia kieliä.

viitataan pääasiassa siihen, johtaako transferi kohdekielen mukaisiin vaiko virheellisiin tuotoksiin[20] (Gass & Selinker 2008: 94). Siirtovaikutuksen tutkimus on keskittynyt pääosin indoeurooppalaisiin kieliin, mutta myös uralilaisisten kielten osalta ainakin vironkielisiä suomenoppijoiden tuotoksia on tarkasteltu siirtovaikutuksen näkökulmasta (esim. Kaivapalu 2005; Spoelman 2013).

Kolmas oppijankielen universaalien piirteiden hypoteesi on kielellinen yksinkertaisuus tai yksinkertaistuminen. Yksinkertaisuus on luonnollinen ominaisuus etenkin alkuvaiheiden oppijankielelle, sillä oppijalla ei voi vielä tuolloin mitenkään olla samoja resursseja kielensä varioimiseen kuin kieltä syntyperäisesti puhuvalla aikuisella. Yksinkertaisuus ei olekaan missään määrin merkki huonoudesta, vaan se kertoo ainoastaan oppijankielen kehitykseen luonnostaan liittyvistä askelmista. Viestintätilanteissa yksinkertaisuus ilmenee erilaisina oppijan käyttäminä strategioina, joihin liittyvät muun muassa kielen pelkistäminen, jolla pyritään esimerkiksi välttämään monimutkaisia rakenteita ja sanastoa sekä virheellisiä muotoja, kiertoilmaukset sekä nonverbaliikka. Tällaiset keinot ilmenevät etenkin puhutussa viestinnässä, vaikka ovat toki käytössä myös kirjoitetussa kielessä, joskin tuolloin keinojen käyttämistä on usein mahdollista pohtia pitempään sekä tukeutua laajemmin apuvälineisiin. (Jantunen 2008: 8–9.)

Viimeinen hypoteesi oppijankielen universaaleista piirteistä on oletus oppijankielen epäkonventionaalisuudesta, jonka voi jakaa kahteen erilaiseen näkökulmaan, joista toinen liittyy yleis- ja puhekielisyyteen ja toinen kontekstuaaliseen epäkonventionaalisuuteen. Kielimuotojen suhteen oppijankielessä voi arvella yleis- ja puhekielen sekoittuvan usein keskenään, koska etenkin oppimisen alkuvaiheessa kielenoppijan on vielä haastavaa hallita eri kielimuotojen konventionaalisia säännöstöjä. Oppijan puhekielen voi myös olettaa pääsääntöisesti olevan natiiveja yleiskielisempää. Mitä tulee kontekstuaaliseen epäkonventionaalisuuteen, on oppijankieli yleensä hapuilevaa syntagmaattisten ja paradigmaattisten valintarajoitusten ja -preferenssien osalta, eli toisin sanoen fraseologiset yksiköt aiheuttavat kielenoppijalle päänvaivaa. Tämä konkretisoituu ennen kaikkea Sinclairin (1991) esittämän vapaan valinnan periaatteen hyödyntämisenä kielituotoksessa idiomiperiaatteen sijaan. (Jantunen 2008: 12–16.) Oppijankielen universaalit huomioidaan tässä tutkimuksissa erilaisia n-grammien piirteitä mahdollisesti selittävinä tekijöinä.

[20] Negatiivisesta siirtovaikutuksesta käytetään nimitystä *interferenssi* (esim. Siivelt & Mustonen 2013: 342).

2.3.2 Fraseologia oppijankielessä

Fraseologian ja kielenkäytön kaavamaisten jaksojen hallitsemisen on todettu olevan merkittävässä yhteydessä oppijankielen natiivinkaltaisen sujuvuuden kanssa (ks. esim. Cowie 1998; Wray 2000; Jantunen 2009b). Kielenoppijoille fraseologian ja siihen lukeutuvien piirteiden omaksumisen on kuitenkin nähty usein olevan haasteellista ja aiheuttavan virheellisyyksiä ja epäidiomaattisuuksia oppijoiden kielenkäytössä (ks. esim. Granger 1998: 158; Nesselhauf 2005; Osborne 2008; Paquot & Granger 2012). Wrayn (2002a: 206) mukaan ongelmat johtuvat ennen kaikkea siitä, että ensikielen jälkeen opittavia kieliä lähestytään hyvin eri tavalla kuin miten ensikieltä opitaan. Siinä missä ensikieli omaksutaan lähtökohtaisesti laajoina, analysoimattomina kielenyksikköinä, kääntyy myöhemmin opittavien kielten kohdalla asia päälaelleen: liikkeelle lähdetäänkin – etenkin luokkahuoneopetuksessa – pienistä yksiköistä, joita pyritään kasvattamaan suuremmiksi. Tämä taas johtaa siihen, että uuden kielen joutuessa koetukselle tosielämässä leimataan se herkästi epäidiomaattiseksi, sillä summittaisesti yhteen liitetyt sanat eivät välttämättä muodostakaan natiivikielelle sen fraseologian suhteen tyypillisiä ilmauksia, vaikka ne olisivatkin täysin ymmärrettäviä ja myös rakenteidensa puolesta päteviä (Pawley & Syder 1983: 194–195).

Wray (2002a: 206–210) havainnollistaa ensikielen ja sitä seuraavien kielten sekä näiden fraseologisten piirteiden oppimista sanayhtymän *major catastrophe* kautta. Hänen mukaansa kyseisen sanayhtymän ensimmäisiä kertoja kohdatessaan ensikielenään englantia puhuva oppii sen kokonaisuutena, jota ei ole tarve sen koommin pilkkoa erikseen analysoitaviin osatekijöihin. Englantia toisena tai vieraana kielenä oppiva taas saattaa analysoida ilmauksen *major catastrophe* koostuvan kahdesta osasta, joille on olemassa synonyymiset vastineensa ’big’ ja ’disaster’, ja varastoida sen täten muistiinsa kahtena erillisenä sanana huomioimatta niiden kiinteää myötäesiintymissuhdetta. Tämän kautta taas jatkossa, kun oppijalle ilmenee tarve kuvailla ilmiötä *major catastrophe*, ei tämä välttämättä muista juuri näiden lekseemien kuuluvan yhteen, jolloin hän voi päätyä poimimaan muististaan vaihtoehtoisia ilmauksia niiden paradigmoista, kuten *big disaster*, *large mishap* tai *considerable tragedy*, joiden natiivinkaltaisuus vaihtelee huomattavasti. Wrayn (mts. 209–210) mukaan vastaavien sana-analysointien kautta voidaan teoreettisesti päätyä jopa tilanteeseen, jossa oppijalla saattaa olla ensikielistä puhujaa laajempi – joskin huomattavasti epäidiomaattisempi – leksikko tämän purkaessa kohtaamansa tekstit aina kattavasti yhden sanan mittaisiin yksiköihin.

Fraseologian suhteen puutteellisen kielitaidon myötä esimerkiksi suomenoppija saattaa siis tulla tuottaneeksi vaikkapa ilmauksen *Ulkona ei ole erittäin kaunis ilma*, joka ei ole

kieliopillisesti väärin, mutta natiivikielelle fraseologisuuden näkökulmasta hyvin epätyypillinen. Fraseologian hallitseminen sekä sujuvoittaa niin puhuttua kuin kirjoitettuakin oppijankieltä että auttaa oppijaa ymmärtämään paremmin vastaanottamaansa kielellistä syötöstä, sillä hän voi tunnistaa siitä valmiita kokonaisuuksia jokaisen yksittäisen sanan merkityksen analysoinnin sijaan. (Jantunen & Brunni 2012: 75–76.) Ajatus on perustaltaan sama kuin kaavamaisen kielenkäytön ja kaavamaisten jaksojen ydinsanoma. Wray (2002a: 143) kiteyttääkin asian toteamalla, että ”tunteaksesi kielen sinun täytyy yksittäisten sanojen lisäksi tietää myös se, miten ne sopivat yhteen.” Fraseologinen tieto auttaa samalla oppijaa samaten erilaisten rekisterien ja tekstityyppien hallinnassa, sillä niitä määrittelee kutakin omanlaisensa fraseologisuus (Jantunen & Brunni 2012: 76).

Natiivipuhujan ja kielenoppijan tuottamia kielimuotoja voidaan suhteuttaa varsin luontevasti Sinclairin (1991) sanojen kontekstuaalisen valinnan periaatteisiin, joista idiomiperiaatteen nähdään ohjaavan natiiveja siinä missä oppijat tukeutuvat todennäköisemmin etenkin kielenoppimisensa alkuvaiheissa vapaan valinnan periaatteeseen (Granger 1998: 145; Wray 2002a; Singleton ym. 2007). Näyttöä on tosin saatu myös siitä, että toisen kielen puhujat käyttäisivätkin kielessään ajateltua enemmän hyödykseen myös idiomiperiaatetta (ks. Vetchinnikova 2014). Samaten Hoeyn (2005) leksikaalisella primingilla on oma roolinsa kielen omaksumisessa. Kuten aiemmin todettiin, priming ei ole pysyvä ominaisuus (mts. 9), ja etenkin oppijankielen osalta voidaan ajatella, että priming on alati muutoksessa, kun kohdekielen sanoja tavataan uusissa ympäristöissä ja niihin liitetään jatkuvasti uusia merkityksiä, assosiaatioita ja käyttöyhteyksiä. Hoey (mts. 11) kuvaileekin, että sanan kuin sanan priming voi murtua (*crack*), ja että yksi tällaista murtumista aiheuttava seikka on opetus: opettaja voi esimerkiksi korjata oppijaa, mikäli tämä on liittänyt tiettyyn sanaan väärän primingin, kuten vaikkapa *to be* -verbin imperfektimuodon *was* persoonapronominiin *you*. Tällaisissa tilanteissa primingilla on potentiaalinen mahdollisuus murtua. Murtumat paikataan joko hylkäämällä vanha primingin uuden tieltä tai samalla vanhalla primingilla uuden hyökkäyksestä välittämättä. Oppijankielessä jonkinasteiseksi ongelmaksi primingin hyödyntämisessä voi nousta se, että oppiessaan uusia sanoja etenkin perinteisesti sanasta sanaan käännettyinä kielenoppija heijastaa usein ensikielensä sanayhtymät suoraan kohdekieleensä (mts. 183–184). Tällainen voi aiheuttaa epätyypillisyyksiä etenkin oppijan käyttämien kollokaatioiden osalta (ks. esim. Nesselhauf 2005).

2.3.3 *Eurooppalaisesta viitekehyksestä* ja sen kynnystasosta (B1)

Tämän tutkimuksen aineistona toimivan *Kansainvälisen oppijansuomen korpuksen* suuren koon (n. miljoona sanaesiintymää) vuoksi aineistoa oli syytä rajata tavalla tai toisella, jotta tutkimuksesta ei tulisi liian laajamittainen. Tutkimus päätettiinkin kohdistaa pelkästään niihin korpuksen teksteistä, jotka on määritelty *Eurooppalaisella viitekehyksellä* arvioituna B1-tasoisiksi. Ne muodostavat noin puolet (51,2 %) kaikista korpuksen teksteistä. *Eurooppalainen viitekehys* (*Common European Framework of Reference for Languages* eli CEFR; jatkossa EVK) on Euroopan maille suunnattu yhteinen malli muun muassa kielten opinto-ohjelmien, opetussuunnitelmien perusteiden, tutkintojen ja oppikirjojen laadintaan. Sen avulla on tarkoitus seurata kielenoppijoiden kielitaidon kehittymistä. EVK:ssa (2003) kuvataan monitahoisesti, mitä kielenoppijan tulisi osata ja oppia tietyllä kielellä voidakseen käyttää sitä viestintään ja mitä tietoja ja taitoja hänen täytyy kehittää, jotta tuo viestintä voisi olla tehokasta. EVK:n kautta määritellään ne kielitaidon tasot, joiden kautta oppijoiden edistymistä kohdekielessä kyetään mittaamaan oppimisen eri vaiheissa oppijan koko elämän ajan. EVK:n on tarkoitus edistää erilaisten kurssien opinto-ohjelmien ja pätevyysvaatimusten läpinäkyvyyttä luomalla yhteisesti jaettu pohja tavoitteiden, sisältöjen ja menetelmien täsmälliselle kuvaukselle, jolloin myös esimerkiksi vaihtelevissa opiskelukonteksteissa hankittujen pätevyyksien ja suoritettujen tutkintojen hyväksyminen eri Euroopan maissa yksinkertaistuu. (EVK 2003: 19.) EVK:ssa on pyritty laaja-alaiseen näkemykseen kielen käytöstä ja oppimisesta, jolloin siinä käytetty lähestymistapa on hyvin toiminnallinen. Siinä kielen kielenkäyttäjät ja -oppijat nähdään ensisijaisesti sosiaalisina toimijoina, joilla on suoritettavinaan erilaisia tehtäviä erilaisissa ympäristöissä ja toimialoilla. Tällöin kielen roolin ei nähdä rajautuvan ainoastaan esimerkiksi yksittäisiin puheenvuoroihin, vaan toiminta puhetilanteessa saa täyden merkityksensä vasta siihen linkittyvästä sosiaalisesta kontekstista. Erilaisten tehtävien suorittamiseksi yksilön tuleekin käyttää taitojaan tarkoituksenmukaisesti. (Mts. 28.) Keskiöön nousee ennen kaikkea se, mitä (kielellisiä) strategioita yksilö käyttää niissä tehtävissä, jotka hänen täytyy suorittaa (mts. 36). Lukuisista hyvistä puolistaan huolimatta EVK on saanut myös kritiikkiä osakseen muun muassa siitä, että viitekehyksestä uupuu teoreettinen tuki SLA-alalta sekä siitä, miten se erottelee taitotasot toisistaan käyttäen ”pystyy tekemään” -tyyppisiä väittämiä, eikä huomioi niinkään sitä, miten, kuinka tarkasti ja millaisia kieliopillisia ja sanastollisia keinoja hyödyntäen oppijat asiat tarkemmin ottaen kykenevät suorittamaan. Ongelmallisena on nähty samaten viitekehyksen mahdollinen väärinkäyttö poliittisissa ja pedagogisissa konteksteissa. (Ks. Alderson 2007; Hawkins & Buttery 2010: 2–3; Callies & Götz 2015: 2; Byrne 2016: 40–44.)

Perinteisimmillään kielitaito on tavattu jakaa kolmelle eri tasolle, jotka ovat perustaso, keskitaso ja edistynyt taso. EVK:ssa kutakin tasoista on laajennettu yhdellä askelmalla, jolloin tuloksena on saatu kirjain-numeroyhdistelmin nimetyt taitotasot A1, A2, B1, B2, C1 ja C2, jotka kattavat eurooppalaisille kieltenoppijoille olennaisen opiskelualueen. A-tasoille sijoittuvat kielitaidoltaan perustasoiset, B-tasoille itsenäiset ja C-tasoille taitavat kielenkäyttäjät. Taitotasoista matalin on täten A1 ja ylin C2. (EVK 2003: 46–47.) Taitotasoista tämän tutkimuksen kannalta olennaisin eli B1-taso asemoituu luokittelussa niin kutsutuksi *kynnystasoksi* (*threshold*). B1-taitotason olennaisimmat piirteet ovat kielenoppijan kyky pitää yllä vuorovaikutusta ja saada viestinsä perille vaihtelevissa tilanteissa sekä taito selviytyä joustavasti arkipäivän ongelmista. B1-tasoinen kielenoppija kykenee esimerkiksi ilmaisemaan itseään A2-tasoista oppijaa monipuolisemmin ja pitämään keskustelua yllä tätä paremmin, mutta B2-tasoinen oppija taas muun muassa pystyy jo B1-tasoista tehokkaampaan argumentointiin ja omasta kielestään tietoisempiin toimintatapoihin. (Mts. 60–62.)

EVK:ssa kuvataan kriteerit sille, mitä kielenoppijan tulisi kullakin kielitaitotasolla kielen neljästä eri osa-alueesta – kuullun ymmärtämisestä, luetun ymmärtämisestä, puhumisesta ja kirjoittamisesta – osata. Koska käsillä olevan tutkimuksen aineistona on kirjoitetun kielen korpus, on osa-alueista luonnollisesti tarkoituksenmukaisinta tarkastella kirjoittamista ja sen B1-vaatimuksia. EVK:ssa ne on jaoteltu kolmeen luokkaan: yleisiin taitoihin, luovaan kirjoittamiseen sekä raportteihin ja kirjoitelmiin, joista jokainen on edustettuna myös *Kansainvälisen oppijansuomen korpuksen* teksteissä. Niiden B1-kriteerit ovat seuraavat:

> **Yleiset taidot**
> Pystyy kirjoittamaan yksinkertaisia, yhtenäisiä tekstejä tavallisista, itseään kiinnostavista aiheista yhdistämällä lyhyempiä, irrallisia, yksinkertaisia ilmauksia yhtenäiseksi tuotokseksi.
>
> **Luova kirjoittaminen**
> Pystyy kirjoittamaan yksinkertaisia, yksityiskohtaisia kuvauksia monista itseään kiinnostavista aiheista. Pystyy kirjoittamaan selostuksia kokemuksista ja kuvaamaan tunteita ja reaktioita yksinkertaisessa, yhtenäisessä tekstissä. Pystyy kirjoittamaan kuvauksia tapahtumista, esimerkiksi joko todellisesta tai kuvitellusta äskettäin tehdystä retkestä. Pystyy kertomaan tarinan.
>
> **Raportit ja kirjoitelmat**
> Pystyy kirjoittamaan hyvin lyhyitä raportteja, jotka noudattavat tavanomaista selosteen rakennetta ja välittävät jokapäiväistä asiatietoa ja toteavat tekojen syyt. Pystyy laatimaan lyhyitä, yksinkertaisia kirjoitelmia itseään kiinnostavista aiheista. Pystyy tekemään tiivistyksiä, raportoimaan ja esittämään suhteellisen hallitusti näkemyksiä kerätystä asiatiedosta, joka koskee oman alan tuttuja rutiininomaisia tai ei-rutiininomaisia asioita. (EVK 2003: 96–97.)

EVK:sta voidaan tehdä joustavampiakin sovelluksia esimerkiksi erilaisten instituutioiden tarpeisiin, ja niissä kukin taitotaso on mahdollista tarpeen mukaan jaotella vielä edellä esitettyä kuusiportaista luokittelua hienojakoisempiin osasiin. Vaikka jaottelun alakategoriat näissä

malleissa lisääntyvät, on tarkoituksena kuitenkin, että yhteys päätavoitteeseen (taitotasoon A1–C2) pysyy selkeänä. (Ks. EVK 2003: 57–59.) Opetushallitus tekikin EVK:n taitotasojaottelusta oman sovelluksensa vuoden 2003 opetussuunnitelman perusteita varten. Vaikka kyseinen opetussuunnitelma onkin jo vanhentunut, on siinä kuvattu taitotasojaottelu edelleen hyödynnettävissä opetuksen arvioinnissa. Jaottelussa tasoja on pilkottu useampaan osaseen, ja B1-taso konkretisoituu siinä luokkina B1.1 (toimiva peruskielitaito) ja B1.2 (sujuva peruskielitaito). (Opetushallitus 2020.) Alla on esitetty Opetushallituksen EVK:n taitotasosovelluksessaan B1.1- ja B1.2-kielitaitotasoille määrittelemät kriteerit kirjoittamistaidon osalta:

> **B1.1 Toimiva peruskielitaito**
> *Pystyy kirjoittamaan ymmärrettävän, jonkin verran yksityiskohtaistakin arkitietoa välittävän tekstin tutuista, itseään kiinnostavista todellisista tai kuvitelluista aiheista.
> *Osaa kirjoittaa selväpiirteisen sidosteisen tekstin liittämällä erilliset ilmaukset peräkkäin jaksoiksi (kirjeet, kuvaukset, tarinat, puhelinviestit). Pystyy välittämään tehokkaasti tuttua tietoa tavallisimmissa kirjallisen viestinnän muodoissa.
> *Osaa useimpien tutuissa tilanteissa tarvittavien tekstien laadintaan riittävän sanaston ja rakenteet, vaikka teksteissä esiintyy interferenssiä ja ilmeisiä kiertoilmaisuja.
> *Rutiininomainen kieliaines ja perusrakenteet ovat jo suhteellisen virheettömiä, mutta jotkut vaativammat rakenteet ja sanaliitot tuottavat ongelmia.
>
> **B1.2 Sujuva peruskielitaito**
> *Osaa kirjoittaa henkilökohtaisia ja julkisempiakin viestejä, kertoa niissä uutisia ja ilmaista ajatuksiaan tutuista abstrakteista ja kulttuuriaiheista, kuten musiikista tai elokuvista.
> *Osaa kirjoittaa muutaman kappaleen pituisen jäsentyneen tekstin (muistiinpanoja, lyhyitä yhteenvetoja ja selostuksia selväpiirteisen keskustelun tai esityksen pohjalta). Osaa esittää jonkin verran tukitietoa pääajatuksille ja ottaa lukijan huomioon.
> *Hallitsee melko monenlaiseen kirjoittamiseen tarvittavaa sanastoa ja lauserakenteita. Osaa ilmaista rinnasteisuutta ja alisteisuutta.
> *Pystyy kirjoittamaan ymmärrettävää ja kohtuullisen virheetöntä kieltä, vaikka virheitä esiintyy vaativissa rakenteissa, tekstin jäsentelyssä ja tyylissä ja vaikka äidinkielen tai jonkin muun kielen vaikutus on ilmeinen. (Opetushallitus 2020.)

Kaiken kaikkiaan taitotasojen kriteereistä välittyy kuva siitä, että B1-taso tarkoittaa kielenoppijan kirjoittamistaitojen osalta sitä, että oppija kykenee jo muodostamaan yhtenäisiä tekstejä, joilla tämä pystyy selviytymään monista arkielämän kirjallisista viestintätilanteista jopa suhteellisen rikasta ilmaisuvarantoa hyödyntäen, vaikkakin aiheet ovat vielä usein itselle tuttuja ja läheisiä. B1-tason onkin jo aiemmissa tutkimuksissa huomattu olevan merkittävä askel oppijankielen kehittymisessä (ks. esim. Kajander 2013: 93–95; Seilonen 2013: 58–61). Brunni, Jantunen ja Skantsi (2019) ovat havainneet, että oppijoiden kielen tarkkuus kehittyy huomattavasti B1-tasolla myös virheiden vähenemisen näkökulmasta, joskin A2- ja B1-tasojen välillä tapahtuu jonkin verran regressiotakin. Tämä johtunee lukuisista uusista B1-tasolla kieleen ilmaantuvista rakenteista, joiden käyttöä harjoiteltaessa myös virheet lisääntyvät hetkellisesti (mts. 297). Opetushallituksen kriteeristöstä on B1.1 ja B1.2 -tasoja vertailemalla myös nähtävissä, että jo

yksinomaan B1-taitotason sisällä tapahtuu paljon kielellistä kehittymistä. Onkin tärkeää huomata, että samalle taitotasollekin arvioidut kielenoppijat muodostavat luonnollisesti kirjavan joukon, jossa sisäinen taitovaihtelu voi olla hyvinkin merkittävää. Ei olekaan täysin yksiselitteistä määritellä, mitä jokainen B1-tasoinen kielenoppija kielellään varmuudella osaa tai ei vielä osaa ilmaista ja tehdä.

Tyydyttävä suullinen ja kirjallinen kielitaito on asetettu yhdeksi edellytykseksi Suomen kansalaisuuden saamiseksi. Tätä voidaan mitata *Yleisellä kielitutkinnolla* (YKI), jonka taitotasoa kolme (keskitaso) tyydyttävä tulos vastaa. (Maahanmuuttovirasto 2020.) YKI:n kuusiportainen taitotasoasteikko taas on yhteismitallinen EVK:n kuuden taitotason kanssa (Jyväskylän yliopisto 2020). Täten YKI:n kolmas, kansalaisuuteen edellytettävä taso, vastaa siis EVK:n B1-kielitaitotasoa. B1-taitotason saavuttaminen on samalla myös tavoite kotoutumiskoulutuksessa, ja sen tulisi riittää yksilön pärjäämisessä koulutuksen jälkeisissä opinnoissaan, työelämässä ja ylipäätään yhteiskunnassa suomen (tai ruotsin) kielellä (ks. OPS 2012: 23–31). Samalla se toimii myös julkishallinnon työtehtävien kynnystasona. Kaikkinensa B1-taitotason voisikin mieltää jopa eräänlaiseksi valtakunnalliseksi kynnystasoksi. (Latomaa, Pöyhönen, Suni & Tarnanen 2013: 176.) Ottaen huomioon seikat niin B1-taitotason merkityksellisyydestä sekä yksilön henkilökohtaisen kielenoppimisen että yhteiskunnassa pärjäämisen ja siihen sopeutumisen kannalta kuin sen, että *Kansainvälisestä oppijansuomen korpuksesta* juuri B1-taitotasolle arvioidut tekstit ovat taitotasoista korpuksessa selvästi eniten edustettuina, on perusteltua, että kaikista EVK:n taitotasoista tämä tutkimus rajataan koskemaan B1-tasoa. Tällä rajauksella on mahdollista kerryttää dataa siitä, millaisia sanakönttiä oppimisensa kynnysvaiheessa olevat, kuitenkin jo oppijansuomellaan itsenäisesti pärjäävät, suomenoppijat ovat jo ottaneet haltuunsa ja mihin ne mahdollisesti pohjaavat.

2.3.4 Edeltävästä oppijankielen n-grammitutkimuksesta

Kuten laajemminkin kielentutkimuksessa myös oppijankielen osalta kiinnostus fraseologiaa kohtaan on viime vuosikymmeninä kasvanut (N. Ellis 2008: 6–7), ja havaittavissa onkin ollut jopa jonkinlainen oppijankielen tutkimuksen fraseologinen buumi (Jantunen 2009b: 360). Oppijankieltä on lähestytty etenkin kollokaatioiden (ks. esim. Nesselhauf 2003; 2005; Grönholm 2007; Akgül 2013; Wang 2016) mutta myös muiden fraseologisten yksikköjen ja piirteiden suhteen sekä Suomessa (ks. esim. Seppälä 2013; Kuuluvainen 2015; Tarvainen 2018) että ulkomailla (ks. esim. Flowerdew 2006; Singleton ym. 2007; Aktas & Cortes 2008; Garner 2016).

Paquot'n (2013: 391) mukaan myös n-grammirakenteita hyödyntävä oppijankielen tutkimus on lisääntynyt huomattavasti 2000-luvulla. Kansainvälisesti oppijankieltä onkin tutkittu n-grammein lukuisista eri näkökulmista. Erilaisten tutkimusten kautta on saatu selville yleisluontoisesti muun muassa se, että kielitaidoltaan matalammalla olevat kielenoppijat näyttäisivät olevan kehittyneempiä oppijoita riippuvaisempia n-grammien käytöstä. Tämä johtunee pääosin alkeisoppijoiden rajoittuneemmasta sanavarastosta, mutta toisaalta vasta alkeita opiskelevat saattavat usein myös kopioida teksteihinsä suoraan esimerkiksi tehtävien ohjeistuksia, jotka päätyvät sitten muodostamaan n-grammeja. Täten siis, kollokaatioista poiketen, n-grammien kokonaismäärä monesti itse asiassa laskee kielitaidon kehittyessä. (Ks. Paquot & Granger 2012: 139.) Aiemmissa n-grammitutkimuksissa on havaittu myös, että oppijat tukeutuvat pienempään määrään erilaisia n-grammeja kuin natiivit ja että ne n-grammit, joita oppijat hyödyntävät, ovat usein ennemminkin puheen kaltaisia kuin akateemisiin tekstilajeihin soveltuvia, joskin tässäkin kehitytään tarkemmaksi kielitaitotason noustessa korkeammalle (ks. Garner 2016: 33–35). On myös todettu, että transferi vaikuttaa tukeutumisessa etenkin sellaisiin n-grammien rakenteisiin ja diskurssifunktioihin, jotka ovat kielenoppijalle tuttuja tämän ensikielestä riippumatta siitä, sopivatko ne tyylillisesti kohdekieleen (ks. Salazar 2014: 36).

Paquot (2014) on selvittänyt, millainen vaikutus ensikielellä, hänen tutkimuksensa tapauksessa ranskalla, on vieraan kielen oppimisessa n-grammien näkökulmasta. Tulosten perusteella yhdeksi oppijankielen universaaleista piirteistäkin hahmotetulla kieltenvälisellä siirtovaikutuksella on roolinsa muun muassa n-grammien kollokationaalisissa ja kolligationaalisissa preferensseissä, mikä voi aiheuttaa kohdekielen, tutkimuksen tapauksessa englannin, näkökulmasta virheellisiä n-grammituotoksia. Peromingo (2012) taas analysoi EVK:lla (2003) B1- ja B2-kielitaitotasoille arvioitujen englanninoppijoiden tuottamien n-grammien määrää ja rakenteita ja Paquot'n (2013; 2014) tapaan oppijoiden äidinkielen mahdollista vaikutusta niihin sekä vertasi oppijoiden n-grammien käyttöä natiiveihin. Tulosten mukaan oppijat tukeutuvat natiiveja enemmän n-grammeihin, joskin niissä ilmenee tiettyjen leksikaalisten elementtien liikakäyttöä ja toisaalta natiiveille tyypillisten n-grammien puutetta. Ensikielen transferin havaittiin tutkimuksessa tulevan suurissa määrin esiin oppijoiden n-grammeissa ja aiheuttavan sekä oikeaa että virheellistä tekstuaalista koheesiota.

Crossley ja Salsbury (2011) ovat taas tehneet pitkittäistutkimusta 2-grammien kehittymisestä puhutun kielen osalta vuoden tarkastelujakson aikana tutkimuskohteinaan englantia toisena kielenään opiskelevat aikuiset. He keskittyivät sellaisiin oppijoiden tuottamiin 2-grammeihin, joita tavataan myös natiivien puheessa. Tulosten perusteella oppijoiden 2-grammeissa tapahtui vuodessa selkeää kehitystä natiivinkaltaisempaan suuntaan sekä niiden pragmaattisten

että syntaktisten funktioiden osalta. Garner (2016) taas lähestyi saksaa ensikielenään puhuvien, viidelle eri EVK:n taitotasoille arvioitujen englanninoppijoiden n-grammeja fraasikehysten kautta ja havaitsi, että ylemmillä taitotasoilla käytettävät fraasikehykset ovat matalampia tasoja vaihtelevampia, ennakoimattomampia ja funktioiltaan kompleksisempia. Byrne (2016) taas teki tutkimuksessaan huomion muun muassa siitä, että erilaisten 3-grammien määrä kasvaa varsin tasaisesti edettäessä EVK:n (2003) taitotasolta B1 tasolle C1, mitä taas ei tapahtunut 4-grammien osalta. N-grammien määrän kasvussa ei kuitenkaan havaittu tilastollisesti merkittävää yhteyttä taitotason kasvun kanssa. Byrnen (2016) aineistona toimi puhekielen korpus, ja korpuksesta löydettävistä n-grammeista useampi keskittyi *think*-verbin ympärille. Etenkin niistä oli mahdollista huomata, että B1-tasoiset oppijat halusivat ennen kaikkea ilmaista mielipiteitään (esim. 4-grammi *I think it's*), siinä missä B2- ja C1-tasoiset oppijat keskittyivät oman näkökantansa esiintuomisen lisäksi myös siihen, mitä muut olivat asioista mieltä (*what do you think*).

Salazarin (2014) väitöstutkimus osuu metodologiansa puolesta jossain määrin lähelle tätä tutkimusta, vaikkakin hänen tutkimuksessaan hyödynnettiinkin frekvenssien lisäksi myös MI-testiä n-grammeja määrittävänä tekijänä. Tutkimuksessaan julkaistujen tieteellisten tekstien n-grammeista Salazar käytti aineistonaan *Health Science* -korpusta, joka on englanninkielisistä biolääketieteen artikkeleista koostettu korpus. Osan korpuksen artikkeleista ovat kirjoittaneet äidinkieleltään englanninkieliset, osan taas ei-natiivit englanninpuhujat. Salazar pyrki tutkimuksellaan selvittämään muun muassa, mitkä ovat kyseisen korpuksen frekventeimmät n-grammit ja mitkä niiden rakenteet ja funktiot ovat kuin myös sitä, miten natiivien ja ei-natiivien kirjoittajien tuottamat n-grammit eroavat frekvensseiltään, rakenteiltaan ja funktioiltaan toisistaan. Hän kokosi listan korpuksen 3-, 4-, 5- ja 6-grammeista, joiden joukosta hän karsi tutkimuksensa kannata epäolennaisimmat pois. Lopulliseen tutkimukseen päätyneistä n-grammeista hän käyttää nimitystä *kohdekimput* (*target bundles*), ja niitä hän lähestyi niiden avainsanat[21] edellä. Avainsanoitetut kohdekimput Salazar ryhmitteli niiden kielellisen rakenteen ja funktion mukaan erilaisiin luokkiin. Tässä tutkimuksessa etsitään samaten 3-, 4-, 5- ja 6-grammeja ja hyödynnetään soveltuvin osin niitä tapoja, joilla Salazarkin karsi tutkimuksensa laajoja n-grammilistauksia.

Oppijansuomea on n-grammien avulla tutkinut laajemmin ainakin Ivaska (2014a; 2015), joka selvitti Turun yliopiston *Edistyneiden suomenoppijoiden korpusta* (Turun yliopiston kieli

[21] *Avainsanalla* Salazar (2014: 53–54) viittaa sanaan, joka kantaa kokonaisen leksikaalisen sekvenssin merkitystä; esimerkiksi *suggest*-avainsanan alle kuuluvat muun muassa n-grammit *results suggest that* ja *these results suggest* (mts. 58). Tämän kaltaisia avainsanoja ei tule sekoittaa sellaisiin sanoihin, jotka esiintyvät yhdessä korpuksessa huomattavan usein vertailukorpukseen nähden ja joista käytetään samaten termiä *avainsana* (ks. Scott & Tribble 2006: 55–59; tämän tutkimuksen luku 5.1).

ja käännöstieteiden laitos 2012) hyödyntämällä muun muassa oppijansuomen rakenteellisia erityispiirteitä sekä sitä, mitkä konstruktiot ovat oppijansuomelle tyypillisiä ja epätyypillisiä suomea ensikielenään käyttäviin verrattuna ja miten eri ensikielet vaikuttavat konstruktioeroihin. Ivaska (2015) tutki n-grammeja ennen kaikkea niiden morfologinen rakenne edellä eikä suoranaisesti sen perusteella, mistä sanoista ne koostuvat, jolloin tutkimuksen näkökulma ei ole aivan samanlainen kuin nyt käsillä olevassa tutkimuksessa. Tutkimuksen mukaan edistyneiden suomenoppijoiden suomessa korostuvat ensikielisiin nähden muun muassa modaaliset verbiketjut. Eri ensikieliset oppijat taas käyttävät esimerkiksi konjunktioita erilaisin tavoin. Myös Jantunen (2017) on sivunnut n-grammeja oppijansuomen yhteydessä selvittämällä, millaisissa 3- ja 4-grammeissa lemma KELLO esiintyy oppijansuomessa useimmin ja verrannut näitä n-grammeja natiivisuomen vastaaviin (tarkemmin tämän tutkimuksen luvussa 5.1). Vaikuttaa kuitenkin siltä, etteivät n-grammit ole saaneet suomalaisessa oppijankielen tutkimuksessa vielä tähän mennessä järin merkittävää jalansijaa, jolloin tämä tutkimus paikkaa osaltaan yhtä tutkimusaukkoa.

2.3.5 Syntaktiset lausetyypit ja verbien tempukset oppijansuomen tutkimuksessa

Tämän tutkimuksen luvussa 6 selvitetään n-grammien rakenteellisista aspekteista etenkin, minkälaista verbien tempustaivutusta ja mitä syntaktisia lausetyyppejä niissä ilmenee. Verbien aikamuotoja on suomen kielessä neljä: preesens, imperfekti, perfekti ja pluskvamperfekti (VISK § 112). Syntaktisia lausetyyppejä taas on *Ison suomen kieliopin* luokittelun mukaan 11. Ne on esitetty alla olevassa asetelmassa (VISK § 891):

	Teema	
Monikäyttöiset lausetyypit		
Transitiivilause	Te	rikoitte ikkunan.
	Minä	pelkään sinua.
Intransitiivilause	Lokit	lentelevät ja kirkuvat.
	Tilastot	valehtelevat ~ valmistuivat.
Kopulalause	Sinä	olet ihana ~ lohtunani.
	Paketti	on Oulusta ~ sinulle ~ kateissa.
Erikoislausetyypit		
Eksistentiaalilause	Pöydällä	on kirjoja.
Omistuslause	Vaarilla	on saari.
Ilmiölause		Syttyi sota.
Tilalause	Ulkona	sataa ja on kylmä.
Kvanttorilause	Syitä	on monenlaisia.
Tuloslause	Meistä	tulee kuuluisia.
Tunnekausatiivilause	Minua	pelottaa.
Genetiivialkuinen	Minun	on sääli häntä.

Etenkin syntaktisia lausetyyppejä ja niiden käyttöä oppijansuomessa on kartoitettu aiemmallakin tutkimuksella. Tutkimuksia on tehty ainakin transitiivilauseesta (Reiman 2011; 2014), eksistentiaalilauseesta (Ivaska 2011; Kajander 2013) ja yleisemmin *olla*-verbistä osana useampia eri lausetyyppejä (Kynsijärvi 2007). Ivaskan ja Siitosen (2011) tutkimuksessa käsiteltiin sekä eksistentiaali- että intransitiivilauseita. Erilaisia lausetyyppejä on sivunnut myös ainakin Seilonen (2013) tutkimuksessaan oppijansuomen epäsuorista henkilöön viittaamisen keinoista sekä Jokela (2017) selvittäessään *se on* -lauseiden käyttöä *Kansainvälisessä oppijansuomen korpuksessa*. Aikamuodoista on tutkittu pro gradu -tutkielmissa menneen ajan tempuksia yleisesti S2-oppijoiden (Ohvo 2008) sekä erikseen unkarilaisten suomenoppijoiden (Valmu 2007) teksteissä. Perfektin ja imperfektin käyttöä ja omaksumista on taas selvitetty ainakin venäjää äidinkielenään puhuvien suomenoppijoiden osalta (Virtanen 2011). Haapala (2008) taas on tutkinut YKI-aineiston eri taitotasojen verbintaivutusta muun muassa tempusten kautta.

Aiempia tutkimuksia esitellään ja tässä tutkimuksessa saatavia tuloksia vertaillaan tarkemmin niiden kanssa soveltuvin ja tarkoituksenmukaisin osin tutkimuksen varsinaisen, n-grammien rakenteita koskevan, analyysin yhteydessä eli luvussa 6. Luvussa käydään samaten tarkemmin läpi sekä *Ison suomen kieliopin* syntaktiset lausetyypit että tempusten määritelmät.

3 TUTKIMUSAINEISTO JA -MENETELMÄ

Tässä luvussa kuvataan tutkimuksessa käytettävä korpusaineisto sekä tutkimuksen aluilleen paneva tutkimusmenetelmä. Ensimmäisessä alaluvussa käsitellään yleisesti (etenkin oppijankielen) korpusten käyttöä kielentutkimuksen kentällä ja luodaan tarkempi katsaus tässä tutkimuksessa hyödynnettävään oppijankielen korpusaineistoon, *Kansainväliseen oppijansuomen korpukseen*. Toisessa alaluvussa käydään läpi kolme mahdollista lähestymistapaa korpustutkimukselle: aineistoesimerkein tuettu, korpuspohjainen ja korpusvetoinen tutkimus. Näistä viimeksi mainittu on menetelmä, johon tässä tutkimuksessa enimmin tukeudutaan. Luvussa 3.3 kuvaillaan, kuinka ICLFI:stä saadaan *AntConc*-korpusohjelman versiota Windows 64-bit 3.5.8 (Anthony 2019) hyödyntäen rajattua tämän tutkimuksen lopullinen tutkimusaineisto, joka sisältää korpuksen kaikkien B1-kielitaitotasolle arvioitujen tekstien frekventeimmät 3-, 4-, 5- ja 6-grammit. Samassa luvussa kerrotaan myös, millaisia valintoja aluksi saatujen n-grammilistausten supistamiseksi täytyi tehdä.

3.1 Oppijankielen korpukset kielentutkimuksessa ja ICLFI-aineisto

Korpukset (*corpus*, mon. *corpora*) ovat lingvistiseen analyysiin tarkoitettuja, useimmiten varsin laajoja, tiettyjen kriteereiden mukaan koostettuja luonnollisen kielen tekstikokoelmia (esim. Hunston 2002: 2; Weisser 2016: luku 2.1). *Korpuslingvistiikalla* viitataan kielitieteen osa-alueeseen, joka hyödyntää tutkimuksessaan näitä yleisimmin tietokoneitse analysoitavia korpusaineistoja (Lounela & Heikkinen 2012: 121–122; McEnery & Hardie 2012: 1). Tietokoneiden tarjoama tuki on olennaista korpuslingvistiikassa, sillä valtavia tekstiaineistoja on lähes mahdotonta tai vähintäänkin erittäin aikaa vievää käydä läpi käsin etenkään täysin virheettömästi (McEnery & Hardie 2012: 2).

Korpukset koostuvat useimmiten joko kirjoitetusta tekstistä, puheesta tai näiden yhdistelmästä (Weisser 2016: luku 2.2). Myös videokorpuksia, joiden avulla voidaan tutkia esimerkiksi viittomakieltä, on kuitenkin olemassa (ks. esim. Salonen, Takkinen, Puupponen, Nieminen & Pippuri 2016). Korpukset voivat olla joko synkronisia eli eräänlaisia näytteitä tai edustuksia tietyn rajatun ajankohdan kielenkäytöstä tai niin sanottuja monitorikorpuksia, joita kasvatetaan jatkuvasti, jolloin ne mahdollistavat kielessä ajan myötä tapahtuvan muuttumisen tutkimisen (McEnery & Hardie 2012: 6–8). Korpusten tarjoamaa dataa voidaan lähestyä esimerkiksi

sanojen esiintymistiheyksien, fraseologian ja kollokaatioiden kautta (Hunston 2002: 3). Korpusten etuna on se, että niiden avulla kyetään useimmiten tekemään luotettavampia yleistyksiä kielestä kuin mihin pelkästään natiivipuhujan henkilökohtainen kieli-intuitio kykenee (mts. 20). Etenkin esiintymistaajuuksia on hankala tiedostaa pelkän intuition pohjalta, mutta frekvenssiperäinen tieto on hyvin merkityksellisessä asemassa kertomassa toisaalta siitä, mitkä ilmiöt kielessä ovat mahdollisia mutta toisaalta myös sitä, mitkä siinä ovat todennäköisiä. Korpukset kykenevätkin paljastamaan myös sellaisia kielellisiä ilmiöitä, joita ei välttämättä osaisi ennalta aavistaakaan. (Granger 2002: 4.) Frekvensseihin tukeutumalla on myös mahdollista päästä kiinni kielen kaavamaisiin jaksoihin (Wray 2002a: 25). Etenkin juuri frekvenssidatansa ansiosta korpuslingvistiikka onkin käytetyin menetelmä nykypäivän fraseologisessa kielentutkimuksessa (Gries 2008: 15–16). Dataa, jota korpukset tuottavat, voidaan tutkia sekä kvantitatiivisella eli määrällisellä että kvalitatiivisella eli laadullisella otteella (Hunston 2002: 2).

Tämän tutkimuksen tutkimusaineistona toimii Oulun yliopistossa vuosina 2007–2013 koottu *Kansainvälinen oppijansuomen korpus* (*The International Corpus of Learner Finnish*; jatkossa ICLFI) (Jantunen, Brunni & Oulun yliopisto 2013). ICLFI on kirjoitetun oppijankielen korpus, jonka koko on reilut miljoona sanaesiintymää eli sanetta[22] (Jantunen & Pirkola 2015: 92). Oppijankielen korpuksella ymmärretään pääsääntöisesti tarkoitettavan korpusta, jonka materiaali koostuu kokonaisuudessaan korpuksen niin kutsuttua kohdekieltä – toisin sanoen kieltä, jolla korpuksen tekstit on kirjoitettu – toisena tai vieraana kielenä opiskelevien henkilöiden tuotoksista (Weisser 2016: luku 2.4.1.2). Oppijankielen korpusten kautta pyritään usein tunnistamaan, missä suhteissa kielenoppijat eroavat kohdekielensä käytöltään sekä toisistaan että kohdekieltä ensikielenään puhuvista. Natiiveihin verrattaessa oppijankorpuksen rinnalle tarvitaan yleensä jonkinlainen vertailukorpu, jonka tulee koostua yksinomaan natiivien tuottamista teksteistä. (Hunston 2002: 15.)

Oppijankielen korpusten avulla on mahdollista päästä käsiksi muun muassa niihin tyypillisiin haasteisiin, joita kielenoppijat oppimisprosessissaan jakavat (Nesselhauf 2004: 126). Niiden avulla voidaan tutkia esimerkiksi oppijoiden tapoja yliyleistää kohdekielen kielioppiseikkoja tai heidän ensikieliensä mahdollista siirtovaikutusta opittavan kielen tuotoksiin (R. Ellis, & Barkhuizen 2005: 343), toisin sanoen siis oppijankielen universaaleiksikin hahmotettuja piirteitä. Oppijankielen korpukset mahdollistavat empiirisesti saatavaa dataa kattavammin useampien erilaisten oppijankielen piirteiden tutkimisen samanaikaisesti. Oppijankielen korpusten kautta onkin saatu uudenlaista dataa niin SLA- kuin FLT-aloillekin, joilla on tukeuduttu

[22] Puheessa tai kirjoituksessa esiintyviä sanoja voidaan nimittää *saneiksi* eli sanaesiintymiksi, toisin sanoen sanan-muodoiksi (Karlsson 2008: 85–86).

perinteisesti pienehköihin empiirisiin otantoihin. (Granger 2002: 5–6.) Paquot ja Granger (2012) näkevät oppijankielen korpukset ideaalisina myös oppijankielen fraseologian tutkimukseen ennen kaikkea kahdesta syystä: ensinnäkin ne koostuvat konteksteistaan irrotettujen sanojen, fraasien ja virkkeiden sijaan katkeamattomista jaksoista oppijoiden puhuttua tai kirjoitettua kieltä. Toisekseen niihin sisällytetyt oppijoiden tuotokset ovat tyypillisesti peräisin sellaisista tekstilajeista ja tehtävänannoista, jotka antavat oppijalle mahdollisuuden sanallistaa asiansa haluamallaan tavalla, pikemminkin kuin edellyttävät tätä tuottamaan jonkin ennalta halutun sanan tai rakenteen. (Mts. 131.) Rajoitteena oppijankielen korpuksille voidaan kuitenkin pitää ainakin sitä, ettei niiden avulla kyetä mitenkään tutkimaan esimerkiksi sitä, tunteeko oppija sellaisia sanoja tai kielen rakenteita, joita tämä ei teksteissään satu käyttämään, tai sitä, kuinka varma oppija on tuottamansa kielen oikeellisuudesta (Nesselhauf 2004: 131–132). Oppijankielen korpuksia on myös kritisoitu autenttisuuden puutteesta, sillä korpusten kielellisten näytteiden ollessa useimmiten peräisin erilaisista oppimistilanteista ja -konteksteista ovat ne aina jollain tasolla keinotekoisia, eivätkä täysin edusta luonnollista kielenkäyttöä[23] (Granger 2002: 8).

Oppijankielen korpuksia on koottu lukuisia ympäri maailmaa. Niitä on listattu kattavasti ainakin Leuvenin yliopiston internetsivuille (Centre for English Corpus Linguistics 2020). Isossa osassa oppijankielen korpuksia kohdekielenä on englanti eli niitä voidaan nimittää oppijanenglannin korpuksiksi. Laajimpiin oppijankielen korpuksiin lukeutuvat *The Cambridge Learning Corpus* (CLC)[24] noin 50 miljoonalla, *The Hong Kong University of Science & Technology* (HKUST) -korpus 25 miljoonalla, *The Longman Learners' Corpus*[25] 10 miljoonalla ja *The International Corpus of Learner English* (ICLE)[26] 3 miljoonalla saneella. Jokaisen edellä mainitun korpuksen kohdekieli on englanti. Suomessa on ICLFI:n lisäksi koottu kuusi muutakin oppijansuomen korpusta. Nämä ovat Jyväskylän yliopiston *Yleisten kielitutkintojen korpus*, *Cefling-korpus*, *Topling-korpus* ja *Dialuki-korpus*, Turun yliopiston *Edistyneiden suomenoppijoiden korpus* sekä Helsingin ja Tallinnan yliopistojen yhteistyöhanke *Long Second* -korpus. Nämä oppijansuomen korpukset on koostettu useissa erilaisissa konteksteissa monien eritasoisten suomenoppijoiden kielestä, ja ne sisältävät vaihtelevasti niin kirjoittua kuin puhuttuakin materiaalia. Osa näistä korpuksista on synkronisia, osa diakronisia ja osa osin molempia. (Ks. Jantunen & Pirkola 2015.)

---

[23] Granger (2002: 8) tosin huomauttaa samassa yhteydessä, että autenttisuudellakin on eri asteensa; siinä määrin kuin esimerkiksi esseen kirjoittamista voi pitää autenttisena luokkahuonetoimintana, voi esseeteksteistä koostuvan oppijakielen korpuksenkin nähdä täten autenttisena tekstidatana.

[24] https://www.cambridge.org/fi/cambridgeenglish

[25] http://www.pearsonlongman.com/dictionaries/corpus/learners.html

[26] https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

ICLFI:n tekstit on kerätty suomenoppijoiden ulkomaisissa yliopistoissa tuottamista teksteistä, eli kyseessä on suomi vieraana kielenä -korpus. Korpus kattaa tekstilajiensa osalta useita erilaisia fiktiivisiä ja ei-fiktiivisiä genrejä. (Jantunen 2011: 86, 89.) Näihin lukeutuvat muun muassa kertomukset, kuvaukset, esseet, päiväkirjat, mielipidekirjoitukset, sähköpostit, kirjeet ja työhakemukset (Jantunen & Pirkola 2015: 92). Teksteistä isoin osa on kuitenkin (kuvaileviksi) esseiksi tulkittavia (Spoelman 2013: 173). On hyvä huomioida, että oppijankielen korpuksiin koostettujen tekstien genret saattavat osaltaan ohjata sitä, minkälaisia kielen makro- ja mikropiirteitä niistä on löydettävissä (R. Ellis & Barkhuizen 2005: 29), joten on sinänsä oppijankielen tutkimisen kannalta otollista, etteivät ICLFI:n tekstit ole sidoksissa ainoastaan yksittäiseen tekstilajiin.

Kaiken kaikkiaan ICLFI:n tekstejä on ollut tuottamassa yhteensä 22:ta eri äidinkieltä puhuvaa kielenoppijaa (Jantunen & Pirkola 2015: 93). Kuhunkin ICLFI:n tekstiin on sisällytetty erilaisia metatietoja muun muassa kulloisenkin tekstin kirjoittajaan, tämän kielenoppimiskontekstiin sekä itse tekstiin liittyen. Nesselhaufin (2004: 131) mukaan mahdollisuus tutkia oppijankielen piirteitä lukuisten taustatietojen valossa onkin yksi oppijankielen korpusten eduista. ICLFI:n tekstien yhteydessä lueteltuihin metatietoihin lukeutuvat muiden ohella kunkin tekstin tuottaneen suomenoppijan arvioitu suomen kielen taitotaso EVK:n (2003) kielitaitotasoilla A1–C2 mitattuna (Jantunen & Pirkola 2015: 94). Tieto on olennainen tämän tutkimuksen kannalta, sillä sen avulla tutkimus kyetään rajaamaan ainoastaan B1-kielitaitotasoarvioinnin saaneisiin teksteihin. Niistä koostetaan tutkimusta varten oma erillinen osakorpuksensa luvussa 3.3 kuvailluin keinoin.

3.2 Korpusvetoisuus lähtökohtana tutkimukselle

Korpuksia tutkimusaineistona hyödynnettäessä niitä voidaan lähestyä kolmella tutkimustavalla, jotka ovat aineistoesimerkein tuettu tutkimus, korpuspohjainen tutkimus ja korpusvetoinen tutkimus[27]. Aineistoesimerkein tuetussa tutkimuksessa korpus toimii lähinnä esimerkkien lähteenä kuvauksen kohteena olevalle ilmiölle. (Jantunen 2009a: 101–102.) *Korpuspohjaisella* (*corpus-based*) tutkimuksella viitataan sen sijaan tutkimukseen, jossa tutkija valitsee ilmiön, jota hän

[27] Myös *aineistopohjaisesta* ja *-vetoisesta* tutkimuksesta voidaan puhua korpuspohjaisen ja -vetoisen tutkimuksen sijaan. Tällöin tullaan huomioineeksi se, että aineistot voivat olla muitakin kielivarantoja kuin pelkkiä korpuksia. (Jantunen 2009a: 102.) Tämä tutkimus toteutetaan kuitenkin korpusmateriaalilla, joten siinä myös viitataan korpusta painottavaan termistöön.

haluaa korpuksesta kartoittaa tukenaan omat huomionsa ja hypoteesinsa aiheeseen liittyen, jo ennen tarkempaa korpukseen perehtymistään. Korpus toimii siis tällöin ennen kaikkea aikaisempien havaintojen ja teorioiden selittäjänä ja vahvistajana. *Korpusvetoisessa* (*corpus-driven*) tutkimuksessa tutkimuksen kohteeksi taas päätyvät ensi sijassa sellaiset kielelliset ilmiöt, jotka nousevat esille korpusaineistosta itsestään ilman, että niitä valikoidaan etukäteen. (Tognini-Bonelli 2001; 2002; Jantunen 2009a; McEnery & Hardie 2012: 5–6.) Korpusvetoinen tutkimus on hyvin induktiivista, jolloin korpus nähdään siinä suurelta osin hypoteesien lähteenä. Korpusvetoisessa tutkimuksessa korpuksen havainnointi johtaa hypoteeseihin, hypoteesit taas yleistyksiin ja yleistykset lopulta yhdistymiseen teoreettisten väittämien kanssa. (Tognini-Bonelli 2001: 84–85.) Joissain tapauksissa korpusvetoisessa lähestymistavassa jopa tutkimuksen tarkempien tutkimuskysymysten muodostaminen voi jäädä täysin aineiston motivoimaksi (Jantunen 2009a: 103). Korpusvetoisessa tutkimuksessa merkittäviä ovat ennen kaikkea frekvenssit, jotka auttavat muodostamaan erilaisia kielellisiä kategorioita. Samalla korpusvetoisessa tutkimuksessa voidaan huomioida myös erilaisten kielen piirteiden aineistosta *uupumisen* mahdollinen potentiaali. (Tognini-Bonelli 2002: 75.) Wrayn (2008: 11) mukaan korpusvetoisella tutkimuksella on mahdollista päästä kiinni muun muassa kielen kaavamaisiin jaksoihin, sillä mikäli jokin sanaketju löytyy tarpeeksi usein korpuksesta, voi sitä tällöin pitää kieleen kuuluvana kaavana. Korpusvetoista tutkimusta on mahdollista ajatella myös tekstintutkimuksen näkökulmasta, jolloin sen avulla voidaan selvittää esimerkiksi tietylle tekstilajille luonteenomaisia piirteitä (Jantunen 2012: 361).

Pelkän n-grammien löytämisen itsessään korpusaineistosta voi jo lähtökohtaisesti katsoa lähestulkoon vaativan korpusvetoista lähestymistapaa (ks. Chen & Baker 2010: 30); korpuksesta etsitään n-grammeja yleensä vailla järin merkittävää intuitiota siitä, minkälaisia tulokset saattavat olla. Esimerkiksi Stubbsin (2007b: 170–171) mukaan natiivienkin kielenpuhujien on erittäin haastavaa luoda pelkkiin intuitiivisiin käsityksiinsä nojaten kattavia listauksia taajaankaan esiintyvien sanojen frekventeimmistä käyttötavoista, vaikkakin toisaalta nähdessään automaattisesti luotuja n-grammilistauksia he tunnistavat niistä hetimiten tavanomaiset ja tutut ilmaisun keinot. Korpuspohjaisestikin n-grammeja on silti mahdollista aineistosta eritellä, ja aiemmassa tutkimuksessa onkin tukeuduttu molempiin edellä esitellyistä keinoista. Esimerkiksi Paquot (2013) hyödynsi osittain korpuspohjaista lähestymistapaa selvittäessään transfer-vaikutusta ranskankielisten englanninoppijoiden n-grammeissa, sillä hän rajasi tutkimuksensa sellaisiin 3-grammeihin, joiden osana oli (pää)verbi. Korpuspohjaisuus oli lähtökohtana myös Csomaylle (2013), joka selvitti kahden eri korpuksen kautta yliopistojen opetustilanteissa ilmeneviä n-grammeja tukeutuen Biberin ym. (2004) artikkelissaan esittämään n-grammien

funktioluokitteluun. Korpusvetoista tutkimustapaa n-grammien kanssa ovat taas hyödyntäneet esimerkiksi Güngör ja Uysal (2016), jotka vertailivat tutkimustaan varten luomastaan korpuksesta ensikielenään englantia käyttävien kirjoittamia tutkimusartikkeleita ja niiden n-grammien rakenteellisia ja funktionaalisia piirteitä ensikielenään turkkia käyttävien vastaaviin vailla valmiita n-grammikategorioita. Myös Biber (2009) käytti korpusvetoista lähestymistapaa etsiessään korpusaineistosta tyypillisimpiä keskustelun ja akateemisen kirjoittamisen monisanaisia jaksoja.

Seuraavassa alaluvussa 3.3 kuvataan, kuinka tässä tutkimuksessa tarkemman tarkastelun ja kielellisen analyysin kohteeksi päätyvät n-grammit nousevat esille *Kansainvälisen oppijansuomen korpuksen* B1-kielitaitotasolle arvioidusta aineistosta ilman, että niitä ennalta valikoidaan tai rajoitetaan. Tutkimus on ainoastaan kohdistettu sellaisiin n-grammeihin, jotka koostuvat vähintään kolmesta ja enintään kuudesta sanasta ja ylittävät esiintymämäärissään kunkin pituiselle n-grammille ennalta määritellyn raja-arvon. Mitään rajoitteita ei tehdä kuitenkaan esimerkiksi sen suhteen, minkä sanaluokkien sanoja n-grammeihin voi sisältyä, eikä n-grammeja pyritä hakemaan esimerkiksi johonkin aiemmin luotuun kategorisointiin tukeutuen. Aineiston kautta ei myöskään yritetä selittää aikaisemmin luotuja teorioita tai hypoteeseja, joskin saatuja n-grammeja verrataan jossain määrin luvuissa 5 ja 6 aiemmin oppijansuomesta tehtyihin tutkimuksiin. Tutkimusta voidaan siis kaiken kaikkiaan pitää korpusvetoisena.

Korpusvetoisuus luo tälle tutkimukselle kuitenkin lopulta ainoastaan raamit, joiden mukaisesti tutkimuksen varsinainen n-grammianalyysi päästään aloittamaan. Korpusvetoisuuden avulla saadaan muodostettua tutkimuksen varsinainen tutkimusaineisto, jonka tutkimista kuitenkin jatketaan vaihtelevin, vielä ennen tutkimuksen toteuttamista tuntemattomin menetelmin. Tämä selittää tutkimuksen toista tutkimuskysymystä. Tutkimuksen luvussa 4 päästään korpusvetoisesti saatavaan frekvenssidataan nojaamalla käsiksi *Kansainvälisen oppijansuomen korpuksen* B1-kielitaitotason tekstien yleisimpiin n-grammeihin. Tarkemmat keinot n-grammien saamiseksi kuvataan alaluvussa 3.3. Tässä yhteydessä kerrytetään siis etenkin määrällistä dataa. Tämän jälkeen luvuissa 5 ja 6 määrällistä analyysia jatketaan laskemalla esiintymämääriä erilaisille kielen ilmiöille n-grammeissa. Määrällisiä tuloksia pohditaan tämän jälkeen vielä myös laadulliselta kantilta. Tutkimusotteeltaan käsillä oleva tutkimus edustaa siis sekä kvantitatiivista että kvalitatiivista tutkimusta. Korpusvetoiselle tutkimusmenetelmälle onkin tyypillistä, että se on pohjimmiltaan kvantitatiivista, ja kvalitatiivisten havaintojen rooliksi jää lähinnä kvantitatiivisten huomioiden selittäminen (Jantunen 2009a: 106).

3.2.1 Tutkijan intuitio korpusvetoisessa tutkimuksessa

Ennen korpuslingvistiikan kehittymistä kielenkäyttäjien intuitiivisten näkemysten kielestä ja sen erilaisista piirteistä ajateltiin olevan varsin tarkkoja (Wray 2002a: 21). Tietokoneiden ja korpusanalyysien myötä kieli kyettiin kuitenkin näkemään uudessa valossa, mikä osoitti esimerkiksi Sinclairin (1991: 4) mukaan sen, ettei intuitio ole hyvä osviitta todellisesta kielenkäytöstä, eikä siitä tällöin ole juurikaan käytännön hyötyä kielentutkimuksessa. Korpustutkimuksin saatavien tulosten tulkinta ja niiden merkitysten pohdinta kielen ja sen tutkimuksen näkökulmasta on kuitenkin aina riippuvaista analysoijan intuitiosta, eivätkä tietokone ja analyysiohjelmat kykene korvaamaan ihmisen intuitiivista näkemystä ja siihen perustuvia päätelmiä tieteenalasta (Jantunen 2009a: 108). Wrayn (2008: 114) mielestä intuition roolia ja potentiaalia työkaluna kielentutkimuksessa ei tulisikaan pyrkiä sivuuttamaan, vaikka objektiivistenkin mittaustapojen kehittäminen on hänenkin mukaansa merkittävää.

Tämän tutkimuksen osalta intuition rooli korostuu ennen kaikkea tutkimuksen rakenneanalyysissa eli luvussa 6. Kuten luvussa 4 selviää, tutkimuksessa käytetyillä hakukriteereillä tutkimuksen analyysivaiheeseen päätyviä n-grammeja kasaantuu lähes tuhat erilaista. Näitä tarkastellaan luvussa 6 rakenteiden puolesta niiden edustamien tempusten ja syntaktisten lausetyyppien osalta. Tällaista tietoa korpusohjelmat eivät osaa automaattisin laskuin tuottaa, joten tutkijan tulee käydä prosessi läpi manuaalisesti. Koska yksittäinenkin n-grammi voi edustaa eri koteksteissaan useampaa erilaista tempusta ja lausetyppiä, ei jokaisen n-grammin kaikkien osumien voi sellaisinaan laskea edustavan tiettyä tempusta tai lausetyyppiä ilman tietoa niiden koteksteista. On selvää, että koska kyseessä on luonnollinen kieli ja vielä sen oppijankielivariantti kynnystasollaan, esille nousee myös poikkeuksellisia rakenteita, joiden yksiselitteinen luokittelu esimerkiksi tiettyyn lausetyyppiin on haastavaa. N-grammiesiintymien suuri määrä hankaloittaa samaten analyysia, sillä jokaisen n-grammin jokaista esiintymää ei ole tarkoituksenmukaista käydä käsillä olevan tutkimuksen puitteissa lävitse. Täten tutkimuksessa joudutaan etenkin lausetyyppien tulkinnan osalta tukeutumaan epäselvemmissä tapauksissa vahvasti tutkijan intuitioon sen suhteen, mitä lausetyypeistä n-grammilla ja sen tietyillä esiintymillä suomenoppija on halunnut ilmaista. Asiaa käsitellään vielä hitusen tarkemmin luvussa 6.1.2. Sitä problematisoidaan lisää myös tutkimuksen onnistumisen arviointia käsittelevässä luvussa 7.2.

3.2.2 *AntConc*-ohjelma korpustutkimuksen työkaluna

Tämän korpusvetoisen tutkimuksen toteuttamiseksi hyödynnettiin *AntConc*-nimistä ohjelmaa, joka on internetistä vapaasti ladattavissa (ks. Anthony 2019). *AntConc* on korpusohjelma, jonka avulla voidaan tutkia tietokoneelle ladattuja korpuksia. *AntConc* mahdollistaa korpusaineistojen lähestymisen muun muassa konkordanssirivien, n-grammien ja sanalistojen kautta. Ohjelmasta käytettiin tässä tutkimuksessa sen versiota Windows 64-bit 3.5.8.

Korpusohjelmien keskiössä on usein niiden konkordanssityökalut (*concordancer*). Niitä voidaan pitää korpustutkimuksen tärkeimpänä työkaluna (McEnery & Hardie 2012: 35). Korpusohjelmia, joista yleensä jokaiseen sisältyy oma konkordanssityökalunsa, on olemassa lukuisia erilaisia. Kaikkien niiden perimmäisenä tarkoituksena on etsiä tutkimuksen kohteena olevasta korpuksesta käyttäjänsä haluamaa sanaa tai fraasia, ja muodostaa hakutuloksista niin sanottuja konkordanssirivejä. *Konkordanssilla* (*concordance*) itsessään tarkoitetaan korpuksesta haettujen sanamuotojen listausta niiden tekstuaalisten esiintymisympäristöjensä kera siten, että noodin eli hakusanan jokainen esiintymä korpuksessa on omalla konkordanssirivillään ruudun keskellä sitä ympäröivän lähimmän tekstikontekstin kanssa (Hunston 2002: 39; Weisser 2016: luku 5.1). Täten konkordanssiohjelmat siis mahdollistavat sanojen tutkimisen suoraan niiden esiintymiskonteksteissaan (McEnery & Hardie 2012: 2). Näistä kontekstiesiintymistä käytetään kuvaavaa nimitystä *key word in context* (KWIC) (mts. 35). *AntConcilla* on mahdollista muodostaa listoja pelkistä n-grammeista ilman niiden kontekseja. Ohjelmassa päästään kuitenkin yksittäistä listan n-grammiosumaa klikkaamalla myös kyseisen n-grammit KWIC-näkymään, joka listaa seuraavan esimerkkikonkordanssin tapaan kaikki n-grammin aineistoesiintymiskontekstit omille konkordanssiriveilleen:

Hit	KWIC			File
1	heti käyn suihkussa. Aamiaiseksi minä	*syön voileipää ja juon*	kahvia. Sitten pukeudun ja lähden	VE0192b.txt
2	eudun ja laitan aamiaisen. Aamiaiseksi	*syön voileipää ja juon*	kahvia. Lähden kotoa puoli kahde	VI0081d.txt
3	un. Sen jälkeen laitan ruokaa. Aamulla	*syön voileipää ja juon*	kahvia. Lähden kotoa vartiia vaill	VI0125b.txt
4	n aamulla puoli yhdeksän. Aamiaiseksi	*syön voileipää ja juon*	kahvia tai teetä. En pitää myslistä.	VI0165b.txt
5	keudun ja syön aamiaisen. Aamiaiseksi	*syön voileipää ja juon*	maitoa. Usein katson aamulla my	VI0361a.txt
6	Keittiössä minä teen aamupalaa. Minä	*syön voileipää ja juon*	mehu. Sitten minä katson vähän te	RU0012a.txt
7	kotona ja katson elokuvaa. Illallistakin	*syön voileipää ja juon*	mehua. Menen nukkumaan yleens	VI0029b.txt
8	amiaisen. Aamiaseksi minä tavallisesti	*syön voileipää ja juon*	teetä. Minä en pidä kahvista. Minä	VE0189a.txt
9	in. Palaan kotiin kello puoli yhdeksältä.	*Syön voileipää ja juon*	teetä. Sen jälkeen minun täytyy vi	VI0193a.txt
10	a. Nousen, käyn suihkussa ja pukeudun.	*Syön voileipää ja juon*	teetä. Lähden kotoa varttia vaille y	VI0198a.txt
11	sta syön kahvilassa jotakin. Tavallisesti	*syön voileipää ja juon*	teetä tai kahvia ja sitten menen	VI0355a.txt
12	an. Lounas on kello yksi tai kello kaksi,	*syön voileipää ja juon*	vettä. Iltapäivällä menen kotiin ja	VI0037a.txt

Konkordanssi 1. 4-grammin *syön voileipää ja juon* esiintymät (n = 12) B1-aineistossa.

Konkordanssissa 1 esitetyssä konkordanssinäkymässä haettu 4-grammi eli noodi on näkymän keskellä lähimpien tekstikontekstiensa kanssa. Siitä voidaan siis jo pelkällä nopealla silmäyksellä nähdä, missä yhteyksissä 4-grammia on korpuksen alkuperäisissä teksteissä käytetty. Noodin vasemmalla ja oikealla puolella näytettävän kontekstin laajuus on muunneltavissa ohjelmassa, ja konkordanssirivit voidaan järjestää listaksi vaihtelevin tavoin. Jokainen Hit-sarakkeen numeroitu konkordanssirivi edustaa omaa erillistä aineistoesiintymänsä. File-sarakkeessa on ilmoitettu se, mistä tiedostosta, ICLFI:n tapauksessa suomenoppijan tekstistä, kukin konkordanssirivi on peräisin. Tiedostonimen yksilöintikoodin kaksi ensimmäistä kirjainta paljastavat tekstin kirjoittajan ensikielen; esimerkiksi tiedoston, jonka nimikoodi alkaa kirjaimilla *VI*, sisältämän tekstin on kirjoittanut viroa ensikielenään puhuva suomenoppija. *RU*-tiedostot ovat taas peräisin ruotsinkielisiltä, *VE*:t venäjänkielisiltä, *SA*:t saksankielisiltä ja niin edelleen. Kirjaimia seuraavat numerot yksilöivät kirjoittajan, ja numeron jäljessä oleva kirjain ilmaisee vielä sen, monesko kirjoittajan kirjoittama teksti on kyseessä.

Mikäli konkordanssinäkymän tarjoama tieto ei riitä haluttujen päätelmien tekemiseen, pääsee *AntConcin* File View -työkalulla kätevästi tarkastelemaan myös niitä kokonaisia tekstejä, joista kukin konkordanssirivi on peräisin. Varsinaisten konkordanssien hyödyntäminen tulee tässä tutkimuksessa tarpeeseen, kun halutaan yllä olevan esimerkin tapaan tarkastella kotekstejä, joissa suomenoppijat tiettyjä n-grammeja käyttävät. Tämä on oleellista etenkin luvussa 6, jossa pyritään selvittämään n-grammien ilmentämiä verbien tempuksia sekä syntaktisia lausetyyppejä. Esimerkiksi 3-grammista *hän ei ollut* ei kyetä ilman laajempaa kotekstia näkemään, käytetäänkö 3-grammia imperfekti- tai perfektilauseissa tai onko se syntaktiselta lausetyypiltään esimerkiksi kopula-, transitiivi- tai intransitiivilause.

AntConc mahdollistaa siis konkordanssinäkymän ohella muun muassa eripituisten n-grammien koostamisen suoraan korpusaineiston kaikista tietyn frekvenssin ylittävistä sanayhtymistä ilman valmiiksi määriteltyjä hakusanoja. Tämä tapahtuu ohjelman Clusters/N-grams-työkalulla. Kun n-grammihakua rajoitetaan tulosten esiintymistaajuuden perusteella, voidaan työkalun avulla hakea korpuksesta esimerkiksi kaikki siinä kaksikymmentä kertaa tai useammin esiintyvät 3-grammit. Aineistosta on tämän tutkimuksen kannalta olennaista rajata pois kaikki tietyn frekvenssimäärän alle jäävät n-grammit, jotta hauilla saadut n-grammit täyttävät määritelmäänsä kuuluvan kriteerin tarpeellisesta toisteisuudesta. Työkalun avulla saadut n-grammitulokset voidaan järjestää muun muassa frekvenssiensä mukaisesti tai aakkostaa n-grammin aloittavan tai lopettavan kirjaimen mukaisesti.

3.3 Lopullisen tutkimusaineiston rajaaminen menetelmällisillä valinnoilla

3.3.1 B1-osakorpuksen kokoaminen ja *AntConcin* hakuasetukset

Kuten jo todettua, tämä tutkimus kohdistettiin ICLFI:n kaikkiin B1-kielitaitotasoarvioinnin saaneisiin teksteihin ja niistä löytyviin n-grammeihin. Jotta ICLFI:stä voitaisiin tehdä hakuja *AntConcilla*, tuli tutkimuksessa käyttää korpuksen ladattavaa versiota. Tutkimuksessa käytettiin korpuksesta sen annotoimatonta eli niin sanottua raakaversiota; toisin sanoen korpuksen teksteihin ei ollut lisätty minkäänlaisia kieliopillisia tunnuksia tai kuvailutietoja[28]. Korpuksen jokaisen tekstin alkuun on kuitenkin sisällytetty metatiedot kulloisenkin tekstin kirjoittajan sekä itse tekstin ja sen tuottamistilanteen taustoista, mitä sivuttiin jo luvussa 3.1. Taustatietoihin lukeutuvat muun muassa kunkin tekstin kirjoittajan syntymävuosi ja sukupuoli, tekstin genre ja tehtävänanto sekä tieto siitä, onko teksti kirjoitettu esimerkiksi kotona tai koetilaisuudessa. Metatiedot jokaisen tekstin kirjoittajasta ovat tämän tutkimuksen kannalta olennaiset, sillä niistä ilmenee myös kirjoittajan EVK:lla (2003) arvioitu kielitaitotaso, jonka tietäminen oli edellytyksenä sille, että tutkimus voitiin rajata juuri B1-kielitaitotasolle.

Tutkimus aloitettiin koostamalla koko ICLFI:stä niin kutsuttu osakorpus, jonka tuli sisältää kaikki korpuksen B1-kielitaitotasolle arvioidut tekstit. Tietokoneelle ladattavassa ICLFI:n versiossa tekstit on jaoteltu omiin alakansioihinsa oppijoiden äidinkielten, tekstien keräyspaikkojen, oppijoiden kielitaitotasojen sekä kirjoitettujen tekstilajien perusteella. Tekemällä koko korpuskansiosta haku “B1” saatiinkin näkyviin kaikki korpuksen B1-kansiot, jotka toisin sanoen sisältävät kaikki korpuksen B1-kielitaitotasolle arvioidut tekstit. Nämä kansiot kopioitiin tutkimusta varten yhteen omaan kansioonsa, josta muodostui täten niin sanottu B1-osakorpus (tästedes B1-aineisto) ja samalla tutkimuksen tarkempirajainen tutkimusaineisto. Tämä B1-aineisto kattaa yhteensä 2 659 tekstiä ja 409 482 sanetta.

B1-aineisto vietiin *AntConc*-ohjelmaan, jonka Clusters/N-grams-työkalulla voidaan siis hakea eripituisia n-grammeja korpuksesta, kuten edellä kuvattiin. Ennen hakuja *AntConcin* oletusasetuksia tuli muokata hieman. Tällä pyrittiin varmistamaan, että korpushauilla saataisiin mahdollisimman osuvia ja tarkoituksenmukaisia n-grammituloksia. Ensinnäkin *AntConcin*

[28] ICLFI:stä on kuitenkin olemassa myös versio, joka on annotoitu kieliopillisesti kokonaan sekä lisäksi osittain sen sisältämien kielivirheiden suhteen (ks. Jantunen, Brunni, Lehto & Airaksinen 2014; virheannotoidusta korpuksesta erikseen ks. Kuuluvainen 2015). En nähnyt annotoidun version käyttöä kuitenkaan tarpeellisena tässä tutkimuksessa, sillä annotointimerkinnät aiheuttivat lähinnä sekaannuksia n-grammeja korpuksesta haettaessa. Annotoitujen korpusten käyttöön on toisaalta suhtauduttu ylipäätään kriittiseltä kantilta korpusvetoisen tutkimusmenetelmän yhteydessä, sillä annotoitimerkintöjen voidaan ajatella ainakin jossain määrin ohjailevan tutkimusta niiden perustuessa aina ennakko-oletuksiin kielellisestä mallista (Mahlberg 2013: 13).

Global Settings -asetusvalikon kautta muutettiin Character Encoding -välilehdellä ohjelman merkistökoodaus tukemaan suomen kielen tarkkeita. Samaisessa Global Settings -valikossa on myös Tags-välilehti, jonka kautta korpuksen sisältämät annotointi- ja metatiedot on mahdollista piilottaa hakutuloksista. Kuten edellä mainittiin, raakakorpuksen tapauksessa merkintöjä oli ainoastaan tekstien alkuihin listatut kirjoittajaa ja tuotettua tekstiä koskevat taustatiedot, mutta nekin oli olennaista saada piilotettua hakutuloksista, sillä muutoin ohjelma lukisi niissäkin olevia sanayhtymiä n-grammeiksi.

Global Setting -valikossa voi määrittää myös sen, mitkä merkit muodostavat "hyväksyttäviä" ja hakutuloksissa näytettäviä saneita. Oletusasetuksena on, että saneet muodostuvat ainoastaan kirjaimista, jolloin ohjelma poista mitä tahansa muita merkkejä sisältävät saneet automaattisesti tuloslistauksista. En kuitenkaan halunnut rajoittaa hakuja pelkkiin kirjaimiin, sillä ICLFI:n teksteissä esiintyy paljon myös esimerkiksi ikiä ja vuosilukuja. Täten hakuehtoihin lisättiin mukaan myös numerot mutta siten, että kaikki teksteissä esiintyvät (numeroin kirjoitetut) luvut korvattiin automaattisesti prosenttimerkillä *%*[29]. Tällöin esimerkiksi 3-grammit *minä olen 20-vuotias* ja *minä olen 21-vuotias*, jotka olisi muutoin laskettu kahdeksi erilliseksi 3-grammiksi, luettiinkin yhdeksi ja samaksi 3-grammiksi *minä olen [%]-vuotias*. Samoin esimerkiksi 5-grammit *hän on 26 vuotta vanha* ja *hän on 44 vuotta vanha* laskettiin 5-grammiksi *hän on [%] vuotta vanha*. Tällä tavoin siis esimerkiksi rakenteen *hän + on + x + vuotta + vanha* frekvenssi saatiin kasvamaan ja näin ollen myös (paremmin) esille hakutuloksiin. Rakenne olisi voinut jäädä jopa kokonaan hakutulosten ulkopuolelle, jos *x*:n paikalla olisi taajaan vaihtelevia lukuja (vrt. luku 2.2.4). Nähdäkseni erilaiset luvut n-grammien sisällä ovat kuitenkin lähinnä sivuseikkoja, jotka eivät muuta n-grammin perusmerkitystä saati rakennetta suuntaan tai toiseen, joten olisi ollut perusteetonta, mikäli tietyt n-grammit olisivat jääneet tämän vuoksi tutkimuksen ulkopuolelle.

Hyväksyttäviin merkkeihin lisättiin myös yhdysviiva (-), jotta yhdysviivan avulla muodostetut yhdyssanat kuten *vapaa-aika* tai *[%]-vuotias* eivät jäisi n-grammilistauksista pois tai sekoittaisi tuloksia. Mikäli merkkiä ei lisätä hakuehtoihin, tulkitsee ohjelma esimerkiksi sanan *vapaa-aika* kahdeksi eri saneeksi, jolloin esimerkiksi *on vapaa-aikaa* olisi ohjelman mukaan 3-grammi, vaikka se muodostuukin vain kahdesta sanasta.

[29] Tämäkin onnistuu Global Settings -valikon kautta. Prosenttimerkinnät on sijoitettu tutkimuksen n-grammiesimerkkien yhteydessä selkeyden vuoksi vielä hakasulkeiden sisään.

3.3.2 B1-aineistosta tehdyt n-grammihaut

Kuten luvussa 2.2.3 mainittiin, Biberin ym. (1999: 992) mukaan n-grammi on ”tarpeeksi toistuva”, mikäli se esiintyy vähintään 10 kertaa miljoonaa sanetta kohden. Tämän tutkimuksen puitteissa tätä kriteeriä kuitenkin tiukennettiin määrällisesti suurten hakutulosten rajoittamiseksi siten, että monisanaisen ketjun tuli esiintyä vähintään 20 kertaa miljoonassa saneessa ollakseen n-grammi. Tälläkin raja-arvolla pitäisi päästä joka tapauksessa käsiksi olennaisimpaan osaan oppijansuomen toistuvia monisanaisia rakenteita. Löyhemmät frekvenssikriteerit mahdollistaisivat vielä lukuisten muidenkin n-grammien, jotka nyt jäävät tutkimuksen ulkopuolelle, tunnistamisen, mutta toisaalta ne samalla paisuttaisivat tutkimusaineiston haastaviin mittasuhteisiin. Samaan 20 esiintymää per miljoona sanetta -raja-arvoon on tutkimuksessaan tukeutunut aiemmin myös ainakin Cortes (2004). Toisaalta esimerkiksi Biber ym. (2004) käyttivät peräti 40 osuman raja-arvoa tutkiessaan n-grammeja yliopiston opetukseen ja oppikirjoihin keskittyvästä korpuksesta.

Yhden n-grammin miljoonaan saneeseen suhteutettu frekvenssi saadaan jakamalla tarkasteltavan n-grammin esiintymämäärä tutkimuksen kohteena olevan korpuksen – tämän tutkimuksen tapauksessa siis B1-aineiston – sanemäärällä ja kertomalla saatu osamäärä miljoonalla. Laskemalla B1-aineistosta suhteellisia frekvenssejä voidaan huomata, että yksittäisen n-grammin tulee esiintyä yhdeksän kertaa tai useammin 409 482 sanetta kohden, jotta se esiintyisi suhteessa yli 20 kertaa miljoonassa saneessa ja olisi näin ollen tämän tutkimuksen kannalta merkitsevä. *AntConcin* Clusters/N-grams-työkalun hakuehdoiksi asetettiin siis 3-grammeja haettaessa seuraavat arvot:

N-gram Size: min. 3, max. 3
Min. Freq.: 9
Min. Range: 5

N-gram Sizen arvot kuvaavat nimensä mukaisesti sitä, monenko sanan mittainen korpuksesta haettava n-grammi saa lyhimmillään ja pisimmillään olla. Range-arvo taas ilmaisee sen, kuinka monen eri tekstin välille yksittäisten n-grammien esiintymien tulee jakaantua; tässä tapauksessa kunkin n-grammin tulee siis löytyä vähintään viidestä eri tekstistä. Tällaista hakukriteeriä käyttämällä pyritään välttämään yhden kielenkäyttäjän kielen ominaispiirteiden heijastuminen hakutuloksiin (Biber ym. 1999: 992–993). 4-grammihaun osalta edellisistä arvoista muutettiin ainoastaan n-grammin minimi- ja maksimikoon arvot kolmesta neljään. 5- ja 6-grammit ovat kuitenkin Biberin ym. (mts. 993) mukaan huomattavasti 3- ja 4-grammeja harvinaisempi,

jolloin niitä haettaessa frekvenssiä oli syytä löysentää. Niiden tapauksessa minimifrekvenssi pudotettiinkin viiteen – suhteutettuna siis 12 esiintymään miljoonassa saneessa. Biber ym. (mts. 992–993) madalsivat tutkimuksessaan 5- ja 6-grammien raja-arvot puoleen 3- ja 4-grammeista eli viiteen esiintymään miljoonassa saneessa. Olettaen kuitenkin, että kriteeri viidessä eri tekstissä esiintymisestä pidetään ennallaan, ei minimifrekvenssiä olisi tässä tutkimuksessa enää viidestä alaspäin hyödyttänyt laskea, sillä esiintyessään alle viidessä eri tekstissä n-grammi ei pääse ylittämään Range-arvoa 5 eikä täten ilmene hakutuloksissakaan.

Arvojen määrittelyn jälkeen aloitettiin varsinaiset n-grammihaut B1-aineistosta, jotta voitaisiin saada vastaus tutkimuksen ensimmäiseen tutkimuskysymykseen eli siihen, mitkä ovat B1-aineiston frekventeimmät n-grammit. Jo ensimmäinen 3-grammihaku paljasti aineistosta nousevan esiin n-grammeja, jotka koostuivat pelkästään tekstien yhteydessä ilmoitetuista metatiedoista kuten alla olevassa kotekstiesimerkissä[30]:

1) käytetty ***apuvälineitä, mitä: kyllä***: sanakirja, kielioppi, oppikirja

Vaikka *AntConcin* asetuksissa olikin määritelty, että erilaisia meta- ja annotointitietoja ei tule huomioida korpushakujen tuloksissa, eivätkä yllä olevat n-grammit kuuluisi täten hakutuloksiin lainkaan, nousi joitain tällaisia kaikesta huolimatta siis esiin. Tarkastelemalla hakutuloksia niiden laajemmissa esiintymiskonteksteissaan File View -työkalun avulla huomattiinkin, että niiden sisältämistä metatiedoista oli jäänyt pois nuolimerkkejä (<, >). Merkit kertovat metatietojen alkamisesta ja loppumisesta. Samalla ne ilmoittavat *AntConcille* sen, ettei niiden sisällä olevia tietoja tule sisällyttää hakutuloksiin mukaan, mikäli metatiedot on asetuksissa määritelty piilotettaviksi. Ongelman ratkaisemiseksi alkuperäiset tekstitiedostot etsittiin B1-aineiston kansioista, ja niihin lisättiin puuttuvat nuolet, jotta edeltävän kaltaiset n-grammit eivät enää ilmenisi aineistossa. Ongelma koski yhteensä viittä eri tekstitiedostoa. Näiden selkeästi tarpeettomien n-grammien eliminoimisen jälkeen haku tehtiin uudelleen edellä mainituin kriteerein 3-grammeille ja tämän jälkeen myös 4-, 5- ja 6-grammeille. Tulokseksi saatiin seuraavat määrät erilaisia n-grammeja:

3-grammit: 983
4-grammit: 158
5-grammit: 154
6-grammit: 51
Yhteensä: 1 346 n-grammia.

[30] Numeroiduissa kotekstiesimerkeissä *AntConcin* ilmoittama n-grammi on tästä eteenpäin selkeyden vuoksi aina sekä *kursivoitu* että **tummennettu**.

3-grammeja oli siis aineiston alustavissa hakutuloksissa huomattavasti eniten, yhteensä noin 73 prosenttia kaikista n-grammeista. Prosentuaalisilta osuuksiltaan saadut n-grammimäärät tukevat pitkälti Biberin ym. (1999: 993) havaintoja, joiden mukaan ainakin sekä keskusteluissa että akateemisissa teksteissä 3-grammeja esiintyisi suunnilleen 10 kertaa enemmän kuin 4-grammeja.

3.3.3 Tulosten karsinta

Jo pelkästään silmämääräisesti n-grammihakujen antamia tuloksia tarkastelemalla pystyttiin havaitsemaan, että saadut n-grammilistaukset pitää käydä vielä manuaalisesti läpi, sillä ne sisälsivät paljon erilaisia epäkohtia, jotka täytyi karsia pois, jotta ne eivät haittaisi tutkimuksen varsinaista analyysia ja jotta hakutulokset suodattuisivat määrällisestikin helpommin käsiteltäviksi. Jokainen *AntConcin* muodostama n-grammilista vietiinkin omana tekstitiedostonaan Microsoftin *Excel*-taulukkolaskentaohjelmaan jatkotarkastelua ja tarpeettomien n-grammien poistoa varten.

Ensimmäisenä listoista poistettiin kaikki sellaiset n-grammit, jotka olivat sekä semanttiselta että syntaktiselta sisällöltään lähes tai täysin olemattomia. Näitä olivat esimerkiksi (lähes) yksinomaan luvuista[31] koostuneet n-grammit, kuten päivämäärät ja kellonajat[32], ja n-grammit, jotka muodostuivat toistensa perään ladelluista sanoista, joilla ei ollut suomenoppijan tuottaman kielen kannalta mitään järin merkittävää funktiota, kuten ”Kirje Joulupukille” -tehtävätyypissä käytetystä Joulupukin postiosoitteesta.[33] Näiden karsintojen jälkeen poistettiin myös kaikki sellaiset n-grammit, jotka koostuivat sanoista, jotka olivat jakaantuneet kahden virkkeen molemmille puolille (esim. 2–4), sillä pitäytymistä yhden virkkeen sisällä käytetään useissa muissakin tutkimuksissa yhtenä n-grammin kriteereistä (ks. esim. Biber ym. 1999: 993; Culpeper & Kytö 2010; Salazar 2014). Toinen samaa luokkaa oleva ilmiö olivat kahden kappaleen rajalle jakaantuvat n-grammit. Pääosin ne olivat oppijan tehtävänä olleen kirjoitelman otsikon ja leipätekstin alun yhdistelmästä koostuvia n-grammeja (5–6). Myös tekstien – usein mitä ilmeisimmin valmiiksi annetuista – otsikoista muodostuneet n-grammit poistettiin (7).

---

[31] Koska kaikki luvut korvattiin prosenttimerkillä, oli esimerkiksi *[%] [%] [%]* – toisin sanoen kolmesta peräkkäisestä mistä tahansa luvusta koostunut 3-grammi – alkuperäisessä listauksessa kaikista 3-grammeista toiseksi frekventein. Tällaisia 3-grammeja muodostui useita esimerkiksi genreltään päiväkirjaksi määriteltävissä teksteissä käytetyistä päivämääristä.

[32] Esim. *[%] joulukuuta [%]*; *kello [%] [%]*.

[33] Tästä muodostui muun muassa 6-grammi *Joulupukin Pääposti, Joulupukin Pajakylä, [%] Napapiiri*.

2) Sen jälkeen laitan ***aamiaisen. Aamiaiseksi syön*** puuroa ja makkaravoileipää.
3) Inna on ***[%] vuotta vanha. Hän*** on paras työskentelija isäntäväen mielestänsä.
4) Hänellä on ruskea ***tukka ja siniset silmät. Hän*** harrastaa maalamista ja on kovin nuorekas.

5) ***Elämäkerta***
Olen syntynyt Tallinnassa.

6) ***Perheeni***
Minun perheeni *ei ole* suuri.

7) ***Kansainvälistyminen uhkaa eri maiden omaa kulttuuriperinnettä***

Aina pelkistä n-grammeista ei voinut välttämättä suoralta kädeltä päätellä, kuuluvatko ne yhteen vai useampaan virkkeeseen; *AntConc* ei näyttänyt antamissaan hakutuloksissa erikseen isoja kirjaimia tai välimerkkejä, jotka olisivat luonnollisesti helpottaneet päättelyä. Isojen kirjainten puute johtui ohjelman Tool Preferences -asetuksissa valitusta kohdasta "Treat all data as lowercase", joka muuttaa kaikki kirjaimet hakutuloksissa automaattisesti pieniksi. Tämän kohdan olisi voinut jättää myös valitsematta, jolloin hakutuloksissa olisi eroteltu isot kirjaimet pienistä, mikä olisi osaltaan nopeuttanut virkerajoilla olevien n-grammien havaitsemista ja tunnistamista. Ongelmaksi tässä olisi kuitenkin muodostunut se, että tällaisessa tapauksessa *AntConc* lukee yksilöllisiksi n-grammeiksi kaikki nekin n-grammit, jotka eroavat toisistaan pelkästään kirjasinkokojen perusteella. Esimerkiksi seuraavat *laitan aamiaisen ja keitän* 4-grammit olisi siis täten luokiteltu kahdeksi eri 4-grammiksi, joilla molemmilla olisi ollut omat isosta ja pienestä *l*-kirjaimesta riippuvat esiintymisfrekvenssinsä:

8) ***Laitan aamiaisen ja keitän*** teetä.
9) Ensiksi ***laitan aamiaisen ja keitän*** kahvia.

Tällainen vääristäisi osaltaan tuloksia sekä esittämällä samoja n-grammeja useilla eri frekvensseillä että poistamalla joitain n-grammeja tutkimuksesta mahdollisesti kokonaan, sillä ne eivät enää useampaan eri luokkaan jaoteltaessa yltäisi määritellyn raja-arvofrekvenssin yläpuolelle. Sama pätee myös niin pilkkujen kuin muidenkin välimerkkien sekä lopetusmerkkien käyttöön hakujen osana: ne huomioimalla esimerkiksi virkkeissä "*Asunnossa on kaksi huonetta.*" ja "*Asunnossa on kaksi huonetta,* keittiö, kylpyhuone ja vessa." olevat (kursivoidut) 4-grammit olisi luokiteltu kahdeksi eri 4-grammiksi, sillä toinen päättyy pisteeseen ja toinen pilkkuun. Toisaalta tämän ominaisuuden avulla olisi toki mahdollista tutkia ja selvittää esimerkiksi yksinomaan virkkeitä aloittavia n-grammeja, kuten Li (2016) teki tutkimuksessaan kiinalaisten, kiinaa ensikielenään puhuvien, ja uusiseelantilaisten, englantia ensikielenään puhuvien, jatko-

opiskelijoiden englanninkielisistä akateemisista teksteistä ja niiden eroista n-grammien näkökulmasta. Tämän kaltaiset seikat jäävät tässä yhteydessä kuitenkin mahdollisen toisen aihetta koskettavan tutkimuksen selvitettäväksi.

Joka tapauksessa virkerajalle kuulumisen suhteen monitulkintaisten n-grammien kohdalla pystyttiin apuna käyttämään *AntConcin* Concordance-työkalua. Työkalun avulla päästiin tarkastelemaan ohjelman hakutuloksissaan listaamia n-grammeja niiden todellisissa esiintymiskonteksteissa (KWIC), eli kyettiin näkemään suoraan, esiintyvätkö ne yhdessä vai kahdessa virkkeessä. Esimerkiksi 3-grammin *kerroksessa meillä on* voisi ajatella kuuluvan osaksi vaikkapa lausetta ”Toisessa *kerroksessa meillä on* kaksi makuuhuonetta”, mutta konkordanssinäkymä paljasti, että tämä 3-grammi esiintyi kuitenkin joka kerta seuraavan kaltaisesti kahden virkkeen rajalla:

10) Asumme kolmenessa ***kerroksessa. Meillä on*** pitkä käytävä, missä on vaatekaapi ja monta peiliä.[34]

Edellä esitettyjen, suhteellisen yksitulkintaisten, karsimisten kautta B1-kielitaitotason n-grammiaineiston n-grammien määrä väheni 1 127:ään. Näillä kriteereillä seulottaessa n-grammeja poistui aineistosta siis yhteensä 219 kappaletta eli noin 16 prosenttia.

Kun yksiselitteisimmät poistot oli tehty, oli seuraavana vuorossa n-grammilistojen karsiminen edelleen siten, että niistä poistettiin vielä sellaiset lyhyemmät n-grammit, jotka toistuivat (lähes) aina täysin samanlaisina pidempien n-grammien osina. Biberin ym. (1999: 993) mukaan pidemmät n-grammit muodostuvat usein yksinkertaisesti lyhyempien n-grammien laajennetuista muodoista ja/tai useamman lyhyemmän n-grammin keskinäisestä yhdistymisestä. Tämänkaltaisista muodostumisista he antavat seuraavan esimerkin:

do you want; you want me; want me to; me to do →
do you want me; you want me to; want me to do →
do you want me to; you want me to do →
do you want me to do (Biber ym. 1999: 993).

Esimerkissä *do you want me to do* -6-grammi pitää siis sisällään useita 3-, 4- ja 5-grammeja ja aiheuttaa niiden kanssa merkittäviä päällekkäisyyksiä. Tämän tutkimuksen aineistossa esimerkiksi 5-grammin *jälkeen laitan aamiaisen ja keitän* frekvenssi oli 5 ja 6-grammin *sen jälkeen laitan aamiaisen ja keitän* frekvenssi oli 5. Tästä voidaan todeta, että 5-grammi *jälkeen laitan aamiaisen ja keitän* esiintyy tässä aineistossa aina osana 6-grammia *sen jälkeen laitan*

[34] Kyseinen n-grammi siis täten poistettiin aineistosta.

aamiaisen ja keitän, jolloin molempien n-grammien sisällyttäminen analysoitavien n-grammien joukkoon olisi jokseenkin merkityksetöntä ja toisi ainoastaan lisäkuormaa analyysille. Lyhyempi n-grammi poistettiinkin täten tutkimusaineistosta turhana toistona. Myös esimerkiksi *on vaalea tukka ja siniset* 5-grammi esiintyi aineistossa seitsemän kertaa kuten *on vaalea tukka ja siniset silmät* 6-grammikin, joten 5-grammi poistettiin jälleen aineistosta.

Käytännössä lyhyempien ja pidempien n-grammien vertailu ja pidemmissä n-grammeissa toistuvien lyhyempien n-grammien karsiminen tapahtui siten, että *Excelissä* vertailtiin keskenään kahta toisiaan n-grammin pituudessa seuraavaa n-grammilistaa aloittaen 3- ja 4-grammilistojen vertailusta. Tarkastelu aloitettiin 4-grammilistan frekventeimmästä 4-grammista, jolle haettiin yhtä sanaa lyhyempiä vastineita 3-grammiaineistosta. Tästä edettiin 4-grammi kerrallaan 4-grammilistan loppuun, minkä jälkeen siirryttiin vertailemaan vastaavalla menetelmällä 4- ja 5-grammilistoja toisiinsa ja tämän jälkeen vielä 5- ja 6-grammeja keskenään. Kun 4-grammina oli siis esimerkiksi *mukava ja onnellinen perhe*, haettiin 3-grammiaineistosta 3-grammeja *mukava ja onnellinen* sekä *ja onnellinen perhe*. Löydettyjen 3-grammien frekvenssejä verrattiin sitten 4-grammin frekvenssiin. Tällä tavoin pystyttiin näkemään, esiintyykö välittömästi 3-grammin *mukava ja onnellinen* jälkeen aina substantiivi *perhe* tai edeltääkö 3-grammia *ja onnellinen perhe* kaikissa tapauksissa adjektiivi *mukava*. Mikäli 3- ja 4-grammien frekvenssi oli sama tai todella lähellä ja mikäli vielä tarpeen vaatiessa tehty konkordanssitarkastelu osoitti n-grammien välillä selkeän ja lähes poikkeuksettoman yhtäläisyyden, karsittiin 3-grammi aineistosta pois.

Esimerkiksi edeltävän esimerkin tapauksessa 3-grammit *mukava ja onnellinen* sekä *ja onnellinen perhe* esiintyivät 3-grammiaineistossa kumpikin 12 kertaa. Myös 4-grammin *mukava ja onnellinen perhe* frekvenssi oli 12, eli 3-grammit poistettiin listalta. 4- ja 5-grammeja keskenään vertailtaessa 5-grammilistalta taas löytyi 5-grammi *oikein mukava ja onnellinen perhe*, jonka frekvenssi oli kuitenkin ainoastaan 7. Täten 4-grammia *mukava ja onnellinen perhe* ei eliminoitu. Konkordanssinäkymän tarkastelu osoitti tässä tapauksessa, että kyseistä 4-grammia edelsi aineistossa *oikein*-astemääritteen lisäksi vaihtelevissa määrin myös astemääritteet *tosi* ja *melko* sekä *olla*-verbi.

Aineistosta ei poistettu kuitenkaan pelkästään tasan yhtä usein pidemmissä n-grammeissa tavattavia lyhyempiä n-grammeja. Esimerkiksi 3-grammi *asunnossa on kaksi* esiintyi aineistossa 12 kertaa, ja 4-grammi *asunnossa on kaksi huonetta* kymmenen kertaa. Siitä huolimatta 3-grammia *asunnossa on kaksi* ei kuitenkaan säilytetty omana n-gramminaan tutkimusaineistoon, sillä konkordanssirivien tarkastelu paljasti, että kaksi poikkeavaa 3-grammin esiintymää

olivat kontekstissaan seuraavanlaiset: ***asunnossa on kaksi** *huoneet*[35] ja ***asunnossa on kaksi** *isoa ikkunaa*. Käytännössä siis yhden muista poikkeavan esiintymän takia olisi ollut tarpeetonta ottaa tutkimukseen mukaan erillisinä sekä 3-grammia *asunnossa on kaksi* että 4-grammia *asunnossa on kaksi huonetta*, joten 3-grammi poistettiin aineistosta. Myös muissa vastaavissa epävarmoissa tapauksissa hyödynnettiin konkordanssirivien tarkastelua. Esimerkiksi 3-grammin *olen syntynyt vuonna* frekvenssi aineistossa oli 36, siinä missä 4-grammin *olen syntynyt vuonna [%]* frekvenssi oli ainoastaan 30. Kuuden ”ylimääräisen” osuman joukossa oli kuitenkin seuraavan kaltaisia fraaseja, jotka edustavat samaa asiaa kuin *olen syntynyt vuonna [%]* -4-grammi, joten 3-grammi poistettiin aineistosta:

11) ***Olen syntynyt vuonna*** tuhatyhdeksänsataakahdeksankymmenenviisi Tartossa.

Tässä alaluvussa on rajattu tutkimusaineisto koko ICLFI-korpuksesta ainoastaan sen B1-taitotason teksteihin ja saatu alustavilla menetelmällisillä valinnoilla koostettua siitä yhdenlaiset n-grammilistat. Toinen tutkija voisi päätyä tekemään erilaisia ratkaisuja esimerkiksi sen suhteen, mitä hakuehtoja n-grammien etsimisessä tulisi hyödyntää tai mitä saaduista n-grammeista pitäisi karsia, joten täysin samanlaista n-grammilistaa ei välttämättä toisessa tutkimuksessa saataisi aikaiseksi. Tässä luvussa on kuitenkin pyritty tukeutumaan sellaisiin metodologisiin valintoihin, jotka olisivat mahdollisimman tarkoituksenmukaisia tämän tutkimuksen kannalta, ja pyritty kuvaamaan ne tarpeeksi yksityiskohtaisesti. Seuraavassa luvussa esitetään, minkälaisiin määrällisiin tuloksiin tässä luvussa kuvatuin B1-aineiston n-grammihauin ja -poistoin päästiin lopulta, eli toisin sanoen kerrotaan, mitkä ovat B1-aineiston frekventeimmät n-grammit.

[35] Asteriski osoittaa kielenvastaista muotoa.

4 OPPIJANSUOMEN B1-TAITOTASON FREKVENTEIMMÄT N-GRAMMIT

Tässä luvussa vastataan tutkimuksen ensimmäiseen tutkimuskysymykseen eli siihen, mitkä ovat kaikkein taajimmin oppijansuomen B1-kielitaitotasolle arvioiduissa teksteissä esiintyvät n-grammit. Niihin päästiin käsiksi edellisessä luvussa kuvatuin menetelmin. Luvussa luodaan myös yleisluontoinen katsaus listaan, joka frekventeimmistä n-grammeista muodostui, ennen seuraaviin lukuihin ja tutkimuksen varsinaiseen, oppijansuomen leksikkoa ja rakennepiirteitä n-grammien kautta kartoittavaan analyysiin siirtymistä.

4.1 Lista tutkimuksen analyysiin päätyneistä n-grammeista

Luvussa 3.3 esitettyjen poistotoimien jälkeen ICLFI:n B1-aineiston lopulliset, tutkimuksen varsinaiseen analyysiin päätyneet n-grammit, näyttävät kappalemäärällisesti ja prosentuaalisesti ilmaistuina seuraavan kaltaisilta:

Taulukko 1. B1-aineistosta varsinaiseen analyysiin päätyneiden n-grammien kappalemäärät ja prosenttiosuudet.

	Yksilölliset n-grammit	**% kaikista n-grammeista**	**N-grammien esiintymät aineistossa**	**% kaikista n-grammiesiintymistä**
3-grammit	804	81	14 502	87,3
4-grammit	102	10,3	1 456	8,8
5-grammit	61	6,2	481	2,9
6-grammit	25	2,5	175	1,0
Yhteensä	992	100	16 614	100

Lopullinen tutkimuksessa analysoitavien uniikkien n-grammien lukumäärä on siis 992. Taulukon 1 *N-grammien esiintymät aineistossa* -sarake kertoo sen, mikä on kunkin pituisten n-grammien kaikkien aineistoesiintymien yhteisfrekvenssi. Esimerkiksi 4-grammi *olen lapsuudesta asti harrastanut* on uniikki 4-grammi, joka esiintyy aineistossa yhteensä 15 kertaa. Täten se kasvattaa yksilöllisten 4-grammien saraketta yhdellä ja 4-grammien esiintymämäärän saraketta 15:llä. Kuten taulukosta nähdään, 3-grammien osuus tutkittavasta aineistosta on poistojen jälkeenkin huomattavan suuri, siinä missä 4-, 5- ja 6-grammit muodostavat yhteen

laskettuinakin ainoastaan 19 prosenttia n-kaikista grammeista ja 12,8 prosenttia kaikista n-grammiesiintymistä.

Seuraavalle sivulle taulukkoon 2 on listattu kaikki ICLFI:n B1-kielitaitotasolle arvioiduissa teksteissä yli 30 kertaa esiintyvät 3-, 4-, 5- ja 6-grammit. Niitä on yhteensä 92 kappaletta. Taulukon *jakauma*-sarakkeen arvot kertovat, monenko eri tekstin välille kunkin n-grammin esiintymät ovat aineistossa jakautuneet. Aineiston laajuuden vuoksi kaikkia saatuja n-grammeja (n = 992) ei olisi mielekästä listata tutkimuksen keskelle, joten tässä yhteydessä on esitetty ainoastaan verrattain taajaan tutkimusaineistossa tavattavat n-grammit. Loput n-grammit löytyvät aivan tutkimuksen lopusta liitteestä 3. Saatua n-grammilistaa tarkastellaan pintapuoleisesti seuraavassa alaluvussa, jossa siitä tehdään joitain yleisluontoisia havaintoja. Saatuihin n-grammeihin perehdytään tarkemmin niiden leksikon ja rakenteiden osalta luvuissa 5 ja 6.

Taulukko 2. Yli 30 kertaa tutkimusaineistossa esiintyvät 3-, 4-, 5- ja 6-grammit esiintymistaajuutensa mukaisesti järjestettyinä frekventeimmästä n-grammista alkaen.

	f	jakauma	n-grammi
1	210	198	minulla ei ole
2	155	112	ja hän on
3	152	129	se ei ole
4	106	95	että se on
5	88	80	koska se on
6	79	74	ja se on
7	79	73	ja sen jälkeen
8	78	77	on pieni mutta
9	77	70	ei ole niin
10	77	73	minulla on kaksi
11	75	73	minulla on myös
12	75	69	mutta se on
13	73	67	mutta hän ei
14	72	71	minulla on yksi
15	71	52	[%] vuotta vanha
16	70	68	minun huone on
17	69	63	mutta se ei
18	67	67	opiskelen tarton yliopistossa
19	64	63	käyn suihkussa ja
20	63	62	ja juon kahvia
21	62	37	hän on [%]-vuotias
22	62	55	minusta se on
23	60	52	että hän on
24	57	48	koska hän on
25	57	53	minä en ole
26	52	45	sen jälkeen menen
27	50	30	on [%] vuotta
28	47	42	koska minulla on
29	47	30	on [%]-vuotias ja
30	46	41	minulla on vielä
31	45	41	hän ei ole
32	45	43	ikkunan vieressä on
33	45	45	voileipää ja juon
34	45	45	vuonna [%] ja
35	44	43	ei ole paljon
36	44	41	meillä ei ole
37	44	44	menen nukkumaan kello
38	43	40	minulla on paljon
39	43	43	minun nimeni on
40	43	42	se on hyvä
41	41	40	mutta minulla on
42	41	40	peseydyn ja pukeudun
43	41	41	pieni mutta viihtyisä
44	40	37	että minulla on
45	40	40	ja juon teetä
46	40	40	ja menen nukkumaan
47	40	38	meillä on myös
48	40	38	menen kotiin ja
49	39	33	että hän ei
50	39	39	huone on pieni
51	39	30	hän on töissä
52	39	39	on kaksi huonetta
53	38	38	aamiaiseksi syön tavallisesti
54	38	31	vuotta vanha ja
55	37	33	hänen nimensä on [name]
56	37	36	kylpyhuone ja vessa
57	37	31	minä ja minun
58	37	37	on kolme huonetta
59	37	37	oven lähellä on
60	37	35	se on niin
61	36	34	ei ole mitään
62	36	34	että se oli
63	36	22	hän on [%]
64	36	34	hän on myös
65	36	27	hän sanoi että
66	36	34	ja hänellä on
67	35	31	hänellä ei ole
68	35	33	koska siellä on
69	35	32	mutta minä en
70	35	35	syntynyt vuonna [%]
71	34	32	en ole ollut
72	34	34	laitan aamiaisen ja
73	34	34	minulla on pieni
74	34	28	minä luulen että
75	33	29	jonka nimi on
76	33	33	nimeni on [name]
77	33	33	on pieni mutta viihtyisä
78	32	31	on iso ja
79	32	31	on kaunis ja
80	32	25	sanoi että hän
81	32	31	se on tosi
82	31	27	alkaa vartin yli
83	31	28	koska hän ei
84	31	30	luulen että se
85	31	29	sitten menen kotiin
86	31	28	sitä mieltä että
87	30	30	aamiaisen ja keitän
88	30	29	ei ole aikaa
89	30	30	minun huone on pieni
90	30	30	olen syntynyt vuonna [%]
91	30	29	siellä ei ole
92	30	28	äiti ja isä

4.2 Yleisluontoisia huomioita laaditusta n-grammilistasta

Kuten jo luvussa 3.3 kävi ilmi, 3-grammit ovat B1-aineistossa muita n-grammeja huomattavasti tyypillisempiä. Saatua B1-aineiston lopullista n-grammilistaa tarkastelemalla voidaan tehdä havainto siitä, että ne hallitsevat vahvasti myös listan frekventeimpien n-grammien päätyä. Edellisen sivun taulukosta 2 nähdään, että 92:sta frekvenssin 30 ylittävästä n-grammista ainoastaan neljä on 4-grammeja. Ne ovat 4-grammit *hänen nimensä on [name]*[36] (f = 37), *on pieni mutta viihtyisä* (f = 33), *minun huone on pieni* (f = 30) ja *olen syntynyt vuonna [%]* (f = 30). Loput taulukon 88 n-grammia ovat kaikki 3-grammeja; 5- tai 6-grammeja ei frekventeimpien 30 n-grammin joukkoon päätynyt siis ensimmäistäkään. Ensimmäinen viidestä sanasta koostuva n-grammi kaikki n-grammit kattavalla listalla (liite 3) on sijalla 204 oleva 5-grammi *minun huone on pieni mutta* frekvenssillä 20. Ensimmäinen 6-grammi taas on *minun huone on pieni mutta viihtyisä*, joka on yksi edeltävän 5-grammin täydennysmahdollisuuksista[37]. Se löytyy frekvenssillä 14 vasta listan sijalta 412.

Toinen yleisemmän luokan huomio, joka saadusta n-grammilistasta voidaan tehdä, on se, että sen sisältämät n-grammit ovat hyvin suurelta osin kielenmukaisia muotoja, eli erilaiset kielivirheet, kömmähdykset tai lipsumiset eivät niissä juurikaan ilmene saati korostu. Tämä on järkeenkäyvääkin: kuten on jo todettu, jotta sanaketju voidaan määritellä n-grammiksi, tulee sen olla usein toistuva, jolloin eri tavoin toisistaan poikkeavat virheelliset sananmuodot tai esimerkiksi vaihtelevalla tavalla erheellisesti muodostetut rakenteet eivät luonnollisestikaan synnytä niin helposti n-grammeja. Virheelliset n-grammit ovatkin siinä määrin hedelmällinen tarkastelun kohde, että niiden kautta on mahdollista saada käsitys sellaisista toistuvista virheistä, joita useampi suomenoppija kirjoitetussa kielessään tekee. Tiettyjen oppijaryhmien vaikeuksien identifiointi onkin merkittävässä asemassa oppijankielen korpuksissa ja niiden hyödyntämisessä (Nesselhauf 2004: 126).

Vaikka niin kutsutulla virheanalyysilla on varsin pitkät – joskaan eivät aivan ongelmattomiksi nähdyt – perinteet yhtenä oppijankielen analyysimetodina (ks. esim. Corder 1981; R. Ellis & Barkhuizen 2005: 50–53), ei tässä tutkimuksessa kuitenkaan perehdytä n-grammien virheellisyyksiin kovinkaan syvällisesti. Niistä merkittävimmät ja toistuvimmat on kuitenkin mainittu alla. On myös hyvä huomata, että onnistuneet ja kielenmukaiset n-grammitkaan eivät

[36] Tunnistettavat nimitiedot oli korvattu merkinnällä *[name]* jo alkuperäisessä ICLFI-aineistossa.
[37] 5-grammin *minun huone on pieni mutta* frekvenssi aineistossa on 20, 6-grammin *minun huone on pieni mutta viihtyisä* taas 14. Vaihtoehtoisia täydennyksiä 5-grammille *minun huone on pieni mutta* ovat adjektiivin *viihtyisä* lisäksi muun muassa adjektiivit *mukava* ja *kodikas*.

väistämättä tarkoita, että niitä olisi käytetty oikein niissä laajemmissa koteksteissaan, joissa ne esiintyvät. Seuraavassa konkordanssissa 2 on esitetty itsessään täysin kielenmukaisen 3-grammin *hän ei tarvitse* esiintymät B1-aineistossa. Konkordanssirivejä tarkastelemalla voidaan kuitenkin havaita, että kyseistä n-grammia ei ole kaikissa tapauksissa käytetty odotuksenmukaisella ja oikealla tavalla. Tarkemmin ottaen virheellinen käyttötapa ilmenee riveillä 2, 3, ja 7, joissa subjektina toimiva *hän*-pronomini on nominatiivimuotoinen, vaikka *tarvita*-verbin avulla muodostettu nesessiivi-ilmaus vaatii subjektin sijaksi genetiivin (ks. VISK § 906)[38].

Hit	KWIC			File
1	o hän Myyn hiihtämään, mutta Myy sanoo että	*hän ei tarvitse*	apua. Helmuti sanoo että hän hal	HO0018.txt
2	si. Hän on yhtä rohkea kuin pohjat, mutta siksi	*hän ei tarvitse*	kiivetä taivaseen. - Viime talvi oli	SA0067.txt
3	inen siirtomies, joka asuu Pariisin keskustassa.	*Hän ei tarvitse*	käydä työssä, koska hän saa rahaa	ES0028a.txt
4	stoksilla, ostamalla muutamien asioiden joiden	*hän ei tarvitse*	ollenkaan, ja sitten hän ajattele hä	IS0002f.txt
5	Hän ei pyydystä hiiriä mutta kehrää koko ajan.	*Hän ei tarvitse*	oma paikka kuin mappea. Svante t	KI0005h.txt
6	yytä hiihtämään, mutta Pikku Myy sanoo että	*hän ei tarvitse*	opetusta. Sitten Hermuli opettaa M	HO0021.txt
7	puutarha, jossa hän vilkelee vihanneksia. Siksi	*hän ei tarvitse*	ostaa paljon ruokaa. Myös minun	SA0009c.txt
8	ngas kengät. Gotlannissa ei ole talvi vielä joten	*hän ei tarvitse*	paksuja vaatteita. Hänen perhe asu	RU0012b.txt
9	. Asunto on todellisesti minun isoisänsä, mutta	*hän ei tarvitse*	sitä tällä hetkellä ja mahdollistaa m	VI0234a.txt

Konkordanssi 2. 3-grammin *hän ei tarvitse* esiintymät (n = 9) B1-aineistossa.

Ensimmäiset eli frekventeimmät B1-aineiston n-grammeista suoraan heijastuvat virheet ovat luokiteltavissa morfosyntaktisiksi virheiksi sanojen viittaussuhteissa (ks. ICLFI-manuaali 2016: 11), sillä ne koskevat genetiivimuotoisia persoonapronomineja seuraavista nomineista uupuvia possessiivisuffikseja. Esimerkki tästä löytyy jo listan sijalla 16 olevasta 3-grammista *minun huone on* (f = 70) (pro *minun huoneeni on*). Muita variaatioita samasta virhetyypistä ovat muun muassa 3-grammit *minun perheessä on* (f = 18) ja *minun herätyskello soi* (f = 12). Possessiivisuffiksin poisjättäminen on toki erityisen tyypillistä puhekielessä (esim. VISK § 1300; Kankaanpää 2009), ja kuten luvussa 2.3.1 esitettiin, Jantunen (2008: 13–15) näkee yleis- ja puhekielen sekoittumisen olevan jopa yksi oppijankielen universaaleja piirteitä. Etenkään suomi toisena kielenä -opetuksessa possessiivisuffiksilla ei edes ole järin keskeisestä asemaa, eikä suomenoppijan ensikielikään usein varsinaisesti tue liitteen oppimista (Siitonen & Mizuno 2010). Kyseessä ei täten ehkä olekaan sellaisenaan kaikkein merkittävin virhetyyppi, vaikka toki yleiskielisiä tekstejä, jollaisiksi ICLFI:n tuotoksetkin on pääosin tarkoitettu, kirjoitettaessa omistusliitteitä tulisi käyttää. Joka tapauksessa omistusliitteen puute ei aiheuta minkäänlaisia vaikeuksia ymmärtää n-grammien merkityksiä ainakaan tämän tutkimuksen n-grammien tapauksissa. Possessiivisuffikseihin voisi silti näiden alustavien n-grammilöydöstenkin

[38] Muiden konkordanssin 2 konkordanssirivien tapauksissa nominatiivi on *hän*-pronominille kielenmukainen muoto johtuen *tarvita*-verbin toisesta, ei-modaalisesta merkityksestä.

perusteella kiinnittää mahdollisesti vielä enemmän huomiota suomen kielen yleiskielisen variantin opetuksessa.

Omistusliitteiden puutetta huomioimatta seuraavaksi frekventein virheellinen n-grammi on frekvenssillään 24 3-grammi *minulla en ole* (pro *minulla ei ole*), jossa siis kieltoverbiä on taivutettu virheellisesti yksikön ensimmäisen persoonan mukaisesti. Tämä johtunee siitä, että lauseen subjektin on ajateltu (virheellisesti) olevan sen teemapaikalla oleva persoonapronomini *minulla*, jota jossain määrin voidaan pitääkin subjektimaisena (ks. VISK § 922). Esimerkkejä tämän n-grammin käytöstä on alla:

12) Minä en ole vielä naimisissa, mutta elän poikaystävän kanssa. Hän on minun-ikäinen, oikein huumorintajuinen ja vilkas poika niin että ***minulla en ole*** koskaan ikävä.
13) ***Minulla en ole*** paljon vapaa aikaa, koska työ ja opiskelu vie melkein kaiken aikani.

Tämän jälkeen seuraavat virheet löytyvät 3-grammeista *ja hänella on* (f = 15) ja *hänen nimensa on* (f = 14). Kummassakin tapauksessa kyse on siis virheestä vokaalisoinnussa. Seuraavat virheelliset n-grammit ovat 3-grammit *huonessa on oikealla* (f = 13) ja *huonessa on vasemalla* (f = 11). Nämäkin 3-grammit sisältävät keskenään samankaltaisen virhetyypin, pitkän vokaalin tai konsonantin lyhenemisen, joka jälkimmäisessä 3-grammissa koskee perätä molempia 3-grammin substantiiveista. Onkin mielenkiintoista, että siitä on kahdesta virheestään huolimatta muodostunut silti oma n-gramminsa. Sama virhetyyppi toistuu vielä muutamissa muissakin n-grammeissa, kuten *minun perheni on* (f = 9) ja *huonessa on oikealla sänky ja* (f = 6). Konsonanttien ja vokaalien kestosuhteet ovatkin tavanomainen haaste suomenoppijoille (esim. Ullakonoja & Dufva 2016: 9), ja näidenkin tulosten perusteella useampi oppija tekee keskenään täysin samanlaisia virheitä kestojen osalta. 3-grammista *peseytyn ja pukeutun* (f = 10), löytyy vielä oma virhetyyppinsä, astevaihteluvirhe, joka jälleen koskettaa kahta 3-grammin sanoista.

Muutoin taulukosta 2 voidaan havaita, että n-grammilistan kärkipäätä näyttäisivät hallitsevan n-grammit, jotka koostuvat pitkälti funktiosanoista ja erilaisista pronomineista. Koko liitteen 3 n-grammilistastakaan ei ole sen sijaan löydettävissä ainuttakaan n-grammia, jonka ainakaan itse voisin tulkita varsinaiseksi idiomiksi. Muutoinkaan n-grammit eivät sinällään vaikuttaisi kielivät erityisestä idiomaattisuudesta, joskaan eivät välttämättä epäidiomaattisuudestakaan. Etenkin 3-grammien osalta on hankala ottaa kantaa niiden idiomaattisuuteen; tähän tarvittaisiin vertailuun jonkin natiiviaineiston n-grammit. Jonkin verran rutiini-ilmauksiksi tai perusfraaseiksi tulkittavia n-grammeja listalta kuitenkin löytyy. Näitä ovat esimerkiksi 3-grammit *minun nimeni on* (f = 43) ja *ja niin edelleen* (f = 23) sekä 4-grammit *hänen nimensä on [name]*

(f = 37), *olen syntynyt vuonna [%]* (f = 30) ja *olen sitä mieltä että* (f = 21). Osassa pitemmistä n-grammeista vaikuttaa esiintyvän myös rakenteita, jotka on mahdollisesti kopioitu suoraan muista tekstiyhteyksistä. Näistä on luvassa nostoja tuonnempana, luvussa 6.1.3.

5 B1-OPPIJANSUOMEN LEKSIKKO N-GRAMMEIN ANALYSOITUNA

Tässä ja seuraavassa luvussa esitetään tutkimuksen varsinainen analyysiosuus. Analyysissa ilmenee myös tarkempia tutkimusmetodeja, joita ei vielä edellisessä luvussa esitelty. Tämä johtuu siitä, että käsillä olevan tutkimuksen tapauksessa menetelmien ja analyysin ero ei hahmotu kaikilta osin täysin selvärajaisena. Vaikka tutkimuksen päämenetelmänä voidaankin pitää korpusvetoista tutkimusta, on se silti, kuten luvussa 3.2 todettiin, lähinnä lähtökohta n-grammiaineiston koostamiseksi ICLFI:stä. Tämän vuoksi niiden tarkempien metodien, joiden kautta tutkimuksen n-grammeja lähestyttiin ja joiden avulla etenkin kvantitatiivista dataa kartoitettiin, kuvaus on yhdistetty analyysin yhteyteen, jotta vältyttäisiin jatkuvilta edestakaisilta viittauksilta analyysiluvuista menetelmälukuun ja päinvastoin. Koska yhtenä tutkimuksen tutkimuskysymyksistä on se, miten erilaiset tutkimusmenetelmälliset valinnat ohjaavat n-grammeista saatavan datan tulkintaa, vastaavat tämä ja seuraava luku siis omilta osin tutkimuksen viimeisen luvun yhteenvedon kanssa samalla tuohon kysymykseen.

Tässä ensimmäisessä analyysiluvussa käydään läpi, millaisin tavoin oppijansuomen B1-kielitaitotason 3-, 4-, 5- ja 6-grammeja, jotka koottiin yhdeksi listaksi luvussa 3 eritellyin keinoin, lähestyttiin tutkimuksessa sanatasolla sekä analysoidaan menetelmien avulla saatavia tuloksia. Luvun ensimmäisessä alaluvussa kerrotaan, miten n-grammeja tutkittiin niiden sisältämien sanojen sananmuotojen osalta, siinä missä toinen alaluku keskittyy lemmamuotojen tarkasteluun. Saatavia tuloksia peilataan samalla joihinkin edeltävistä oppijansuomen tutkimuksista ja niiden tuloksista.

5.1 Sananmuotoanalyysi

Varsinaisen tutkimuksen kohteeksi päätyneiden 3-, 4-, 5- ja 6-grammien yhdeksi listaksi koostamisen jälkeen (liite 3) listasta päätettiin tutkia ensimmäiseksi tarkemmin sitä, millaista sanastoa oppijansuomen B1-kielitaitotason n-grammit pitävät sisällään. Tämän suhteen lähdettiin liikkeelle n-grammien sanojen sananmuodoista, sillä sananmuodot voidaan nähdä n-grammeista suoraan ilman, että sanoja tulee ensin esimerkiksi lemmata, kuten jäljempänä tässä luvussa tehdään. Listaus n-grammeissa esiintyvistä sananmuodoista voidaan koostaa *AntConcin* Word List -toiminnolla. Sen avulla on mahdollista järjestää tutkittavan korpuksen saneet

sanalistaksi erilaisin tavoin, esimerkiksi sananmuotojen frekvenssien mukaisesti. *Sanalista* on siis yhdenlainen tapa ottaa teksti tai kokoelma tekstejä ja esittää se tai ne perinteisen lineaarisen muodon sijaan uudessa valossa, kokonaisten virkkeiden sijaan yksittäisiä sanoja sisältävänä luettelona (Scott & Tribble 2006: 12). Jantusen (2012) mukaan sanalistan tuottaminen on yksi helpoimpia tapoja aloittaa korpusvetoinen analyysi. Sanalistat kielivät muun muassa siitä, millaiset sanat esiintyvät aineistossa taajaan ja mitä teemoja aineistoon lukeutuvat tekstit käsittelevät. (Mts. 361.) Sanalistoille ei ole olemassa yhtä tavanomaista käyttötapaa, vaan tutkija voi hyödyntää niitä omien prioriteettiensa mukaisesti. Yleensä niillä pyritään kuitenkin lisäämään ymmärrystä käsittelyn alla olevan korpuksen tai muun tekstiaineiston leksikosta. (Scott & Tribble 2006: 30–31.)

Jotta sanalista voitiin luoda *AntConcilla*, tuli lopullisesta B1-aineiston n-grammilistasta koota oma ”korpuksensa” erilliseen tekstitiedostoon[39]. Tekstitiedosto vietiin *AntConciin*, joka laski kunkin siinä olevan sananmuodon kaikki esiintymät yhteen ja listasi sananmuodot frekvenssiensä mukaisesti laskevaan järjestykseen. Frekvenssin mukaan järjestetyllä sanalistalla kyetään tuomaan esiin tiettyjä tekstin erityispiirteitä (Scott & Tribble 2006: 15). N-grammien sananmuodoista koostettu sanalista paljastaa, että B1-aineiston 992 uniikissa n-grammissa on käytetty yhteensä 444 erilaista sananmuotoa. Seuraavan sivun taulukossa 3 on esitetty B1-aineiston n-grammeissa jokainen 10 kertaa tai useammin esiintyvä sananmuoto (n = 66) tiheimmin tavattavimmasta alkaen. Laajempi, kaikki vähintään viisi kertaa n-grammeissa esiintyvän sananmuodon kattava lista on nähtävissä liitteestä 1.

Niin alla olevan kuin liitteen 1 sanalistankaan tapauksessa frekvenssi ei siis tässä yhteydessä ilmaise sitä, kuinka monta kertaa tietty sananmuoto esiintyy yhteensä koko 16 599 n-grammiesiintymää kattavassa aineistossa, vaan ainoastaan sen, kuinka monessa yksilöllisessä n-grammissa sananmuoto esiintyy kaikista tutkimukseen päätyneistä 992 n-grammista. Vaihtelevia lukuja sekä tunnistettavia nimitietoja osoittavat *[%]* ja *[name]* -merkinnät on jätetty pois sekä tästä sananmuotolistasta että luvun 5.2 lemmalistasta.

[39] Tämä tekstitiedosto – näennäisesti siis korpus – oli toisin sanoen pelkkä listaus, joka sisälsi kaikki tutkimuksen 992 n-grammia.

Taulukko 3. B1-aineiston n-grammeissa vähintään 10 kertaa esiintyvät sananmuodot.

	f	sananmuoto			
1	391	on	34	16	noin
2	285	ja	35	16	tukka
3	88	hän	36	15	tai
4	79	ei	37	15	tavallisesti
5	74	se	38	15	vuotta
6	70	että	39	13	hänellä
7	66	ole	40	13	keittiö
8	63	minulla	41	13	kotiin
9	47	mutta	42	13	käyn
10	43	minä	43	13	ollut
11	41	minun	44	13	sänky
12	35	en	45	12	he
13	33	syön	46	12	huone
14	31	menen	47	12	huonetta
15	31	olen	48	11	alkaa
16	29	oli	49	11	kahvia
17	28	sen	50	11	kirjoituspöytä
18	27	kaksi	51	11	kolme
19	27	koska	52	11	laitan
20	26	jälkeen	53	11	nimi
21	26	paljon	54	11	oikealla
22	26	pieni	55	11	siellä
23	25	vuonna	56	11	silmät
24	24	kello	57	11	suihkussa
25	24	yksi	58	11	vanha
26	22	juon	59	11	vielä
27	21	iso	60	11	vieressä
28	19	myös	61	10	kun
29	18	puuroa	62	10	monta
30	17	meillä	63	10	nukkumaan
31	17	niin	64	10	samana
32	17	sitten	65	10	voi
33	17	voileipää	66	10	yli

Sanalistan kärkipäästä voidaan nähdä, että n-grammeissa käytetyimpien sananmuotojen frekvensseissä on suuria pudotuksia listalla sananmuodosta toiseen siirryttäessä. Kuten huomataan, ainoastaan kaksi frekventeintä sananmuotoa ylittää sadan esiintymisen rajan: *on* esiintyy yli kolmasosassa kaikista n-grammeista ja *ja* noin 29 prosentissa. Sananmuotoja, jotka tavataan ainoastaan yksittäisessä n-grammissa, on yhteensä 165 kappaletta. N-grammien sananmuotojen frekvenssien jakautuminen vaikuttaisikin toteuttavan *Zipfin lakina* tunnettua jakaumaa. Zipfin lain mukaan (luonnollisessa) kielessä kuin kielessä pieni määrä sanoja saa osakseen huomattavan paljon esiintymiä, siinä missä suuri määrä sanoja taas esiintyy kielessä vain harvoin. Näitä niin kutsuttuja zipfiläisiä jakaumia tavataan myös lukuisilla muilla elämänaloilla, ja ne voidaan kuvata matemaattisesti. (Ks. Zipf 1949; Scott & Tribble 2006: 26–29.) Sanalistan kärkipäässä on myös lukuisia funktiosanoja, kuten *ja*, *hän*, *että*, *koska*, *jälkeen*, *niin* ja *sitten*. Warrenin (2011: 159) mukaan onkin erittäin tavanomaista, että korpuksista luotujen sanalistojen kärkipäitä hallitsevat sekä kieliopilliset että funktiosanat.

Listaa silmämääräisesti tarkastelemalla pystytään toteamaan myös se, että B1-oppijansuomen tekstit näyttäisivät n-grammien sananmuotojen perusteella käsittelevän usein oppijoita itsejään; persoonapronomini *minä* esiintyy koko sananmuotolistassa viidessä eri taivutusmuodossa ja yhteensä 160 n-grammissa, ja myös lukuisat listan verbit näyttävät taipuneen yksikön ensimmäisen persoonan mukaan. Jotta tarkempaan dataan siitä, kuinka usein n-grammeihin sisältyy yksikön ensimmäisen persoonan finiittiverbi, päästään käsiksi, voidaan sanalistan sananmuodot järjestää frekvenssiensä sijaan aakkosten mukaisesti sanan lopettavien kirjainten perusteella. Siinä missä tavanomaiseen tapaan aakkostettu sanalista mahdollistaa luonnollisestikin frekvenssilistaa nopeammin tiettyjen kiinnostuksen kohteena olevien sanojen löytämisen listalta (Scott & Tribble 2006: 15), tarjoaa Jantusen (2012: 363–364) mukaan sanan lopun perusteella aakkostettu sanalista eli *käänteissanalista* potentiaalisen keinon tarkastella esimerkiksi eri taivutusmuotojen ja suffiksien käyttöä tutkittavassa aineistossa. Käänteissanalistalla voidaankin saada helposti numeerista tietoa siitä, kuinka usein B1-aineiston n-grammeissa käytetään yksikön ensimmäisessä persoonassa taipuneita verbejä. Huomio tulee tässä tapauksessa siis keskittää niihin verbien sananmuotoihin, jotka päättyvät persoonapäätteeseen *-n*. Ne kaikki on esitetty alla olevassa taulukossa 4.

Taulukko 4. Käänteissanalistan avulla löydetyt, kaikki B1-aineiston n-grammeissa esiintyvät persoonapäätteen *-n* sisältävät finiittiverbit frekvensseineen.

sananmuoto	**f**		
palaan	1	kirjoitin	6
laitan	11	otin	4
ostan	2	uin	2
en	35	uskon	1
lähden	5	katson	4
opiskelen	4	juon	22
kuuntelen	1	toivon	5
ajattelen	1	pukeudun	7
olen	31	puhun	1
luulen	7	asun	7
menen	31	pukeutun	1
pesen	4	peseydyn	3
nousen	3	peseytyn	1
luen	3	käyn	13
tein	1	tiedän	2
olin	4	pidän	3
menin	4	keitän	7
voin	3	herään	1
nousin	1	syön	33
pääsin	9	**Yhteensä**	**284**

Taulukon 4 kautta kyetään laskemaan, että yli neljäsosassa – noin 29 prosentissa – B1-aineiston n-grammeista esiintyy finiittiverbi, joka on taipunut yksikön ensimmäisessä persoonassa.[40]

Edeltävässä luvussa todettiin, että possessiivisuffiksien puute näyttäisi olevan jokseenkin toistuva virhetyyppi B1-taitotasoisten suomenoppijoiden n-grammeissa. Sananmuodon viimeisen kirjaimen perusteella aakkostetun sanalistan kautta onkin mahdollista tarkastella samalla myös sitä, kuinka usein n-grammeihin sisältyvissä sanoissa käytetään mitäkin omistusliitettä. Yksikön ensimmäisen persoonan omistusliitteen *-ni* sisältävät sananmuodot esiintymämäärineen on listattu taulukkoon 5.

Taulukko 5. Käänteissanalistassa esiintyvät *-ni*-possessiivisuffiksin sisältävät sananmuodot frekvensseineen.

sananmuoto	**f**
hampaani	2
vanhempani	3
huoneessani	1
perheeni	2
huoneeni	1
perheni	1
nimeni	3
rakkauteni	1
äitini	6
siskoni	1
isäni	6
mielestäni	3
ystäväni	1
Yhteensä	**31**

Sanalistalta ei löydy yllä olevan taulukon 5 sananmuotoja sekä kahta yksikön kolmannen persoonan omistusliite-esiintymää (*nimensä* ja **nimensa*) lukuun ottamatta muita nomineja, joissa olisi käytetty possessiivisuffikseja. Vaikka tällaisen sanalistan avulla ei voidakaan päästä suoraan käsiksi siihen, kuinka usein omistusliitteet jäävät suomenoppijoiden n-grammeista nimenomaan uupumaan, nähdään listasta silti, ettei possessiivisuffiksi sisälly järin moneen - yhteensä vain noin 3,3 prosenttiin – B1-aineiston n-grammeista ottaen huomioon, että esimerkiksi sananmuodon *minun* frekvenssi on jo yksinään 41. Tässä esitetyt käänteissanalistat ovatkin vain yksittäisiä, esimerkinomaisia mahdollisuuksia monista muista tutkia n-grammien leksikkoa sen

[40] Hakemalla käänteissanalistalta pelkästään *-n*-persoonasuffiksiin päättyviä verbejä ei tulla kuitenkaan huomioineeksi sitä mahdollisuutta, että verbi ilmentää yksikön ensimmäistä persoonaa, mutta päättyykin johonkin liitepartikkeliin, kuten *-pA*, *-kO*, tai *-kin*. Näitä etsittiinkin sanalistalta erikseen hakutoiminnolla, mutta niitä ei kuitenkaan löytynyt yksikön ensimmäisen persoonan verbien yhteydestä yhtäkään.

sananmuotojen kautta. Käänteisen sanalistan avulla olisi persoonapäätteiden ja omistusliitteiden ohella mahdollista kartoittaa vielä esimerkiksi siitä, mitkä suomen kielen sijamuodoista saavat eniten edustusta n-grammeissa.

Etenkin funktiosanojen suuren määrän vuoksi substantiiveja ja adjektiiveja yleisimpien sananmuotojen joukossa on vain vähän. Substantiiveista frekventeimmät ovat *vuonna*, *kello*, *puuroa* ja *voileipää* ja adjektiiveista *pieni*, *iso* ja *vanha*. *Vuonna* ja *kello* -sananmuodot liittyvät selkeästi n-grammeihin, joilla ilmaistaan jollain tapaa aikaa tai kerrotaan esimerkiksi, milloin tietty henkilö syntyi. Ehkä hieman yllättävätkin *puuroa* ja *voileipää* -sananmuodot kuuluvat sen sijaan poikkeuksetta omasta päivästä kertoviin teksteihin, kuten seuraavien esimerkkien tapauksissa, joista kahdessa (15–16) molemmat sananmuodoista ovat mukana samassa n-grammissa.

14) ***Aamiaiseksi syön tavallisesti puuroa*** ja juon maitoa.
15) Aamulla syön ***puuroa ja voileipää*** ja juon teetä.
16) Syön aamiaiseksi yleensä ***puuroa tai voileipää ja juon*** teetä.

Sanalistan frekventeimpien sananmuotojen päässä olevaa sananmuotoa *kello* (f = 24) on aiemmin oppijansuomesta tutkinut Jantunen (2017) tämän tutkimuksen tapaan ICLFI:iin tukeutuen. Hänen tutkimuksensa aikaan ICLFI kattoi yhteensä 730 000 sanetta. Tutkimuksessaan Jantunen koosti kaikista ICLFI:n teksteistä avainsanalistan löytääkseen aineistosta oppijankielen sanastollisia yliedustumia. *Avainsanat* ovat sellaisia sanoja, joiden esiintymistaajuus tutkimuksen alla olevassa korpuksessa on huomattavasti suurempi kuin samantyyppisessä vertailukorpuksessa (Scott & Tribble 2006: 55–56). Jantunen (2009a: 104) toteaa avainsana-analyysin nostavan esiin ”muun muassa oppijankielelle tyypillisiä sanoja (tarkemmin sananmuotoja), siis sanoja, jotka esiintyvät oppijansuomessa huomattavasti useammin kuin natiivisuomessa. Tällaisia ovat muun muassa *koska*, *paljon* ja *kello*.” Näiden kolmen esimerkinomaisen sananmuodon voidaan huomata löytyvät korkealta myös B1-aineiston n-grammien sananmuotojen listalta.

Vertailukorpus vaikuttaa omalta osaltaan siihen, millaisia avainsanoja aineistosta nousee esille (ks. Scott & Tribble 2006: 63–65). Jantunen (2017) käytti vertailukorpuksenaan tutkimuksessaan *Käännössuomen korpusta*, ja hänen luomansa avainsanalistan 20 merkitsevintä avainsanaa on listattu taulukkoon 6. Listaus on tämän tutkimuksen kannalta mielenkiintoinen, sillä siinä voidaan nähdä selkeät yhtymäkohdat B1-aineiston n-grammien sananmuotolistan kanssa.

Taulukko 6. ICLFI:n 20 merkitsevintä avainsanaa *Käännössuomen korpukseen* verrattaessa Jantusen (2017: 259) tutkimuksen mukaan.

	sananmuoto		
1	on	11	syön
2	koska	12	kielen
3	paljon	13	minusta
4	minun	14	me
5	minä	15	opiskelen
6	olen	16	suomea
7	menen	17	täytyy
8	kello	18	he
9	minulla	19	yliopistossa
10	pidän	20	asun

Taulukon 6 avainsanalistalla olevista sananmuodoista jokaista on käytetty B1-aineiston n-grammeissa, eli ne löytyvät myös grammien sananmuotolistalta. Peräti 11 niistä (*on*, *koska*, *paljon*, *minun*, *minä*, *olen*, *menen*, *kello*, *minulla*, *syön*, *he*) ylittää taulukossa 3 esitetyn kymmenen esiintymän raja-arvon.

Jantunen (2017) valitsi siis tutkimuksessaan merkittävimmistä avainsanoista tarkasteltavakseen sananmuodon *kello*. Tutkimuksen mukaan *kello*-sananmuodon yliedustumista oppija-aineistossa selittää pitkälti se, että oppijat käyttävät sitä ilmaisemaan aikaa, ja näissä yhteyksissä he tuppaavat tukeutumaan sanan yksinkertaisimpaan muotoon eli nominatiiviin. Tämä kävi tutkimuksessa ilmi sellaisista 3-grammirakenteista, joiden osaksi sananmuoto *kello* lukeutui. Toisin sanoen tutkimuksen perusteella suomenoppijat käyttävät todennäköisemmin ilmausta *Tulen syömään kello kaksi* kuin *Tulen syömään kahdelta*. Vertailussa natiiviaineiston kanssa jälkimmäinen tapa ilmaista kellonaikaa on natiiveille huomattavasti tyypillisempi. *Kellolla* näyttää olevan aivan samanlaista käyttöä myös B1-aineiston n-grammeissa, mistä on esimerkkejä alla. B1-aineiston n-grammien ajanilmauksista on luvassa vielä hieman lisää huomioita luvussa 6.2.

17) Minä ***nousen tavallisesti kello*** yhdeksän aikoihin.
18) Noin ***kello kaksitoista syön*** kevyen lounaan kotona tai kahvilassa, välillä otan eväät kotoa mukaan.
19) Tavallisesti ***menen nukkumaan kello kaksitoista***, mutta viikonloppulla myöhemmin.
20) Bussi saapuu Tartoon ***kello puoli yhdeksän*** ja sitten ma menen kotiin.

Edellä esitelty, erilaisia lekseemien sananmuotoja kattava listaus kykenee paljastamaan joka tapauksessa vain yhden puolen n-grammein lähestyttävästä B1-oppijansuomen leksikosta. Sanaston tutkimista olikin syytä laajentaa tarkastelemalla sitä myös n-grammien sanojen lemmattujen muotojen osalta.

5.2 Lemma-analyysi

5.2.1 Sanojen lemmamuotojen listaus ja huomioita listasta

Edeltävän sananmuotolistan kaltaisesti myös sanojen lemmamuodoista voidaan tehdä *AntConcin* WordList-työkalulla oma listansa, joka voidaan samaten järjestää muun muassa lemmojen frekvenssien tai aakkosten mukaiseen järjestykseen. Lemmalistan koostamiseksi B1-aineiston n-grammeihin sisältyvien sanojen sananmuodot tuli ensin kuitenkin lemmata[41] eli palauttaa lekseemimuotoihinsa. Tämä onnistui luomalla erillinen tekstitiedosto, johon eri lekseemimuotojen yhteyteen koottiin kaikki niiden aineistosta löytyvät taipuneet sananmuotovastineensa. Tämä lemmat taivutusparadigmoineen sisältävä lista ladattiin tämän jälkeen *AntConciin*, joka siihen tukeutuen laski n-grammien sisältämät yhden lekseemin eri sananmuotoesiintymät yhteen ja loi niiden pohjalta lemmalistan. Listan kautta selvisi, että aineiston 444 erilaista sananmuotoa kuuluvat yhteensä 307 lemmaan. Seuraavan sivun taulukkoon 7 on koottu kaikki aineistossa yli 10 kertaa esiintyvät lemmamuodot frekvenssiensä mukaisesti järjestettyinä. Kuten sananmuotojenkin osalta, myös lemmojen tapauksessa frekvenssimäärä kertoo ainoastaan sen, monessako erillisessä n-grammissa (n = 992) tietty lemma esiintyy, ei sitä, monestiko kyseinen lemma esiintyy koko B1-aineiston kaikissa n-grammiesiintymissä (n = 16 614). Kaikki n-grammien lemmat kattava lista on nähtävissä liitteestä 2.

[41] Termi *lemma* viittaa yhteen sananmuotoon, joka edustaa kaikkia kyseisen sanan taivutusmuotoja. *Lemmaus* taas on prosessi, jossa päätetään tai ratkaistaan tiettyjen sananmuotojen kuuluminen samaan lemmaan. (Karlsson 2008: 187–188.)

Taulukko 7. B1-aineiston n-grammeissa vähintään kymmenen kertaa esiintyvät lemmat esiintymämääriensä mukaisesti järjestettyinä frekventeimmästä lemmasta alkaen.

	f	**lemma**			
1	546	olla	36	15	keittiö
2	285	ja	37	15	tai
3	160	minä	38	15	tavallisesti
4	123	ei	39	15	voida
5	114	se	40	14	äiti
6	107	hän	41	13	hyvä
7	70	että	42	13	suomi
8	47	mutta	43	12	isä
9	42	vuosi	44	12	kahvi
10	40	mennä	45	12	sama
11	39	huone	46	12	tarto
12	33	syödä	47	12	vartti
13	28	paljon	48	11	nimi
14	27	kaksi	49	11	alkaa
15	27	koska	50	11	asua
16	26	jälkeen	51	11	kirjoituspöytä
17	26	koti	52	11	kolme
18	26	pieni	53	11	laittaa
19	25	iso	54	11	oikea
20	24	kello	55	11	sanoa
21	24	yksi	56	11	siellä
22	23	juoda	57	11	silmä
23	23	me	58	11	suihku
24	20	käydä	59	11	täytyä
25	19	aamiainen	60	11	vanha
26	19	myös	61	11	vielä
27	18	puuro	62	11	vieressä
28	17	niin	63	11	yliopisto
29	17	sitten	64	10	kun
30	17	voileipä	65	10	moni
31	16	nimi	66	10	nousta
32	16	noin	67	10	nukkua
33	16	sänky	68	10	opiskella
34	16	tukka	69	10	pitää
35	15	he	70	10	tosi
			71	10	yli

Taulukon kärkipäätä hallitsevat odotetustikin monet jo sananmuotolistalla olleiden saneiden lemmatut muodot. Listalta löytyvät myös esimerkiksi samat funktiosanat, kuten *ja*, *että* ja *koska*. Niiden taivutusparadigmaan ei luonnollisesti muita muotoja kuulukaan, joten niillä on lemmalistalla samat frekvenssit kuin sananmuotolistallakin. Sekä taulukon 7 typistetyn että liitteen 2 koko lemmalistan perusteella vaikuttaa siltä, että sananmuotolistan tapaan myös lemmalista noudattaa Zipfin lakia: vain kourallinen lemmoista esiintyy esimerkiksi yli sadassa eri n-grammissa; OLLA[42] frekvenssillään 546 tavataan jopa yli puolessa n-grammeista. Sen sijaan lemmoja, jotka esiintyvät korkeintaan kolmessa n-grammissa, on liitteen 2 lemmalistan mukaan

[42] Sanojen lemmamuodot on kirjoitettu leipätekstissä kapiteelein.

yhteensä 158 eli yli puolet kaikista lemmoista. Seuraavassa luvussa 5.2.2 on esitetty lisää yleisluontoisia päätelmiä listasta heijastellen sitä natiivisuomen tavallisimpiin lemmoihin.

Lemmalistalta löytyy paljon sellaista kielenainesta, jonka käyttöä on kartoitettu jo aiemmin ICLFI:stä tai muista oppijansuomen aineistoista tehdyissä tutkimuksissa. Esimerkiksi Tervo (2013) on tutkinut ICLFI:stä löytyviä frekventtejä intensiteettimääritteitä *hyvin*, *oikein*, *todella*, *tosi* ja *erittäin*. B1-aineiston n-grammien lemmoista näistä löytyvät TOSI (f = 10), OIKEIN (f = 9), HYVIN (f = 7) ja ERITTÄIN (f = 2). Tervon tutkimus kartoitti niin alkeis- ja keskitason kuin edistyneidenkin suomenoppijoiden tekstien määritteitä, ja tutkimuksessa *hyvin* oli kaikista yleisimmin koko korpuksesta löydetty astemäärite. N-grammeissa se ei vaikutakaan seitsemällä esiintymällään olevan aivan yhtä tavanomainen. Jantusen (2015) tutkimuksen mukaan myös *oikein*-astemääritettä suositaankin ainakin nimenomaan B1-taitotasolla siinä missä *hyvin*-määritettäkin.

Jantunen (2016) on toisessa tutkimuksessaan selvittänyt tarkemmin *oikein*-astemääritteen käyttöä ICLFI:ssä. Tutkimuksen mukaan suomenoppijat käyttävät *oikein*-määritettä usein rakenteellisesti pätevissä mutta fraseologialtaan epätyypillisissä sanojen yhtymissä. Vastaavia tapauksia löytyy myös B1-taitotason n-grammeista, kuten seuraavista esimerkeistä voidaan huomata. Näiden kaltaista käyttöä *oikein*-määrite ei yleisesti natiivien tuottamissa teksteissä saa.

21) Myöskin hän ***on oikein kaunis***, ja kaikki miehet, jotka näkevät Turandotin, rakaistuvat häneen.
22) Isoäitini on ***oikein mukava ja*** melkein aina hyvällä tuulella, siksi olen siellä mielelläni kylässä.
23) Minulla ei ole vielä verhettä eikä mattoa, mutta ***pidän oikein paljon*** uudesta huonesta.

Ylipäänsä astemääritteistä tiedetään se, että kielenoppijat käyttävät niistä ainoastaan pientä osaa ja että esimerkiksi juuri *oikein* on huomattavasti yleisemmässä käytössä suomenoppijoilla kuin ensikielenään suomea puhuvilla (ks. Jantunen 2015: 113; 2016). Niiden on nähtykin olevan suomenoppijoille niin sanottuja *leksikaalisia nallekarhuja* (*lexical teddy bears*) eli sanoja, jotka ovat oppijoille tuttuja ja turvallisia ja joita he täten toistavat usein, mikä toisinaan johtaa myös virheisiin niiden käytössä (Jantunen 2015; käsitteestä ks. Hasselgren 1994). Tieto vähäisestä määrästä käytettyjä astemääritteitä tukee oppijankielen universaaliksi piirteeksi hahmoteltua ajatusta oppijankielen yksinkertaisuudesta.

Astemääritteistä lemmalistalta löytyy myös PALJON, joka esiintyy yhteensä 28 n-grammissa. *Paljon*-adverbin käyttöä oppijansuomessa on aiemmin kartoittanut laajemmin Kallioranta (2009). Kalliorannan aineistona oli samaten ICLFI, joka oli tosin tuolloin kooltaan vain reilu kymmenesosa nykyisestä, yhteensä 102 137 sanetta. Kallioranta lähestyi aineistoa tämän tutkimuksen tapaan korpusvetoisesti. Juuri *paljon*-adverbi nousi hänen tutkimuksessaan

tarkastelun kohteeksi, koska aineistosta luoto avainsanalista paljasti, että *paljon* oli aineistossa huomattavan yliedustettu sanamuoto, eli sen frekvenssi oli natiivikieleen verrattuna poikkeuksellisen suuri. Kallioranta vertasi kollokaatioanalyysilla, mitä epätyypillisyyksiä adverbin *paljon* käytössä on oppijansuomen ja natiivisuomen välillä *Käännössuomen korpusta* vertailukorpuksenaan käyttäen. Tutkimuksessa selvisi, että natiivisuomessa *paljon* saa huomattavan määrä sellaisia tilastollisesti merkitseviä kollokaatteja sekä vasemmalle että oikealle puolelleen, joita ei ICLFI:stä löydy. Sen sijaan oppijansuomessa *paljon* saa sekä vasemmassa että oikeassa kotekstissaan ainoastaan muutamia pelkästään tälle kielimuodolle tyypillisiä kollokaatteja. Näitä ovat vasemmassa kotekstissa lemmat PITÄÄ, PUHUA ja KESKUSTELLA ja oikeassa KOSKA, ASIA, ERILAINEN, RUOKA, JUHLA, LUMI ja TIETO. Esimerkiksi sellaiset natiiviaineistolle hyvin tyypilliset kollokaatit, kuten vasemman kotekstin YHTÄ ja MITEN sekä oikean kotekstin PALJON[43] ja KUIN, uupuivat oppijansuomesta.

Kalliorannan (2009) tekemä kollokaatioanalyysi paljasti siis, etteivät suomenoppijat hallitse *paljon*-adverbin kaikkia kollokationaalisia ominaisuuksia. Natiivisuomelle epätyypilliset kollokaatit kertovat, että *paljon* on mukana useassa natiivisuomelle poikkeuksellisessa ilmauksessa, mikä samalla selittää sanan yliedustumista, sillä mikäli *paljon* korvattaisiin jollakin muulla sanalla, voisi sen yliedustus pienetä. Kuten ylempänä todettiin, Kalliorannan (2009) tutkimus perustui huomattavasti pienempään osuuteen ICLFI:stä kuin mitä korpus tänä päivä kattaa, mutta ottaen huomioon, kuinka yleinen sananmuoto ja lemma *paljon* on n-grammeissakin, voidaan olettaa, että sen jonkinasteinen liikakäyttö on myös jatkunut korpuksen koon karttuessa. Tarkastelemalla tämän tutkimuksen listaa kaikista tutkimuksen n-grammeista (liite 3), voidaan huomata, että n-grammeja, jotka sisältävät sananmuodon *paljon* on yhteensä 26 erilaista. Niiden yhteisfrekvenssi on 381. Useissa näistä *paljon* on n-grammin viimeinen sana, ja kyseiset n-grammit kuvaavat, mitä jollakin omistajalla tai jossakin paikassa on tai ei ole paljoa. Tällaisia ovat esimerkiksi 3-grammit *minulla on paljon* (f = 43) ja *suomessa on paljon* (f = 9). Kalliorannan (2009) löytämistä epätyypillisistä kollokaateistakin n-grammiaineistosta löytyy pari esiintymää, nimittäin vasemman kotekstin lemmat PITÄÄ ja PUHUA. PITÄÄ-lemma on osana 3-grammeja *pidän erittäin paljon* (24–25) ja *pidän oikein paljon* (26–27):

24) Kuuntelen mielelläni musiikkia, ***pidän erittäin paljon*** folkista ja maailman musiikkista.
25) Eräänä iltana kävimme oopperassa, koska minä ***pidän erittäin paljon*** oopperasta.
26) ***Pidän oikein paljon*** töistä ja koulusta.
27) Entiseen tapan ***pidän oikein paljon*** kasvisruuasta.

[43] Tutkimuksen lemmamuotoisissa kollokaateissa PALJON-lemmaan lukeutuivat myös sen komparaatiojohdokset *enemmän* ja *eniten* (Kallioranta 2009: 36).

PUHUA-lemma esiintyy *paljon*-astemääritteen kanssa 3-grammissa *luennossa puhuttiin paljon*, jonka frekvenssi on 9. Alla olevassa konkordanssissa 3 on esitetty *luennossa puhuttiin paljon* -3-grammin konkordanssinäkymä. Itse 3-grammista voidaan tehdä pelkää kieli-intuitiotakin hyödyntäen huomio, että LUENTO-lekseemistä olisi inessiivin sijaan tyypillisempää käyttää tässä yhteydessä adessiivimuotoa (ks. VISK § 1239–1241). Aiemmassa tutkimuksessa onkin havaittu, että inessiiviä käytetään enemmän alemmilla tasoilla ja että sen käyttö vakiintuu niin konkreettisissa kuin abstrakteissakin käytöissä adessiivia aiemmin (Siivelt & Mustonen 2013).

Hit	KWIC			File
1	äiväkirja %.%.% Tässä viikon historian	*luennossa puhuttiin paljon*	erilaisia asioita. Ensiksi oli S	KI0010t.txt
2	Päiväkirja Kulttuuri Luento %.% Tässä	*luennossa puhuttiin paljon*	Kalevalasta ja suomen mytol	KI0010k.txt
3	Päiväkirja Kulttuuri Luento %.% Tässä	*luennossa puhuttiin paljon*	Sibeliuksen elämästä ja musii	KI0010o.txt
4	Päiväkirja Kulttuuri Luento %.% Tässä	*luennossa puhuttiin paljon*	suomalaisista maalauksista.	KI0010n.txt
5	Kulttuuri Luento Päiväkirja %.% Tässä	*luennossa puhuttiin paljon*	suomen maanviljelijän töistä.	KI0010i.txt
6	Kulttuuri Luento Päiväkirja %.% Tässä	*luennossa puhuttiin paljon*	suomen luonnosta ja luonnon	KI0010j.txt
7	Kulttuuri Luento %.% Tänään kulttuuri	*luennossa puhuttiin paljon*	suomen kansallismusiikista.	KI0010m.txt
8	Kulttuuri Luento Päiväkirja %.% Tässä	*luennossa puhuttiin paljon*	suomen juhlista. Tätä teema	KI0010q.txt
9	Kulttuuri Luento Päiväkirja %.% Tässä	*luennossa puhuttiin paljon*	suomen kirjallisuudesta. Ja ti	KI0010s.txt

Konkordanssi 3. 3-grammin *luennossa puhuttiin paljon* esiintymät (n = 9) B1-aineistossa.

Konkordanssin 3 konkordanssiriveissä huomio kiinnittyy ensimmäisenä siihen, että n-grammin vasemmanpuoleinen koteksti on jokaisessa lähes toistaan vastaava. Sen informaatio tulee paljastaneeksi, että tekstien tehtävänantoina on mitä ilmeisimmin ollut luentopäiväkirjan kirjoittaminen. File-sarakkeesta nähdään, että jokaisen tekstin tiedostonimi on tyyppiä KI0010*x*. Tämä ei ainoastaan kerro, että tekstintuottaja on kiinaa ensikielenään puhuva suomenoppija, vaan toistuva numerosarja 0010 ilmaisee samalla, että tekstit ovat peräisin yhdeltä ja samalta suomenoppijalta. Numeron perässä oleva vaihtuva kirjain kertoo sen, monesko oppijan kirjoittama teksti on kyseessä. Toisin sanoen vaikka tutkimuksen n-grammihauissa range-arvoksi asetettiin 5, jotta vältyttäisiin yksittäisen kielenkäyttäjän idiolektin korostumiselta n-grammeissa, tultiin tässä yhteydessä paljastaneeksi, että vaikka n-grammit esiintyvätkin eri teksteissä, voivat ne siitä huolimatta olla samalta suomenoppijalta peräisin. Idiolektilta välttyminen ei siis olekaan taattua pelkästään tätä hakukriteeriä käyttämällä, ainakin mikäli jakauma-arvo jätetään turhan pieneksi.

Lemmojen kautta on mahdollista päästä käsiksi myös siihen, minkä sanaluokkien sanoja n-grammeissa käytetään eniten. Tässä utkimuksessa lemmat listattiin *Excel*-tiedostoon, jossa jokainen niistä käytiin läpi ja jaoteltiin omaan sanaluokkaansa. Laskennan tulokset on esitetty taulukossa 8.

Taulukko 8. B1-aineiston n-grammien sanojen lemmamuotojen sanaluokkajakauma.

Sanaluokka	**Frekvenssi**	**%**
Verbit	1 028	32,0
Substantiivit	711	22,2
Pronominit	458	14,2
Adjektiivit	181	5,7
Numeraalit	91	2,8
Partikkelit	552	17,2
Adverbit	113	3,5
Adpositiot	78	2,4
Yhteensä	3 212	100

Kaikista B1-aineiston n-grammien sanoista verbejä on siis 32 prosenttia, siinä missä nomineja eli substantiiveja, adjektiiveja, pronomineja ja numeraaleja on yhteensä noin 45 prosenttia. Taipumattomat tai vaillinaisesti taipuvat sanat eli partikkelit, adpositiot ja adverbit taas muodostavat reilun 23 prosentin osuuden n-grammien lemmoista; näistä partikkeleja on selvästi eniten. Vertailemalla sanaluokkataulukkoa frekventeimpiin n-grammien lemmoihin voidaan huomata, että verbeistä reilu puolet (n = 546; n. 53,1 %) on *olla*-verbin esiintymiä. Verbin *olla* onkin havaittu yliedustuvan oppijansuomessa, sillä sitä saatetaan käyttää esimerkiksi vaikeampien verbien kiertämiseen ja korvaamiseen (ks. esim. Grönholm 1993: 144–145).

Oppijankielen adjektiivien osalta esimerkiksi Nieminen (2001: 97–106) on havainnut oppijoiden olevan taipuvaisia tiettyjen adjektiivien liikakäyttöön natiiveihin nähden, mikä johtuu oppijoiden suppeammasta sanavarastosta. Hän toteaa, että oppijat käyttävät usein myös niin sanottuja likiarvoja eli käyttötilanteeseen sopivimpia tuntemiaan ilmauksia, koska he eivät tunne tai muista spesifimpää ilmausta. Likiarvoja käyttäessään he tulevat usein kategorisoineeksi suuren määrän kielellisiä ilmiöitä yhden tällaisen likiarvon alle. Adjektiiveista esimerkiksi *hyvä*, *huono*, *kaunis*, *ruma*, *iso* ja *pieni* ovat likiarvoja, jotka voivat pitää sisällään hyvin suuren joukon erilaisia ilmiöitä ja jotka ovat täten oppijoille erittäin käyttökelpoisia. (Mts. 81–84). Edeltävistä adjektiiviesimerkeistä *hyvä*, *kaunis*, *iso* ja *pieni* löytyvät myös n-grammien lemmamuotojen listalta (liite 2). Adjektiivien käyttöä oppijansuomessa on aiemmin kartoittanut myös ainakin Akgül (2013), jonka tutkimus keskittyi nimenomaisesti erittäin frekventteihin adjektiiveihin *hyvä*, *pieni* ja *iso*. Taulukon 7 lemmalistalta nähdään, että samat kolme adjektiivia ovat edustettuina merkittävissä määrin myös B1-aineiston n-grammeissa: lemman PIENI frekvenssi on 26, ISO-lemman 25 ja HYVÄ-lemman 13. Akgülin (2013) tutkimuksessa toiseksi

frekventein suomenoppijoiden käyttämä adjektiivi oli PIENI, ja etenkin sen käytön havaittiin motivoituvan oppijoiden tekstien tehtävänannoista. Sama voidaan huomata myös n-grammeista, joihin lemma PIENI lukeutuu. Alla on esitetty otos 3-grammin *on pieni mutta* (f = 78) esiintymistä B1-aineistossa.

Hit	KWIC			File
1	Minun perhe	*on pieni, mutta*	erittäin tärkea minulle. Minun vanh	VI0443b.txt
2	a sinne toiseen laitaan asti ” Minusta kaupunki	*on pieni, mutta*	heidän mielistä, se on suuri. Kertoj	VE0144.txt
3	igassa. Veiko opiskelee siellä taloutta. Minulla	*on pieni, mutta*	hyvin mukava perhe.	VI0028.txt
4	on siellä ensimäisessä kerroksessa yksiö. Koti	*on pieni, mutta*	hyvin mukava. Siellä on keittiö, olo	VI0196.txt
5	en Tarton Yliopistossa kirjallisuutta. Perheeni	*on pieni, mutta*	hyvin mukava.	VI0292b.txt
6	lä pelaan lentopalloa ja teen käsitöitä. Minulla	*on pieni, mutta*	hyvä perhe. Meidän perhe asuu Tartos	VI0055f.txt
7	len lyhyt. Minä pidää tanssiasta. Minun perhe	*on pieni, mutta*	iloinen.	VI0116a.txt
8	Minun perheni Minun perheeni	*on pieni, mutta*	iso sukuni. Minulla ei ole ei siskoa	VI0241b.txt
9	on meillä kahden kesken paljon tilaa. Huoneni	*on pieni, mutta*	kaikki tarpeellinen on olemassa. Huo	VI0237.txt
10	a kolme makuuhuoneta. Minun makuuhuoneni	*on pieni, mutta*	kaunis, seinä on punainen. Huonessa	PU0008f.txt
11	lli-kadulla asunnossa. Minun mielestäni Pärnu	*on pieni, mutta*	kaunis vanha kaupunki. Asunnossa o	VI0259a.txt
12	nto meren ja keskustan lähellä. Meidän asunto	*on pieni, mutta*	kaunis. Asunossa on kaksi huonetta, a	VI0365.txt
13	n suuri ruskea sohva ja vanha televisio. Keittiö	*on pieni mutta*	keittiö riitää. Se on valkoinen. Keittiö	SA0145.txt
14	Perheni Minun perheni	*on pieni, mutta*	kiva. Minun vanhemmat on erronneet	VI0063c.txt
15	nnossa, missä on kaksi huonetta. Minun huone	*on pieni, mutta*	kodikas. Huonessa on sohva, pöytä ja	VI0303a.txt
16	en pariin. Meillä on yksi huone. Meidän huone	*on pieni mutta*	melko kodikas. Huonessa on vasemm	VI0366.txt

Konkordanssi 4. Otos 3-grammin *on pieni mutta* esiintymistä (n = 78) B1-aineistossa.

Konkordanssista 4 havaitaan, että tarkastellun 3-grammin kotekstit käsittelevät poikkeuksetta joko perhettä, huonetta, asuntoa tai kaupunkia ja ovat keskenään varsin lähellä toisiaan. Sama ilmiö toistuu myös muun muassa n-grammien *pieni mutta viihtyisä* (f = 41), *huone on pieni* (f = 39) ja *minulla on pieni perhe* (f = 12) koteksteissa. PIENI-adjektiivin käyttö vaikuttaa siis n-grammienkin perusteella hyvin tehtävänantomotivoituneelta.

5.2.2 Pintapuolista vertailua natiivisuomen sanastoon

Alla olevan taulukon vasemmanpuoleisella palstalla on esitetty uudelleen edeltävän lemmalistan (taulukko 7) 30 frekventeintä lemmaa. Oikeanpuoleiseen palstaan on taas liitetty vertailun vuoksi Suomen sanomalehtikielen taajuussanaston (CSC – IT Center for Science 2004) 30 frekventeintä lemmaa, jotta oppijansuomen n-grammien sanastoa voitaisiin vertailla edes pintatasolla natiivisuomen kanssa. Suomenoppijoiden ja natiivikielenkäyttäjien sanastojen vertailu on toki pelkästään näihin aineistoihin ja tekstilajeihin tukeutumalla varsin mielivaltaista, mutta sillä voidaan silti saada jonkinlaista kuvaa näiden kahden kielimuodon eroista.

Taulukko 9. B1-aineiston n-grammien 30 frekventeintä lemmaa vertailtuna natiivisuomen aineiston (Suomen sanomalehtikieli) frekventeimpien lemmojen kanssa.

OPPIJANSUOMI (N-GRAMMIT)			**NATIIVISUOMI**	
	f	**lemma**		**lemma**
1	546	olla	1	olla
2	285	ja	2	ja
3	160	minä	3	ei
4	123	ei	4	se
5	114	se	5	että
6	107	hän	6	joka
7	70	että	7	vuosi
8	47	mutta	8	hän
9	42	vuosi	9	myös
10	40	mennä	10	saada
11	39	huone	11	mutta
12	33	syödä	12	tämä
13	28	paljon	13	voida
14	27	kaksi	14	tulla
15	27	koska	15	suomi
16	26	jälkeen	16	tehdä
17	26	koti	17	kun
18	26	pieni	18	pitää
19	25	iso	19	mukaan
20	24	kello	20	uusi
21	24	yksi	21	jo
22	23	juoda	22	kuin
23	23	me	23	hyvä
24	20	käydä	24	ne
25	19	aamiainen	25	sanoa
26	19	myös	26	kaikki
27	18	puuro	27	markka
28	17	niin	28	nyt
29	17	sitten	29	suuri
30	17	voileipä	30	kertoa

Taulukosta 9 voidaan huomata, että listauksissa ensimmäisistä kymmenestä lemmasta seitsemän löytyy molemmista listoista. Lemmojen sijoitukset ovat myös varsin samanlaiset: OLLA ja JA ovat kummassakin ensimmäisinä, ja viiteen yleisimpään lemmaan kuuluvat molemmissa aineistoissa lisäksi lemmat SE ja EI. Molemmilta listoilta löytyvät myös samat funktiosanat ETTÄ ja MUTTA. Molempien lemmalistojen yleisin substantiivi on VUOSI. VUOSI-lemman käyttö vaikuttaa oppijansuomessa juontuvan useimmiten lauseista, joissa kerrotaan, kuinka monta vuotta vanha tietty henkilö on (28) tai tapauksista, joissa kuvaillaan, mitä jonakin vuonna tapahtui (29–30). Sanomalehtikielessä VUOSI-lemmalla voisi olettaa olevan etenkin jälkimmäinen funktio.

28) ***Hän on [%] vuotta vanha***, mutta näyttää paljon nuorilta, eniten [%]-vuotiaalta.
29) Hän oli kapteeni maajoukkuin ja tuli maailmanmestareksi ***vuonna [%] ja*** Euroopan mestareksi vuonna [%].

30) Ylioppilaaksi ***kirjoitin vuonna [%]*** ja samana vuonna pääsin Tarton yliopistoon opiskelemaan viron kirjallisuutta.

Toiseksi yleisin lemmasubstantiivi oppijansuomessa on ehkä hieman yllättäenkin HUONE. Tämä kuitenkin selittyy sillä, että useiden oppijansuomen kirjoitelmien tehtävänantona on ollut kirjoittaa teksti omasta huoneesta (31), mistä saatiin jo viitteitä edellä PIENI lemman yhteydessä. Kolmanneksi yleisin substantiivi KOTI liittyy samankaltaiseen tehtävänantoon, jossa huoneen sijaan on vain pitänyt kirjoittaa omasta kodista (32). Usein KOTI mainitaan kuitenkin myös teksteissä, joissa oppijoiden tehtävänä on ollut kirjoittaa päivänsä kulusta (33–34). Samaan kategoriaan näiden lemmojen kanssa kuuluvat myös esimerkiksi lemmat AAMIAINEN, PUURO ja VOILEIPÄ.

31) ***Minun huone on pieni mutta*** viihtyisä, tarpeeksi suuri minulle.
32) ***Minun koti on*** korkeassa ruskeassa kerrostalossa.
33) ***Sitten menen kotiin*** tai tapaan asiakkaan tai menen kauppaan, jos on tarpeen.
34) Sen jälkeen pukeutun ja ***lähden kotoa puoli*** kahdeksalta.

Neljäs substantiivi oppijansuomen listalla on KELLO, jonka toistuva käyttö selittyy ajankohtaa kuvailevista n-grammeista, kuten *menen nukkumaan kello* ja *luento alkaa kello kaksitoista* (tästä lisää luvussa 6.1.3). Tällaisia n-grammeja näyttää samaten useimmiten muodostuvan teksteissä, joiden tehtävänantona on ollut kirjoittaa omasta päivästä. Sanomalehtikielessä KOTI löytyy sijalta 397, KELLO taas sijalta 473 ja HUONE vasta sijalta 1 841. Mikäli n-grammien lemmoista luotaisiin avainsanalista käyttäen jotain natiivisuomen korpusta vertailukorpuksena, voisi tällaisten substantiivien olettaakin nousevan listan kärkipäähän oppijansuomea natiivisuomesta erottavina sanastollisina piirteinä.

Verbeistä merkillepantavaa on, että oppijansuomen n-grammien 30 yleisintä lemmaa sisältävät kieltoverbin ja OLLA-lemman ohella ainoastaan verbilemmat MENNÄ, SYÖDÄ, JUODA ja KÄYDÄ. Sanomalehtikielessä MENNÄ ja SYÖDÄ eivät ole edustettuina ensimmäisten 30 lemman joukossa lainkaan (vaan sijoilla 88 ja 920), mutta tältä listalta löytyvät kieltoverbin ja OLLA-lemman ohella kuitenkin modaaliverbit SAADA ja VOIDA sekä TULLA, SANOA ja KERTOA. B1-aineiston n-grammeissa VOIDA-lemman frekvenssi on 15, SANOA-lemman 11 ja TULLA- ja KERTOA-lemmojen 2, siinä missä SAADA-lemmaa ei löydy n-grammeista laisinkaan. SAADA ei itsessään ole ICLFI:ssä mitenkään harvinainen lemma, ja Tarvainen (2018) onkin verrannut SAADA-lemman käyttöä ICLFI:n ja natiivitekstien välillä kohdistaen tutkimuksensa kuitenkin ylemmille EVK:n (2003) taitotasoille B2–C2. B1-kielitaitotason n-grammeihin SAADA ei pääsekään osaksi, mikä voi johtua siitä, että SAADA ei ole

ylipäänsä niin taajaan käytössä vielä B1-taitotason teksteissä tai mahdollisesti siitä, että sitä käytetään vaihtelevilla tavoilla ja vaihtelevissa yhteyksissä, jolloin se ei päädy muodostamaan n-grammirakenteita.

Modaaliverbeihin liittyvissä oppijansuomen tutkimuksissa Ivaska (2014b) on aiemmin havainnut, että *voida* on yleisin mahdollisuudenilmaus edistyneessä oppijansuomessa. Myös Haltia (2015) on tutkinut oppijansuomen modaaliverbejä aineistonaan ICLFI:n EVK:n (2003) taitotasoille A2–C2 arvioidut tekstit. Hänen tutkimuksessaan *voida*-verbi oli juuri B1-tasolla modaaliverbeistä reilusti yleisin. Haltia huomioi, että se sekoittuu toisinaan *osata* ja *saada* -verbien kanssa, mistä on viitteitä myös joissain VOIDA-lemman sisältävien n-grammien alla olevissa esimerkkiesiintymissä (35–37). Biber ym. (1999: 996–997; 1001–1003) lukevat modaaliverbin sisältävät n-grammit osaksi ”persoonapronomini + pääverbi” -yläkategoriaa, joka on heidän mukaansa nimenomaisesti keskustelun eikä niinkään akateemisten tekstien n-grammien luokka. Keskustelun n-grammien luokista se on kaikkein hallitsevin. Modaaliverbillisistä n-grammeista he antavat muun muassa esimerkit *I can’t do it* ja *I have to go*.

35) [– –] minä haluan käännää suomalainen kirjaa kiinaksi ja myös kiinalainen kirjaa suomeksi. Uskon, ***että minä voin*** tehdä sitä hyvin.
36) ***En voi sanoa***, missä se oli rikki koska oli ehjän näköinen.
37) Kotona Tartossa minulla ei ole kotieläimiä, koska yliopisto-asuntolassa ***ei voi olla*** kotieläimiä.

B1-aineiston n-grammien ja Suomen sanomalehtikielen osalta yhdeksi merkittävimmistä eroista yleisimpien lemmojen osalta voidaan huomata vielä se, että oppijansuomen kolmanneksi yleisin lemma on MINÄ, joka on sanomalehtikielen listalla vasta sijalla 118. Tämä kielii entisestään siitä, että suomenoppijoiden kirjoittamat tekstit käsittelevät useimmiten tekstin kirjoittajaa itseään ja hänen kokemuksiaan, mikä taas ei ole yhtä usein asian laita sanomalehtikielessä.

6 B1-OPPIJANSUOMEN RAKENTEET N-GRAMMEIN ANALYSOITUINA

Tässä n-grammien rakenneanalyysiin painottuvassa luvussa kuvataan, kuinka B1-aineiston n-grammeja lähestyttiin tutkimuksessa niiden tempukset ja syntaktiset lausetyypit pääkohteina ja millaisia tuloksia n-grammien perusteella täten saatiin B1-tasoisten suomenoppijoiden käyttämistä rakenteista. Rakenneanalyysi päätettiin kohdistaa niihin aineiston n-grammeihin, joihin sisältyy verbi, verbiliitto tai edes osa verbiliitosta. Tämä valinta taas osaltaan ohjasi päätöstä tutkia näistä n-grammeista nimenomaan tempuksia ja lausetyyppejä. Täysin vailla verbiä olevat n-grammit – jotka lopulta muodostivat B1-aineiston n-grammeista ainoastaan noin 15 prosentin osuuden – käydään luvun lopuksi pintapuolisesti läpi sen suhteen, millaisia toimintoja niillä vaikuttaisi B1-aineistossa olevan. Käsillä olevassa luvussa myös verrataan saatavia tuloksia soveltuvin osin edeltäviin oppijansuomesta tehtyihin tutkimuksiin.

6.1 Finiittiverbi tai verbiliitto lähtökohtana rakenneanalyysille

Miltei tuhannen erilaisen n-grammin ja yli 16 000 n-grammiesiintymän analysoiminen aukottomasti niiden kielellisen rakenteen perusteella olisi ollut ainakin tämän tutkimuksen puitteissa liian massiivinen tehtävä, joten tutkimusta varten oli päätettävä, mihin n-grammien rakenteessa keskitytään ensisijaisesti, jotta niiden analysointi pysyisi tarkoituksenmukaisena. Tutkimuksen lopulliseen B1-aineiston n-grammilistaukseen (liite 3) tarkemmin perehtymällä oli pian huomattavissa, että suurin osa löydetyistä n-grammeista näyttää pitävän sisällään pääverbin tai vähintäänkin jonkin osan verbiliitosta. Tarkempi laskeminen paljastikin, että tutkimukseen päätyneistä 992 erilaisesta n-grammista yhteensä 842 sisältää verbin tai verbiliiton osan, eli täysin verbittömiä n-grammeja on B1-aineistossa yhteensä ainoastaan 150 kappaletta (n. 15,5 prosenttia). Esiintymämäärissä tämä tarkoittaa sitä, että siinä missä kaikkien B1-aineiston n-grammien yhteisfrekvenssi on 16 614, on verbillisten n-grammien esiintymien määrä tästä 14 263 eli noin 85,9 prosenttia. Tutkimuksen B1-oppijansuomen rakenteiden tarkasteluun päätettiin ottaa nämä n-grammit, joiden osaksi lukeutuu verbi, verbiliitto tai osa verbiliitosta[44] ja keskittyä niissä etenkin siihen, mitä tietoa verbit antavat n-grammien rakenteellisesta käytöstä.

[44] Näihin n-grammeihin viitataan tutkimuksessa tästedes kokoavalla nimityksellä *verbilliset n-grammit*.

Päätökseen valita juuri verbilliset n-grammit tutkimuksen keskiöön vaikutti verbin erittäin keskeinen rooli (suomen) kielessä ja syntaksissa. Persoonamuotoisen eli finiittiverbin katsotaan lauseopissa kuuluvan täydennyksineen lauseen ydinjäseniin. Lauseella on yleensä ominaisuuksia, jotka johtuvat nimenomaan sen finiittiverbistä. Ominaisuuksista merkittävimpiin lukeutuu mahdollisuus saada subjekti, jonka kanssa verbi kongruoi persoonassa ja luvussa. Verbin morfologisesta rakenteesta on luettavissa myös paljon lauseen tehtävästä tekstissä; verbi paljastaa lauseen kuvaaman asiatilan suhteutumisen ajallisesti puhehetkeen, puhujan suhtautumisen esittämäänsä asiantilaan sekä lauseen puhefunktion. Lauseenjäsenenä finiittinen verbi tai verbiketju tunnetaan predikaattina. (VISK § 864–866.)

Verbillisissä n-grammeissa päätettiin tutkimuksessa keskittyä pääosin siihen, mitä tempuksia ja syntaktisia lausetyyppejä ne edustavat. Samalla kartoitettiin myös, kuinka usein ne ovat osana myöntö- tai kieltolauseita ja mitä moduksia niissä ilmenee. Erilaisia lukumääriä listattiin *Excel*-tiedostoihin. Aiemmassa n-grammitutkimuksessa verbillisiä n-grammeja yhtenä n-grammien rakenneluokkana ovat esimerkiksi fraasikehystutkimuksissa käyttäneet ainakin Gray ja Biber (2013) sekä Garner (2016). Näissä tutkimuksissa verbipohjaiset fraasikehykset (*verb based frames*) on määritelty kehyksiksi, joihin lukeutuu yksi tai useampi modaali-, apu- tai pääverbi (esim. *can speak * languages*, *is * my house*, *I * like to*) (Gray & Biber 2013; Garner 2016: 40).

Seuraavissa alaluvuissa on kerrottu, miten analyysi aikamuotojen ja syntaktisten lausetyyppien perusteella eteni tässä tutkimuksessa ja miten tulokset peilautuvat joihinkin aiempiin oppijansuomesta tehtyihin tutkimuksiin.

6.1.1 Tempusanalyysi

Yksi asia, josta voidaan saada viitteitä n-grammeista lähestyttäessä niitä niihin sisältyvien verbien kautta, on n-grammien ilmentämä tempus eli aikamuoto. Suomen kielessä verbit voivat taipua neljässä tempuksessa, jotka ovat preesens, imperfekti, perfekti ja pluskvamperfekti. Näistä preesens ja imperfekti ovat finiittisen verbinmuodon tempuksia ja perfekti ja pluskvamperfekti niin sanottuja liittotempuksia. Liittotempukset muodostuvat apuverbistä ja pääverbin NUT- tai TU-partisiipista. (VISK § 112.) Liittotempukset ovat haastavia tämän tutkimuksen kannalta, koska kyseiset aikamuodot eivät kaikissa tapauksissa ole yksiselitteisesti ja suoraan luettavissa jokaisesta verbillisestä n-grammista, joihin ne sisältyvät. Etenkin kaikki n-grammit, jotka päättyvät *olla*-verbiin tai kieltoverbiin, ovat lähtökohtaisesti aikamuodoiltaan enemmän tai vähemmän tulkinnanvaraisia. Esimerkiksi 3-grammit *että hän ei ole* ja *en ole vielä*, voivat

edustaa tempuksiltaan joko preesensiä tai perfektiä, siinä missä muun muassa 3-grammit *hän ei ollut* ja *mutta hän oli* voivat kieliä joko imperfektin tai pluskvamperfektin käytöstä. Toisaalta taas esimerkiksi pelkkään kieltoverbiin päättyvä 3-grammi *koska hän ei* voi taas olla osana joko preesens-, imperfekti-, perfekti- tai pluskvamperfektilausetta. Tällaista ongelmaa ei ole finiittisten verbinmuotojen osalta, sillä esimerkiksi 3-grammin *laitan ruokaa ja* voi suoraan todeta olevan preesens- ja 6-grammin *[%] ja samana vuonna pääsin yliopistoon* taas imperfektimuotoinen. Aikamuotoanalyysin kannalta oli kuitenkin olennaista selvittää todelliset ja mahdollisimman tarkat lukemat kunkin aikamuodon edustukselle, jolloin epäselvien n-grammien tapauksessa apuna tuli käyttää *AntConcin* Concordance-työkalua ja sen KWIC-näkymää. Konkordanssissa 5 on esimerkkinä esitetty *AntConcin* antama konkordanssinäkymä jokaiselle 3-grammin *koska en ole* osumista B1-aineistossa.

Hit	KWIC			File
1	ta kymmesen illalla. En syö iltapalaksi mitään,	*koska en ole*	aikaa. Työn jälkeen olen väsynyt	VI0042a.txt
2	omessa. Mitä minä tiedän, olen kuullut toisilta,	*koska en ole*	koskaan ollut Suomessa Juhannuk	HO0011w.txt
3	koulua Suomessa. Haluan heille kylään mennä,	*koska en ole*	koskaan siellä käynyt. He puhuvat	VI0037b.txt
4	aa, menen tänä kesänä Espanjaan tai Ranskaan,	*koska en ole*	käynyt näissä maihin.	VI0087.txt
5	matkustan kotiin tunturista, olen ihan väsynyt	*koska en ole*	nukku paljon. Tarvitsen suihkussa	RU0011d.txt
6	n paikaan hyllyyn. Otan rikki monta pakkausta,	*koska en ole*	riittävä varovainen kun avaan kart	RU0013i.txt
7	matkustaisin jossakin, esimerkiksi Ranskaan,	*koska en ole*	siellä ollut. Aikaisin alkaa tenttika	VI0098.txt
8	kirjoittaa päiväkirjaa. En pidä siitä ajatuksesta,	*koska en ole*	systemaattinen, eikä ahkera henki.	PU0007b.txt
9	dä. Mutta minulta olisi kivaa valaa tinaa ensin,	*koska en ole*	tehnyt sitä ennen. No siis, mennää	SA0173.txt
10	hta formaldehydestä oli erikoisesti kiinostava,	*koska en ole*	tiennyt vaaraja, jotka seuraavat po	SA0058c.txt
11	ämpimä ja ei sataa vettä. Olen vähän surullista,	*koska en ole*	ulkona, vain opiskelija-asunnossa	VI0052a.txt

Konkordanssi 5. 3-grammin *koska en ole* esiintymät (n = 11) B1-aineistossa.

Esimerkin 3-grammin tapauksessa voidaan nähdä, että *koska en ole* esiintyy aineistossa seitsemässä tapauksessa perfektissä (rivit 2–5, 7, 9–10) ja neljässä preesensissä (1, 6, 8, 11). Näin ollen aikamuotoja laskettaessa *Excel*-taulukkoon merkittiin 3-grammin *koska en ole* yhteyteen neljä osumaa preesensille ja seitsemän perfektille. Konkordanssista voidaan huomata, että sen kotekstiin sisältyy myös kielenvastaisia muotoja (1, 5, 6, 10). Kyseisen *koska en ole* -3-grammin tapauksessa virheellisistäkin muodoista näkee silti, mitä tempusta ne edustavat, mutta joitain tulkinnanvaraisempiakin tapauksia aineistossa on. Niiden osalta tarkasteltiin koko tekstiä sekä hyödyttiin tarvittaessa myös intuitiota sen päättelyssä, mitä tempusta suomenoppija on n-grammilla halunnut ilmaista.

Kuten edellä sivuttiin, tutkimuksen n-grammeista etenkin finiittiverbin sisältävät ovat kaikesta huolimatta sellaisia, joista aikamuoto voidaan nähdä suoraan ilman tarvetta tukeutua KWIC-näkymään laisinkaan. Esimerkiksi 3-grammit *sanoi että hän* ja *hän kertoi minulle* ovat yksiselitteisesti imperfektimuotoja, kun taas n-grammit *matka kestää noin* ja *sitten käyn suihkussa ja* edustavat preesensiä. Kotekstittomista n-grammeista ei toki kyetä näkemään

minkäänlaisia tietoja niiden esiintymisympäristöistä, ja tarkempi konkordanssirivien tarkastelu voisikin osoittaa esimerkiksi jonkinlaisia lipsahduksia, joissa alkuperäisessä tekstissä olisikin mahdollisesti haluttu – tai olisi ainakin pitänyt – käyttää eri tempusta, kuin mitä n-grammi itsessään ilmaisee. Tällaiset mahdollisuudet ovat kuitenkin todennäköisesti marginaalitapauksia, ja tutkimuksen aineiston koon huomioon ottaen ei olisi ollut mielekästä tarkastella jokaisen verbillisen n-grammin jokaista konkordanssiriviä pelkästään varmuuden varalta. Täten siis niistä verbillisistä n-grammeista, joiden finiittiverbistä kävi aikamuoto selkeästi ilmi, ei lähtökohtaisesti tarkasteltu niiden kotekstiesiintymiä laisinkaan.

Taulukossa 10 on esitetty laskelmat kaikkien B1-aineiston verbillisten n-grammien edustamista aikamuodoista frekvensseineen sekä prosenttiosuuksineen kaikista verbillisistä n-grammeista.

Taulukko 10. B1-aineiston verbillisten n-grammien ilmentämät tempukset.

Aikamuoto	**f**	**%**
Preesens	12 639	88,6
Imperfekti	1 115	7,8
Perfekti	460	3,2
Pluskvamperfekti	49	0,4
Yhteensä	14 263	100

Kuten taulukosta voidaan lukea, preesens on, ehkä varsin odotetustikin, n-grammien perusteella selvästi yleisin B1-aineiston aikamuodoista. Toiseksi frekventein – ja samalla mennen ajan tempuksista käytetyin – on imperfekti, mutta sitäkin tavataan verbillisissä n-grammeissa silti yli kymmenen kertaa harvemmin kuin preesensiä. Perfektiä taas käytetään yli puolet vähemmän kuin imperfektiä. Pluskvamperfekti on B1-aineiston verbillisissä n-grammeissa hyvin harvinainen tempus sen esiintyessä ainoastaan 49 kertaa yhteensä 17 erillisessä n-grammissa. Haapala (2008) on pro gradu -tutkielmassaan samaten laskenut tempusten jakautumista oppijansuomessa, tarkemmin ottaen YKI-testien kirjoitetuissa teksteissä, ja hänen tutkimuksen keskitason tekstien aikamuotojakaumat näyttävät hyvinkin samankaltaisilta ICLFI:n B1-aineiston n-grammien tempusten kanssa. Ne on esitetty seuraavassa taulukossa 11.

Taulukko 11. Tempusten jakauma YKI-testien keskitason teksteissä Haapalan (2008: 30) tutkimuksen mukaan.

Aikamuoto	**%**
Preesens	89,3
Imperfekti	6,8
Perfekti	3,8
Pluskvamperfekti	0,1
Yhteensä	100

Yksi selittävä tekijä suomenoppijoiden n-grammeissaan käyttämille aikamuodoille lienee kirjoituksille määrätyt tehtävänannot, jotka jo itsessään ohjaavat pitkälti teksteihin valittavia aikamuotoja. Preesens on yleinen n-grammien tempus, kun kirjoitetaan esimerkiksi kuvailevia tekstejä omasta huoneesta (38), perheestä (39) tai päivästä (40), esseitä (41) tai vaikkapa mielipidekirjoituksia (42).

38) ***Huoneen keskellä on*** keinutuoli ja pieni sohvapöydä.
39) ***Minun isäni on*** myös neljäkymmentäkuusivuotias, hänellä oli myös vastikään syntumäpäivä.
40) Illalla syön jotakin keitiössä ***ja käyn suihkussa***.
41) ***Luulen että se*** on kiinteä juoni, koska toiminnan vaiheet seuraavat toisiaan ja johtuvat loogisesti.
42) ***Minun mielestäni on*** riittävä kun joku on vegetaari tai toinen ehdotus voi olla että ihmiset vain söisivät vähemmän lihaa.

Imperfektimuotoiset n-grammit näyttäisivät taas usein olevan peräisin teksteistä, joiden tehtävänantona on ollut esimerkiksi elämäkerran tai kertomuksen (43–46), mutta yhtä lailla myös omasta päivästä (47) tai esimerkiksi kesälomasta (48) kirjoittamisen.

43) ***Samana vuonna pääsin*** yliopistoon opiskelemaan kirjallisuusta ja kansanrunousta.
44) Vuonna % perheni muutti Rakveressa ***ja minun täytyi*** vaihtaa koulua ja musiikkikoulua.
45) ***Hän sanoi että hän*** halusi että minä lähdin hänen kanssa.
46) Hänen vanhemmat ovat lähteneet kesämökiin ***ja hän oli*** sen takia yksin kotona.
47) Me tulimme Uumajaan ja ***sitten me menimme*** autolla kotiin.
48) Melkein joka päivä kävin meressä uimassa ***ja otin aurinkoa***.

Perfektimuotoisilla n-grammeilla sen sijaan ilmaistaan kyseiselle tempukselle tyypillisen käytön tapaan (ks. VISK § 1534) päättynyttä tilannetta painottamaan tilanteen vaikutusta tai tulosta (49) sekä aiemmin alkaneen tilanteen jatkuvuutta (50–52). Pluskvamperfektillä ilmaistaan samaten kokonaan päättynyttä, mutta kuitenkin imperfektiä aiemmin tapahtunutta, tilannetta (53–54). Useat sen osumista ovat myös peräisin kertomuksiksi luokiteltavista teksteistä (55–56).

49) ***Olen ollut töissä*** kaupassa. Nykyisin olen ravintolassa tarjoilija.
50) ***Olen lapsuudesta asti*** harrastanut laulamista.
51) Koska ***olen opiskellut suomea***, on mahdollinen, että menen Suomeen.
52) En kerää mitään, ***en ole koskaan ollut*** kiinnostunut tästä.
53) Hän oli oikeastaan ruotsikielinen, ***mutta hän oli*** oppinut suomea ja päätti, että kielen täytyi olla myös kirjakieli.
54) Elokuva alkaa, kun Elina voi jälleen mennä kouluun, ***koska hän oli*** ollut sairas
55) Yhtäkkiä tuo nainen otti Karinin laukku! Karin rupesi huutamaan mutta ***hän oli jo*** mennyt.
56) Isästä Matti on maanpetturi, koska ***hän ei ollut*** taistellut sodissa.

Kuten luvussa 2.3.5 mainittiin, tempuksia oppijansuomen osalta ovat aiemmin kartoittaneet – edellä mainitun Haapalan (2008) lisäksi – ainakin Valmu (2007), Ohvo (2008) ja Virtanen (2011). Ohvon (2008) tutkimus keskittyi YKI-testien perustason, keskitason ja ylimmän tason kirjoitelmissa käytettäviin menneen ajan aikamuotoihin. Eri taitotasojen tempusten käyttöä sekä niissä ilmeneviä virhetyyppejä vertailtiin tutkimuksessa keskenään. Tutkimuksen tulokset ovat varsin samassa linjassa n-grammeissa ilmenevän tempuskäytön kanssa: menneistä aikamuodoista käytetyin on imperfekti, joka on erityisen suosittu YKI-testien kirjoitetuissa teksteissä etenkin perus- ja keskitasolla. Imperfektiä käytetään myös huomattavan paljon oikein, joskin sitä korvataan myös tasaisesti muilla aikamuodoilla silloinkin, kun tekstiyhteyden perusteella sitä kuuluisi käyttää. Etenkin imperfektin ja perfektin välistä eroa on keskitasolla toisinaan vaikea hahmottaa. Saman havainnon teki myös Valmu (2007) tutkimuksessaan unkarilaisten suomenoppijoiden tempusten käytöstä sekä Virtanen (2011) venäjänkielisten suomenoppijoiden osalta. Sama ilmiö on mahdollista huomata myös verbillisistä n-grammeista ilmenevistä aikamuodoista. Esimerkiksi seuraavissa esimerkeissä imperfekti on korvattu perfektillä, mikä saa virkkeet kuulostamaan epäidiomaattisilta; ne kuvaavat selvästi ennen puhehetkeä päättynyttä tilanne, mitä suomessa ilmaistaan imperfektillä tai pluskvamperfektillä (ks. VISK § 1530; § 1540). Esimerkissä (57) on myös mukana ylimääräinen *olla*-verbi.

57) Ennen tuntia mina ***en ole ollut*** tietänyt suomen uskonnosta.
58) [– –] Nanin, oli opettaja koulussa Kitessä. Hän tykäsi opettaa siis hän päätti mennä töihin sinne. ***Kun he ovat*** olleet Kiteessä, Aune kasvoi nopeasti ja kauniisti ja jonkin ajan kuluttui hän suostui mennä Joensuuhun lopettamaan koulua.

N-grammien aikamuodoissa itsessään ei ehkäpä edellä esitetyn valossa ole varsinaisesti yllättävää uutta tietoa. Toisaalta saadut tulokset tukevat sitä, että preesensmuodot ovat ylivoimaisen suosittuja ainakin kielitaidoltaan keskitasoisten suomenoppijoiden tekstituotoksissa ja että menneen ajan aikamuodot menevät helposti sekaisin keskenään.

6.1.2 Lausetyyppianalyysi

Lauseita voidaan tyypitellä eri luokkiin joko sen mukaan, mitä jäseniä niissä verbin lisäksi on tai millaisia puhefunktioita ne ilmentävät (VISK § 886; § 891). Tässä tutkimuksessa lausetyypillä viitataan ja lausetyyppianalyysissa keskitytään nimenomaan niin sanottuihin syntaktisiin eli erilaisten lauseopillisten rakenteiden kautta analysoituihin lausetyyppeihin (VISK § 891–906) erotuksena modaalisista, puhefunktioita ilmaisevista, lausetyypeistä eli niin kutsutuista väite-, kysymys-, käsky- ja huudahduslauseista (ks. VISK § 886–890). Suomen kielen lausetyyppejä on kuvailtu ja luokiteltu vuosien saatossa useammassa eri kieli- ja lauseopissa hieman vaihtelevin tavoin (ks. esim. Setälä 1952 [1880]; Hakulinen & Karlsson 1995; Vilkuna 2003; White 2008). Vuoden 2004 *Isossa suomen kieliopissa* (VISK § 891) eritellään 11 erilaista syntaktista lausetyyppiä – transitiivi-, intransitiivi-, kopula-, eksistentiaali-, omistus-, ilmiö-, tila-, kvanttori-, tulos-, tunnekausatiivi- ja genetiivialkuinen lause – joista esitettiin asetelma luvussa 2.3.5. Tässä tutkimuksessa tukeudutaan tähän syntaktisten lausetyyppien jaottelutapaan.

B1-aineiston verbillisistä n-grammeista selvitettiin tutkimuksessa siis, mitä yllä mainituista 11 syntaktisesta lausetyypistä ne edustavat. Kuten aikamuotojen myös lausetyyppien toteaminen tukeutumalla ainoastaan verbillisiin n-grammeihin itsessään vailla tietoa niiden käyttökoteksteista onnistuu yksiselitteisesti vain tiettyyn pisteeseen asti: useita verbillisistä n-grammeista on mahdollista käyttää vaihtelevien lausetyyppien yhteyksissä, joten tempusanalyysin tapaan yksittäisten n-grammien esiintymät jakaantuvat monesti useamman eri lausetyypin alle. Tapauksissa, joissa lausetyyppiä ei voitu nähdä suhteellisen yksiselitteisesti suoraan n-grammista, tuli aikamuotojen tapaan siis tarkastella n-grammin eri aineistoesiintymiä KWIC-näkymässä.

Syntaktisten lausetyyppien kartoittaminen aloitettiin kuitenkin tulkinnan kannalta yksiselitteisimmin tiettyjä lausetyyppejä ilmentävistä n-grammeista. Tässä apuna käytettiin hakutoimintoa *Excelissä*, johon kaikki B1-aineiston n-grammit kattava lista oli viety. Toiminnon avulla pystyttiin tekemään n-grammilistalta hakuja erilaisin kielenaineksin, jotka voivat suoraan kieliä tietyistä lausetyypeistä. Näitä ovat esimerkiksi inessiivin ja adessiivin päätteet *-ssA* ja *-llA*, jotka kuuluvat erilaisissa lausetyypeissä teemapaikan ottavien nominien osiksi. Päätteet auttoivat omistuslausetta ilmentävien n-grammien, kuten *minu**lla** on aikaa* ja *mei**llä** on myös*, niin kuin myös eksistentiaalilauseiden, kuten *toise**ssa** kerroksessa on* ja *oven lähe**llä** on*, löytämisessä. Tällaisten ilmeisen selkeästi tiettyä lausetyyppiä edustavien n-grammien kaikki osumat merkittiin suoraan kyseisen lausetyypin alle *Excelissä*. Täten siis esimerkiksi 3-grammin *minulla on paljon*, jonka frekvenssi on 43, merkittiin edustavan jokaisella 43 esiintymällään

omistuslausetta ilman, että n-grammin jokainen esiintymä tarkistettiin erikseen KWIC-näkymässä. Samoin esimerkiksi 4-grammin *sen jälkeen laitan aamiaisen* (f = 11) esiintymät merkittiin jokaisen olevan peräisin transitiivilauseista, 3-grammin *huoneessa on vasemmalla* (f = 13) eksistentiaalilauseista, 3-grammin *minun täytyy opiskella* (f = 25) genetiivialkuisista lauseista ja niin edelleen. Tarkempi konkordanssien tutkiminen saattaisi tässäkin yhteydessä paljastaa, että n-grammeja on käytetty paikoitellen oletusten vastaisisilla tavoilla ja että joillain niistä onkin haluttu ilmaista toista lausetyyppiä kuin miltä päälle päin näyttää. Tällaiset tapaukset ovat kuitenkin todennäköisesti jälleen yksittäisiä, eikä konkordanssirivien suuren määrän vuoksi jokaista niistä ole mahdollista käydä läpi vain asiasta varmistumisen vuoksi.

Taivutuspäätteitä hakusanoina käyttämälläkään ei voitu löytää sellaisia verbillisiä n-grammeja, jotka kuuluvat osiksi vaikkapa juuri omistus- tai eksistentiaalilauseita, mutta joista itsestään ei voida lukea niiden teemaa eli esimerkiksi omistajaa tai paikkaa ilmaisevaa sanaa, kuten 3-grammien *ei ole aikaa* (esim. omistuslause ”Hänellä *ei ole aikaa* hoitaa Sophia”) ja *on kaksi sänkyä* (esim. eksistentiaalilause ”Poikien huoneessa *on kaksi sänkyä* ja vaatekaappi”) tapauksissa. Eniten konkordanssitarkasteluja aiheuttivat jälleen *olla*- ja kieltoverbilliset n-grammit, jotka voivat olla osana useita erilaisia lausetyyppejä. Esimerkiksi 3-grammeista *ei ole vielä* (f = 29), *on niin paljon* (f = 23) ja *ei ole ollut* (f = 17) ei voida sellaisinaan yksiselitteisesti nähdä, mitä syntaktista lausetyyppiä ne edustavat, sillä niistä ei voida lukea lauseen teemapaikkaa. Tällaisia n-grammeja katsottiin KWIC-näkymässä, jonka perusteella niiden ilmentämät lausetyypit pääteltiin. Seuraavassa on esitetty esimerkin vuoksi 3-grammin *on neljä huonetta* konkordanssinäkymä, joka kattaa kyseisen 3-grammin kaikki esiintymät aineistossa.

Hit	KWIC			File
1	rrostalossa ensimäisessä kerroksessä. Meillä	*on neljä huonetta.*	Asun viihtyisässä huonessa sis	VI0004.txt
2	alon toisen kerroksen asunnossa. Asunnossa	*on neljä huonetta.*	Huoneessa on vasemmalla sän	VI0050a.txt
3	sa. Asunnonsa on hyvin suuri. Asunnollansa	*on neljä huonetta.*	Keittiö, olonuone ja kaksi mak	TS0016d.txt
4	ä kerroksessa. Talossa on myös hissi. Meillä	*on neljä huonetta,*	keittiö, kylpyhuone ja vessa. V	VI0033a.txt
5	olmannessa kerroksessa. Meidän asunnossa	*on neljä huonetta,*	keittiö, kylpyhuone ja toaletti.	VI0048.txt
6	e sijaitsee kolmannessa kerroksessa ja siinä	*on neljä huonetta,*	keittiö, vessa, kylpyhuone, etei	VI0072d.txt
7	ni, mutta kaunis vanha kaupunki. Asunnossa	*on neljä huonetta,*	keittiö, eteinen, kylpyhuone, ve	VI0259a.txt
8	veli viihtyymme maalla. Meidän asuntossa	*on neljä huonetta,*	keittiö, kylpyhuone ja vessa se	VI0302a.txt
9	ntolassa, ensimmäisessä kerroksessa. Meillä	*on neljä huonetta.*	Kotona asuvat minun vanhemp	VI0198.txt
10	:)]. Asuntoni on tarpeeksi suuri. Asuntossa	*on neljä huonetta,*	kylpyhoune, vessa, keittiö ja pa	VI0126a.txt
11	sun kolmannessa kerroksessa. Huoneistossa	*on neljä huonetta.*	Mielestäni minulla on aika kau	VI0180b.txt
12	Koiran niemi on Fritzi. Toisessa kerroksessa	*on neljä huonetta:*	minun huoneeni, veljeni huone	SA0195d.txt
13	summe neljännessä kerroksessa. Asunnossa	*on neljä huonetta,*	yksi keittiö, vessa, kyplpyhuon	VI0061d.txt

Konkordanssi 6. 3-grammin *on neljä huonetta* esiintymät (n = 13) B1-aineistossa.

Konkordanssista voidaan nähdä, että kyseinen 3-grammi kuuluu B1-aineistossa osaksi sekä eksistentiaali- (rivit 2, 5–8, 10–13) että omistuslauseita (1, 4, 9). Rivin 3 esiintymä on tulkinnanvaraisempi, sillä siinä on piirteitä kummastakin lausetyypistä. Tällaiset tapaukset eivät

ole mitenkään tavattomia, ja Ivaska (2011) onkin todennut (oppijan)kielestä ja lausetyyppien sekoittumisesta siinä seuraavaa:

> Todellisen kielen tekstilauseet eivät ole millään lailla selvärajaisia kategorioita, vaan niissä sekoittuu usein piirteitä useista eri lausetyypeistä. Oppijankielessä ilmenevä epäidiomaattisuus voidaan kuitenkin monesti nähdä koko rakenteen epäidiomaattisuutena, joka puolestaan juontaa eri lausetyypeille ominaisten elementtien yhdistymisestä yhdessä lauseessa. (Ivaska 2011: 65.)

Ei-niin-selvärajaisissa tapauksissa lausetyypin päättämisessä tulikin tukeutua jossain määrin intuitioon. Joissain tapauksissa n-grammi voitiin merkitä lausetyypiltään kokonaan tulkitsemattomaksi, mikäli lausetyyppi ei ollut edes jossain määrin selkeästi hahmotettavissa ja suhteellisen pienin ponnistuksin pääteltävissä; konkordanssirivien suuresta määrästä johtui jälleen se, ettei yksittäisen n-grammin lausetyypin pohtimiseen ja päättämiseen ollut mielekästä käyttää ylettömästi aikaa. Edeltävän esimerkin tapauksessa rivin 3 lauseen tulkittiin kallistuvan enemmän eksistentiaalilauseeseen, sillä *asunnon* ei nähty edustavan tyypillistä, elollista omistajaa. 3-grammin *on neljä huonetta* osalta *Exceliin* merkittiin siis 10 esiintymää eksistentiaalilauseelle ja kolme omistuslauseelle.

Tulkinnan kannalta epäselviä tapauksia aiheuttivat myös esimerkiksi sellaiset verbillisistä n-grammeista, jotka saattavat itsessään toimia muiden kotekstissaan olevien lauseiden objekteina. Tällaisiksi havaitut n-grammit merkittiin lausetyypiltään tulkitsemattomiksi. Kahden lauseen rajalla olevien n-grammien lausetyypit taas pääteltiin niissä ensimmäisenä olevan verbin perusteella. Esimerkiksi 4-grammi *juon kahvia ja syön* (f = 9) on kahden lauseen rajalla, ja siinä on kaksi pääverbiä, joista ensimmäinen on transitiivinen mutta toinen voisi periaatteessa olla myös intransitiivinen, mikäli siihen ei ole liitetty objektia. Tällaisten n-grammien lausetyyppi joka tapauksessa siis pääteltiin n-grammin ensimmäisen verbin perusteella, toisin sanoen edeltävän 3-grammin kaikkien esiintymien tulkittiin edustavan transitiivilausetta riippumatta siitä, saiko *syödä*-verbi jokaisessa kotekstissaan objektin osakseen. Ylipäänsä kaikkien transitiiviverbin sisältävien n-grammien, esimerkiksi 3-grammin *kotiin ja syön* (f = 18), kaikkien esiintymien merkittiin automaattisesti edustavan transitiivilausetta ilman KWIC-tarkastelua, vaikkakin siitä voisi ilmetä myös objektittomia vastineita n-grammille. Ivaska (2011) näkee lauseenjäsenten puuttumisen edistävän lausetyyppien mahdollista sekoittumista keskenään, ja puuttuvien lause-elementtien olettamisen tai olettamatta jättämisen olevan hyvin tilannekohtaista. Tämän analyysin yhteydessä oli luontevampaa olettaa, että esimerkiksi transitiiviverbillisiin n-grammeihin sisältyy aina objekti, jolloin ne edustavat lähtökohtaisesti aina transitiivilauseita.

Taulukossa 12 on esitetty laskelmat n-grammien edustamista syntaktisista lausetyypeistä frekvensseineen ja prosenttiosuuksineen. Taulukon arvot eivät voi edellä mainituitten joustojen vuoksi edustaa absoluuttisia totuuksia B1-aineiston verbillisten n-grammien ilmentämistä lausetyypeistä, mutta antavat kuitenkin suhteellisen tarkan osviitan siitä, missä määrin mikäkin syntaktinen lausetyyppi on mukana muodostamassa verbillistä n-grammia.

Taulukko 12. B1-aineiston verbillisten n-grammien jakauma syntaktisiin lausetyyppeihin.

Lausetyyppi	**Frekvenssi**	**%**
Kopulalause	5 075	35,5
Omistuslause	2 595	18,2
Intransitiivilause	2 358	16,5
Transitiivilause	2 330	16,3
Eksistentiaalilause	1 561	10,9
Genetiivialkuinen lause	183	1,3
Tilalause	36	0,3
Tuloslause	5	~0,1
Tulkitsemattomat	125	0,9
Yhteensä	14 268	100

B1-aineiston n-grammien edustamista lausetyypeistä selvästi yleisin on siis kopulalause, jota seuraavat omistus-, intransitiivi- ja transitiivilauseet. Yhteensä n-grammeja, jotka edustavat *Ison suomen kieliopin* (VISK § 891) mukaisia monikäyttöisiä lausetyyppejä on peräti 68 prosenttia kaikista verbillisistä n-grammeista. Erikoislausetyypeistä omistuslauseen ohella myös eksistentiaalilause saa runsaasti esiintymiä osakseen, siinä missä B1-aineiston verbillisistä n-grammeista ei sen sijaan yksikään ilmennä ilmiö-, kvanttori- tai tunnekausatiivilausetta. Genetiivialkuiset lauseet sekä tila- ja tuloslauseetkin jäävät lähinnä marginaalitapauksiksi. Seuraavaksi esitellään tarkemmin verbillisissä n-grammeissa eniten ilmenneiden lausetyyppien määritelmät, esimerkkejä niiden esiintymistä B1-aineistossa sekä yhteyksiä asiasta aiemmin saatuun tutkimustietoon.

Monikäyttöisiin lausetyyppeihin kuuluvat *Ison suomen kieliopin* mukaan siis transitiivi-, intransitiivi- ja kopulalauseet. Näistä transitiivilauseella tarkoitetaan lähtökohtaisesti lausetta, jonka verbin valenssiin kuuluu objekti. Intransitiivilause on päinvastaisesti objektiton lausetyyppi. Transitiivilause ilmaisee siis tyypillisimmillään tekoa, joka pyrkii tai johtaa tulokseen, intransitiivilause taas suuntautumatonta toimintaa. Kopulalause on sen sijaan lausetyyppi, jonka verbinä on *olla*. Kopulalauseen jäseniin lukeutuvat subjektin lisäksi predikatiivi tai adverbiaali.

Aspekteiltaan kopulalauseet ovat tilankuvauksia eli ne kuvaavat yhtäjaksoisia tilanteita, joilla ei ole luontaista päätepistettä. Tilassa tyypillisimmillään joku tai jokin on jonkinlainen tai jossakin, kuuluu tai sisältyy johonkin. (VISK § 891; § 1502.) Alla on varsin prototyyppisiä esimerkkejä niin n-grammiesiintymien ilmaisemista transitiivi- (59–60), intransitiivi- (61–62) kuin kopulalauseistakin (63–64).

59) Illala ***laitan ruokaa ja*** sitten minun täytyy opiskella.
60) Minun koulutus***: olen käynyt peruskoulun ja lukion*** Viljandissa.
61) Minulla ei ole vielä verhettä eikä mattoa, mutta ***pidän oikein paljon*** uudesta huonesta.
62) Sitten menen keittiöön ja juon kahvia. ***Sen jälkeen pukeudun*** ja laitan koulutavarat valmiiksi.
63) ***Hän on iloinen***, puhelias ja sosiaalinen.
64) En seuraa painasi eikkä ole painovartija, ***en ole ollut*** dietillä.

Aiemmin transitiivilausetta oppijansuomessa sekä sen käytön kehittymistä EVK:n (2003) taitotasoilla on tutkinut ainakin Reiman niin aikuisten (2011) kuin yläkouluikäistenkin (2014) suomenoppijoiden osalta. Hänen tutkimuksensa (2014) mukaan aikuisten suomenoppijoiden transitiivi-ilmausten määrä kasvaa eniten B1-tasolla, jonka jälkeen transitiivin käyttö alkaa vakiintua. Erilaisten lausetason ilmiöiden yhteydessä etenkin passiivinen (esim. *Lippu on jo maksettu*) ja infiniittinen (*Menen hakemaan paketin*) transitiivilauseen käyttö lisääntyy tasaisesti B-tasoilla. Passiivisesta transitiivilauseen käytöistä ei näy viitteitä B1-aineiston verbillisissä n-grammeissa, mutta infiniittisestä käytöstä on jonkin verran osumia ainakin *opiskella*-verbin yhteydessä (65). *Olla*-verbin käytöstä oppijansuomessa, kopulalauseet mukaan lukien, tutkimusta on tehnyt ainakin Kynsijärvi (2007). Hänen tutkimuksessaan YKI-aineiston eri taitotasojen kirjoitelmista lähes puolet *olla*-verbin kopulakäytöstä juontaa ominaisuuksien kuvailusta, mikä vaikuttaisi olevan yleisin käyttö myös kopulalausetta ilmaiseville B1-aineiston verbillisille n-grammeille (66–67). Samoin luokittelua eli luokan määräämistä tai nimeämistä Kynsijärven tulkitsemissa kopulalauseissa oli mukana runsaasti, mistä on samaten viitteitä verbillisissä n-grammeissakin (68–69), tosin huomattavasti ominaisuuksien kuvailua vähemmän.

65) [%] kirjoitin ylioppilaaksi ***ja samana vuonna pääsin yliopistoon opiskelemaan*** hammaslääketiedettä.
66) ***Hän on hyvin*** ammattimainen ja hän rakastaa hänen työnsää.
67) Jos mä ***en ole vielä*** liian väsynyt lähden baari tai kylässä.
68) ***Äitini on opettaja***.
69) ***Hän on ollut*** kirurgi jo yli % vuotta ja hän tekee vuodessa kaksisata leikkausta.

Erikoislausetyyppeihin sisältyy useampia vaihtelevia lausetyyppejä, ja ne ovat monikäyttöisiä lausetyyppejä selkeämmin rakenteita, jotka ovat vakiintuneet tiettyihin merkityksiin. Tunnekausatiivilausetta lukuun ottamatta ne ovat käytännössä intransitiivisia. Eksistentiaalilause (e-lause) on erikoislausetyypeistä keskeisimpiä, ja omistus- ja tilalauseet voidaan

(ilmiölauseen ohella) luokitella sen alatyypeiksi (VISK § 891). Prototyyppisimmillään e-lauseen verbinä toimii *olla*-verbi ja teemapaikassa on paikanilmaus, jolla lause alkaa. Subjektin paikka on verbin jäljessä. E-lauseessa subjektin ollessa jaollistarkoitteinen on se yleensä partitiivissa, kuten lauseessa *Lasissa on maitoa*. (VISK § 893.) Yleisesti sellaisia lauseita, joissa subjekti on tai voisi olla partitiivissa, onkin suomen kieliopissa pitkään nimitetty e-lauseiksi (Vilkuna 2003: 116). Omistuslauseella on samat ominaisuudet kuin e-lauseella, mutta siinä omistettavaa ilmaiseva nominilauseke (NP) on usein määräinen, ja siinä asemassa oleva persoonapronomini saa sijakseen epäsubjektimaisen akkusatiivin kuten lauseessa *Onneksi minulla on sinut*. Omistuslauseen omistaja on usein elollinen, jolloin sen voi nähdä subjektimaisemmaksi kuin e-lauseen paikanilmauksen. (VISK § 895.) Alla on muutamia esimerkkejä B1-aineiston n-grammiesiintymien ilmaisemista eksistentiaali- (70–71) ja omistuslauseista (72–73).

70) Olohuoneessa on musta sohva, tv-pöytä ja sohvapöytä. Siellä ***on myös iso*** kirjahylly ja pieni vitriinikaappi.
71) ***Minun perheessä on*** viisi henkilöä.
72) ***Jos sinulla on*** liikennekortti,sinä maksat halvemman hinnan.
73) Minä en ole naimisissa ja ***minulla ei ole vielä*** lapsia.

Tutkimusta e-lauseesta oppijansuomessa on tehnyt laajemmassa mittakaavassa ainakin Kajander (2013), joka väitöstutkimuksessaan perehtyi CEFLING-korpuksen eri EVK:n (2003) taitotasoille arvioitujen kirjoitelmien e-lauseisiin. Kajanderin tutkimuksessa e-lauseisiin määriteltiin tosin kuuluviksi muun muassa omistus- ja ilmiölauseet, joita tässä tutkimuksessa pidetään omina lausetyyppeinään, joten tutkimusten suoranainen vertailu on hankalaa. Kajanderin tutkimuksessa havaittiin kuitenkin yleisesti B1-tasolla e-lauseissa subjektin tarkkuuden kasvavan ja persoonapronominin käytön omistajanilmauksena sekä *olla*-verbin käytön vähenevän merkittävästi.

E-lauseen alakategorioista tilalause on persoonaton lausetyyppi, jossa verbi on yksikön kolmannessa persoonassa ja teemapaikassa on ajan tai paikan adverbiaali. Lause ilmaisee tyypillisesti säätiloja, vuorokaudenaikoihin liittyviä muutoksia ja muita ulkoisen ympäristön aistittavia ominaisuuksia. Tilalauseita ovat esimerkiksi lauseet *Täällä tuulee* ja *Huoneessa oli siistiä*. (VISK § 900.) Säätilojen ilmaisuja sisältyi muutamaan verbilliseen n-grammiin (74). Genetiivialkuisia lausetyyppejä on olemassa kolmenlaisia: nesessiivirakenteita, *tuli tehtyä* -rakenteita sekä kokijan genetiivirakenteita (VISK § 906). Näistä ainoastaan ensin mainitulle löytyy edustuksia verbillisistä n-grammeista (75–76).

74) Tänään ***on kaunis ilma***, aurinko paistaa ja taivas on sininen.

75) ***Minun täytyy opiskella*** paljon ja siitä syystä on aina paljon tekemistä.
76) Ehkä ***minun täytyy mennä*** hämmäslääkärille vielä kerran, mutta minulla ei ole rahaa ja vierailu on liian kallis.

B1-aineiston n-grammeista löytyviä nesessiiviverbejä *täytyä* ja *pitää* on aiemmin ICLFI:stä ja sen eri taitotasoilta tutkinut ainakin Haltia (2015). Hänen tutkimuksensa perusteella *täytyä* on koko korpuksessa selvästi yleisin välttämättömyyttä ilmaiseva modaaliverbi. *Täytyä* on frekvenssillä 11 samaten yleisin B1-aineiston n-grammien nesessiiviverbeistä (ks. taulukko 7), joskin *pitää*-verbikin saa 10 esiintymää osakseen. *Pitää* tosin esiintyy B1-aineiston n-grammeissa usein modaalisen merkityksensä sijaan merkityksessä 'tykätä'. Haltian tutkimuksen perusteella *täytyä*- ja *pitää*-verbeillä on B1-tasolla selkeä työnjako, jossa *täytyä* toimii henkilöä koskevien velvollisuuksien ilmaisussa, siinä missä *pitää*-verbillä taas ilmaistaan kirjoittajan tekemiä toimintaehdotuksia ja kannanottoja. Myös B1-aineiston verbillisissä n-grammeissa *täytyä* toteuttaa samaa tehtävää (77), mutta *pitää*-verbille Haltian kuvailemaa käyttöä ei ole havaittavissa. Sitä käytetään samoihin funktioihin kuin *täytyä*-verbiäkin (78).

77) Koulun jälken menen kirjastoon tai kotiin ***koska minun täytyy*** lukea paljon.
78) ***En halua nousta, mutta nousta pitää***, kosko koulu odottaa.

Kaiken kaikkiaan voidaan todeta, että tässä esitetty tutkimuksen syntaktisten lausetyyppien analyysi jäi B1-aineiston verbillisten n-grammien runsaan määrän vuoksi varsin pintapuoleiseksi. Myös lausetyypeistä tehtyjä nostoja voidaan pitää jossain määrin sattumanvaraisina poimintoina. Lausetyyppikartoituksella saatiin kuitenkin muodostettua kuvaa siitä, mitä syntaktisista lausetyypeistä suomenoppijat missäkin määrin B1-taitotasolla suosivat ja mitä tyypillisiä käyttötapoja niille on B1-aineiston n-grammeissa.

6.1.3 Muita huomioita verbillisistä n-grammeista

Samalla kun B1-aineiston verbillisten n-grammien tempuksia ja syntaktisia lausetyyppejä kartoitettiin, tarkasteltiin niitä myös päällisin puolin muiltakin kanteilta. Kuten luvussa 5.1 käänteisen sanalistan avulla huomattiin, n-grammeja, joiden finiittiverbi on taipunut yksikön ensimmäisessä persoonassa, on yhteensä 284 erilaista. Tästä voidaan laskea, että yhteensä kolmasosa (33,7 %) kaikista verbillisistä n-grammeista on siis n-grammeja, joita suomenoppijat käyttävät lähtökohtaisesti oman toimintansa kuvaamiseen.

Verbillisistä n-grammeista laskettiin myös, kuinka suureen osaan niistä sisältyy negaatio eli kielto. Täten saatiin tieto siitä, kuinka suuri osa verbillisistä n-grammeista kuuluu osaksi

kielto- ja kuinka moni osaksi myöntölauseita. Koska kieltolauseeseen lukeutuu apuverbinä toimiva kieltosanasta ja lauseen pääverbin kieltomuoto (VISK § 1615), oli negaatiota ilmaisevien n-grammien havaitseminen n-grammilistasta yksiselitteisempää ja yksikertaisempaa kuin tempusten ja lausetyyppien tulkinta, eikä vaatinut erillisiä konkordanssitarkasteluja. Yksittäisestä n-grammista tuli ainoastaan katsoa, onko siinä mukana joko apuverbinä toimiva kieltosana tai kieltomuotoinen pääverbi, mahdollisesti jopa molemmat. B1-aineiston kielteisistä n-grammeista lähes jokaiseen sisältyy varsinainen kieltosana; aineistossa on ainoastaan muutama tapaus, joissa n-grammi edustaa kieltoa, vaikkei varsinainen kieltosana kuulukaan osaksi sitä. Tällainen on esimerkiksi 4-grammi *ole ollut vaihto-oppilaana*. Laskelmien kautta selvisi, että kielteisiä n-grammeja on B1-aineistossa yhteensä 125 kappaletta ja että niiden yhteisfrekvenssi on 2 547. Toisin sanoen verbillisistä n-grammeista yhteensä 12,6 prosenttia on kielteisiä. Kielteiset n-grammit taas muodostavat 15,3 prosentin osuuden kaikista verbillisten n-grammien esiintymistä aineistossa.

Samassa yhteydessä selvitettiin, mitä verbien moduksia eli tapaluokkia verbillisissä n-grammeissa on edustettuina. Modusten tehtävänä on ilmaista modaalista merkitystä ja suhteuttaa puheena oleva tilanne todellisuuteen. Moduksia on suomessa neljä: indikatiivi on tunnukseton muoto, konditionaali ilmaisee asiantilaa vaihtoehtona, imperatiivi taas deonttista ja potentiaali episteemistä modaalisuutta. (VISK § 111, 115, 1590.) Hienoisena yllätyksenä tuli se, että B1-aineistosta löytyi ainoastaan yksi n-grammi, jonka finiittiverbin modus ei ole indikatiivi. Kyseinen n-grammi on konditionaalimuotoinen 3-grammi *jos minulla olisi*, jonka frekvenssi on 11. Sitä käytetään alla olevien esimerkkien tapaan kuvaamaan, mitä kirjoittaja tekisi, jos hänellä olisi jotain substanssia, pääosin aikaa tai rahaa, enemmän tai riittävästi.

79) Loin jo esitelmääni suomen kielen tunnille, mutta en osaa sitä vielä ulkoa, lisäksi tiedän että ***jos minulla olisi*** enemmän aikaa se olisi parempi.
80) ***Jos minulla olisi*** riittävästi rahaa, menisin Itävaltaan ja Sveitsiin, jossa EM-kilpailu tapahtuu.

Potentiaali on hyvin harvinainen modus natiivisuomessakin, eikä imperatiiville ehkäpä ICLFI:n tekstilajit ja tehtävänannot huomioon ottaen ole välttämättä tarvetta ainakaan siinä määrin, että samanlaiset imperatiivirakenteet toistuisivat useasti ja muodostaisivat täten n-grammeja. Myös ainakin Haapala (2008: 56) on laskenut tutkimuksessaan oppijansuomen moduksia, ja hänen havaintojensa perusteella YKI-testien aineiston keskitason teksteissä käytetään moduksista indikatiivia 90 prosenttia, konditionaalia 8 prosenttia ja imperatiivia loput kaksi prosenttia. Tämän tutkimuksen tutkimushypoteesin osalta indikatiivin odotettiinkin edustuvan

mitä suurimmissa määrin n-grammeissa, mutta jossain määrin yllättävää oli kuitenkin se, ettei konditionaalia edustavia n-grammeja ollutkaan B1-aineistossa yksittäistä osumaa enempää.

Etenkin pitkissä, viidestä ja kuudesta sanasta koostuvissa verbillisissä n-grammeissa voidaan havaita vielä ainakin yksi kiinnostava ilmiö: niissä hyödynnetään jonkin verran rakenteita, jotka vaikuttavat olevan suoraan muista teksteistä kopioituja. Tällaisia ovat ainakin alla olevien esimerkkien n-grammit, joista yksi (81) voidaan tunnistaa *Punahilkka*-sadusta[45] ja kahden muun (82–83) voidaan ajatella olevan mahdollisesti peräisin muunlaisista yhteyksistä.

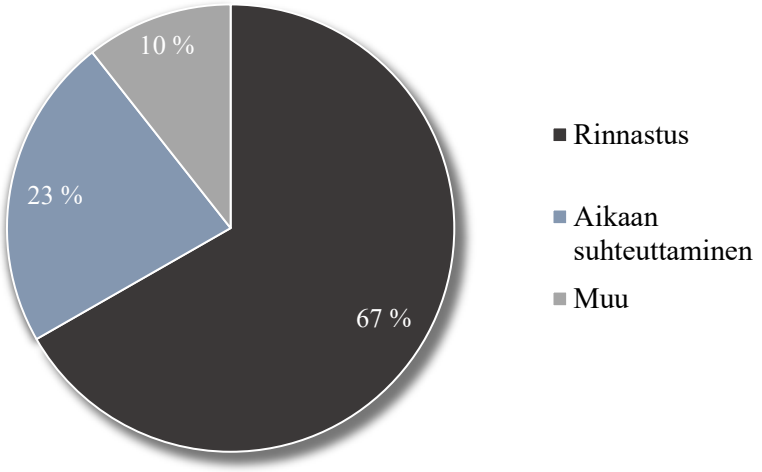
81) Ja sitten kysyi: ”Mummo, ***miksi sinulla on niin isot korvat***?” ”Se koska voisin kuulla paremmin sinut.”
82) Mitän ***voin käyttää hyväksi kurssin tietoja*** kääntämisessä?
83) Ilmat ovat lämpimämpiä ja elämä hauskempaa, yöt eivät ole enää niin pimeitä ja pitkiä, lumi sulaa, ***puihin tulee vähitellen vihreitä lehtiä***.

Näyttää siis siltä, että osa pitemmistä n-grammeista on mahdollisesti poimittu suoraan kopioiden joistain muista tekstiyhteydestä. *Punahilkka*-esimerkin kanssa tämä on varsin selvää. Esimerkin (82) tapauksessa n-grammi taas muodostuu esseetekstiä jäsentävästä kysymyksestä, joka toistuu kääntämisen opiskelijoiden teksteissä tehtävänannon motivoimana, todennäköisesti valmiiksi annettuna. Esimerkki (83) on siinä määrin kiintoisa, että kyseinen 5-grammi ei vaikuta mitenkään odotuksenmukaiselta saati tavanomaiselta sanayhtymältä, mutta silti se esiintyy B1-aineistossa yhteensä viisi kertaa ja viiden eri kirjoittajan teksteissä. KWIC-näkymä ja sen File-palkki paljastavat tällä kertaa, että kaikki kyseisen 5-grammin esiintymät ovat peräisin virolaisilta suomenoppijoilta. Itse tekstitiedostojen metatiedoista nähdään lisäksi, että ensinnäkin tehtävänantona jokaiselle näistä teksteistä on ollut vapaasti valittava aihe, joka on kuitenkin pitänyt valita oppikirjan teemojen mukaan, ja toiseksi, että jokainen tekstit tuottaneista suomenoppijoista on käyttänyt teosta *Suomi selväksi* (Kuusk 1999) suomen kielen opiskelussaan. Olettamukseni onkin, että kyseinen *puihin tulee vähitellen vihreitä lehtiä* -rakenne esiintyisi esimerkiksi jossain kyseisen oppikirjan tekstikappaleista, ja oppijat ovat yksinkertaisesti kopioineet sen kappaleesta käytettäväksi kirjoitelmissaan. Tätä hypoteesia ei tosin ole varmennettu itse teoksesta.

[45] Tekstien, joista esimerkin (81) tapaisia n-grammeja löytyi, tehtävänantona oli ollut *Punahilkka*-sadun kertominen suomeksi.

6.2 Verbittömistä n-grammeista

Tässä luvussa on keskitytty tähän asti yksinomaan niihin B1-aineiston n-grammeihin, jotka sisältävät verbin, verbiliiton tai verbiliiton osan. On silti tarpeellista sivuta vähintäänkin maininnan tasolla myös sitä, minkälaisia ovat vastaavasti ne B1-oppijansuomen n-grammit, joissa verbiä ei ole lainkaan ja millaisia funktioita niillä vaikuttaisi B1-tasoisten suomenoppijoiden kielenkäytössä olevan. Kuten luvun alussa kerrottiin, verbittömiä n-grammeja B1-aineiston kaikista n-grammeista (n = 992) on yhteensä 151 kappaletta eli noin 15,5 prosenttia. Näissä n-grammeissa näyttäisi toistuvan ainakin kaksi selkeämpää funktiota ylitse muiden: asioiden rinnastaminen toisiinsa konjunktioin sekä toiminnan suhteuttaminen aikaan. Kuviossa 2 on esitetty verbittömien n-grammien jakautuminen erilaisten funktioiden ilmentämisen mukaan.



Kuvio 2. B1-aineiston verbittömien n-grammien jakauma funktioiden mukaan.

Rinnastuskonjunktiolliset (verbittömät) n-grammit ovat yksinkertaisesti sellaisia n-grammeja, joihin sisältyy jokin rinnastuskonjunktioista, ja niitä on siis yhteensä kaksi kolmasosaa (n = 100) verbittömistä n-grammeista. Käytetyin rinnastuskonjunktio näissä n-grammeissa on *ja* (joka on myös toiseksi frekventein n-grammien lemma, ks. taulukko 7), jota hyödynnetään suurilta osin asioiden listaamisessa (84–86). Samassa roolissa toimii myös *tai*-konjunktio (87). Listaamisen ohella *ja* toimii verbittömissä n-grammeissa myös lauseiden rinnastajana (88) *mutta*-konjunktion ohella (89–90).

84) Sängyn vieressä on pieni ***kirjoituspöytä ja tuoli***.
85) Aamiaiseksi syön tavallisesti voileipää sekä juon ***kahvia kerman ja sokerin kanssa***.
86) Hänellä on ***tumma tukka ja harmaat silmät***.

87) Tavallisesti minulla on ***kaksi tai kolme*** luennoa päivässä ja niiden jälkeen menen suoraan kotiin ja tekemään lounasta.
88) Odottin vähän aikaa ***ja sitten hän*** ilmestyi marketin pihalla.
89) Hän oli sihteeri, ***mutta nyt hän*** on hieroja.
90) Kanerva aloi nauraa, ***mutta kun hän*** lakasi nauramasta, hän otti pienen palasen.

Viimeinen kolmasosa verbittömistä n-grammeista jakautuu aikaan suhteuttamisen ja muunlaisten n-grammien välille. Aikaan suhteuttamista esiintyy yhteensä 34 erilaisessa verbittömässä n-grammissa. Näistä useampi on kellonaikojen ilmaisua vaihtelevin rakentein sekä sanan *kello* kanssa (91–93), että ilman sitä (94–95). Aikaan suhteuttaviksi laskettiin myös niin 3-grammit *kaksi kertaa viikossa* (96), *melkein joka päivä* (97) ja *[%] vuotta sitten* (98), 4-grammi *vuodesta [%] vuoteen [%]* (99), kuten erilaiset *sen jälkeen* -rakenteetkin (100–101). Muutaman verbittömistä n-grammeista voisi lukea sekä aika- että rinnastusryhmään kuuluviksi (esim. 3-grammi *maanantaisin ja keskiviikkoisin*), mutta tässä yhteydessä kaikki rinnastuskonjunktiolliset verbittömät n-grammit lajiteltiin kuitenkin ensisijassa rinnastusluokkaan. *Kello*-tyyppisistä ajanilmauksista on huomattavissa yhtymäkohta Jantusen (2017) tutkimukseen: tämänkin tutkimuksen n-grammien perusteella suomenoppijat vaikuttavat suosivan ainoastaan *kellon* ja nominatiivin yhdistelmää, eivätkä käytä ablatiivista ajanilmaisutapaa.

91) ***Noin kello [%]*** kävimme asemalla ja sitten he meni Hangzhouhun junalla
92) Yleensä herään ***kello puoli seitsemän***.
93) Palaan ***kotiin noin kello*** neljä tai kello kuusi.
94) Seuraava luento on mikrobiologia, joka alkaa ***vartin yli kymmenen*** ja kestää myös puolitoista tuntia.
95) Syön aamiaisen ja lähden ***kotoa varttia vaille*** kahdeksalta.
96) Olen kurssilla ***kaksi kertaa viikossa***, maanantaisin ja torstaisin.
97) Syksy on minusta tylsin vuodenaika vuodessa, koska sitten sataa ***melkein joka päivä*** vettä sekä välillä voi tulla räntääkin.
98) Minä ja Mikko, minun mieheni, tapasimme ***[%] vuotta sitten*** milloin hän muutti Turkuun.
99) ***Vuodesta [%] vuoteen [%]*** opiskelin viulunsoittoa musiikkikoulussa.
100) ***Sen jälkeen me*** matkustimme takaisin Saksalle.
101) Vuonna [%] saattiin sähköinen valaistus, ***sen jälkeen kun*** Helsinki oli ottanut puhelimen käyttöön jo vuonna [%].

Myös Varis (2010) on kartoittanut ajanilmauksia oppijansuomessa EVK:n (2003) taitotasoilla aineistonaan Cefling-hankkeen nuorten suomenoppijoiden kirjoitelmat. Hänen tutkimuksensa perusteella B1-taitotasolla oppijat ilmaisevat aikaa eniten *kun*-lauseella (esim. *Kun minä olin lapsi*) ja adverbilla *sitten*, joista kumpiakin löytyy siis myös tämän tutkimuksen verbittömistä n-grammeista. Ajanilmausten hallitsemisen tarkkuus on Variksen (2010) tutkimuksessa B1-tasolla jo hyvin korkea, yli 83 prosenttia, eikä verbittömien n-grammien ajanilmauksissakaan ole havaittavissa juurikaan virheellisyyksiä.

Loput verbittömistä n-grammeista (n = 16) eivät vaikuttaisi muodostavan mitään järin suuria yhtenäisiä ryhmiä, joten ne laskettiin muu-luokkaan kuuluviksi. Niihin lukeutuu niin

sanaliittoja (102), luettelon osia ilman rinnastuskonjunktiota (103) kuin paikan ja mielipiteen ilmaisujakin (104–105).

102) Sitten minulla oli englannin kielen tunti ja sen jälkeen ***suomen kielen tunti***.
103) Asunnossa on ***kolme huonetta, keittiö***, eteinen, vessa ja kylpyhuone.
104) Vierashuonessa on ***oikealla ikkunan vieressä*** kaksi nojatuolia ja vasemalla sähkötakka.
105) Kotiseutumuseon varajohtaja Denis Kuznetsov on ***sitä mieltä, että*** maailman kulttuuri köyhtyy, kun joku kansan kulttuuri kuolee.

B1-aineiston verbittömät n-grammit näyttävät siis toteuttavan etenkin asioiden listaamisen ja rinnastamisen sekä tapahtumien ja tekemisen aikaan suhteuttamisen funktioita. Vaikkakin nämä n-grammien tehtävät vaikuttavatkin jokseenkin selkeiltä, tulee silti pitää mielessä, että yhdellä ja samallakin n-grammilla voi olla vaihtelevia funktioita sen käyttökonteksteista riippuen, kuten Biber (2006: 139) toteaa.

7 JOHTOPÄÄTÖKSIÄ JA POHDINTAA

7.1 Yhteenvetoa ja päätelmiä – oppijansuomi n-grammien valossa

Maisterintutkielmassani olen tarkastellut, minkälaisia suhteellisen usein toistuvia kolmesta, neljästä, viidestä tai kuudesta perättäisestä sanasta koostuvia fraseologisia yksikköjä eli n-grammeja suomenoppijoiden EVK:n (2003) B1-taitotasoarvioinnin saaneissa kirjoitetuissa teksteissä esiintyy. Tutkimukseni aineistona toimi noin miljoonan saneen ICLFI-korpus (Jantunen, Brunni & Oulun yliopisto 2013) ja sen kaikki B1-taitotasolle arvioidut tekstit. Tämä niin kutsuttu B1-aineisto kattoi yhteensä 2 659 tekstiä ja 409 482 sanetta. Lähestyin aineistoa päättämättä etukäteen, mitä kielen piirteitä haluan siitä n-grammien kautta tutkia eli korpusvetoisesti. Loin tutkimuksessani B1-aineistosta *AntConc*-korpusohjelman (Anthony 2019) avulla listan siinä esiintyvistä, ennalta määrittelemäni raja-arvot ylittävistä, 3-, 4-, 5- ja 6-grammeista. Typistin listaa jonkin verran karsimalla siitä n-grammeja, joiden en katsonut tuovan tutkimukselle juurikaan lisäarvoa. Lopulliseen listaukseen kertyi yhteensä 992 erilaista n-grammia, joiden yhteisfrekvenssi B1-aineistossa oli 16 614. Listan kokoamisen jälkeen tarkastelin ja analysoin n-grammilöydöksiä niiden sanaston ja erilaisten rakennepiirteiden suhteen sekä vertailin n-grammien tarjoamaa tietoa joihinkin edeltäviin oppijansuomen tutkimuksiin.

Tavoitteenani oli selvittää tutkimuksellani etenkin, minkälaisia B1-taitotason suomenoppijoiden tuottamat n-grammit ovat ja mitä ne voivat kertoa heidän suomen kielen käytöstään leksikon ja rakenteiden osalta. Tutkimuksen leksikaalinen analyysi toi ilmi muun muassa, että n-grammien leksikko painottaa yksikön ensimmäistä persoonaa: esimerkiksi lemma MINÄ on n-grammien kolmanneksi frekventein lemma (f = 160) ja verbejä, jotka ovat taipuneet yksikön ensimmäisessä persoonassa käytetään yhteensä 284 n-grammissa (eli n. 28,6 %:ssa n-grammeista). Pintapuoleisen tarkastelun perusteella suomenoppijoiden n-grammeissaan tyypillisimmin käyttämä sanasto on natiivisuomessakin hyvin frekventtiä, joskin eräät lemmat, kuten HUONE, KOTI ja AAMIAINEN saavat huomattavan paljon esiintymiä. Tämä selittyy suomenoppijoiden kirjoittamien tekstien tehtäväannoilla, joissa tehtävänä on ollut kirjoittaa esimerkiksi omasta huoneesta tai päivän kulusta. N-grammien sanojen lemmatut muodot painottuvat vahvimmin nomineihin (45 %) ja verbeihin (32 %), joskin partikkeleillakin (17,2 %) on niissä suurehko osansa. Sekä sananmuotojen että lemmojen osalta n-grammien voidaan havaita toteuttavan niin kutsuttua Zipfin lakia (ks. Zipf 1949), jolla viitataan siihen, että pientä määrää

sanoja tavataan kielessä hyvin tiheästi, siinä missä suurimpaan osaan sanoista törmätään vain harvakseltaan.

Tutkimuksen rakenteita analysoivassa osassa päätin keskittyä pääosin niihin B1-aineiston n-grammeihin, joihin sisältyi verbi, verbiliitto tai osa verbiliitosta. Tämän päätöksen tein sekä näiden niin kutsuttujen verbillisten n-grammien kokonaismäärän että verbin lauseopillisen merkityksellisyyden vuoksi. Analyysissa selvisi, että B1-tasoiset suomenoppijat tukeutuvat teksteissään n-grammien perusteella suurilta osin indikatiivimuotoisiin preesenslauseisiin, jotka ovat useimmiten syntaktiselta lausetyypiltään kopulalauseita (35,5 %). Myös omistus- (18,2 %), intransitiivi- (16,5 %), transitiivi- (16,3 %) ja eksistentiaalilauseet (10,9 %) ovat hyvin usein muodostamassa n-grammeja. Muunlaisten lausetyyppien esiintymät ovat lähinnä marginaalitapauksia. Verbilliset n-grammit ovat useimmiten myöntömuotoisia ja yhtä poikkeusta lukuun ottamatta indikatiivimuotoisia. Tutkimuksen perusteella taas niiden n-grammien, joiden osaksi ei kuulu lainkaan verbiä, selkeimpinä tehtävinä on rinnastaa asioita toisiinsa sekä suhteuttaa tapahtumia aikaan.

Vertailemalla löydettyjä sanasto- ja rakennepiirteitä joihinkin edeltävistä oppijansuomen tutkimuksista huomasin, että n-grammien kautta tehtävät havainnot B1-suomenoppijoiden suomen kielen käytöstä ovat pitkälti linjassa aiemmin selvitetyn kanssa. Tämä korostuu etenkin sellaisten piirteiden suhteen, jotka on aiemmissa tutkimuksissa huomattu erityisen yleisiksi suomenoppijoiden tuotoksissa. Näitä ovat esimerkiksi runsas preesensin ja indikatiivin käyttö sekä tiettyjen adjektiivien ja astemääritteiden suosiminen. Aiemmissa tutkimuksissa osoitetut, mutta niissä luonteeltaan harvinaislaatuisemmiksi todetut seikat eivät välttämättä lopulta heijastuneet lainkaan n-grammeihin, mikä on luonnollista n-grammien toisteisuuden vaatimuksen vuoksi.

Tiivistetysti voidaan siis todeta, että B1-tasoiset suomenoppijat painottavat n-grammeissaan jonkin verran itseään ja puhehetkeen sidoksissa olevia tekemisiään, ja tässä he käyttävät apunaan tavanomaisimpia suomen kielen lausetyyppejä. Monia n-grammeista voi pitää hyvin tehtävänantokeskeisinä. Vaikka tietyllä tapaa jokaisen tämänkaltaisen korpusaineiston n-grammin voidaan ajatella olevan jossain määrin tehtävänannon motivoima – onhan korpuksen teksteistä jokainen tuotettu nimenomaan tiettyyn ennalta määrättyyn tehtävään vastaamiseksi – voi eräistä, etenkin nominipainotteisista n-grammeista, kuten *minun huone on* (f = 70), *laitan aamiaisen ja keitän kahvia* (f = 16), *keittiö, kylpyhuone ja vessa* (f = 16) ja *mukava ja onnellinen perhe* (f =12) lukea korostetun selkeän tehtävänannon vaikutuksen niiden muodostumisessa. Tekstilajin on todettu vaikuttavan siihen, minkälaisia n-grammeja kielestä on löydettävissä (ks. esim. Biber ym. 1999), ja koska tehtävänanto usein ohjaa tekstilajinkin valintaa, voidaan todeta, että myös tehtävänannon vaikutus n-grammeihin on ilmeinen.

N-grammien perusteella on vaikea tehdä varmoja tai yleispäteviä päätelmiä siitä, missä määrin suomenoppijat oppijansuomessaan ja n-grammituotoksissaan mahdollisesti nojautuvat Sinclairin (1991) esittämiin idiomi ja/tai vapaan valinnan periaatteisiin. Täten on hankala myös sanoa varmuudella, miten Hoeyn (2005) esittämä leksikaalinen priming näkyy n-grammeissa. Mikäli dataa olisi myös suomea ensikielenään käyttävien n-grammeista, voitaisiin yrittää vertailla natiivi- ja oppijansuomen eroja sekä n-grammien idiomaattisuudessa että leksikaalisissa primingeissa. N-grammien kautta päästään mielestäni kuitenkin käsiksi ainakin jossain määrin niihin leksikaalisiin primingeihin, joita suomenoppijat keskenään jakavat. N-grammien perusteella voidaan tehdä joitain olettamuksia siitä, että kieltä on opittu ainakin joissain suhteissa könttärakenteina, jotka perustuvat sanojenvälisiin primingeihin. Tästä kertoo nähdäkseni ainakin kielenmukaisten n-grammien valtava edustus B1-aineistossa. Yhtenä esimerkkinä tästä on koko B1-aineiston yleisin n-grammi *minulla ei ole* (f = 210). Kuten tutkimuksessa huomattiin, sille löytyy n-grammeista myös kielenvastainen muoto **minulla en ole* (f = 24), joka on kuitenkin lähes yhdeksän kertaa harvinaisempi kuin oikein muodostettu vastineensa. N-grammeissa on myös useita muita *minulla ei (ole)* -tyyppisiä rakenteita, muttei enempää **minulla en (ole)* -tyypin esiintymiä. Mielestäni tuo virheellinen muoto kieliikin siitä, että kyseisen n-grammin tuottaneet oppijat ovat analysoineet kieltä ”liian pitkälle”, jolloin he ovat tulleet virheelliseen päätelmään siitä, että omistuslauseen subjekti on *minä*. Täten he ovat luonnollisesti päätyneet taivuttamaan kieltoverbin vastaamaan yksikön ensimmäistä persoonaa, kuten kuuluisikin, mikäli subjekti todella olisi *minä*. Näenkin, että sellaisen oppijan, joka on ottanut kieltä – muiden muassa *minulla ei ole* -rakenteen – haltuunsa ennemmin valmiina könttänä, ei tarvitse pohtia kielen sääntöjä yhtä pitkälle, vaan hänen mielessään *-llA ei ole* on aina erottamaton osa kieltomuotoista omistuslausetta.

Mielestäni myös esimerkiksi oikeiden paikallissijojen valinta vaikkapa n-grammien *huoneessa on oikealla* (f = 17), *seinällä on kirjahylly* (f = 14) ja *lattialla on punainen matto* (f = 9) tapauksissa voivat puhua sen puolesta, että rakenteet on omaksuttu yhtenäisinä sanakimppuina. Sen sijaan taas *luennossa puhuttiin paljon* (f = 9) -3-grammia (vaikkakin se on B1-aineiston tapauksessa vain yhdeltä suomenoppijalta peräisin; ks. luku 5.2.1) ei todennäköisesti ole opittu valmiina könttänä, vaan inessiivistä ja adessiivista oppija on ehkäpä päätynyt valitsemaan omasta mielestään loogisemman vaihtoehto, mikä ei kuitenkaan välttämättä vastaa natiivipuhujan intuitiivista kielitajua. *Luento*-sana ei siis ole oppijan mielessä vielä saavuttanut niin kutsuttua morfologista primingiaan (ks. Jantunen & Brunni 2012) *-lla*-päätteelle. Myös osittaisesta fraseologisten seikkojen hallinnan puutteesta ja täten vapaan valinnan periaatteeseen

nojautumisesta puhuvat muutamat natiivikielelle epätyypilliseltä vaikuttavat astemääritteiset n-grammit, kuten *on oikein kaunis* (f = 25) ja *pidän oikein paljon* (f = 11).

Oppijankielen universaaleista piirteistä (Jantunen 2008) tutkimuksessa analysoitujen n-grammien kautta vaikuttaisivat käyvän hyvin ilmi kielenainesten epätyypilliset frekvenssit sekä kielen yksinkertaisuus. N-grammeissa voidaan katsoa yliedustuvan tiettyjen leksikaalisten elementtien, kuten *olla*-verbin ja persoonapronominin *minä*. Jälkimmäistä selittävät pitkälti tehtävätyypit, ensin mainittua taas mahdollisesti se, että oppijat ovat taipuvaisia kiertämään hankalampia verbejä tukeutumalla niiden sijaan tuttuun *olla*-verbiin. *Olla*-verbin liikakäyttö linkittyykin samalla kielen yksinkertaisuuteen, sillä sen kautta pyritään mahdollisesti välttelemään virheitä kielentuotoksessa (ks. Grönholm 1993: 144–145; Aalto 2000). Eräänlaista n-grammituotosten yksinkertaisuutta selittää myös n-grammin perusluonteeseen kuuluva toisteisuuden vaatimus, jolloin järin monimutkaiset sanayhtymät eivät ole taipuvaisia muodostamaan n-grammeja. Muista oppijankielen universaaleista piirteistä n-grammeista löytyy myös linkki yleis- ja puhekielen sekoittumiseen, joskin tämä ilmenee lähinnä n-grammeista, kuten *minun huone on pieni* (f = 30) ja *minun herätyskello soi* (f = 12), uupuvien possessiivisuffiksien kautta. Virheelliset muodot tai oppijankielen kontekstuaalinen epäkonventionaalisuus eivät sen sijaan suuremmin korostuneet n-grammeissa, vaikka niidenkin puolesta saatiin muutamia viitteitä.

Tutkimuksessa havaittiin, että pidemmät n-grammit eivät välttämättä kerro aina pelkästään jo opitun soveltamisesta, vaan ehkä pikemminkin oppijoiden keinoista kopioida valmiita rakenteita suoraan omaan käyttöönsä. Tämän huomion kautta ei ole tarkoitus arvottaa oppijoita tai oppimisprosesseja, sillä kielenoppimiseen kuuluu luonnollisena ja väistämättömänä osana muilta ja muualta jäljitteleminen. Voidaan kuitenkin otaksua, että mikäli tutkittaisiin edistyneempiä suomenoppijoita, saattaisivat suoraan kopioidut n-grammit ja täten n-grammin kokonaiskirjokin vähentyä, mistä on tehty huomioita aiemmassakin tutkimuksessa (ks. Paquot & Granger 2012: 139).

Tutkimuksen toisessa tutkimuskysymyksessä kysyttiin, miten erilaiset metodologiset valinnat ohjaavat n-grammeista saatavaa dataa ja sen tulkintaa. Tässä tutkimuksessa tarkoituksena oli siis kokeilla joitain tapoja n-grammien lähestymiseksi ja katsoa, mihin suuntaan ne vievät tutkimusta. Jokainen menetelmällisistä valinnoista vaikutti osaltaan sekä itse aineistoon että siitä tehtyihin tulkintoihin. Ensimmäinen metodologinen valinta tehtiin, kun tutkimus päätettiin toteuttaa korpusvetoisesti. Mikäli tutkimuksesta olisi tehty korpuspohjainen, olisi n-grammihakujen tuloksista voitu rajata suoraan epäolennaiset, eli sellaiset, jotka eivät sisällä haluttua kielen elementtiä, pois.

Pelkkä korpusvetoisuus ei siis vielä varsinaisesti rajannut aineistoa mitenkään. ICLFI-aineistoa tuli kuitenkin supistaa sen suuren koon vuoksi joillain tavoilla, joten ensimmäiseksi päätin kohdistaa tutkimuksen ainoastaan B1-tasoisiin teksteihin. Tämä rajasi aineiston yli puolet pienemmäksi alkuperäisestä. Seuraavat menetelmälliset valinnat koskivat sitä, millaisin kriteerein n-grammeja etsitään aineistosta. Lähtökohdaksi päätin ottaa frekvenssin, ja tässä tukeuduin pitkälti Biberin ym. (1999) esittämiin arvoihin sekä n-grammien esiintymämääristä, tekstienvälisestä jakaumasta että sanapituudesta, joita myös esimerkiksi Salazar (2014) tutkimuksessaan sovelsi. N-grammiksi luokiteltavan sanaketjun tuli tässä tutkimuksessa olla vähintään kolme ja enintään kuusi sanaa pitkä, ja 3- ja 4-grammien tuli esiintyä suhteutettuna vähintään 20 ja 5- ja 6-grammien vähintään 12 kertaa miljoonaa sanetta kohden. Sanaketjun tuli löytyä vähintään viidestä eri tekstistä, jotta se voitiin hyväksyä n-grammiksi. Erilaisia valintoja tehtiin myös sen suhteen, millaiset sanat ovat hyväksyttäviä korpusohjelman hakutuloksissa. Saatuja n-grammilistoja myös karsittiin tukeutuen joihin Salazarin (2014) käyttämistä tavoista, jolloin esimerkiksi semanttiselta sisällöltään olemattomat tai virkerajalle jakaantuneet n-grammit poistettiin n-grammiaineistosta. Näillä valinnoilla saatiin koottua tutkimuksen lopullinen analysoitavien n-grammien lista, joka on siis vahvasti riippuvainen niistä valinnoista, jotka itse päätin tehdä. Toinen tutkija päätyisi omilla valinnoillaan todennäköisesti saamaan ainakin jossain määrin erilaisen n-grammilistan B1-aineistosta: hän voisi ensinnäkin käyttää hyvinkin erilaisia raja-arvoja ja toisekseen myös karsia saatuja n-grammeja tästä tutkimuksesta poikkeavin tavoin.

Itse n-grammien lähestymisessä kokeilin tutkimuksessa muun muassa sanalistaa, niin tavanomaista kuin käänteistäkin, sekä n-grammien erottelua verbillisiin ja verbittömiin. Otin tutkimuksessani osittain myös n-grammien kotekstiesiintymät huomioon, eli n-grammeja ei tulkittu pelkästään sellaisinaan, vailla tietoa niiden esiintymisympäristöistä. Tutkimuksessa valitut lähestymistavat eivät tietenkään olleet ainoita mahdollisia vaihtoehtoja, ja ne valinnat, joita tein esimerkiksi n-grammien jakamisesta verbillisiin ja verbittömiin sekä päätöksestä tarkastella juuri tempuksia ja syntaktisia lausetyyppejä ja tarkemmin ainoastaan joitakin n-grammien leksikaalisista elementeistä, saattavat näyttäytyä jossain määrin satunnaisina. Perustelen valintani kuitenkin tutkimuksen n-grammiaineistolla, joka antoi virikkeitä viedä tutkimus juuri näille urille; esimerkiksi verbillisten n-grammien valintaa motivoi niiden suuri määrä kaikista B1-aineiston n-grammeista. Samalla ne vaikuttivat verbin merkityksellisyyden vuoksi myös kiinnostavilta analyysin kohteilta.

Näenkin, että siitä, mitkä kielen ilmiöt kulloistakin tutkijaa kiinnostavat, riippuu pitkälti se, minkälaisia metodeja tämän tulisi n-grammitutkimuksessaan hyödyntää. N-grammeja

voidaan lähestyä esimerkiksi tietty sanaluokka, sijamuoto tai lauseopillinen rooli edellä, niistä voidaan luoda tavanomaisia tai käänteisiä sanalistoja, niiden KWIC-esiintymiä voidaan tutkia tai olla tutkimatta, niistä voidaan perehtyä ainoastaan sellaisiin, jotka sisältävät ennalta valittuja rakenteita tai sanastoa, tai niitä voidaan sovittaa ennalta annettuun tai tutkimuksen aikana kehitettävään kategorisointiin ja niin edelleen. Kaikki nämä valinnat vaikuttavat osaltaan siihen, minkälaiseksi n-grammeista saatava datamassa muodostuu ja miten sitä kyetään tulkitsemaan ja soveltamaan. Itse n-grammit eivät siis sido tutkimusta tiettyyn lähestymistapaan, vaan – etenkin tapauksissa, jossa niitä on hyvin runsaasti – on tutkijan itse kyettävä päättämään, mihin niissä hän haluaa tarkemman huomionsa keskittää. Tämä päätös voidaan tehdä jo etukäteen, jolloin tutkimus ottaa askeleen korpuspohjaisempaan suuntaan, tai katsoa, millaiset metodit tuntuvat saatujen n-grammien perusteella kaikkein tarkoituksenmukaisimmilta, kuten tässä tutkimuksessa tehtiin.

7.2 Tutkimuksen onnistumisen arviointia

Tutkimukseni tavoitteena oli ennen kaikkea kartoittaa oppijansuomea ja sen piirteitä lähtökohtana toistaiseksi jokseenkin vähäiselle huomiolle suomalaisessa kielentutkimuksessa jäänyt fraseologinen yksikkö, n-grammi. Tutkimus lähti liikkeelle korpusvetoisesti, jolloin etukäteen ei ollut tiedossa, miten tutkimus tulee lingvistisen analyysinsa osalta n-grammihauilla saatavien tulosten myötä etenemään. Ennakkoon ei ollutkaan järin helppoa esittää hypoteeseja tutkimustuloksista tai ideoita siitä, miten tutkimuksessa löydettäviä n-grammeja kannattaisi ylipäänsä lähestyä. Tutkimuskirjallisuudesta oli jonkin verran apua, vaikkakin se painottuu (oppijan)englannin tutkimukseen, jolloin sen linkittäminen tämän tutkimuksen kanssa ei ollut kaikilta osin luontevaa. Koska oppijansuomen n-grammikartoitusta ei ole aiemmin tehty paljoakaan, tehtiinkin tässä tutkimuksessa lopulta ainoastaan jonkinlaista peruskartoitusta aiheesta, mikä osaltaan selittää tutkimuksen kokeellisia menetelmävalintoja sekä paikoitellen ehkä hivenen poukkoilevalta vaikuttavaa luonnetta.

Tutkimuksen avulla onnistuttiin kuitenkin saamaan vastaukset tutkimuksen tutkimuskysymyksiin, eli siltä osin se täytti tavoitteensa. Vastaukset eivät ehkä ole aukottomia saati kerro kaikkea mahdollista esimerkiksi oppijansuomen B1-tason n-grammien sanasto- ja rakennepiirteistä, mutta sellaista ei voinut odottaakaan etenkään enää sen jälkeen, kun tutkimuksen lopullinen analysoitavien n-grammien runsas määrä paljastui. Ensimmäiseen tutkimuskysymykseen

saatiin kuitenkin tutkimuksella selkeä vastaus, joka on luettavissa suoraan liitteestä 3. Ensimmäisen tutkimuskysymyksen kahteen alakysymykseen taas tarjottiin erilaisia näkymiä, jotka valottavat joitain oppijansuomen B1-taitotason sanasto- ja rakennepiirteistä. Toisen tutkimuskysymyksen osalta kokeiltiin erilaisia menetelmävalintoja ja katsottiin, mihin tietoon ja millaisiin tulkintoihin ne johtavat. Mielestäni tutkimus toteuttikin onnistuneesti nimenomaan korpusvetoista luonnettaan: aineisto ohjasi nyt pitkälti sitä, mihin suuntaan tutkimus lopulta vietiin. Korpusvetoisuus osoittautui mielestäni toimivaksi vaihtoehdoksi, sillä kuten ylempänä mainitsin, ennalta oli vaikea arvella, minkälaisia n-grammeja tutkimuksessa saadaan käsiteltäviksi, joten tietyn kieliseikan valinta korpuspohjaisesti tutkittavaksi etukäteen olisi ollut hankalaa. Myös frekvenssiperustainen lähestymistapa n-grammeihin toimi varsin hyvin tällaisessa osittaisessa pioneerikartoituksessa aiheesta, joskin tilastollisiin yhteyksiin perustuvien MI- ja t-testien kautta n-grammeihin olisi mahdollisesti voitu saada heijastumaan paremmin kielenoppijoiden fraseologista tietoutta, kuten joissain aiemmissa tutkimuksissa on ensikielenään englantia käyttävien osalta havaittu (ks. Garner ym. 2020: 55–56).

Mielestäni sekä tutkimuksen vahvuutena että toisaalta myös sen ehkä suurimpana heikkoutena on se, että sen avulla onnistuttiin kartoittamaan runsas määrä kvantitatiivista dataa lukuisista erilaisista B1-oppijansuomen piirteistä. Tässä yhteydessä se osoitti aineistosta n-grammitutkimuksen monet mahdollisuudet. Kuitenkin kvantitatiivisen ja samalla tarkempaa kvalitatiivista analyysia kaipaavan n-grammidatan paisuessa jäi varsinainen n-grammien analysointi osittain pintatasoiseksi. Tämä johtuu ennen kaikkea siitä, ettei kaikkeen mahdolliseen saatuun määrälliseen dataan olisi voinutkaan perehtyä kvalitatiivisesti kaikenkattavasti tällaisen tutkielman puitteissa tai välttämättä muutoinkaan. Tutkimuksessa onnistuttiin käsittelemään joitain oppijansuomen leksikon ja rakenteiden erityispiirteistä, mutta täydellistä kokonaiskuvaa se ei tietenkään voi antaa, mihin ei tutkimuksella toisaalta pyrittykään.

Toisena heikkoutena tutkimukselleni näen siinä käytetyt tutkimusmenetelmälliset joustot: etenkin lausetyyppianalyysissa jouduin vetämään mutkia suoriksi laskemalla useita n-grammeista tiettyjen lausetyyppien edustajiksi tarkastelematta niiden jokaisen kaikkia esiintymiskoteksteja. Kotekstiesiintymiä oli kuitenkin niin paljon, ettei niistä jokaista ollut mahdollista tarkastella sillä tarkkuudella, kuin mitä ne olisivat ehkä osikseen vaatineet. Tämä aiheuttaa myös sen, että laskennoissa jotkin n-grammeista on nyt hyvin mahdollisesti tulkittu väärien lausetyyppien alle, sillä jokaisen n-grammin jokaista kotekstiesiintymää ei ole tarkistettu. Toinen tutkija saattaisikin tarkemmalla perehtymisellä sekä toisaalta myös omilla tulkinnoillaan, intuitiollaan ja kielitajullaan saada hieman tästä tutkimuksesta poikkeavat lukemat tuloksiksi verbillisten n-grammien edustamista lausetyypeistä. Täten n-grammien perusteella tehtyä

lausetyyppianalyysia ei voi luonnehtia täysin reliaabeliksi. Uskon kuitenkin, että vaikka laskenta teetätettäisiinkin uudelleen, suhteet erilaisten lausetyyppien välillä pysyisivät silti varsin samanlaisina, eikä kovin monen prosenttiyksikön horjumista suuntaan tai toiseen tapahtuisi.

Mikäli tekisin tällaisen samantyyppisen tutkimuksen uudestaan, muokkaisin sitä vähintäänkin niin, että käyttäisin siinä hieman tästä tutkimuksesta poikkeavia aineiston rajauksia. Mielestäni tutkimuksen kohdistaminen ainoastaan yhdelle EVK:n (2003) taitotasoista oli perusteltu ja toimivakin valinta, mutta vaihtoehtoisesti – tai taitotasorajauksen lisäksi – tutkimuksen olisi voinut rajata myös vaikkapa tekstilajien perusteella käsittämään esimerkiksi pelkistä esseeteksteistä löydettävät n-grammit. Useissa aiemmissa n-grammitutkimuksista tarkastelu on suunnattu nimenomaan akateemisiin teksteihin (ks. esim. Cortes 2004; 2008; Ivaska 2015; Li 2016). Tarkastelemalla yksinomaan akateemisen oppijansuomen n-grammeja olisi tästä tutkimuksesta voitu vetää mahdollisesti joitain päätelmiä aiempiin n-grammitutkimuksiinkin verraten. Mahdollista olisi ollut myös suoraan n-grammihakuja tehtäessä rajata tutkimus esimerkiksi niihin n-grammeihin, joissa on mukana verbi, ja karsia muut pois.

Hypoteettisessa uusintatutkimuksessa, kohdistuisipa se sitten edelleen koko B1-taitotasoon tai esimerkiksi vain sen esseeteksteihin, nostaisin myös raja-arvoa, jolla sanaketju määritellään tarpeeksi toistuvaksi ja täten n-grammiksi. Raja-arvot olisivat voineet tässä tutkimuksessa olla peräti kaksi kertaa suuremmat; 3- ja 4-grammien osalta siis vasta esiintyminen suhteutettuna 40 kertaa miljoonassa sanassa loisi hyväksytyn n-grammin. Tätä raja-arvoa on käyttänyt esimerkiksi Biber (2006). Samalla myös kriteeriä siitä, kuinka monen eri tekstin välille n-grammien esiintymien tulee jakautua, tulisi nostaa, sillä tämän tutkimuksen viiden tekstin raja-arvolla ei vältytty yksittäisen suomenoppijan idiolektin heijastumiselta saatuihin n-grammeihin, kuten luvussa 5.2.1 huomattiin. Rajausten puolesta mahdollista olisi myös jättää 3-grammit kokonaan sivuun ja tutkia vaikkapa molempien B-taitotasojen neljä- ja useampisanaisia n-grammeja. Pitempisanaisia n-grammeja on huomattavasti 3-grammeja vähemmän, ja etenkin tarpeeksi suuria raja-arvoja hyödyntämällä olisi tutkittava massa pysynyt pelkkien pitempien n-grammien kanssa todennäköisesti tätä tutkimusta paremmin kasassa.

Suuremmilla raja-arvoilla n-grammeja olisi toisaalta saatu vähemmän, mutta niiden kvalitatiivinen tutkiminen olisi voinut olla pitemmälle vietyä ja perustavanlaatuisempaa. Tässä tutkimuksessa käytetyin raja-arvoin etenkin n-grammilistan loppupäähän jäi pidemmissä n-grammeissa toistuvien lyhyempien n-grammien poistoista huolimatta vielä paljon sellaisia n-grammeja, jotka ovat informaatiosisällöiltään ja rakenteiltaan pitkälti toisiaan vastaavia. Näitä ovat esimerkiksi 5-grammi *samana vuonna pääsin yliopistoon opiskelemaan* (f = 10) ja 6-grammi *ja samana vuonna pääsin yliopistoon opiskelemaan* (f = 6), 5-grammit *tumma tukka ja harmaat*

silmät (f = 5) ja *tumma tukka ja ruskeat silmät* (f = 5) sekä 3-grammit *minä herään kello* (f = 9) ja *minä nousen kello* (f = 9). Näitä n-grammeja olisi ehkä voitu niputtaa eräänlaisten kohdekimppujen alle muun muassa Cortesin (2004) Salazarin (2014) tutkimusten tapaan. Tällöin esimerkiksi 5-grammit *matka kestää noin kymmenen minuuttia* (f = 6) ja *matka kestää noin viisitoista minuuttia* (f = 6) olisi voitu hahmotella yhdeksi, rakenteen *matka kestää noin x minuuttia* -tapaiseksi, kohdekimpuksi. Vaihtoehtoisesti n-grammeja olisi voitu myös tutkia luvussa 2.2.4 esitellyin fraasikehyksin tai skip-, flex- tai konkgrammein, jotka pystyvät niputtamaan ainakin osan tällaisista rakenteista yhteen. Tutkimuksessa olisi voitu myös keskittyä pelkästään joko leksikkoon tai rakenteisiin, ja tehdä valitun puoliskon osalta kokonaisvaltaisempi analyysi.

Tutkimukseni jää mielestäni vielä hieman liian yleiselle tasolle, sillä siinä pyritään kurottamaan hieman kaikkialle, jolloin sen suoranainen vertailu muihin, selkeämpirajaisiin n-grammitutkimuksiin on haastavaa. Rajaamalla tutkimukseni esimerkiksi pelkkiin ”akateemisiin teksteihin”, toisin sanoen esimerkiksi esseisiin, olisin voinut mahdollisesti saada selkeämpi vertailukohtia aiemmin tutkittuun. Vertailua tosin hankaloittaa ennen kaikkea varsinaisen oppijansuomen n-grammitutkimuksen vähäinen määrä. Toisaalta, kuten Paquot ja Granger (2012: 138–139) huomauttavat, n-grammirakenteita koskettavien tutkimusten tuloksia on ylipäänsä varsin hankala verrata keskenään johtuen etenkin tutkimuksissa käytetyistä vaihtelevista n-grammien pituuksista ja raja-arvoista. Vertailua voi nähdä vaikeuttavan entisestään sen, ettei erilaisten n-grammitutkimusten välillä ole useinkaan yhteisymmärrystä muun muassa sen suhteen, mitkä ovat tarkkarajaisemmat n-grammeja määrittävät piirteet, mitä metodologioita niiden tunnistamiseksi tulisi käyttää ja tulisiko tutkimuksessa keskittyä vain pieneen joukkoon tärkeitä n-grammeja vaiko niiden koko kirjoon (ks. Biber ym. 2004: 372).

Voin joka tapauksessa todeta tutkimuksestani tiivistetysti sen, että se vahvistaa uusin lähestymistavoin todennettuna aiemmin tiedettyä oppijansuomesta. Vaikka useat tässä tutkimuksessa n-grammien kautta tehdyt kielelliset löydökset onkin havaittu jo edeltävissä tutkimuksissa, on uutta kaikki se informaatio, jota B1-oppijansuomen n-grammeista itsessään on saatu, sillä vastaavaa kartoitusta ei ole aiemmin ollut olemassa. Tutkimus vahvistaakin käsitystä siitä, että n-grammeihin heijastuu vahvasti ennen kaikkea tutkimuksen alla olevan kielenkäyttäjäryhmän kielen kaikkien tyypillisimmät piirteet. Tutkimus siis tukee jo tutkittua, mutta samalla osoittaa, että samanlaisiin tuloksiin on mahdollista päästä myös jossain määrin tuntemattomammankin tutkimustavan, n-grammianalyysin, avulla. Tämän vuoksi pidän tutkimustani lopulta päällisin puolin onnistuneena.

7.3 Jatkotutkimusmahdollisuuksia

Kuten tässä tutkimuksessa on jo muutamaan kertaan tähdennetty, tähänastinen n-grammien tarkastelu oppijansuomesta on ollut vähäistä. Siksipä näenkin milteipä kaikenlaisen toistuviin sanaketjuihin keskittyvän lisätutkimuksen olevan tervetullutta laajentamaan fraseologisen oppijansuomen tutkimuksen kenttää muun muassa verrattain usein tehtävän kollokaatiotutkimuksen rinnalle. Erittelen tässä alaluvussa kuitenkin vielä muutamia hivenen tarkempirajaisempia aihioita mahdolliselle jatkotutkimukselle.

Koska käsillä oleva tutkimus kohdennettiin ICLFI:n EVK:lla (2003) B1-kielitaitotasoarvioinnin saaneisiin suomenoppijoiden teksteihin ja niiden n-grammeihin, olisi yksi selkeä jatkotutkimuksen paikka selvittää vielä ICLFI:n B2-, A- ja/tai C-tasoille arvioiduissa teksteissä esiintyviä n-grammeja ja niistä vaikkapa tämän tutkimuksen tapaan leksikkoa tai rakenteita. Näistä taitotasoista kullekin arvioituja tekstejä on korpuksessa huomattavasti B1-taitotason tekstejä vähemmän, joten yhdessä tutkimuksessa voitaisiin mahdollisesti tutkia samalla esimerkiksi molempia A- tai C-taitotasoista. N-grammihakujen raja-arvoja voisi myös nostaa esimerkiksi 40 esiintymään miljoonassa saneessa, jolloin aineistoista saataisiin koottua määränsä puolesta helpommin käsiteltäviä n-grammilistauksia. Tällaisista tutkimuksista saatavia tuloksia voitaisiin sitten vertailla soveltuvin osin tässä tutkimuksessa saatuihin tuloksiin.

Yksi mahdollisuus lähestyä oppijansuomen n-grammeja olisi myös kontrastiivisen oppijankielen analyysin (*Contrastive Interlanguage Analysis* eli CIA) (ks. Granger 1996; 2002: 12–13) kautta. Sen avulla on mahdollista vertailla joko natiivikielen ja oppijankielen tai oppijankielen eri varianttien keskinäisiä eroja kielentuotoksissa (Granger 1996: 44). Ensin mainittua tutkimusta ovat aiemmin n-grammien osalta tehneet muun muassa Salazar (2014) sekä Li (2016). Jälkimmäisen kaltaisessa tutkimuksessa on taas kunnostautunut esimerkiksi Paquot (2013; 2014). Muun muassa Peromingo (2012) ja Ivaska (2015) ovat hyödyntäneet tutkimuksissaan kumpaakin tulokulmaa vertailuun. Mahdollisessa jatkotutkimuksessa voitaisiin siis selvittää johonkin natiivisuomen aineistoon tukeutuen ensikielenään suomea puhuvien tuottamia n-grammeja ja verrata niitä suomenoppijoiden n-grammituotoksiin tai vertailla esimerkiksi vironkielisten tuottamia n-grammeja muunkielisten vastaaviin. Ensin mainitun kautta kyettäisiin näkemään esimerkiksi, missä suhteissa natiivi- ja oppijasuomessa tukeudutaan n-grammeihin ja kuinka idiomaattisia suomenoppijoiden n-grammit ovat natiiveihin verrattuna. Jälkimmäisen avulla voitaisiin taas päästä kiinni mahdolliseen transferin, eli vielä yhden oppijankielen universaalin piirteen, heijastumiseen n-grammeissa. Vertailua voitaisiin tehdä myös suomea

toisena ja vieraana kielenä opiskelevien välillä hypoteesina esimerkiksi se, että suomi toisena kielenä -oppijoiden tuottamissa n-grammeissa olisi vieraana kielenä suomea opiskelevia enemmän natiivinkaltaisuutta mukana. Tällaiseen vertailuun tarvittaisiin siis lisäksi korpus, johon on kerätty nimenomaisesti S2-oppijoiden tuotoksia, sillä esimerkiksi ICLFI koostuu ainoastaan suomea vieraana kielenä opiskelevien oppijoiden teksteistä.

N-grammitutkimusta voisi lähteä viemään myös korpuspohjaisempaan suuntaan valitsemalla etukäteen yhden tai useamman lekseemin, saneen, rakenteen tai kielenyksikön, jonka tai joiden käyttöä oppijansuomessa pyrittäisiin n-grammein selvittämään. Esimerkiksi Jantunen (2017) tutki *kello*-sananmuodon käyttöä ICLFI:ssä n-grammien avulla. Toisaalta jatkotutkimuksissa olisi mahdollista pitäytyä korpusvetoisuudessa, mutta – kuten yllä olevassa kritiikkiosiossa jo sivuttiin – rajata tutkimus käsittelemään vaikkapa vain yhtä ICLFI:n tekstilajeista. Tällöin voitaisiin selvittää esimerkiksi, mitkä ovat suomenoppijoiden esseeteksteissään useimmin toistamat n-grammit sekä niiden rakenteet ja funktiot. Oppijansuomen n-grammeja voisi lähestyä myös luvussa 2.2.4 esiteltyjen skip-, flex- ja konkgrammien kautta, joiden avulla pystyttäisiin ehkä luomaan yleispätevämpiä n-grammilistoja, joissa toisiaan lähellä olevat n-grammit saataisiin niputettua yhteisten kokonaisuuksien alle, eikä jokaista vain hieman toisesta poikkeavaa sanayhtymää olisi välttämätöntä tarkastella omana n-gramminaan.

N-grammit itsessään viitoittavat joka tapauksessa tietä lukuisiin eri tapoihin lähestyä kieltä ja sen piirteitä etenkin kvantitatiiviselta kantilta, mistä tämä tutkimus toimii toivon mukaan jonkinasteisena todisteena. Jatkotutkimuksissa paikallaan voisi ollakin keskittyminen kokonaisvaltaisemmin yksittäiseen tai korkeintaan pariin kielen ilmiöön tai elementtiin, eikä tämän tutkimuksen tapaan pyrkiä kartoittamaan hieman kaikkea kerralla.

LÄHTEET

Aalto, Marjo 2000: Suomen frekventit verbit oppijankielessä. – Sanna Martin & Helena Sulkala (toim.), *Tutkielmia oppijankielestä* s. 91–110. Oulu: Oulun yliopiston suomen ja saamen kielen ja logopedian laitos.

Akgül, Anna 2013: *Frekventtien adjektiivien kollokaatiot oppijansuomessa ja natiivisuomessa*. Pro gradu -tutkielma. Oulun yliopisto.

Aktas, Rahime Nur & Cortes, Viviana 2008: Shell nouns as cohesive devices in published and ESL student writing. – *Journal of English for Academic Purposes* 7 s. 3–14.

Alderson, J. Charles 2007: The CEFR and the need for more research. – *The Modern Language Journal* 91 (4) s. 659–663.

Altenberg, Bengt 1998: On the phraseology of spoken English: The evidence of recurrent word combinations – Anthony Paul Cowie (toim.), *Phraseology. Theory, analysis and applications* s. 101–122. Oxford: Oxford University Press.

Anthony, Laurence 2019: *AntConc (Windows 64-bit 3.5.8).* Tokyo: Waseda University. – http://www.laurence-anthony.net/software/ 20.2.2020.

Ball, Philip 2005: A new kind of alchemy. *New Scientist* 2495 s. 30–33.

Banerjee, Satanjeev & Pedersen, Ted 2003: The design, implementation, and use of the Ngram Statistics Package. – Alexander F. Gelbukh (toim.), *Computational linguistics and intelligent text processing: 4th International Conference, CICLing 2003 Mexico City, Mexico, February 16–22, 2003 Proceedings* s. 370–381. New York: Springer.

Bednared, Monika 2012: "Get us the hell out of here": Key words and trigrams in fictional television series – *International Journal of Corpus Linguistics* 17 (1) s. 35–63.

Bestgen, Yves & Granger, Sylviane 2014: Quantifying the development of phraseological competence in L2 English writing: An automated approach. – *Journal of Second Language Writing* 26 s. 28–41.

Biber, Douglas 2006: *University language. A corpus-based study of spoken and written registers*. Studies in corpus linguistics 23. Amsterdam: John Benjamins.

——— 2009: A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and writing. – *International Journal of Corpus Linguistics* 14 (3) s. 275–311.

Biber, Douglas & Conrad, Susan 1999: Lexical bundles in conversation and academic prose. – Hilde Hasselgård & Signe Oksefjell (toim.), *Out of corpora. Studies in honour of Stig Johansson* s. 181–190. Amsterdam: Rodopi.

Biber, Douglas, Conrad, Susan & Cortes, Viviana 2003: Lexical bundles in speech and writing: an initial taxonomy. – Andrew Wilson, Paul Rayson & Tony McEnery (toim.), *Corpus linguistics by the Lune: a festschrift for Geoffrey Leech* s. 71–92. Frankfurt: Peter Lang.

——— 2004: *If you look at…*: Lexical bundles in university teaching and textbooks. – *Applied Linguistics* 25 (3) s. 371–405.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward 1999: *Longman grammar of spoken and written English*. Harlow: Longman.

Brunni, Sisko, Jantunen, Jarmo & Skantsi, Valtteri 2019: Korpusavusteinen virheanalyysi tarkkuuden kehityksestä EVK:n taitotasoilla A2–B2. – *Puhe ja kieli* 39 (3) s. 275–304.

Byrne, Shelley 2016: *An examination of successful language use at B1, B2 and C1 level in UCLanESB speaking tests in accordance with the Common European Framework of References for Languages*. Preston: University of Central Lancashire.

Callies, Marcus & Götz, Sandra 2015: Learner corpora in language testing and assessment: prospects and challenges. – Marcus Callies & Sandra Götz (toim.), *Learner corpora in language testing and assessment* s. 1–9. Amsterdam: John Benjamins.

Centre for English Corpus Linguistics 2020: *Learner corpora around the world.* Louvain-la-Neuve: Université catholique de Louvain. – https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html 29.4.2020.

Chen, Yu-Hua & Baker, Paul 2010: Lexical bundles in L1 and L2 academic writing. – *Language Learning & Technology* 14 (2) s. 30–49.

Cheng, Winnie, Greaves, Chris & Warren, Martin 2006: From n-gram to skipgram to concgram. – *International Journal of Corpus Linguistics* 11 (4) s. 411–433.

Chomsky, Noam 2006: *Language and mind*. Cambridge: Cambridge University Press.

Cook, Vivian 2016: *Second language learning and language teaching*. New York: Routledge.

Corder, Stephen Pit 1981: *Error analysis and interlanguage*. Oxford: Oxford University Press.

Cortes, Viviana 2004: Lexical bundles in published and student disciplinary writing: Examples from history and biology. – *English for Specific Purposes* 23 (4) s. 397–423.

——— 2008: A comparative analysis of lexical bundles in academic history writing in English and Spanish. – *Corpora* 3 (1) s. 43–57.
——— 2012: *The purpose of this study is to*: Connecting lexical bundles and moves in research article introductions. – *Journal of English for Academic Purposes* 12 s. 33–43.
——— 2015: Situating lexical bundles in the formulaic language spectrum. Origins and functional analysis developments. – Douglas Biber, Eniko Csomay & Viviana Cortes (toim.), *Corpus-based research in applied linguistics: studies in honor of Doug Biber* s. 197–216. Amsterdam: John Benjamins.
Cowell, Henry 1996 [1930]: *New musical resources*. Cambridge: Cambridge University Press.
Cowie, Anthony 1998: Introduction. – Anthony Cowie (toim.), *Phraseology: Theory, analysis, and applications* s. 1–20. Oxford: Oxford University Press.
——— 2006: Phraseology. – Keith Brown (toim.), *Encyclopedia of Language & Linguistics. Second edition* s. 579–585. Amsterdam: Elsevier.
Crossley, Scott & Salsbury, Thomas Lee 2011: The development of lexical bundle accuracy and production in English second language speakers. – *International Review of Applied Linguistics in Language Teaching* 49 (1) s. 1–26.
CSC – IT Center for Science 2004: *Suomen sanomalehtikielen taajuussanasto* [tekstikorpus]. Kielipankki. – http://urn.fi/urn:nbn:fi:lb-201405272 8.4.2020
Csomay, Eniko 2013: Lexical bundles in discourse structure: a corpus-based study of classroom discourse. – *Applied Linguistics* 34 (3) s. 369–388.
Culpeper, Jonathan & Kytö, Merja 2010: *Early modern English dialogues. Spoken interaction as writing*. Studies in English language. Cambridge: Cambridge University Press.
Ebeling, Jarle, Ebeling, Signe Oksefjell & Hassegård, Hilde 2013: Using recurrent word-combinations to explore cross-linguistic differences. – Karin Aijmer & Bengt Altenberg (toim.), *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson* s. 177–199. Amsterdam: John Benjamins.
Ellis, Nick C. 1997: Vocabulary acquisition: Word structure, collocation, grammar, and meaning. – Michael McCarthy & Norbert Schmidt (toim.), *Vocabulary: description, acquisition and pedagogy* s. 122–139. Cambridge: Cambridge University Press.
——— 2008: Phraseology. The periphery and the heart of language. – Fanny Meunier & Sylviane Granger (toim.), *Phraseology in foreign language learning and teaching* s. 1–13. Amsterdam: John Benjamins.
Ellis, Nick C., O'Donnell Matthew Brook & Römer, Ute 2013: Usage-based language: Investigating the latent structures that underpin acquisition. – *Language Learning* 63 (1) s. 25–51.
Ellis, Nick C. & Ogden, David C. 2017: Thinking about multiword constructions. Usage-based approaches to acquisition and processing. – *Topics in Cognitive Science* 9 s. 604–620.
Ellis, Rod & Barkhuizen, Gary 2005: *Analysing learner language*. Oxford applied linguistics. Oxford: Oxford University Press.
EVK 2003 = *Eurooppalainen viitekehys. Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys 2003*. Helsinki: WSOY.
Firth, John Rupert 1968 [1957]: A synopsis of linguistic theory, 1930–55. – Frank Robert Palmer (toim.), *Selected papers of J. R. Firth 1952–59* s. 168–205. London: Longmans.
Fletcher, William H. 2006: "Phrases in English" Home. – http://phrasesinenglish.org/ 22.3.2020.
Flowerdew, John 2006: Use of signalling nouns in a learner corpus. – *International Journal of Corpus Linguistics* 11 (3) s. 345–362.
Frankenberg-Garcia, Ana, Flowerdew, Lynne & Aston, Guy 2011: Introduction. – Ana Frankenberg-Garcia, Lynne Flowerdew & Guy Aston (toim.), *New trends in corpora and language learning* s. 153–166. London: Continuum.
Garner, James R. 2016: A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. – *International Journal of Learner Corpus Research* 2 (1) s. 31–68.
Garner, James, Crossley, Scott & Kyle, Kristopher 2020: Beginning and intermediate L2 writer's use of n-grams: an association measures study. – *International Review of Applied Linguistics in Language Teaching* 58 (1) s. 51–74.
Gass, Susan M. & Selinker, Larry 2008: *Second language acquisition – An introductory course*. New York: Taylor & Francis.
Granger Sylviane 1996: From CA to CIA and back: an integrated contrastive approach to computerized bilingual and learner corpora. – Karin Aijmer, Bengt Altenberg & Mats Johansson (toim.), *Languages in contrast. Text-based cross-linguistic studies* 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: Oxford University Press.
——— 2002: A bird's-eye view of learner corpus research. – Stephanie Petch-Tyson, Joseph Hung & Sylviane Granger (toim.), *Computer learner corpora, second language acquisition, and foreign language teaching* s. 3–33. Amsterdam: John Benjamins.

——— 2005: Pushing back the limits of phraseology: How far can we go? – Christelle Cosme, Céline Gouverneur, Fanny Meunier & Paquot Magali (toim.), *Pre-proceedings of phraseology 2005. The many faces of phraseology. An interdisciplinary conference. Louvain-la-Neuve Belgium, 13–15 October 2005* s. 165–168.
——— 2011: From phraseology to pedagogy: Challenges and prospects. – Thomas Herbst, Susen Faulhaber & Peter Uhrig (toim.), *The phraseological view of language. A tribute to John Sinclair* s. 123–146. Berlin & New York: Mouton de Gruyter.
——— 2014: A lexical bundle approach to comparing languages: Stems in English and French. – *Languages in Contrast* 14 (1) s. 58–72.
Granger, Sylviane & Meunier, Fanny 2008: Introduction. The many faces of phraseology. – Sylviane Granger & Fanny Meunier (toim.), *Phraseology. An interdisciplinary perspective* s. xix–xxviii. Amsterdam: John Benjamins.
Granger, Sylviane & Paquot, Magali 2008: Disentangling the phraseological web. – Sylviane Granger & Fanny Meunier (toim.), *Phraseology. An interdisciplinary perspective* s. 27–49. Amsterdam: John Benjamins.
Gray, Bethany & Biber, Douglas 2013: Lexical frames in academic prose and conversation. – *International Journal of Corpus Linguistics* 18 (1) s. 109–136.
Gries, Stefan Th. 2008: Phraseology and linguistic theory. – Sylviane Granger & Fanny Meunier (toim.), *Phraseology. An interdisciplinary perspective* s. 3–25. Amsterdam: John Benjamins.
Grönholm, Maija 1993: *Tv on pang pang – verbisanaston kehitys toisen kielen kirjoittamisessa.* Rapporter från Pedagogiska fakulteten vid Åbo Akademi 4. Vasa: Åbo Akademi.
——— 2007: Idiomien ja kollokaatioiden oppimisjärjestys suomen kielessä (L2). – Olli-Pekka Salo, Tarja Nikula & Paula Kalaja (toim.), *AFinLAn vuosikirja 2007* s. 269–286. Jyväskylä.
Guthrie, David, Allison, Ben, Liu, Wei, Guthrie, Louise & Wilks, Yorick 2006: A closer look at skipgram modelling. – *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* s. 1222–1225.
Güngör, Fatih & Uysal, Hacer Hande 2016: A comparative analysis of lexical bundles used by native and non-native scholars. – *English Language Teaching* 9 (6) s. 176–188.
Haapala, Terhi 2008: *Finiittiverbeistä verbiketjuihin: verbiytimien kompleksistuminen S2-oppijoiden kielessä.* Pro gradu -tutkielma. Tampereen yliopiston kieli- ja käännöstieteiden laitos.
Hakulinen, Auli & Karlsson, Fred 1995: *Nykysuomen lauseoppia.* Helsinki: Suomalaisen Kirjallisuuden Seura.
Haltia, Heidi 2015: *Modaaliverbien käyttö oppijansuomessa Eurooppalaisen viitekehyksen taitotasoilla A2–C2.* Pro gradu -tutkielma. Helsingin yliopiston suomen kielen, suomalais-ugrilaisten ja pohjoismaisten kielten ja kirjallisuuksien laitos.
Hasselgren, Angela 1994: Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. – *International Journal of Applied Linguistics* 4 (2) s. 237–258.
Hawkins, John A. & Buttery, Paula 2010: Criterial features in learner corpora: Theory and illustrations. – *English Profile Journal* 1 (1) s. 1–23.
Hoey, Michael 2004: The textual priming of lexis. – Guy Aston, Silvia Bernardini & Dominic Stewart (toim.), *Corpora and language learners* s. 21–41. Amsterdam: John Benjamins.
——— 2005: *Lexical priming. A new theory of words and language.* London: Routledge Falmer.
——— 2007: Lexical priming and literary creativity. – Michael Hoey, Michaela Mahlberg, Michael Stubbs & Wolfgang Teubert (toim.), *Text, discourse and corpora. Theory and analysis* s. 7–29. London: Continuum.
——— 2009: Corpus-driven approaches to grammar. The search for common ground. – Ute Römer & Rainer Schulze (toim.), *Exploring the lexis-grammar interface* s. 33–47. Amsterdam: John Benjamins.
Hunston, Susan 2002: *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
ICLFI-manuaali 2016 = *Ohjeita ICLFI-korpuksen käyttöön*. Oulun yliopisto. – https://www.oulu.fi/suomitoisena-kielena/node/16078 4.5.2020.
Ivaska, Ilmari 2011: Lausetyyppien sekoittuminen edistyneessä oppijansuomessa – näkökulmana eksistentiaalilause. – *Lähivõrdlusi = Lähivertailuja* 21 s. 65–85.
——— 2014a: Edistyneen oppijansuomen avainrakenteita. Korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin. – *Virittäjä* 118 (2) s. 161–193.
——— 2014b: Mahdollisuuden ilmaiseminen S1-suomea ja edistynyttä S2-suomea erottavana piirteenä. – *Lähivõrdlusi = Lähivertailuja* 24 s. 47–80.
——— 2015: *Edistyneen oppijansuomen konstruktiopiirteitä korpusvetoisesti: avainrakenneanalyysi*. Turun yliopiston julkaisuja 409. Turku: Painosalama Oy.
Ivaska, Ilmari & Siitonen, Kirsi 2011: Avainrakenneanalyysi. Tapa tutkia oppijankielen lauserakennetta korpusvetoisesti. – Esa Lehtinen, Sirkku Aaltonen, Merja Koskela, Elina Nevasaari & Mariann Skog-Södersved (toim.), *AFinLA-e Soveltavan kielitieteen tutkimuksia 3* s. 35–47.
Jantunen, Jarmo Harri 2001: "Tärkeä seikka" ja "keskeinen kysymys": Mitä korpuslingvistinen analyysi paljastaa lähisynonyymeista? – *Virittäjä* 105 (2) s. 170–192.

——— 2004: *Synonymia ja käännössuomi. Korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin*. Joensuun yliopiston humanistisia julkaisuja 35. Joensuu: Joensuun yliopisto.
——— 2008: Haasteita oppijankielen korpusanalyysille: oppijankielen universaalit. – Pille Eslon (toim.), *Õppijakeele analüüs: võimalused, probleemid, vajadused* s. 1–26. Tallinn: Tallinna Ülikool Kirjastus.
——— 2009a: Ei pelkästään mielikuvituksen puutteen vuoksi – kieliaineistojen systemaattinen käyttö kielentutkimuksessa. – *Virittäjä* 113 (1) s. 101–113.
——— 2009b: ”Minulla on aivan paljon rahaa” – fraseologiset yksiköt suomen kielen opetuksessa. – *Virittäjä* 113 (3) s. 356–381.
——— 2011: Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. – *Lähivõrdlusi = Lähivertailuja* 21 s. 86–105.
——— 2012: Korpusvetoinen tekstilajianalyysi: sanalistat ja genreavainsanat. – Vesa Heikkinen, Eero Voutilainen, Petri Lauerma, Ulla Tiililä & Mikko Lounela (toim.), *Genreanalyysi – tekstilajitutkimuksen käsikirja* s. 360–371. Helsinki: Gaudeamus.
——— 2015: Oppimiskontekstin vaikutus oppijanpragmatiikkaan: astemääritteet leksikaalisina nallekarhuina. – *Lähivõrdlusi = Lähivertailuja* 25 s. 105–136.
——— 2016: Corpora, phraseology and dictionaries: How does corpus research intersect language teaching and learning? – Begoña Sanromán Vilas (toim.), *Collocations cross-linguistically: Corpora, dictionaries and language teaching* s. 97–119. Uusfilologinen Yhdistys.
——— 2017: Lexical and morphological priming: a holistic phraseological analysis of the Finnish time expression *kello*. – Michael Pace-Sigge & Katie Patterson (toim.), *Lexical priming. Applications and advances* s. 254–272. Amsterdam: John Benjamins.
Jantunen, Jarmo Harri & Brunni, Sisko 2012: Morfologinen priming ja fraseologia vieraan kielen oppimisessa: Korpustutkimus oppijansuomesta. – *Lähivõrdlusi = Lähivertailuja* 22 s. 71–100.
Jantunen, Jarmo Harri, Brunni, Sisko, Lehto, Liisa-Maria & Airaksinen, Valtteri 2014: Oppijankieliaineistojen annotointi – esimerkkinä ICLFI:n annotoinnin prosessit, ongelmat ja ratkaisut. – Maarit Mutta, Pekka Lintunen, Ilmari Ivaska & Pauliina Peltonen (toim.), *AFinLA-e. Soveltavan kielitieteen tutkimuksia* 7 s. 60–80.
Jantunen, Jarmo Harri & Pirkola, Silja 2015: Oppijansuomen sähköiset tutkimusaineistot. Nykytilanne. – *Virittäjä* 119 (1) s. 88–103.
Jarema, Gonia & Libben, Gary 2007: Introduction: matters of definition and core perspectives. – Gonia Jarema & Gary Libben (toim.), *The mental lexicon: core perspectives* s. 1–6. Amsterdam: Elsevier.
Jokela, Hanna 2017: *Se*-pronomini muodollisena subjektina suomenoppijoiden teksteissä. – *Lähivõrdlusi = Lähivertailuja* 27 s. 107–131.
Jyväskylän yliopisto 2020: *Yleistä YKIstä*. – https://www.jyu.fi/hytk/fi/laitokset/solki/yki/yleista 24.4.2020.
Kaivapalu, Annekatrin 2005: *Lähdekieli kielenoppimisen apuna*. Jyväskylä studies in humanities 44. Jyväskylä Jyväskylän yliopisto.
Kajander, Mikko 2013: *Suomen eksistentiaalilause toisen kielen oppimisen polulla*. Jyväskylä studies in humanities 220. Jyväskylä: Jyväskylän yliopisto.
Kallioranta, Otto 2009: Paljon-*adverbin kollokointi oppijansuomessa – korpusvetoinen tutkimus*. Pro gradu -tutkielma. Oulun yliopisto.
Kangas, Akseli 2018: *Adverbien* täysin *ja* kokonaan *kollokaatit ja semanttinen prosodia.* Maisterintutkielma. Jyväskylän yliopiston kieli- ja viestintätieteiden laitos.
Kangasniemi, Heikki 1997: *Sana, merkitys, maailma: katsaus leksikaalisen semantiikan perusteisiin*. Helsinki: Finn Lectura.
Kankaanpää, Salli 2009: *Omistusliitteet kuuluvat yleiskieleen.* – https://www.kielikello.fi/-/omistusliitteet-kuuluvat-yleiskieleen 29.4.2020.
Karlsson, Fred 2008: *Yleinen kielitiede*. Helsinki: Gaudeamus.
Kemppanen, Hannu 2008: *Avainsanoja ja ideologiaa. Käännettyjen ja ei-käännettyjen historiatekstien korpuslingvistinen analyysi*. Joensuun yliopiston humanistisia julkaisuja 51. Joensuu: Joensuun yliopisto.
Kuiri, Kaija 2012: *Johdatus semantiikkaan*. Helsinki: Finn Lectura.
Kuuluvainen, Helena 2015: *Fraseologiset virheet Kansainvälisessä oppijansuomen korpuksessa*. Pro gradu -tutkielma. Oulun yliopisto.
Kuusk, Margit 1999: *Suomi selväksi. Soome keele õpik.* Tartu: Eesti Keele Sihtasutus.
Kynsijärvi, Taru 2007: Se johtuu siitä, että minulla oli muistinmenetys – Olla-*verbirakenteiden kehkeytyminen oppijankielessä.* Pro gradu -tutkielma. Jyväskylän yliopiston kielten laitos.
Latomaa, Sirkku & Tuomela, Veli 1993: Suomi toisena vai vieraana kielenä? – *Virittäjä* 97 (2) s. 238–245.
Latomaa, Sirkku, Pöyhönen, Sari, Suni, Minna & Tarnanen, Mirja 2013: Kielikysymykset muuttoliikkeessä. – Tuomas Martikainen, Pasi Saukkonen & Minna Säävälä (toim.), *Muuttajat. Kansainvälinen muuttoliike ja suomalainen yhteiskunta* s. 163–183. Helsinki: Gaudeamus.

Lehto, Liisa-Maria 2018: *Korpusavusteinen diskurssianalyysi japaninsuomalaisten kielipuheesta*. Oulu: Oulun yliopisto.
Li, Liang 2016: *Sentence initial bundles in L2 thesis writing: A comparative study of Chinese L2 and New Zealand L1 postgraduates' writing.* Waikato: The University of Waikato.
Liu, Dilin 2012: The most frequently-used multi-word constructions in academic written English: A multi-corpus study. – *English for Specific Purposes* 31 (1) s. 25–35.
Lounela, Mikko & Heikkinen, Vesa 2012: Korpus. – Vesa Heikkinen, Eero Voutilainen, Petri Lauerma, Ulla Tiililä & Mikko Lounela (toim.), *Genreanalyysi – tekstilajitutkimuksen käsikirja* s. 120–127. Helsinki: Gaudeamus.
Maahanmuuttovirasto 2020: *Kielitaito*. – https://migri.fi/kielitaito 24.4.2020.
Mahlberg, Michaela 2013: *Corpus stylistics and Dickens's fiction*. Routledge advances in corpus linguistics 14. New York: Routledge.
Martin, Maisa 1999: Suomi toisena ja vieraana kielenä. – Kari Sajavaara & Arja Piirainen-Marsh (toim.), *Kielenoppimisen kysymyksiä* s. 157–178. Jyväskylä: Jyväskylän yliopisto, soveltavan kielentutkimuksen keskus.
Mauranen, Anna 2000: Strange strings in translated language. A study on corpora. – Maeve Olohan (toim.), *Intercultural faultlines. Research models in translation studies I. Textual and cognitive aspects* s. 119–141. Manchester: St. Jerome.
McCauley, Stewart M. & Christiansen, Morten H. 2015: Computational investigations of multiword chunks in language learning. – *Topics in Cognitive Science* 9 s. 637–652.
McEnery, Tony & Hardie, Andrew 2012: *Corpus linguistics. Method, theory and practice*. Cambridge: Cambridge University Press.
Moisl, Hermann 2015: *Cluster analysis for corpus linguistics*. Berlin: De Gruyter.
Mollin, Sandra 2009: 'I entirely understand' is a Blairism: the methodology of identifying idiolectal collocations. – *International Journal of Corpus Linguistics* 14 (3) s. 367–392.
Mueller, Scott 2003: *Upgrading and repairing PCs, 14th edition*. Indianapolis: Que.
Mustonen, Sanna 2015: *Käytössä kehittyvä kieli: paikat ja tilat suomi toisena kielenä -oppijoiden teksteissä*. Jyväskylä studies in humanities 255. Jyväskylä: Jyväskylän yliopisto.
Nekrasova, Tatiana M. 2009: English L1 and L2 speakers' knowledge of lexical bundles. – *Language Learning* 59 (3) s. 647–686.
Nesi, Hilary & Basturkmen, Helen 2006: Lexical bundles and discourse signalling in academic lectures. – *International Journal of Corpus Linguistics* 11 (3) s. 283–304.
Nesselhauf, Nadja 2003: The use of collocations by advanced learners of English and some implications for teaching. – *Applied Linguistics* 24 (2) s. 223–242.
——— 2004: Learner corpora and their potential for language teaching. – John Sinclair (toim.), *How to use corpora in language teaching* s. 125–152. Amsterdam: John Benjamins.
——— 2005: *Collocations in a learner corpus*. Studies in corpus linguistics 14. Amsterdam: John Benjamins.
Nieminen, Taija 2001: Kuvailun keinot oppijankielessä – merkitysten joustaminen ja intensiteetin ilmaiseminen espanjankielisten puhumassa suomessa. – Taija Nieminen (toim.), *Kakkoskieli 3. Vuorovaikutus ja suomen kielen oppiminen* s. 76–126. Helsinki: Helsingin yliopiston suomen kielen laitos.
Nissilä, Leena 2011: *Viron kielen vaikutus suomen kielen verbien ja niiden rektioiden oppimiseen*. Oulu: Oulun yliopisto.
Ohvo, Maija 2008: *Menneen ajan aikamuotojen käyttö S2-oppijoiden kirjoitelmissa*. Pro gradu -tutkielma. Tampereen yliopiston kieli- ja käännöstieteiden laitos.
Opetushallitus 2020: Taitotasoasteikko vuoden 2003 opetussuunnitelman mukaan. – https://www.oph.fi/fi/koulutus-ja-tutkinnot/taitotasoasteikko-vuoden-2003-opetussuunnitelman-mukaan 7.6.2020
OPS 2012 = *Aikuisten maahanmuuttajien kotoutumiskoulutuksen opetussuunnitelman perusteet.* Helsinki: Opetushallitus. https://www.oph.fi/fi/koulutus-ja-tutkinnot/aikuisten-maahanmuuttajien-kotoutumiskoulutus
Osborne, John 2008: Phraseology effects as a trigger for errors in L2 English: The case of more advanced learners. – Fanny Meunier & Sylviane Granger (toim.), *Phraseology in foreign language learning and teaching* s. 67–83. Amsterdam: John Benjamins.
Pace-Sigge, Michael T.L. 2013: The concept of Lexical Priming in the context of language use. – *ICAME Journal* 37 s. 149–173.
Papi, Mostafa & Teimouri, Yasser 2014: Language learner motivational types: a cluster analysis study. – *Language Learning* 64 (3) s. 493–525.
Paquot, Magali 2007: *EAP vocabulary in native and learner writing: From extraction to analysis. A phraseology-oriented approach*. Louvain: Universite catholique de Louvain.
——— 2008: Exemplification in learner writing. A cross-linguistic perspective. – Fanny Meunier & Sylviane Granger (toim.), *Phraseology in foreign language learning and teaching* s. 101–119. Amsterdam: John Benjamins.
——— 2013: Lexical bundles and L1 transfer effects. – *International Journal of Corpus Linguistics* 18 (3) s. 391–417.

——— 2014: Cross-linguistic influence and formulaic language: Recurrent word sequences in French learner writing. – Leah Roberts, Ineke Vedder & Jan Hulstijn (toim.), *EUROSLA Yearbook 14* s. 216–237. Amsterdam: John Benjamins.

Paquot, Magali & Granger, Sylviane 2012: Formulaic language in learner corpora. – *Annual Review of Applied Linguistics* 32 s. 130–149.

Pawley, Andrew & Syder, Frances Hodgetts 1983: Two puzzles for linguistic theory: nativelike selection and nativelike fluency. – Jack C. Richards & Richard W. Schmidt (toim.), *Language and communication* s. 191–226. New York: Longman.

Peromingo, Juan Pedro Rica 2012: Corpus analysis and phraseology: transfer of multi-word units. – *Linguistics and the Human Sciences* 6 (1–3) s. 321–343.

Pirkola, Silja 2016: *Synonyymien EHKÄ ja MAHDOLLISESTI kollokaatit ja semanttiset preferenssit*. Maisterintutkielma. Jyväskylän yliopiston kielten laitos.

Reiman, Nina 2011: Transitiivikonstruktio ikkunana syntaksin kehitykseen: Infiniittiset rakenteet ja passiivi taidon indikaattoreina S2-oppijoiden teksteissä. – *AFinLA-e: Soveltavan kielitieteen tutkimuksia* (3) s. 142–157.

——— 2014: Yläkoulun S2-oppilaiden transitiivi-ilmausten käyttö Eurooppalaisen viitekehyksen taitotasoilla. – *Lähivõrdlusi = Lähivertailuja* 24 s. 183–220.

Renouf, Antoinette & Sinclair, John 1991: Collocational frameworks in English. – Karin Ajimer & Bengt Altenberg (toim.), *English corpus linguistics. Studies in honour of Jan Svartvik* s. 128–143. Harlow: Longman.

Salazar, Danica 2014: *Lexical bundles in native and non-native scientific writing: applying a corpus-based study to language teaching*. Studies in corpus linguistics 65. Amsterdam: John Benjamins.

Salonen, Juhana, Takkinen, Ritva, Puupponen, Anna, Nieminen, Henri & Pippuri, Outi 2016: Creating corpora of Finland's sign languages. – Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, & Johanna Mesch (toim.), *Workshop proceedings: 7th workshop on the representation and processing of sign languages: Corpus Mining / Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* s. 179–184. Paris: European Language Resources Association (ELRA).

Saussure, Ferdinand de 2014: *Yleisen kielitieteen kurssi*. Suomentanut Tommi Nuopponen. Tampere: Vastapaino.

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: Jyväskylän yliopisto.

Selinker, Larry 1972: Interlanguage. – *International Review of Applied Linguistics* 10 s. 209–241.

Seppälä, Tanja 2013: Oppijansuomen kolligaatit ketjuuntuvissa verbirakenteissa. – *Lähivõrdlusi = Lähivertailuja* 23 s. 315–340.

Setälä, Emil Nestor 1952 [1880]: *Suomen kielen lauseoppi*. Helsinki: Otava.

Shibuya, Yoshikata & Jensen, Kim Ebensgaard 2015: Mining for constructions in texts using n-gram and network analysis. – *Globe: A Journal of Language, Culture and Communication* 2 s. 23–54.

Siitonen, Kirsti & Mizuno, Manami 2010: Suomen monitahoinen possessiivisuffiksi ja suomenoppija. – *Lähivõrdlusi = Lähivertailuja* 19 s. 136–159.

Siivelt, Keaty & Mustonen, Sanna 2013: Lähdekielen vaikutus ja kielitaitotasot: paikallissijojen kehitys oppijansuomessa. – *Lähivõrdlusi = Lähivertailuja* 23 s. 341–370.

Sinclair, John 1991: *Corpus, concordance, collocation*. Oxford: Oxford University Press.

——— 2004: *Trust the text. Language, corpus and discourse*. London: Routledge.

Sinclair, John, Jones, Susan & Daley, Robert (toim.) 2004: English collocation studies: the OSTI report. London: Continuum.

Singleton, David, Leśniewska, Justyna & Witalisz, Ewa 2007: Open choice versus the idiom principle in L2 lexical usage. – Mirosław Pawlak (toim.), *Exploring focus on form in language teaching* s. 209–224. Poznań: Adam Mickiewicz University.

Spoelman, Marianne 2013: *Prior linguistic knowledge matters: The use of the partitive case in Finnish learner language*. Oulu: Oulun yliopisto.

Stubbs, Michael 1995: Collocations and cultural connotations of common words. – *Linguistics and Education* 7 s. 379–390.

——— 1996: *Text and corpus analysis. Corpus-assisted studies of language and culture*. Oxford: Blackwell.

——— 2007a: An example of frequent English: phraseology: distributions, structures and functions. –Roberta Facchinetti (toim.), *Corpus linguistics 25 years on* s. 89–105. Amsterdam: Rodopi.

——— 2007b: Quantitative data on multi-word sequences in English: the case of the word world. – Michael Hoey, Michaela Mahlberg, Michael Stubbs & Wolfgang Teubert (toim.), *Text, discourse and corpora. Theory and analysis* s. 163–189. London: Continuum.

——— 2009: Technology and phraseology. – Ute Römer & Rainer Schulze (toim.), *Exploring the lexis-grammar interface* s. 15–31. Amsterdam: John Benjamins.

Tarvainen, Jenny 2018: *SAADA-verbin fraseologiaa: vertaileva korpustutkimus oppijan- ja natiivisuomesta*. Maisterintutkielma. Jyväskylän yliopiston kieli- ja viestintätieteiden laitos.
Tervo, Anne 2013: Intensiteettisanojen frekvenssit ja kollokaatiot oppijansuomessa. Pro gradu -tutkielma. Vaasan yliopiston filosofinen tiedekunta.
Tieteen termipankki 2020: *Nimitys: korpus*. – https://www.tieteentermipankki.fi/wiki/Nimitys:korpus 8.3.2020.
Tognini-Bonelli, Elena 2001: *Corpus linguistics at work*. Studies in corpus linguistics 6. Amsterdam: John Benjamins.
——— 2002: Functionally complete units of meaning across English and Italian: Towards a corpus-driven approach. – Sylviane Granger & Bengt Altenberg (toim.), *Lexis in contrast: Corpus-based approaches* s. 73–95. Amsterdam: John Benjamins.
Turun yliopiston kieli- ja käännöstieteiden laitos 2012: *Edistyneiden suomenoppijoiden korpus* [tekstikorpus]. Kielipankki. – http://urn.fi/urn:nbn:fi:lb-201407167 11.9.2020.
Turunen, Ulla 2012: *Suomi toisena kielenä -oppijoiden possessiivisuffiksin käyttö.* Pro gradu -tutkielma. Jyväskylän yliopiston kielten laitos.
Ullakonoja, Riikka & Dufva, Hannele 2016: Toisen ja vieraan kielen ääntämisen oppimisen haasteet. – *Oppimisen ja oppimisvaikeuksien erityislehti* 26 (2) s. 4–18.
Valmu, Jenni 2007: *Suomen menneen ajan tempusten semantiikka unkarilaisilla kielenoppijoilla*. Pro gradu -tutkielma. Tampereen yliopiston kieli- ja käännöstieteiden laitos.
van Gompel, Maarten & van den Bosch, Antal 2016: Efficient n-gram, skipgram and flexgram modelling with Colibri Core. – *Journal of Open Research Software* 4 (1) s. 1–10.
Varis, Klára 2010: *Ajanilmaukset Cefling-hankkeen koululaisaineistossa.* Pro gradu -tutkielma. Jyväskylän yliopiston kielten laitos.
Vetchinnikova, Svetlana 2014: *Second language lexis and the idiom principle*. Helsinki: Helsingin yliopisto.
Vilkuna, Maria 2003: *Suomen lauseopin perusteet*. Helsinki: Edita.
Virtanen, Veera 2011: Minä lienen tullut joskus Suomessa vielä? *Venäjänkielisten suomi toisena ja suomi vieraana kielenä -oppijoiden perfektin ja imperfektin omaksumisen ongelmista.* Pro gradu -tutkielma. Jyväskylän yliopiston kielten laitos.
VISK = Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen & Irja Alho 2004: *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura. – http://scripta.kotus.fi/visk
Wang, Ying 2016: *The idiom principle and L1 influence. A contrastive learner-corpus study of delexical verb + noun collocations*. Studies in corpus linguistics 77. Amsterdam: John Benjamins.
Warren, Martin 2011: Learning and teaching of phraseological variation. – Ana Frankenberg-Garcia, Lynne Flowerdew & Guy Aston (toim.), *New trends in corpora and language learning* s. 153–166. London: Continuum.
Weisser, Martin 2016: *Practical corpus linguistics*. *An introduction to corpus-based language analysis*. Chichester, England: Wiley Blackwell.
White, Leila 2008: *Suomen kielioppia ulkomaalaisille*. Helsinki: Finn Lectura.
Williams, Geoffrey C. 2008: The Good Lord and his works. A corpus-driven study of collocational resonance. – Sylviane Granger & Fanny Meunier (toim.), *Phraseology. An interdisciplinary perspective* s. 159–173. Amstedam: John Benjamins.
Wray, Alison 2000: Formulaic sequences in second language teaching: principle and practice. – *Applied Linguistics* 21 (4) s. 463–489.
——— 2002a: *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
——— 2002b: Formulaic language in computer-supported communication: theory meets reality. – *Language Awareness* 11 (2) s. 114–131.
——— 2008: *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.
Zipf, George Kingsley 1949: *Human behavior and the principle of least effort. An introduction to human ecology*. Cambridge: Addison-Wesley Press.

Aineistolähde

ICLFI = Jantunen, Jarmo, Brunni, Sisko & Oulun yliopisto, suomen kielen oppiaine 2013: *Kansainvälinen oppijansuomen korpus* [tekstikorpus]. Kielipankki. – http://urn.fi/urn:nbn:fi:lb-20140730163

LIITE 1

B1-aineiston n-grammeissa vähintään viisi kertaa esiintyvät sananmuodot esiintymämääriensä mukaisesti järjestettyinä frekventeimmästä sananmuodosta alkaen.

	f	sananmuoto			
1	391	on	58	11	vanha
2	285	ja	59	11	vielä
3	88	hän	60	11	vieressä
4	79	ei	61	10	kun
5	74	se	62	10	monta
6	70	että	63	10	nukkumaan
7	66	ole	64	10	samana
8	63	minulla	65	10	voi
9	47	mutta	66	10	yli
10	43	minä	67	9	aamiaiseksi
11	41	minun	68	9	aamiaisen
12	35	en	69	9	kylpyhuone
13	33	syön	70	9	nyt
14	31	menen	71	9	oikein
15	31	olen	72	9	pääsin
16	29	oli	73	9	suuri
17	28	sen	74	9	täytyy
18	27	kaksi	75	9	vessa
19	27	koska	76	8	huoneessa
20	26	jälkeen	77	8	kirjaa
21	26	paljon	78	8	siinä
22	26	pieni	79	8	suomen
23	25	vuonna	80	7	asun
24	24	kello	81	7	hyvin
25	24	yksi	82	7	keitän
26	22	juon	83	7	kirjahylly
27	21	iso	84	7	kotoa
28	19	myös	85	7	kymmenen
29	18	puuroa	86	7	lattialla
30	17	meillä	87	7	luulen
31	17	niin	88	7	minusta
32	17	sitten	89	7	pukeudun
33	17	voileipää	90	7	tartossa
34	16	noin	91	7	tiedä
35	16	tukka	92	7	tosi
36	15	tai	93	7	vartin
37	15	tavallisesti	94	7	äiti
38	15	vuotta	95	6	hyvä
39	13	hänellä	96	6	ikkunan
40	13	keittiö	97	6	isäni
41	13	kotiin	98	6	kaunis
42	13	käyn	99	6	kirjoitin
43	13	ollut	100	6	lamppu
44	13	sänky	101	6	me
45	12	he	102	6	melko
46	12	huone	103	6	minulle
47	12	huonetta	104	6	minuuttia
48	11	alkaa	105	6	mukava
49	11	kahvia	106	6	neljä
50	11	kirjoituspöytä	107	6	nousta
51	11	kolme	108	6	ovat
52	11	laitan	109	6	puoli
53	11	nimi	110	6	sanoi
54	11	oikealla	111	6	seinällä
55	11	siellä	112	6	sisko
56	11	silmät	113	6	teetä
57	11	suihkussa	114	6	televisiota

	f	sananmuoto
115	6	tietokone
116	6	tärkeä
117	6	töissä
118	6	viihtyisä
119	6	yliopistoon
120	6	ylioppilaaksi
121	6	äitini
122	5	aika
123	5	eivät
124	5	halua
125	5	hänen
126	5	isä
127	5	jos
128	5	kielen
129	5	kotoisin
130	5	kotona
131	5	liian
132	5	luento
133	5	lähden
134	5	matto
135	5	miksi
136	5	siniset
137	5	sinulla
138	5	tarton
139	5	toivon
140	5	tumma
141	5	vaalea
142	5	varttia
143	5	verho

LIITE 2

B1-aineiston n-grammien sanojen lemmamuodot esiintymämääriensä mukaisesti järjestettyinä frekventeimmästä lemmasta alkaen.

	f	lemma
1	546	olla
2	285	ja
3	160	minä
4	123	ei
5	114	se
6	107	hän
7	70	että
8	47	mutta
9	42	vuosi
10	40	mennä
11	39	huone
12	33	syödä
13	28	paljon
14	27	kaksi
15	27	koska
16	26	jälkeen
17	26	koti
18	26	pieni
19	25	iso
20	24	kello
21	24	yksi
22	23	juoda
23	23	me
24	20	käydä
25	19	aamiainen
26	19	myös
27	18	puuro
28	17	niin
29	17	sitten
30	17	voileipä
31	16	nimi
32	16	noin
33	16	sänky
34	16	tukka
35	15	he
36	15	keittiö
37	15	tai
38	15	tavallisesti
39	15	voida
40	14	äiti
41	13	hyvä
42	13	suomi
43	12	isä
44	12	kahvi
45	12	sama
46	12	tarto
47	12	vartti
48	11	nimi
49	11	alkaa
50	11	asua
51	11	kirjoituspöytä
52	11	kolme
53	11	laittaa
54	11	oikea
55	11	sanoa
56	11	siellä
57	11	silmä
58	11	suihku
59	11	täytyä
60	11	vanha
61	11	vielä
62	11	vieressä
63	11	yliopisto
64	10	kun
65	10	moni
66	10	nousta
67	10	nukkua
68	10	opiskella
69	10	pitää
70	10	tosi
71	10	yli
72	9	aika
73	9	kirja
74	9	kylpyhuone
75	9	nyt
76	9	oikein
77	9	päästä
78	9	suuri
79	9	tiedä
80	9	työ
81	9	vessa
82	8	luulla
83	8	perhe
84	8	pukeutua
85	8	sisko
86	8	veli
87	7	haluta
88	7	hyvin
89	7	ikkuna
90	7	keittää
91	7	kieli
92	7	kirjahylly
93	7	kymmenen
94	7	lattia
95	7	sinä
96	7	televisio
97	6	aurinko
98	6	kaunis
99	6	kirjoittaa
100	6	lamppu
101	6	luento
102	6	melko
103	6	minuutti
104	6	mukava
105	6	neljä
106	6	puoli
107	6	ruskea
108	6	seinä
109	6	tee
110	6	tietokone
111	6	tärkeä
112	6	vasen
113	6	viihtyisä
114	6	ylioppilas

	f	lemma
115	5	asunto
116	5	joka
117	5	jos
118	5	kotoisin
119	5	lapsi
120	5	liian
121	5	lähteä
122	5	matto
123	5	mieli
124	5	miksi
125	5	sininen
126	5	syntyä
127	5	toivoa
128	5	tumma
129	5	vaalea
130	5	verho
131	5	vihreä
132	4	aina
133	4	harmaa
134	4	kaksitoista
135	4	kanssa
136	4	katsoa
137	4	kerros
138	4	kiva
139	4	kuin
140	4	oma
141	4	ottaa
142	4	peseytyä
143	4	pestä
144	4	ruoka
145	4	tarpeeksi
146	4	vaikea
147	4	vaille
148	4	vain
149	3	asti
150	3	hammas
151	3	hetki
152	3	jalka
153	3	jo
154	3	kahdeksan
155	3	kaikki
156	3	kerta
157	3	kestää
158	3	koskaan
159	3	koulu
160	3	kurssi
161	3	lapsuus
162	3	lukea
163	3	lyhyt
164	3	lähellä
165	3	matka
166	3	mitä
167	3	naimisissa
168	3	peruskoulu
169	3	pöytä
170	3	tallinna
171	3	tarvita
172	3	tehdä
173	3	toinen
174	3	tuoli
175	3	vaatekaappi
176	3	vanhempi
177	3	viikko
178	3	viisitoista
179	2	aamu
180	2	alla
181	2	aste
182	2	englanti
183	2	enää
184	2	erittäin
185	2	harrastaa
186	2	iloinen
187	2	kertoa
188	2	kuulua
189	2	lukio
190	2	maanantai
191	2	mielenkiintoinen
192	2	onnellinen
193	2	opettaja
194	2	opiskelija
195	2	ostaa
196	2	paistaa
197	2	pitkä
198	2	puhua
199	2	punainen
200	2	rauhallinen
201	2	sijaita
202	2	siksi
203	2	sokeri
204	2	talo
205	2	torstai
206	2	tuli
207	2	tulla
208	2	tämä
209	2	uida
210	2	varma
211	2	vähän
212	2	väsyä
213	2	yksitoista
214	1	ajatella
215	1	anssi
216	1	edelleen
217	1	edessä
218	1	ehkä
219	1	ensimmäinen
220	1	firma
221	1	hauska
222	1	herätyskello
223	1	herätä
224	1	hyväntuulinen
225	1	ihan
226	1	ilma
227	1	jutta
228	1	kauan
229	1	kauppa
230	1	kaupunki
231	1	keltainen
232	1	kerma
233	1	keskellä
234	1	keskiviikko
235	1	kesäkuu
236	1	kevyt
237	1	kiina
238	1	komea
239	1	korva
240	1	kukaan
241	1	kuolla
242	1	kuunnella
243	1	kuusi
244	1	kylmä

	f	lemma
245	1	käyttää
246	1	lehti
247	1	loppua
248	1	lounas
249	1	luonne
250	1	luonto
251	1	lämmin
252	1	maito
253	1	marja
254	1	mehu
255	1	melkein
256	1	mikään
257	1	musiikki
258	1	nojatuoli
259	1	näyttää
260	1	omakotitalo
261	1	osata
262	1	ostos
263	1	ovi
264	1	palata
265	1	peili
266	1	peite
267	1	poika
268	1	poikaystävä
269	1	puu
270	1	puutarha
271	1	päivä
272	1	päälle
273	1	radio
274	1	raha
275	1	rakkaus
276	1	seitsemän
277	1	sellainen
278	1	sieni
279	1	sohva
280	1	soida
281	1	takia
282	1	tarkoittaa
283	1	taulu
284	1	tenttikausi
285	1	terveinen
286	1	tieto
287	1	tiistai
288	1	tilava
289	1	tunti
290	1	tuntua
291	1	tuuli
292	1	tyyny
293	1	tähti
294	1	urheilu
295	1	uskoa
296	1	vaate
297	1	vaihto-oppilas
298	1	vaikka
299	1	valitettavasti
300	1	vapaa
301	1	viime
302	1	viro
303	1	vähitellen
304	1	yhdeksän
305	1	ymmärtää
306	1	ystävä
307	1	ystävällinen

LIITE 3

Tutkimuksessa löydetyt, *Kansainvälisen oppijansuomen korpuksen* B1-kielitaitotasolle arvioitujen tekstien 3-, 4-, 5- ja 6-grammit (n = 992; f = 16 614) esiintymämääriensä mukaisesti järjestettyinä frekventeimmästä n-grammista alkaen.[46]

	f	jakauma	n-grammi
1	210	198	minulla ei ole
2	155	112	ja hän on
3	152	129	se ei ole
4	106	95	että se on
5	88	80	koska se on
6	79	74	ja se on
7	79	73	ja sen jälkeen
8	78	77	on pieni mutta
9	77	70	ei ole niin
10	77	73	minulla on kaksi
11	75	73	minulla on myös
12	75	69	mutta se on
13	73	67	mutta hän ei
14	72	71	minulla on yksi
15	71	52	[%] vuotta vanha
16	70	68	minun huone on
17	69	63	mutta se ei
18	67	67	opiskelen tarton yliopistossa
19	64	63	käyn suihkussa ja
20	63	62	ja juon kahvia
21	62	37	hän on [%]-vuotias
22	62	55	minusta se on
23	60	52	että hän on
24	57	48	koska hän on
25	57	53	minä en ole
26	52	45	sen jälkeen menen
27	50	30	on [%] vuotta
28	47	42	koska minulla on
29	47	30	on [%]-vuotias ja
30	46	41	minulla on vielä
31	45	41	hän ei ole
32	45	43	ikkunan vieressä on
33	45	45	voileipää ja juon
34	45	45	vuonna [%] ja
35	44	43	ei ole paljon
36	44	41	meillä ei ole
37	44	44	menen nukkumaan kello
38	43	40	minulla on paljon
39	43	43	minun nimeni on
40	43	42	se on hyvä
41	41	40	mutta minulla on
42	41	40	peseydyn ja pukeudun
43	41	41	pieni mutta viihtyisä
44	40	37	että minulla on
45	40	40	ja juon teetä
46	40	40	ja menen nukkumaan
47	40	38	meillä on myös
48	40	38	menen kotiin ja
49	39	33	että hän ei
50	39	39	huone on pieni
51	39	30	hän on töissä
52	39	39	on kaksi huonetta
53	38	38	aamiaiseksi syön tavallisesti
54	38	31	vuotta vanha ja
55	37	33	hänen nimensä on [name]
56	37	36	kylpyhuone ja vessa
57	37	31	minä ja minun
58	37	37	on kolme huonetta
59	37	37	oven lähellä on
60	37	35	se on niin
61	36	34	ei ole mitään
62	36	34	että se oli
63	36	22	hän on [%]
64	36	34	hän on myös
65	36	27	hän sanoi että
66	36	34	ja hänellä on
67	35	31	hänellä ei ole
68	35	33	koska siellä on
69	35	32	mutta minä en
70	35	35	syntynyt vuonna [%]
71	34	32	en ole ollut
72	34	34	laitan aamiaisen ja
73	34	34	minulla on pieni
74	34	28	minä luulen että
75	33	29	jonka nimi on
76	33	33	nimeni on [name]
77	33	33	on pieni mutta viihtyisä
78	32	31	on iso ja
79	32	31	on kaunis ja
80	32	25	sanoi että hän
81	32	31	se on tosi
82	31	27	alkaa vartin yli
83	31	28	koska hän ei
84	31	30	luulen että se
85	31	29	sitten menen kotiin
86	31	28	sitä mieltä että
87	30	30	aamiaisen ja keitän
88	30	29	ei ole aikaa
89	30	30	minun huone on pieni
90	30	30	olen syntynyt vuonna [%]
91	30	29	siellä ei ole
92	30	28	äiti ja isä
93	29	28	ei ole vielä
94	29	28	ja minulla on
95	29	23	se ei ollut
96	29	27	se on hyvin
97	28	24	että se ei
98	28	28	kaksi kertaa viikossa
99	28	27	mutta se ei ole
100	28	28	mutta se oli

[46] Sarakkeen *jakauma* arvot ilmaisevat, monenko eri tekstin välille kunkin n-grammin esiintymät ovat aineistossa jakautuneet.

	f	jakauma	n-grammi
101	27	24	ja hän ei
102	27	26	ja minä olen
103	27	27	kaksi tai kolme
104	27	27	keittiö, kylpyhuone ja
105	27	23	kun hän oli
106	27	21	kun minä olin
107	27	27	laitan aamiaisen ja keitän
108	27	27	matka kestää noin
109	27	27	syön tavallisesti puuroa
110	26	19	[%] vuotta vanha ja
111	26	17	[%]-vuotias ja hän
112	26	23	että hän oli
113	26	26	huone on pieni mutta
114	26	22	ja se oli
115	26	26	kello puoli yhdeksän
116	26	25	koska minulla ei
117	26	26	minä asun tartossa
118	26	26	minä en pidä
119	26	24	minä en voi
120	26	25	on kirjahylly ja
121	25	25	minulla on yksi sisko
122	25	25	minulla on äiti
123	25	25	minun täytyy opiskella
124	25	25	mutta ei ole
125	25	24	mutta hän on
126	25	25	on oikein kaunis
127	25	25	puuroa ja juon
128	25	25	pöydällä on lamppu
129	25	23	siellä on myös
130	25	25	syön puuroa ja
131	25	25	syön voileipää ja
132	24	23	en ole koskaan
133	24	22	hän ei voi
134	24	19	hän on työssä
135	24	24	ja juon maitoa
136	24	23	ja niin edelleen
137	24	23	ja siellä on
138	24	24	lähden kotoa puoli
139	24	24	menen nukkumaan noin
140	24	24	minulla en ole
141	24	24	mutta en ole
142	24	17	noin kello [%]
143	24	23	on kaksi kerrosta
144	24	24	samana vuonna pääsin
145	24	24	se on myös
146	24	23	se voi olla
147	24	22	sen jälkeen minä
148	24	24	syön tavallisesti voileipää
149	24	23	tukka ja siniset
150	23	22	hänen nimi on [name]
151	23	22	jälkeen menen kotiin
152	23	23	koska minulla ei ole
153	23	7	miksi sinulla on niin
154	23	23	minulla on kolme
155	23	23	on mukava ja
156	23	23	on myös pieni
157	23	22	on niin paljon
158	23	23	on noin [%]
159	23	23	on pieni ja
160	23	23	on äiti, isä
161	23	22	se on oikein
162	23	23	sitten käyn suihkussa
163	23	22	tukka ja siniset silmät
164	23	21	vaalea tukka ja
165	23	20	voin sanoa että
166	22	22	ei ole suuri
167	22	22	hän ei ollut
168	22	20	hänellä on lyhyt
169	22	22	ja isäni on
170	22	22	ja keitän kahvia
171	22	21	juon kahvia ja
172	22	21	kun minulla on
173	22	21	minun mielestäni on
174	22	21	minun täytyy mennä
175	22	22	nousen aamulla kello
176	22	22	olen kotoisin tallinnasta
177	22	14	on [%]-vuotias ja hän
178	22	21	on aina paljon
179	22	22	on iso sänky
180	22	22	on monta kirjaa
181	22	20	on myös kaksi
182	22	21	vessa ja kylpyhuone
183	21	21	että hän voi
184	21	21	ja minä en
185	21	19	ja se ei
186	21	21	kirjoitin vuonna [%]
187	21	20	kirjoituspöytä ja tuoli
188	21	19	koska hän oli
189	21	16	minulla ei ollut
190	21	18	minusta se oli
191	21	18	minusta tuntuu että
192	21	20	minä en tiedä
193	21	21	mutta en tiedä
194	21	19	olen sitä mieltä että
195	21	21	on vain yksi
196	21	21	ylioppilaaksi vuonna [%]
197	20	17	[%] ja [%]
198	20	19	hän on hyvin
199	20	16	ja harmaat silmät
200	20	20	ja minulla ei ole
201	20	19	ja siinä on
202	20	19	ja sitten menen
203	20	20	jälkeen menen ostoksille
204	20	20	minun huone on pieni mutta
205	20	20	minun äitini on
206	20	20	olen lapsuudesta asti
207	20	20	on oikealla sänky
208	20	17	on yksi poika
209	20	19	opiskelee tarton yliopistossa
210	20	20	peruskoulun ja lukion
211	20	20	se on totta
212	20	19	se on vaikea
213	20	14	suomen kieli on
214	20	20	tarpeeksi suuri minulle
215	20	19	tumma tukka ja
216	20	19	voi sanoa että
217	19	18	alkaa kello [%]
218	19	19	aurinko paistaa ja
219	19	19	ehkä se on
220	19	19	ei ole lapsia
221	19	19	en ole vielä
222	19	17	että se ei ole
223	19	16	hän ei halua
224	19	18	ja heillä on
225	19	17	ja lattialla on
226	19	19	ja luulen että
227	19	18	jos minulla on
228	19	15	jos sinulla on

	f	jakauma	n-grammi
229	19	19	jälkeen peseydyn ja pukeudun
230	19	18	kahvia ja syön
231	19	17	kolme kertaa viikossa
232	19	19	menen jalan koska
233	19	17	minulla on vähän
234	19	18	minun isäni on
235	19	17	mutta nyt hän
236	19	19	on yksi veli
237	19	16	se on ihan
238	19	19	se on pieni
239	19	19	yksi sisko ja
240	19	19	ylioppilaaksi kirjoitin vuonna [%]
241	18	18	heillä on kaksi
242	18	18	huoneen keskellä on
243	18	17	hän on syntynyt
244	18	12	ja hän asuu
245	18	18	ja katson televisiota
246	18	17	ja käyn suihkussa
247	18	15	ja sanoi että
248	18	16	ja sitten hän
249	18	14	koska minulla oli
250	18	18	koska se oli
251	18	18	kotiin ja syön
252	18	17	kun minä olen
253	18	18	laitan ruokaa ja
254	18	18	luulen että se on
255	18	17	minulla on aikaa
256	18	18	minulla on vain
257	18	18	minulla on yksi veli
258	18	17	minun huoneessa on
259	18	18	minun perheessä on
260	18	17	mutta luulen että
261	18	17	mutta minulla ei
262	18	18	nyt asun tartossa
263	18	17	on melko pieni
264	18	18	on tosi hyvä
265	18	18	on vaatekaappi ja
266	18	17	peite ja tyyny
267	18	16	se oli tosi
268	18	16	siellä on paljon
269	18	18	vaikka se on
270	18	18	äiti, isä ja
271	17	17	aamiaiseksi syön tavallisesti puuroa
272	17	17	ei ole ollut
273	17	12	ei ole rahaa
274	17	15	että hänellä on
275	17	17	huoneessa on oikealla
276	17	17	hänellä on paljon
277	17	15	ja sitten minä
278	17	17	ja toivon että
279	17	16	kello puoli seitsemän
280	17	17	koulun jälkeen menen
281	17	17	meillä on kaksi
282	17	16	meillä on kolme
283	17	17	meillä on kotona
284	17	17	minulla on iso
285	17	17	minun ei tarvitse
286	17	17	mitä sinulle kuuluu
287	17	17	niin paljon kuin
288	17	17	olen varma että
289	17	12	on [%] vuotta vanha ja
290	17	17	on iso puutarha
291	17	16	on kaunis ilma
292	17	17	on kirjoituspöytä ja
293	17	16	on myös iso
294	17	17	on televisio ja
295	17	17	pesen hampaat ja
296	17	14	se oli hyvä
297	17	17	se tarkoittaa että
298	17	17	sitten käyn suihkussa ja
299	17	16	sitten me menimme
300	17	14	tukka ja harmaat silmät
301	17	17	vartin yli kymmenen
302	17	17	äitini on opettaja
303	16	11	[%]-vuotias ja hän on
304	16	16	aamiaiseksi syön tavallisesti voileipää
305	16	16	ei ole niin paljon
306	16	16	en tiedä mitä
307	16	16	että he ovat
308	16	16	että minä olen
309	16	14	että suomi on
310	16	15	heillä ei ole
311	16	16	hetkellä asun tartossa
312	16	16	huone on pieni mutta viihtyisä
313	16	11	hän on [%] vuotta vanha
314	16	16	ja he ovat
315	16	16	ja kaksi tuolia
316	16	16	ja yksi veli
317	16	16	keittiö, kylpyhuone ja vessa
318	16	16	keittiö, vessa ja
319	16	16	kirjoitin ylioppilaaksi vuonna [%]
320	16	16	koska he eivät
321	16	16	kotoa varttia vaille
322	16	16	laitan aamiaisen ja keitän kahvia
323	16	16	lähden kotoa kello
324	16	16	minä ajattelen että
325	16	13	minä olen ollut
326	16	15	mutta hän oli
327	16	16	nyt hän on
328	16	15	nyt minä olen
329	16	16	on aika iso
330	16	16	on kotona vielä
331	16	16	on rauhallinen ja
332	16	15	on se että
333	16	14	on tärkeä että
334	16	16	pöydällä on tietokone
335	16	10	sanoo että hän
336	16	16	sen jälkeen hän
337	16	16	sen jälkeen peseydyn ja pukeudun
338	16	16	siellä on kaksi
339	16	15	siellä on monta
340	16	15	sohvan edessä on
341	16	14	suomen kielen kurssi
342	16	16	tavallisesti menen nukkumaan
343	16	16	tavallisesti voileipää ja
344	16	16	vartin yli kahdeksan
345	16	16	vieressä on iso
346	16	14	voi olla että
347	16	16	voileipää ja juon teetä
348	16	16	yksi tai kaksi
349	16	16	äidin nimi on
350	15	14	[%] vuotta sitten

	f	jakauma	n-grammi
351	15	15	aamiaiseksi syön puuroa
352	15	15	ei ole iso
353	15	15	ei voi olla
354	15	15	en halua nousta
355	15	15	en tiedä miksi
356	15	15	he eivät ole
357	15	15	hän on kotoisin
358	15	15	iso kirjoituspöytä ja
359	15	15	ja hän oli
360	15	14	ja hän pitää
361	15	12	ja hänella on
362	15	15	ja laitan aamiaisen
363	15	7	ja samana vuonna pääsin
364	15	15	katson televisiota ja
365	15	15	käyn suihkussa ja pesen
366	15	15	lähden kotoa varttia
367	15	14	maanantaisin ja keski-viikkoisin
368	15	15	minulla on yksi sisko ja
369	15	15	minun perhe on
370	15	15	mutta he eivät
371	15	15	mutta minulla ei ole
372	15	15	mutta siellä on
373	15	14	noin kello kaksitoista
374	15	14	olen lapsuudesta asti harrastanut
375	15	15	on myös yksi
376	15	15	on sänky ja
377	15	15	samana vuonna pääsin yliopistoon
378	15	14	se on tärkeä
379	15	14	sen jälkeen menin
380	15	15	sen nimi on
381	15	13	sen vieressä on
382	15	15	sitten pukeudun ja
383	15	15	syön tavallisesti voileipää ja
384	15	15	vieressä on kirjoituspöytä
385	15	15	voileipää ja juon kahvia
386	15	15	äitini ja isäni
387	14	14	[%] vuotta ja
388	14	14	ei ole enää
389	14	13	ei ole liian
390	14	13	että minä en
391	14	12	hän on iloinen
392	14	14	hän on oikein
393	14	13	hän on tosi
394	14	14	hänellä on oma
395	14	11	hänen nimensa on
396	14	14	iso sänky ja
397	14	14	ja kuuntelen musiikkia
398	14	14	ja ostan ruokaa
399	14	14	ja sitten me
400	14	14	jo [%] vuotta
401	14	14	kello puoli kahdeksan
402	14	14	koska minun täytyy
403	14	14	kylpyhuone, vessa ja
404	14	14	käyn suihkussa ja pukeudun
405	14	14	luento alkaa vartin yli
406	14	14	menen nukkumaan kahdeltatoista
407	14	14	mielestäni se on
408	14	14	minulla ei ole lapsia
409	14	14	minulla ei ole paljon
410	14	13	minulla on [%]
411	14	14	minulla on aika
412	14	14	minun huone on pieni mutta viihtyisä
413	14	13	minun perheeni on
414	14	14	minun suuri rakkauteni on
415	14	14	minun täytyy tehdä
416	14	13	on kaksi lasta
417	14	14	on kaksi sänkyä
418	14	14	on melko suuri
419	14	14	on myös paljon
420	14	13	on oma huone
421	14	14	on pieni huone
422	14	14	on vielä yksi
423	14	14	se oli niin
424	14	13	se on kaunis
425	14	14	se on minun
426	14	14	seinällä on kirjahylly
427	14	13	seinällä on taulu
428	14	14	sen jälkeen laitan
429	14	14	sen jälkeen menen kotiin
430	14	14	suomen kielen tunti
431	14	14	sängyn vieressä on
432	14	14	sänky ja oikealla
433	14	14	tai katson televisiota
434	14	14	talossa on kaksi
435	14	14	toisessa kerroksessa on
436	14	11	viisitoista minuuttia yli
437	13	12	en tiedä vielä
438	13	13	ensimmäinen luento alkaa
439	13	13	että kaikki on
440	13	13	huoneessa on vasemmalla
441	13	13	huonessa on oikealla
442	13	13	hyvin englantia ja
443	13	11	hänellä on myös
444	13	12	ja aurinko paistaa
445	13	13	ja juon mehua
446	13	13	ja kaksi veljeä
447	13	6	ja miksi sinulla on niin
448	13	13	keittiö on pieni
449	13	13	keitän kahvia ja
450	13	9	kello [%] ja
451	13	13	kertaa viikossa käyn
452	13	13	koska minä en
453	13	13	kotiin noin kello
454	13	13	käy vielä koulua
455	13	13	luen kirjaa tai
456	13	13	luento alkaa kello
457	13	13	lähden kotoa varttia vaille
458	13	12	meillä on iso
459	13	13	menen nukkumaan kello kaksitoista
460	13	9	menin kotiin ja
461	13	11	minulla oli paljon
462	13	12	minulla on oma
463	13	11	minulla on vapaa
464	13	10	minun täytyi mennä
465	13	13	minun vanhempani asuvat
466	13	12	minä opiskelen suomea
467	13	13	mutta en pidä
468	13	12	mutta kun hän
469	13	13	mutta me emme

	f	jakauma	n-grammi
470	13	12	mutta sen jälkeen
471	13	13	nousen tavallisesti kello
472	13	9	on [%]-vuotias ja hän on
473	13	12	on hyvin tärkeä
474	13	13	on iso kirjoituspöytä
475	13	13	on neljä huonetta
476	13	13	on niin tärkeä
477	13	12	on vielä kaksi
478	13	13	pesen hampaani ja
479	13	12	pääsin tarton yliopistoon
480	13	13	se on liian
481	13	12	se on mielenkiintoinen
482	13	12	sen jälkeen kun
483	13	12	tällä hetkellä olen
484	13	12	vaalea tukka ja siniset silmät
485	12	12	aamulla syön tavallisesti
486	12	12	alkaa suomen kielen
487	12	11	ei ole hyvää
488	12	12	ei ole vain
489	12	10	ei ollut niin
490	12	12	en ole naimisissa
491	12	12	en pidä kahvista
492	12	12	että hänellä oli
493	12	12	että meillä on
494	12	12	hän ei voinut
495	12	10	hän on pitkä
496	12	12	hän sanoi että hän
497	12	10	ikkunan alla on
498	12	10	ja käy töissä
499	12	12	ja lähden kotoa
500	12	11	ja seinällä on
501	12	11	ja sen takia
502	12	12	ja vihreät silmät
503	12	10	ja vuonna [%]
504	12	12	juon teetä tai
505	12	12	kanssa ja sitten
506	12	12	katson televisiota tai
507	12	10	kello vartin yli
508	12	11	koska he ovat
509	12	10	kun he ovat
510	12	9	kun olin pieni
511	12	12	käyn suihkussa ja menen
512	12	12	luulen että hän
513	12	12	meillä on pieni
514	12	12	melkein joka päivä
515	12	12	menen kotiin ja syön
516	12	12	menen nukkumaan noin kello
517	12	12	menen nukkumaan puoli
518	12	12	menen nukkumaan yhdeltätoista
519	12	12	minulla on pieni perhe
520	12	11	minun asunto on
521	12	12	minun herätyskello soi
522	12	12	minun nimeni on [name]
523	12	12	minusta se ei ole
524	12	11	minä en halua
525	12	11	minä tiedän että
526	12	12	minä uskon että
527	12	12	mukava ja onnellinen perhe
528	12	12	mutta hänellä on
529	12	12	mutta nyt asun
530	12	8	nimi on [name]
531	12	11	näyttää siltä että
532	12	12	olen opiskellut suomea
533	12	12	on iso peili
534	12	12	on lamppu ja
535	12	12	on myös tärkeä
536	12	12	on oikealla sänky ja
537	12	12	on opettaja ja
538	12	12	on paljon työtä
539	12	12	on punainen matto
540	12	12	on tietokone ja
541	12	12	on vaalea tukka ja
542	12	12	on yksi huone
543	12	12	pidän erittäin paljon
544	12	12	puuroa ja voileipää
545	12	12	pääsin yliopistoon opiskelemaan
546	12	12	ruskea tukka ja
547	12	10	sanoi minulle että
548	12	12	se ei ole niin
549	12	8	se oli hyvin
550	12	12	se oli oikein
551	12	12	se on aika
552	12	12	se on mukava
553	12	12	sen jälkeen pukeudun
554	12	12	siellä on iso
555	12	11	sitten minulla on
556	12	12	syön kevyen lounaan
557	12	12	syön puuroa ja juon
558	12	12	syön voileipää ja juon
559	12	12	tällä hetkellä asun
560	12	12	verhot ja lattialla
561	12	12	vuonna [%] ja samana vuonna pääsin
562	12	11	vuonna [%] mutta
563	12	11	ystäväni tuli käymään
564	11	11	aamiaiseksi syön puuroa ja
565	11	11	asun tartossa ja
566	11	11	asunto on pieni
567	11	11	en ole varma
568	11	11	enemmän ja enemmän
569	11	10	että en ole
570	11	11	että hänen täytyy
571	11	11	että minun täytyy
572	11	9	että minä voin
573	11	10	huonessa on vasemalla
574	11	11	hän oli jo
575	11	11	hän on hyvä
576	11	11	hän on komea
577	11	11	hän on minun
578	11	11	hän sanoo että
579	11	8	hänellä ei ollut
580	11	9	hänellä on yksi
581	11	11	isän nimi on
582	11	11	isäni nimi on
583	11	11	isäni on [%]-vuotias
584	11	10	ja ei ole
585	11	9	ja että hän
586	11	11	ja he eivät
587	11	9	ja hän on [%]-vuotias
588	11	8	ja hän on töissä
589	11	9	ja hän sanoi
590	11	11	ja meillä on
591	11	11	ja minun täytyi
592	11	11	ja minun täytyy
593	11	11	ja otin aurinkoa
594	11	11	ja sen tähden

	f	jakauma	n-grammi
595	11	11	ja sokerin kanssa
596	11	11	ja toinen on
597	11	10	ja yksi tuoli
598	11	10	jonka nimi oli
599	11	10	jos minulla olisi
600	11	10	kaksi veljeä ja
601	11	10	keittiö, vessa ja kylpyhuone
602	11	11	kello kaksitoista syön
603	11	11	kirjoituspöytä ja kaksi
604	11	11	koska en ole
605	11	11	koska hänellä on
606	11	11	koska meillä on
607	11	11	koska se ei
608	11	10	kun hän on
609	11	11	kun hän tuli
610	11	11	käyn suihkussa ja pesen hampaani
611	11	11	maanantaisin ja torstaisin
612	11	10	marjassa ja sienessä
613	11	11	menen kauppaan ja ostan
614	11	10	menen keittiöön ja
615	11	11	menen kouluun jalan
616	11	5	miksi sinulla on niin isot
617	11	10	minulla ei ole vielä
618	11	11	minulla on kotona
619	11	11	minulla on melko
620	11	11	minulla on monta
621	11	11	minun huoneeni on
622	11	11	minun huoneessani on
623	11	11	minä olen [name]
624	11	11	minä olen opiskelija
625	11	11	mutta meillä on
626	11	11	mutta minä olen
627	11	11	neljä alkaa suomen
628	11	7	noin [%] astetta
629	11	10	noin kello kymmenen
630	11	11	oikein mukava ja
631	11	11	olen kotoisin tallinnasta mutta
632	11	11	olen kotoisin virosta
633	11	11	olen ollut töissä
634	11	11	olen väsynyt ja
635	11	11	on aika pieni
636	11	11	on kirjahylly ja siinä on
637	11	11	on kylmä ja
638	11	11	on liian paljon
639	11	9	on liian pieni
640	11	11	on melko iso
641	11	5	on niin isot
642	11	11	on pieni kaupunki
643	11	11	on pitkä ja
644	11	11	on toisessa kerroksessa
645	11	11	on tosi kiva
646	11	11	on tosi tärkeä
647	11	9	on tumma tukka
648	11	10	on yksi iso
649	11	11	palaan kotiin noin
650	11	11	pidän oikein paljon
651	11	11	se ei ole totta
652	11	9	se oli vaikea
653	11	11	se on iso
654	11	10	se on melko
655	11	11	sen jälkeen käyn
656	11	11	sen jälkeen laitan aamiaisen
657	11	11	sen jälkeen me
658	11	11	sen jälkeen olin
659	11	11	sen jälkeen syön
660	11	11	siinä on monta
661	11	9	siskoni ja minä
662	11	10	sitten minä menen
663	11	11	syntyi vuonna [%]
664	11	10	tiedän että se
665	11	11	tiistaisin ja torstaisin
666	11	10	toivon että se
667	11	11	vanhempani asuvat tallinnassa
668	11	11	vartin yli neljä
669	11	11	varttia vaille kuusi
670	11	11	varttia vaille kymmenen
671	11	11	vieressä on pieni
672	10	10	aamiaiseksi minä syön
673	10	10	asunnossa on kaksi huonetta
674	10	10	ei enää ole
675	10	10	ei ole televisiota
676	10	10	ei ole vaikea
677	10	10	en ole liian
678	10	10	en tiedä paljon
679	10	10	en voi sanoa
680	10	10	että he voivat
681	10	7	että hän haluaa
682	10	9	että siellä on
683	10	7	että Suomessa on
684	10	10	huoneessa on kaksi
685	10	10	huonetta ja keittiö
686	10	8	hän kertoi minulle
687	10	9	hän käy töissä
688	10	9	hän on aina
689	10	10	hän on erittäin
690	10	10	hän on naimisissa
691	10	9	hän on ollut
692	10	10	ikkunan vieressä on kirjoituspöytä
693	10	10	ikkunassa on vihreät verhot
694	10	10	ja en halua
695	10	9	ja juomme kahvia
696	10	10	ja monta kirjaa
697	10	10	joka ei ole
698	10	10	kanssa ja juon
699	10	10	keittiö, vessa, kylpyhuone
700	10	6	kiinan kieli on
701	10	10	kolme tai neljä
702	10	9	koska hän haluaa
703	10	10	koska ne ovat
704	10	10	kotoa puoli kahdeksalta
705	10	10	kotona laitan ruokaa
706	10	9	lattialla on matto
707	10	7	luonteeltaan hän on
708	10	7	luulen että on
709	10	10	lähellä on vaatekaappi
710	10	10	meillä on kolme huonetta
711	10	10	meillä on omakotitalo
712	10	10	meillä on paljon
713	10	9	meillä on vielä
714	10	10	minulla ei ole aikaa
715	10	10	minulla on kaksi siskoa
716	10	10	minulla on myös kaksi
717	10	10	minulla on myös yksi

	f	jakauma	n-grammi
718	10	10	minulla on nyt
719	10	10	minulla on poikaystävä
720	10	10	minulla on yksi sisko ja yksi
721	10	10	minun mielestäni se
722	10	10	minun äiti on
723	10	9	minä menin kotiin
724	10	10	minä olen [%]-vuotias
725	10	10	minä olen kotoisin
726	10	10	minä olen syntynyt
727	10	9	minä olen väsynyt
728	10	7	minä toivon että
729	10	10	mutta ei se
730	10	10	mutta he ovat
731	10	10	mutta minun täytyy
732	10	10	niin kauan kuin
733	10	9	noin [%] vuotta
734	10	5	nousin kello [%]
735	10	9	nyt minulla on
736	10	10	nyt se on
737	10	9	oikealla ikkunan vieressä
738	10	9	oikealla on iso
739	10	10	olen [%] vuotta vanha
740	10	10	olen iloinen että
741	10	10	olen käynyt peruskoulun
742	10	10	on kaksi veljeä
743	10	10	on lamppu, tietokone
744	10	10	on oikein hyvä
745	10	10	on paljon kirjoja
746	10	10	on paljon töitä
747	10	8	on parempi kuin
748	10	10	on sama kuin
749	10	10	on yksi sisko ja yksi veli
750	10	10	otin aurinkoa ja
751	10	10	perheeni ei ole
752	10	10	peseytyn ja pukeutun
753	10	10	puhun hyvin englantia
754	10	10	ruokaa ja syön
755	10	10	samana vuonna pääsin yliopistoon opiskelemaan
756	10	10	se on hauska
757	10	10	se on tosi hyvä
758	10	10	se on vain
759	10	10	se on vähän
760	10	10	seinällä on iso
761	10	10	seinällä on kirjahylly ja
762	10	10	sellainen on minun
763	10	10	sen jälkeen he
764	10	9	sen jälkeen on
765	10	10	siinä on paljon
766	10	9	siksi hän on
767	10	10	siksi se on
768	10	10	sitten menen kotiin ja
769	10	10	syön aamiaisen ja
770	10	9	talossa on kaksi kerrosta
771	10	9	toivon että kaikki
772	10	10	valitettavasti minulla ei
773	10	10	vuodesta [%] vuoteen [%]
774	10	10	yli neljä alkaa
775	10	10	ystävällisin terveisin [name]
776	10	10	äitini nimi on
777	10	10	äitini, isäni ja
778	9	9	[%] astetta lämmintä
779	9	9	aina hyvällä tuulella
780	9	7	anssi ja jutta
781	9	9	asunnossa on kolme huonetta
782	9	9	asuvat minun vanhempani
783	9	9	ei ole koskaan
784	9	9	ei ole tarpeeksi
785	9	9	ei ollut paljon
786	9	9	ei tarvitse mennä
787	9	8	ei voi tehdä
788	9	9	en ole koskaan ollut
789	9	8	en tiedä jos
790	9	9	että ei kukaan
791	9	9	että he eivät
792	9	9	että hän ei ole
793	9	9	että hän tulee
794	9	9	että minulla ei ole
795	9	8	että sinä olet
796	9	9	harmaat silmät ja
797	9	9	huone on melko
798	9	9	huoneessa on myös
799	9	8	huoneessa on oikealla sänky
800	9	9	huoneessa on vasemmalla sänky
801	9	9	hän ei pidä
802	9	9	hän ei tarvitse
803	9	9	hän luulee että
804	9	9	hän oli hyvin
805	9	9	hän oli myös
806	9	9	hän on [%]-vuotias ja
807	9	9	hän on jo
808	9	9	hän on rauhallinen
809	9	9	ja asun tartossa
810	9	8	ja hän käy
811	9	7	ja hän on myös
812	9	9	ja kaikki on
813	9	9	ja luen kirjaa
814	9	8	ja me emme
815	9	9	ja menen keittiöön
816	9	8	ja minusta se
817	9	9	ja ruskeat silmät
818	9	9	ja se ei ole
819	9	9	ja sen vieressä
820	9	6	ja sitten menin
821	9	9	ja sitten syön
822	9	9	ja syön aamiaista
823	9	8	ja yksi sisko
824	9	9	joka on myös
825	9	9	jos en ole
826	9	9	juon kahvia ja syön
827	9	9	kaksi voileipää ja
828	9	9	kaunis ja mukava
829	9	9	keittiö ja kylpyhuone
830	9	9	kello kymmenen menen
831	9	8	kertoi minulle että
832	9	9	kesäkuussa oli tenttikausi
833	9	9	kirjoitin ylioppilaaksi vuonna [%] ja
834	9	9	kirjoituspöytä ja yksi
835	9	9	kolme huonetta ja
836	9	9	kolme huonetta, keittiö
837	9	8	koska me emme
838	9	9	koska minä olen
839	9	9	koska minä pidän

	f	jakauma	n-grammi
840	9	9	kuoli vuonna [%]
841	9	9	käyn suihkussa ja menen nukkumaan
842	9	9	laitan aamiaisen ja keitän teetä
843	9	9	lamppu, tietokone ja
844	9	9	lattialla on punainen matto
845	9	9	lattialla on ruskea
846	9	9	lattialla on vihreä matto
847	9	8	loppuu kello [%]
848	9	9	luennossa puhuttiin paljon
849	9	9	meillä on kotona vielä
850	9	9	meillä on yksi
851	9	9	menen kotiin ja laitan
852	9	7	minulla on aina
853	9	9	minulla on oma huone
854	9	9	minulla on suomen
855	9	9	minulla on suuri
856	9	9	minulle kuuluu hyvää
857	9	8	minun huonessa on
858	9	7	minun koti on
859	9	9	minun nimi on
860	9	8	minun perheni on
861	9	9	minusta hän on
862	9	8	minä en osaa
863	9	9	minä herään kello
864	9	9	minä nousen kello
865	9	9	mitä tein viime
866	9	9	monta kirjaa ja
867	9	9	mutta en voi
868	9	8	mutta hän ei ole
869	9	9	mutta minusta se
870	9	9	mutta nousta pitää
871	9	9	mutta nyt olen
872	9	9	noin [%] minuuttia
873	9	9	noin kello yksitoista
874	9	9	noin kymmenen minuuttia
875	9	9	nousta mutta nousta
876	9	9	ole ollut vaihto-oppilaana
877	9	9	oli [%] vuotta vanha
878	9	9	oli kaunis ja
879	9	8	oli oikein kiva
880	9	8	oli vaikea ymmärtää
881	9	8	on hyväntuulinen ja
882	9	9	on iso vaatekaappi
883	9	7	on kaksi huonetta, keittiö
884	9	9	on keltaiset verhot
885	9	9	on mielenkiintoinen ja
886	9	9	on myös minun
887	9	9	on nojatuoli ja
888	9	9	on paljon aikaa
889	9	9	on ruskea matto
890	9	9	on siniset verhot
891	9	9	on suomen kielen
892	9	9	on suuri ja
893	9	9	on tarpeeksi suuri
894	9	9	on tilava ja
895	9	9	on totta että
896	9	9	on vanha ja
897	9	9	opiskelija tarton yliopistossa
898	9	9	puuroa ja juon teetä
899	9	9	puuroa tai voileipää
900	9	9	se oli kiva
901	9	9	se on hyvää
902	9	9	se on hänen
903	9	9	se on kiva
904	9	9	se on paras
905	9	8	se on sama
906	9	9	se on suuri
907	9	9	sen jälkeen menimme
908	9	9	sen täytyy olla
909	9	9	sijaitsee aika lähellä
910	9	8	Suomessa on paljon
911	9	9	syön ja juon
912	9	9	syön puuroa tai
913	9	9	sängyn alla on
914	9	9	sänky, kirjoituspöytä ja
915	9	9	tai luen kirjaa
916	9	9	tai voileipää ja
917	9	9	tartossa koska opiskelen
918	9	9	tavallisesti puuroa tai
919	9	6	toivon että minä
920	9	9	tukka ja ruskeat
921	9	9	tukka ja vihreät silmät
922	9	9	uin, otin aurinkoa
923	9	9	vaatteet päälle ja
924	9	8	veljen nimi on
925	9	9	verhot ja lattialla on
926	9	9	vessa ja keittiö
927	9	9	äitini on [%]-vuotias
928	8	6	hän on [%] vuotta vanha ja
929	8	8	ikkunan vieressä on iso kirjoituspöytä ja
930	8	8	ja minulla ei ole lapsia
931	8	8	minulla on äiti, isä ja
932	8	8	olen käynyt peruskoulun ja lukion
933	7	7	[%] ja samana vuonna pääsin yliopistoon
934	7	7	aamiaiseksi syön puuroa ja juon
935	7	7	aamiaiseksi syön tavallisesti puuroa ja
936	7	7	huoneessa on vasemmalla sänky ja
937	7	5	ja siinä on monta kirjaa
938	7	7	kirjahylly ja siinä on monta kirjaa
939	7	7	kirjoitin vuonna [%] ja samana vuonna
940	7	7	minulla ei ole niin paljon
941	7	7	oikein mukava ja onnellinen perhe
942	7	6	on [%] vuotta vanha ja hän
943	7	7	on kirjahylly ja siinä on monta
944	7	7	on vaalea tukka ja siniset silmät
945	6	6	[%] vuotta vanha ja hän on
946	6	6	aamiaiseksi syön tavallisesti voileipää ja juon
947	6	6	en halua nousta mutta nousta pitää
948	6	6	huonessa on oikealla sänky ja

	f	jakauma	n-grammi
949	6	6	hän on töissä isossa firmassa
950	6	6	ikkunan vieressä on kirjoituspöytä ja
951	6	6	ja samana vuonna pääsin yliopistoon opiskelemaan
952	6	6	kahvia kerman ja sokerin kanssa
953	6	6	luento alkaa vartin yli kymmenen
954	6	6	luento alkaa viisitoista minuuttia yli
955	6	6	matka kestää noin kymmenen minuuttia
956	6	6	matka kestää noin viisitoista minuuttia
957	6	6	olen lapsuudesta asti harrastanut urheilua
958	6	6	on lyhyt vaalea tukka ja
959	6	6	on vasemalla sänky ja oikealla
960	6	6	pöydällä on lamppu, tietokone ja
961	6	6	seinällä on kirjahylly ja siinä on
962	6	6	sen jälkeen menen kotiin ja
963	6	6	sitten käyn suihkussa ja pukeudun
964	6	6	syön puuroa ja juon kahvia
965	6	6	ylioppilaaksi kirjoitin vuonna [%] ja samana
966	5	5	asunto on pieni mutta viihtyisä
967	5	5	ei ole radiota eikä televisiota
968	5	5	huonetta, keittiö, kylpyhuone ja vessa
969	5	5	hänellä on lyhyt tumma tukka ja
970	5	5	käyn suihkussa ja sen jälkeen
971	5	5	menen jalan koska yliopisto sijaitsee
972	5	5	miksi sinulla on niin isot korvat
973	5	5	minulla on äiti, isä ja kaksi
974	5	5	naimisissa ja minulla ei ole lapsia
975	5	5	neljä alkaa suomen kielen kurssi
976	5	5	olin paljon luonnossa, uin, otin aurinkoa
977	5	5	on kolme huonetta ja keittiö
978	5	5	on lamppu, tietokone ja monta
979	5	5	on oikealla sänky ja vasemmalla
980	5	5	on pieni mutta siellä on
981	5	5	puihin tulee vähitellen vihreitä lehtiä
982	5	5	puuroa tai voileipää ja juon
983	5	5	sen jälkeen laitan aamiaisen ja keitän
984	5	5	syön tavallisesti puuroa ja juon
985	5	5	syön tavallisesti puuroa tai voileipää
986	5	5	tai voileipää ja juon teetä
987	5	5	tartossa koska opiskelen tarton yliopistossa
988	5	5	tumma tukka ja harmaat silmät
989	5	5	tumma tukka ja ruskeat silmät
990	5	5	viihtyisä ja tarpeeksi suuri minulle
991	5	5	voin käyttää hyväksi kurssin tietoja
992	5	5	ylioppilaaksi vuonna [%] ja samana vuonna
	16614		