Ulkomaisten yliopistojen suomenoppijoiden leksikaalisen diversiteetin kehittyminen intensiivikurssilla Suomessa

Karoliina Kuusinen

Pro gradu -tutkielma

Suomen kieli ja suomalais-ugrilainen kielentutkimus, suomen kieli

Kieli- ja käännöstieteiden laitos

Humanistinen tiedekunta

Turun yliopisto

Toukokuu 2021

TURUN YLIOPISTO
Kieli- ja käännöstieteiden laitos / Humanistinen tiedekunta
KUUSINEN, KAROLIINA: Ulkomaisten yliopistojen suomenoppijoiden leksikaalisen diversiteetin kehittyminen intensiivikurssilla Suomessa

Pro gradu -tutkielma, 60 s., 4 liites.
Kielen oppimisen ja opettamisen tutkinto-ohjelma, suomen kieli
Toukokuu 2021

– –

Tässä pro gradu -tutkielmassa tarkastellaan leksikaalisen diversiteetin eli sanastollisen monimuotoisuuden kehittymistä kahdentoista suomea vieraana kielenä opiskelevan oppijan kirjallisissa ja suullisissa kertomuksissa. Tutkimuksen tavoitteena on selvittää, miten näiden suomenoppijoiden leksikaalinen diversiteetti kehittyy neljän viikon intensiivikurssin aikana, kun he opiskelevat kokoaikaisesti Suomessa eli kohdekielisessä ympäristössä. Tutkimuksessa tarkastellaan sekä kirjallisia että suullisia kertomuksia erikseen, jotta pystytään selvittämään, millaisia muutoksia leksikaalisen diversiteetin kehityksessä on nähtävissä eri tuottotapojen kertomuksissa. Lisäksi tavoitteena on selvittää, onko kirjallisten ja suullisten kertomusten leksikaalisen diversiteetin kehittymisessä havaittavissa eroavaisuuksia tuottotapojen välillä.

Tutkimuksen informantteina olevat suomenoppijat, ovat osallistuneet Kansainvälisen liikkuvuuden yhteistyön keskuksen CIMO:n (nyk. Opetushallitus) järjestämälle Suomen kielen ja kulttuurin kesäkurssille. Tutkimusaineistona käytetään oppijoiden kurssin alussa ja lopussa tuottamia kirjallisia ja suullisia kertomuksia, joissa he ovat kertoneet kuvasarjojen avulla erilaisista vapaa-ajan tilanteista.

Tutkimuksessa on käytetty pääasiassa kvantitatiivisia menetelmiä. Monimuuttuja-analyysia hyödyntämällä on pyritty selvittämään, miten intensiivikurssilla opiskelu on vaikuttanut leksikaalisen diversiteetin osa-alueiden – vaihtelevuuden, erityisyyden, tasaisuuden, harvinaisuuden ja sironnan – kehittymiseen, kun vertaillaan alussa ja lopussa tuotettuja kertomuksia. Tarkastelemalla kunkin kertomuksen saamia arvoja eri osa-alueiden osalta, on pyritty myös selvittämään, millaisia eroja ja yhtäläisyyksiä kirjallisten ja suullisten kertomusten leksikaalisen diversiteetin kehittymisessä on.

Tutkimustulosten perusteella voidaan päätellä, että leksikaalinen diversiteetti kehittyy oppijoilla intensiivikurssin aikana sekä kirjallisissa että suullisissa kertomuksissa. Tulosten perusteella voidaan todeta, että kehitys on nähtävissä yhdenmukaisena suurella osalla oppijoista, mutta jonkin verran yksilökohtaista vaihtelua on. Mielenkiintoisena tuloksena voidaan pitää sitä, että tiettyjen osa-alueiden – tasaisuuden, sironnan ja erityisyyden – osalta leksikaalinen diversiteetti kehittyy oppijoiden suullisissa kertomuksissa enemmän kuin kirjallisissa kertomuksissa. Toinen mielenkiintoinen tulos on se, että oppijat käyttävät alussa harvinaisempia sanoja kuin lopussa, vaikka oletettavaa olisi, että kurssin aikana opitaan uutta ja harvinaisempaa sanastoa.

Kokonaisuudessaan voidaan todeta, että intensiivikurssin aktiivisella opiskelulla ja kohdekielisessä oppimisympäristössä oppimisella on positiivisia vaikutuksia tarkastelun kohteena olleiden oppijoiden leksikaaliseen diversiteetin ja sen myötä myös kielitaidon kompleksistumiseen. Kurssin aikana oppijat ovat oppineet hyödyntämään sanoja monipuolisemmin ja pystyvät rakentamaan kertomuksiaan jäsennellymmin kurssin loppuvaiheessa. Jatkotutkimusta ajatellen olisi kiinnostavaa tutkia kaikkien intensiivikurssin oppijoiden kertomuksia ja sitä kautta selvittää laajemminkin vieraan kielen oppijoiden leksikaalisen diversiteetin kehittymisestä lyhyen aikavälin sisällä.

Asiasanat: leksikaalinen diversiteetti, sanastollinen monimuotoisuus, suomenoppija, sanasto

Sisällys

1. Johdanto

Pro gradu -tutkielmassani tarkastelen leksikaalista diversiteettiä eli sanastollista monimuotoisuutta yhden suomenoppijaryhmän kirjoitetuissa ja puhutuissa kertomuksissa. Tutkimuksessa keskityn siihen, miten oppijoiden leksikaalinen diversiteetti kehittyy neljän viikon intensiivisen kesäkurssin aikana ja pohdin, onko kirjoitettujen ja suullisten kertomusten sanastollisessa monimuotoistumisessa havaittavissa yhdensuuntaista kehitystä vai kehittyykö leksikaalinen diversiteetti näissä eri tavoin.

Tutkimuskysymykseni ovat seuraavat:

1. Miten oppijoiden leksikaalinen diversiteetti kehittyy kirjallisissa ja suullisissa kertomuksissa intensiivikurssin aikana, kun tarkastellaan kurssin alussa ja lopussa tuotettuja kertomuksia?
2. Millaisia yhteyksiä tai eroja kirjoitettujen ja suullisten kertomusten leksikaalisessa diversiteetissä on nähtävissä?

Tutkimusaineistonani käytän kertomuksia, jotka on kerätty Kansainvälisen liikkuvuuden yhteistyön keskuksen CIMO:n (nyk. Opetushallitus) järjestämällä Suomen kielen ja kulttuurin kesäkurssilla. Kurssille on osallistunut opiskelijoita useista yliopistoista ympäri maailman, ja heidät on jaettu kahdelle kurssille kielitaidon taitotason mukaan. Aineisto koostuu alku- ja lopputesteistä, joissa opiskelijat ovat tuottaneet kirjallisia ja suullisia kertomuksia kuvasarjojen avulla. Sekä kirjallisiin että suullisiin kertomuksiin on ollut omat kuvasarjansa eli opiskelijat ovat kertoneet kirjallisesti eri kuvasarjoista kuin suullisesti. Molemmissa kertomistavoissa lopputestissä käytettiin samaa kuvasarjaa kuin alkutestissä. Kokonaisuudessaan alku- ja lopputestauksiin on osallistunut 67 opiskelijaa, mutta tässä tutkimuksessa tarkastelen vain yhden ryhmän eli kahdentoista oppijan sanaston kehittymistä intensiivikurssilla. Yhden kokonaisen ryhmän valinta perustuu siihen, että tulosten vertailtavuus samalla taitotasolla olevien opiskelijoiden kertomusten välillä on mahdollista ja opiskelijat ovat ryhmänä osallistuneet samoille oppitunneille sekä kurssin aktiviteetteihin, joten opetus ja oppimisympäristöt ovat olleet kaikilla samat.

Koska intensiivikurssilla on tarkoitus oppia mahdollisimman paljon ja monipuolisesti kielen eri osa-alueita, ensimmäinen hypoteesini on, että opiskelijoiden leksikaalinen diversiteetti kehittyy yhtä lailla muun kielitaidon kanssa. Intensiivikurssilla keskitytään kielen rakenteisiin, ymmärtämiseen ja tuottamiseen sekä suomalaiseen kulttuuriin, jolloin opiskelijoille tarjoutuu tilaisuus tutustua myös uuteen sanastoon. Kesällä 2018 pääsin opetusharjoittelijana osallistumaan intensiivikurssin

toimintaan ja näin, miten erilaisia asioita opetettiin ja opittiin. Esimerkiksi vapaa-ajan ohjelma, kuten pesäpallo ja saunominen, sidottiin vahvasti oppitunteihin, joilla käytiin läpi historiaa, kulttuuria ja sanastoa aiheisiin liittyen. Vapaa-ajan ohjelmissa oppitunneilla kerrottua tietoa ja sanastoa päästiin hyödyntämään konkreettisissa tilanteissa. Tästä syystä oletankin, että oppijan sanasto kehittyy ja monimuotoistuu, kun oppimisympäristö ja opittavat asiat tukevat toisiaan.

Aiemmissa tutkimuksissa on osoitettu, että usein kirjoitettujen tekstien leksikaalinen diversiteetti kehittyy enemmän verrattuna puhuttuihin teksteihin (ks. esim. Honko 2017). On kuitenkin todettu, että tietyissä yhteyksissä leksikaalinen diversiteetti voi näyttäytyä hyvinkin samalla tasolla riippumatta siitä, onko tekstit tuotettu kirjoittaen vai suullisesti (ks. esim. Yu 2009, 251–253). Toisena hypoteesinani esitän, että opiskelijoiden leksikaalisen diversiteetin kehittyminen on nähtävissä paremmin suullisissa kertomuksissa kirjallisten kertomusten sijaan. Hypoteesini perustuu siihen, että osallistuessaan intensiivikurssille opiskelijat pääsevät osallistumaan vuorovaikutukseen suomenkielisessä ympäristössä ja käyttämään suomea aktiivisesti päivittäin, minkä voi ajatella vahvistavan erityisesti suullista kielitaitoa. Suomen ulkopuolella opiskelevilla opiskelijoilla suullinen kielen käyttö usein rajautuu pitkälti vain opiskeluympäristöön, koska mahdollisuuksia puhua opiskeltavalla kielellä ei muualla useinkaan ole, kun taas kirjallista tuottamista pystytään vahvistamaan paikasta riippumatta. Koska intensiivikurssilla keskitytään hyvin vahvasti suulliseen tuottamiseen sekä oppitunneilla että niiden ulkopuolella, voidaan olettaa, että suullinen taito vahvistuu entisestään, kun oppijoilla on mahdollisuus käyttää kieltä ja sen sanastoa monipuolisesti erilaisissa tilanteissa.

Leksikaalista diversiteettiä on tutkittu paljon, erityisesti englantia joko toisena tai vieraana kielenä opiskelevien opiskelijoiden kohdalla (ks. esim. Johansson 2008; Yu 2009; Jarvis 2013a, 2013b, 2017). Myös suomenoppijoiden leksikaalista diversiteettiä ja sen arviointia on tutkittu, erityisesti kirjoitetusta kielestä (ks. esim. Honko 2013), mutta tutkimusta on vähemmän kuin englannin kielestä. Tästä syystä onkin tärkeää tarkastella suomenoppijoiden leksikaalisen diversiteetin kehittymistä sekä puhutussa että kirjoitetussa kielessä. Näin voidaan saada lisätietoa siitä, miten erityisesti ulkomaisten yliopistojen opiskelijat voivat laajentaa suomen kielen taitojaan kohdemaassa järjestettävillä intensiivikursseilla. Lisäksi on oleellista tutkia sekä kirjoitettuna että puhuttuna tuotettuja tekstejä, sillä vertaamalla näitä keskenään, voidaan myös pohtia erilaisten tilanteiden vaikutuksia kielen tuottamiseen ja siihen, millaisia valintoja kielenoppija tekee tuottaessaan kieltä.

Tutkimuksessani tarkastelen oppijoiden leksikaalisen diversiteetin kehittymistä hyödyntäen pääasiassa kvantitatiivisia menetelmiä. Näkökulmana on se, että leksikaaliseen diversiteettiin sisältyy monenlaisia asioita eikä se koostu vain yhdestä asiasta. Tavoitteena onkin selvittää, millaisia muutoksia leksikaalisessa diversiteetissä tapahtuu, kun otetaan huomioon useampia osa-alueita, jotka

kertovat sanastollisesta monimuotoisuudesta. Hyödynnän tässä apuna monimuuttuja-analyysia ja pohdin sen avulla leksikaalisen diversiteetin kehittymistä kurssin aikana.

Tutkimukseni alussa, luvussa 2, esittelen tutkimukseni kannalta tärkeimpiä käsitteitä ja tarkastelen kielitaidon osa-alueita kiinnittäen erityistä huomiota sanaston oppimiseen ja sanojen osaamiseen sekä käyttöön. Luvussa 3 pohdin puhutun ja kirjoitetun kielen suhdetta toisiinsa sekä sitä, vaikuttaako kielen tuottamisen tapa mahdollisesti sanastoon ja sen käyttöön. Luvussa 4 selvitän, miten leksikaalinen diversiteetti nähdään nykytutkimuksessa, miten sitä on aiemmin tutkittu ja arvioitu sekä selvennän, millaisia leksikaalisen diversiteetin mittareita tutkimuksissa on aiemmin käytetty. Luvussa 5 esittelen tarkemmin käyttämäni aineiston ja analyysimenetelmät. Luvuissa 6–7 käyn läpi kirjallisia ja suullisia kertomuksia sekä perehdyn siihen, millaisia yhtäläisyyksiä ja eroja kertomuksista löytyy. Lopuksi luvussa 8 teen yhteenvedon tuloksista ja pohdin mahdollisuuksia jatkotutkimukselle.

2. Kielitaidon kehittyminen L2-oppijoilla

2.1. Keskeisiä käsitteitä

Kielentutkimuksessa ja kielen oppimisen tutkimuksessa käytetään paljon käsitteitä, joiden merkitykset saattavat vaihdella tutkimuskohtaisesti. Monelle käsitteelle löytyykin useita, hieman toisistaan poikkeavia määritelmiä, ja siksi on tärkeä selventää ajatuksia kunkin käsitteen taustalla. Tästä syystä määrittelen tässä luvussa omassa tutkimuksessani käyttämäni käsitteet.

Tutkimuksessani tarkastelen suomen kielen oppimista ja kielitaidon kehittymistä intensiivikurssin aikana ja käytän informanteista nimitystä **suomenoppija** tai **oppija**. Suomenoppija on henkilö, joka oppii ja opiskelee suomea missä tahansa oppimisympäristössä joko ulkomailla tai Suomessa. Intensiivikurssille osallistuneet suomenoppijat ovat opiskelleet pääsääntöisesti muualla kuin Suomessa, minkä vuoksi heidän voidaan ajatella perinteisen jaottelun mukaisesti opiskelleen suomea **vieraana kielenä**. Osallistuessaan intensiivikurssille opiskelija jatkaa suomen kielen opiskelua vieraana kielenä, vaikka opintojakso suoritetaankin Suomessa. Käsitettä vieraana kielenä opiskelu käytetään usein rinnakkain **toisen kielen** opiskelun kanssa. Toisen kielen opiskelu ja oppiminen tarkoittaa pääasiassa sitä, että opiskeltavaa kieltä opiskellaan ja omaksutaan joko varsinaisissa opetus- tai kommunikaatiotilanteissa maassa, jossa opiskeltavaa kieltä puhutaan valtakielenä. (Sajavaara 1999, 75; Martin 2003, 75–78; Pietilä–Lintunen 2014, 12.) Kielentutkimuksessa kielenoppimiseen ja opeteltavaan kieleen viitataan myös termillä **L2** (*target language*), jolla tarkoitetaan **kohdekieliä** eli niitä kieliä, joita oppija opiskelee (Sajavaara 1999, 75). Omassa tutkimuksessani hyödynnän käsitteinä sekä kohdekieltä että suomen kieltä puhuessani suomenoppijoiden opiskeltavasta kielestä.

Tutkimukseni kannalta oleellisia käsitteitä ovat **sana**, **lekseemi** ja **sane**. Ongelmallisin määriteltävä näistä on sana*,* sillä se voidaan määritellä monella tavoin, esimerkiksi ortografisesti, morfologisesti, fonologisesti, syntaktisesti tai semanttisesti (Singleton 1999, 10–14; Koivisto 2013, 33–35). Sanojen avulla voimme kuvata ympäröivää maailmaa ja ilmaista tarpeellisia asioita. Sana onkin sinällään sekä arkikäsite että kielitieteellinen käsite. (Koivisto 2013, 25–26.) Omassa tutkimuksessani sanan määritelmään liittyy sekä ortografinen että osiltaan morfologinen ja fonologinen näkökulma. Ortografisesti määriteltynä sana on kirjoitetussa tekstissä esiintyvä yhtenäinen tekstinosa, jota edeltää ja jonka jäljessä on sanaväli, kappaleen alku tai välimerkki (Koivisto 2013, 28, 35). Koska opiskelijoiden kertomukset ovat kirjallisessa muodossa, niin niistä sanat ovat helposti tunnistettavissa tämän määritelmän mukaisesti omiksi yksiköikseen, sillä kutakin sanaa edeltää ja sen jälkeen on sanaväli, kappaleen alku tai välimerkki. Koska mukana on myös suullisia kertomuksia, sanan

määritelmässä täytyy ottaa huomioon morfologinen ja fonologinen määritelmä. Morfologisesti sana voidaan nähdä morfologisena yksikkönä, joka koostuu sanavartalosta eli yhdestä leksikaalisesta morfeemista tai leksikaalisesta morfeemista ja siihen liittyvistä sidonnaisista morfeemeista. Toisaalta täytyy ottaa huomioon myös fonologinen määritelmä eli se, että sana erottuu puhutusta kielestä sekä fonologisten (eli fonotaksia noudattavien) että prosodisten piirteiden, kuten painon, intonaation ja kvantiteetin avulla. Se voi olla yksittäinen foneemijono *kissa* tai niin sanottu syntagmaattinen sulauma, jossa kaksi sanaa on sulautunut toisiinsa, kuten *onkse ”onko se”, ootsä ”oletko sinä”, ettei ”että ei”*. (Koivisto 2013, 28–34.) Kun suulliset aineistot litteroitiin kirjalliseen muotoon (ks. luku 5.1.), niin siinä pyrittiin ottamaan huomioon se, mitä puhuja sanoo ja miten. Näin on pyritty erottelemaan toisistaan kaikki sanat, joita kertomuksissa on tuotettu. Sanoista on huomioitu sekä niiden morfologinen rakenne että fonologinen tuottotapa. Litteraatit ovat kirjallisessa muodossa, joten niistä sanoja voidaan tarkastella myös ortografisina yksiköinä.

Hyvin tyypillisesti sanan käsitteestä erotetaan lekseemi ja siihen läheisesti liittyvänä käsite **sananmuoto**. Lekseemi (*type*) voidaan yksinkertaistetusti määritellä kielen sanaston sanaksi, jolla on oma leksikaalinen merkityksensä. Lekseemillä on siis tietty merkitys sekä fonologinen ulkoasu eli sanavartalo, joka koostuu foneemeista ja yhdestä tai useammasta morfeemista. (Koivisto 2013, 33.) Lekseemi on sanavaraston abstrakti yksikkö (KS 2020, s.v. *lekseemi*), joka näyttäytyy sen sanamuotojen avulla (Koivisto 2013, 33). Suomessa lekseemeistä puhutaan usein käyttäen sanan perusmuotoa: nomineista yksikön nominatiivia ja verbeistä A-infinitiivin perusmuotoa. Lekseemi voi edustua kuitenkin sekä yhtenä (*kissa*) että useampana hieman vaihtelevana vartalovarianttina (*pöytä-*, *pöydä-*). Vaihtelusta huolimatta saman lekseemin eri vartalovariantit lasketaan kuitenkin yhdeksi. (Singleton 1999, 10; Nation 2001, 7–8; Koivisto 2013, 35.) Lekseemin toteumia kutsutaan sananmuodoiksi ja ne voidaan katsoa kieliopillisiksi sanoiksi. Nämä sananmuodot edustavat siis erilaisia kieliopillisia muotoja sekä taivutusmuotoja. Esimerkkinä lekseemi KISSA voi edustua sananmuotoina *kissa+n*, *kisso+i+na,* jne., mutta vaikka niiden sanavartalo vaihtelee, ne edustavat kuitenkin laskennallisesti yhtä ja samaa lekseemiä. (Koivisto 2013, 40–45.) Lekseemin lisäksi kielitieteellisessä tutkimuksessa puhutaan usein lemmoista. **Lemmalla** (*lemma*) tarkoitetaan sanan perus- tai hakumuotoa ja siihen liittyviä taivutusmuotoja. Lemman ja lekseemin käsitteet ovatkin hyvin lähellä toisiaan, mutta lemman kohdalla on oleellista, että eri sanaluokkien sanat edustavat omia lemmojaan, jolloin sanojen merkitykset voidaan erotella toisistaan. (Nation 2001, 7–8; Honko 2013, 56–57.) Omassa tutkimuksessani ero lekseemin ja lemman välillä ei ole relevantti, sillä huomio kohdistuu vain varsinaisiin sanaesiintymiin, ei sanojen sisältämiin merkityksiin. Käytän siis tutkimuksessani selkeyden vuoksi käsitettä lekseemi.

Sanan ja lekseemin lisäksi voidaan puhua saneista (*token*) eli puheessa tai kirjoituksessa esiintyvistä konkreettisista tekstisanoista, jotka edustavat lekseemejä (Koivisto 2013, 33). Kielitoimiston sanakirja määrittelee perusmuotoisen tai taivutetun saneen (KS 2020 s.v. *sane*) sanaksi puheen tai tekstin osana. Sane katsotaan myös laskennalliseksi yksiköksi, jos halutaan laskea tekstistä kaikkien sen sisältämien sanojen määrä. Vaikka sama sana siis esiintyisi tekstissä useita kertoja, se katsotaan aina omaksi saneekseen. (Read 2005, 18.) Lekseemien kohdalla puolestaan huomioidaan lekseemin toteutumat eli sananmuodot ja niiden sisältämät sanavartalot, jotka edustavat vain yhtä lekseemiä kerrallaan, eikä niitä näin ollen lasketa moneen kertaan. (Singleton 1999, 10; Nation 2001, 7–8; Koivisto, 2013, 32–33.) Esimerkiksi virkkeessä "Me ostimme uuden kissan ja päätimme, että kissan nimeksi tulee Sisu" on yksitoista sanetta, mutta kymmenen lekseemiä, sillä saneet *kissa* ja *kissan* katsotaan yhdeksi lekseemiksi.

Tutkimuksessani hyödynnän pääsääntöisesti sanan käsitettä esimerkiksi kuvatessani oppijan sanojen ja sanaston oppimista ja osaamista. Aineiston käsittelyssä ja tarkastelussa hyödynnän tarpeen mukaan käsitteitä lekseemi ja sane, kun pyrin määrittelemään esimerkiksi kertomusten sanastollista vaihtelevuutta ja sitä, kuinka paljon kokonaisuudessaan teksteissä esiintyy sanoja.

Kielen oppimisen ja omaksumisen kannalta on oleellista huomioida myös käsitteet **leksikko**, **sanasto** ja **sanavarasto**. Kuten sanan määrittelyssä, myös näiden käsitteiden määrittelyssä on monenlaisia lähtökohtia; välillä käsitteitä saatetaan käyttää synonyymisina, mutta välillä niiden käytössä on selkeitä merkityseroja. Esimerkiksi Hongon (2013, 20) mukaan toisistaan erotetaan usein yksilön sanasto käyttämällä termiä sanavarasto ja koko kielen sanasto käyttämällä termiä leksikko. Puron (2002, 7) mukaan sanasto voidaan nähdä sanoja ja kielitaidon eri osa-alueita yhdistävänä kokonaisuutena, jossa sekä sanat että osa-alueet ovat vuorovaikutuksessa keskenään. Koivisto (2013, 48–56) puolestaan on määritellyt sanastoa useasta eri näkökulmasta. Hänen mukaansa lekseemit muodostavat sanaston, sillä sanasto on se, jonka avulla voimme muodostaa lausekkeita, lausumia, lauseita ja tekstejä. Koivisto toteaakin, että yleistäen sanasto voidaan nähdä jonkin kielen tai kielimuodon sanastona, esimerkiksi suomen kielen sanastona. Ongelmallista on se, että sanasto sinällään poikkeaa jo puhujienkin kesken, sillä molemmat, kirjoitettu ja puhuttu kieli, sisältävät omia varianttejaan ja murteet ja slangit lisäävät sanaston monimuotoisuutta (Koivisto 2013, 48). Monesti tutkimuksessa leksikko ja sanasto liitetään myös siihen, miten sanat ovat järjestyneet oppijan mielessä. Erityisesti psykolingvistiikassa käytetään termiä mentaalileksikko, jonka pohjana on ajatus kielestä dynaamisena järjestelmänä, joka on varastoitunut mieleen (Puro 2002, 3).

Kun tarkastellaan termien sanakirjamääritelmiä, Kielitoimiston sanakirja määrittelee sanavaraston yksilön hallitsemiksi sanoiksi (KS 2020 s.v. *sanavarasto*). Sanastolla puolestaan on useita merkityksiä: 1) sanat, sanavarat, 2) suppeahko sanakirja, 3) sanahakemisto, jossa on esitetty

hakusanojen merkitykset sekä käännökset jollakin toisella kielellä (KS 2020 s.v. *sanasto*). Leksikko saa merkitykset sanasto tai sanavarat (KS 2020 s.v. *leksikko*). Kaikkien kolmen käsitteen määritelmät limittyvät siis Kielitoimiston sanakirjan perusteella toisiinsa. Omassa työssäni käytän termiä sanasto, erityisesti viitatessani opiskelijan osaamiin sanoihin tai hänen sanavarastoonsa eli siihen sanastolliseen suomen kielen osaamiseen, joka suomenoppijoilla on ja joka aineistona olevista kertomuksista nousee esiin.

2.2. Kielitaidon osa-alueet

Perinteisen kielitaitoajattelun pohjana toimii niin sanottu kielitaidon kämmenmalli, jossa on viisi osa-aluetta: puhuminen, ymmärtäminen / kuunteleminen, lukeminen, kirjoittaminen sekä sanasto ja rakenne. Kämmenmallissa ajatuksena on se, että peukalo edustaa sanastoa ja rakennetta, jotka toimivat yhdessä muiden kielitaidon osa-alueiden kanssa, kuten peukalo toimii kaikkien sormien kanssa. (Nissilä–Martin–Vaarala–Kuukka 2006, 41.) Eurooppalaisen viitekehyksen – Common European Framework of Reference for Languages – kielitaidon tasojen (A1-C2) kuvausasteikossa esitetään myös samankaltaiset kategoriat: kuullun ymmärtäminen, puhuminen, luetun ymmärtäminen ja kirjoittaminen. Sanaston ja rakenteen osaamista ei ole erikseen jaoteltu, mutta ne sisältyvät kaikkiin kategorioihin. (Euroopan neuvosto 2012.)

Sekä kuullun että luetun ymmärtäminen katsotaan reseptiivisiksi eli vastaanottaviksi kielitaidon osa-alueiksi, kun taas puhuminen ja kirjoittaminen ovat produktiivisia eli tuottavia osa-alueita. Reseptiivisissä taidoissa olennaista on, että otamme toisilta ihmisiltä vastaan syötöstä (*input*) kuuntelemalla ja lukemalla, kun taas produktiivisissa taidoissa tuotamme kieltä (*output*) puhumalla ja kirjoittamalla, jolloin välitämme tietoa ja merkityksiä toisille. Jaottelusta huolimatta produktiiviset ja reseptiiviset taidot limittyvät, sillä kuunnellessamme ja lukiessamme tuotamme myös merkityksiä ymmärtääksemme. (Nissilä ym. 2006, 45; Nation 2011, 24–27.)

Kämmenmallin rinnalle on noussut käsitys sujuvuudesta (*fluency*), tarkkuudesta (*accuracy*) ja kompleksisuudesta (*complexity*). Kompleksisuus, tarkkuus ja sujuvuus ovat nousseet mukaan erityisesti kielitaidon arvioinnin yhteydessä. (Housen–Kuiken–Vedder 2012, 2.) Yksinkertaistetusti kielen sujuvuudella tarkoitetaan sitä, että kielenkäyttäjä pystyy ilmaisemaan itseään, omia ajatuksiaan ja viestejään johdonmukaisesti ja ymmärrettävästi. Tarkkuus puolestaan viittaa “virheettömyyteen” eli norminmukaisuuteen ja kohdekielen mukaiseen viestimiseen. Kompleksisuudella taas tarkoitetaan yksinkertaistetusti oppijankielen monimutkaistumista eli kielenkäyttäjän kykyä ilmaista itseään monin eri tavoin. (Nissilä ym. 2006, 45–46.)

Oma tutkimukseni liittyy vahvimmin juuri kielen kompleksisuuden tarkasteluun kirjoitetussa ja puhutussa kielessä, sillä sanaston kehittyminen ja laajeneminen ovat erityisen tärkeitä monipuolisen ilmaisun kannalta. Martinin (2003, 85) mukaan pelkästään kaikkein tavallisimmilla sanoilla kommunikoiminen on haastavaa ja jossain määrin puuduttavaa, eikä oppija välttämättä koe kuuluvansa kieliyhteisöön, jos hän ei pysty kommunikoimaan monipuolisesti. Kompleksisessa kielenkäytössä oppija siis osaa hyödyntää tilanteisiin soveltuvia sanavalintoja, lauserakenteita ja viittaussuhteita sekä ymmärtää kielen variaatiota. Etenkin opintojensa alkuvaiheessa oppija selviytyy perussanastolla, jonka avulla hän tulee ymmärretyksi, mutta mitä pidemmälle oppija etenee kielenoppimisessaan, sitä enemmän hän tarvitsee tarkempia ja kuvaavampia sanoja. (Nissilä ym. 2006, 136–162.)

Kompleksisuutta on kielen oppimisen tutkimuksessa määritelty monin eri tavoin, sillä ei ole olemassa vain yhdenlaista kompleksisuutta. Bram Bulté ja Alex Housen (2012, 23) esittelevät artikkelissaan oppijankielen kompleksisten rakenteiden luokittelun, joka paljastaa, kuinka laajasta asiasta kompleksisuudessa on kyse. Kompleksisuuteen liittyvät niin kielelliset (*linguistic*), diskurssi-vuorovaikutukselliset (*discourse-interactional*) kuin ehdotukselliset (*propositional*) aspektit, joista erityisesti kielellisen kompleksisuuden voidaan katsoa jakautuvan erilaisiin alakategorioihin (Bulté–Housen 2012, 23).

Kielellinen kompleksisuus voidaan jakaa kahteen suurempaan ryhmään, systeemiseen (*systemic complexity*) ja rakenteelliseen (*structure complexity*) kompleksisuuteen. Systeemisessä kompleksisuudessa tarkastellaan oppijan kielisysteeminen dynaamisuutta eli sitä, miten mutkikasta kielenkäyttäjän kieli on. Rakenteellinen kompleksisuus viittaa puolestaan kielenoppijan kielellisten yksiköiden, rakenteiden ja sääntöjen hallintaan, joista oppijan kielisysteemi koostuu. Oman tutkimukseni osalta keskityn pitkälti juuri systeemiseen kompleksisuuteen, jossa tarkastellaan oppijan osaamien erilaisten rakenteiden ja yksiköiden määrää, vaihtelua, monimuotoisuutta tai rikkautta eli esimerkiksi sitä, millaisen määrän oppija osaa eri sanoja tai kieliopillisia rakenteita ja kykeneekö oppija käyttämään näitä kaikkia vai vain osaa kielen äänteistä. (Bulté–Housen 2012, 23–25.) Koska omassa tutkimuksessani keskityn sanastolliseen monimuotoisuuteen, tarkastelen sanojen määrän lisäksi sanojen harvinaisuutta, vaihtelevuutta, erityisyyttä, tasaisuutta ja sirontaa (ks. tarkemmin luku 5.3.). Tarkastelemalla sanastoa monipuolisesti paljastuu myös se, muuttuuko oppijan sanasto kompleksisemmaksi intensiivikurssin kuluessa.

2.3. Oppijan sanasto

Kuten jo kielitaidon osa-alueiden yhteydessä pohdin, sanasto kuuluu oleellisena osana kielitaitoon. Ilman sanoja on vaikea tuottaa tai ymmärtää vierasta kieltä, ja siksi sanaston osaaminen on tärkeää. Sanojen tunnistamisen lisäksi oppijan tuleekin ymmärtää sanojen merkityksiä ja osata liittää niitä erilaisiin yhteyksiin, sillä tuottaessaan kieltä oppija joutuu tekemään monenlaisia valintoja saadakseen viestinsä ymmärrettävästi näkyviin (Chafe–Danielewicz 1987, 86).

Monesti tutkimuksissa on keskitytty siihen, kuinka laaja oppijan sanasto on eli kuinka monta sanaa oppija osaa. Tämä tarkastelutapa on sinänsä ollut rajallinen, koska sanan osaamiseen liittyy vahvasti monia muitakin asioita kuin pelkästään esimerkiksi se, että sana osataan kääntää kohdekielelle tai kohdekielestä. Sanaston laajuuden lisäksi onkin tärkeää tarkastella myös sanaston syvyyttä eli sitä, miten hyvin oppija osaa sanojen merkityksiä, pystyykö hän käyttämään sanoja erilaisissa yhteyksissä ja osaako hän liittää sanoihin erilaisia merkityksiä. (Schwartz–Katzir 2012, 1948–1949.) Sanastolliseen syvyyteen liittyy siis se, että käsite osataan yhdistää merkitykseen, mutta oppija pystyy liittämään siihen myös muita tarkoituksia sekä ymmärtää sanaan liittyviä paradigmaattisia ja syntagmaattisia suhteita, kuten taivutusparadigmaa tai sitä, millaisten sanojen kanssa käytettävä sana useimmiten esiintyy. Olennaista on myös se, että mitä syvemmälle sanan osaamisessa mennään, niin sen tarkemmin oppija osaa hahmottaa myös sanastollisen yksikön syntaktisia ja morfologisia rajoituksia ja piirteitä. (Henriksen 1999, 305–306.)

Psykolingvistiikassa kielen oppimisen yhteydessä puhutaan usein mentaalisesta leksikosta eli yksilöllisestä sanavarastosta, joka muodostuu oppijan mielessä (Niitemaa 2014, 145). Jokaisella nähdään olevan oma mentaalinen leksikkonsa, jonka avulla sanat pidetään järjestyksessä, ja josta haetaan sanoja ja poimitaan niitä käyttöön. Mentaalileksikko vaikuttaa myös siihen, miten sanoja vastaanotetaan ja tulkitaan. (Aitchison 1994, 12–14.) Voidaan ajatella, että oppijan mentaalinen leksikko rakentuu sekä tuottamiseen että ymmärtämiseen linkittyvistä sanastoista, jotka limittyvät ja kietoutuvat myös muiden kielen osa-alueiden ja kognition kanssa. Mentaalileksikko ei ole muuttumaton, vaan sanat muuttuvat ja tukevat toisiaan. (mt.: 12–13.)

Sanaston oppimisessa voidaan jaotella sekä produktiivinen (tuottava) että reseptiivinen (vastaanottava) sanasto, mikä on olennaista erityisesti kielen oppimisen tutkimuksen kannalta. Oppimisen alkuvaiheessa on usein niin, että reseptiivinen ja produktiivinen sanasto ovat saman laajuisia, mutta usein sanaston kehittyessä ero näiden välillä kasvaa (Niitemaa 2014, 143). Käyttääkseen reseptiivistä sanastoa oppijan on lukiessa tai kuunnellessa osattava havainnoida sanamuotoja, ja sen perusteella palautettava mieleensä sanan merkityksiä. Produktiivista sanastoa

käyttäessään eli puhuessaan tai kirjoittaessaan oppijan täytyy puolestaan olla halukas ilmaisemaan merkityksiä, jotta voi tuottaa sopivan sanan. Sanaston käyttö vaatii siis sanan tunnistamisen lisäksi sen merkityksen ja käyttötapojen ymmärtämistä. (Nation 2011, 24–27.) Aitchisonin (1994, 35) mukaan ihmisen on tiedettävä sanasta kolme merkittävää asiaa, jotta sanaa voi ylipäänsä käyttää: sanan merkitys, tehtävä lauseessa sekä se, miltä sana kuulostaa. Nationin (2011, 26) lähtökohta on samankaltainen, sillä hänen mukaansa sanasta tulee tietää sen muoto, merkitys ja käyttö.

Myös Puro (1999) on lähtenyt siitä ajatuksesta, että sanasto koostuu siihen kuuluvista sanoista, mutta myös sanoihin liittyvästä sanastollisesta tiedosta eli fonologiasta, morfologiasta, syntaksista, semantiikasta, pragmatiikasta sekä tiedosta sanojen esiintymistodennäköisyydestä. Oleellista on huomata se, että vaikka oppijan tuleekin ymmärtää sanojen merkityksiä, hänellä pitää olla myös kyky käyttää sanaa eri yhteyksissään ja osata liittää se oikeisiin asioihin. (mt.: 7.) Voidaankin todeta, että sanan oppimiseen liittyy moniosainen prosessi, jossa oppija syventää sanan osaamistaan. Ensin ymmärretään sanan muoto ja merkitys, sen jälkeen ymmärrys siitä kasvaa, millaisia asioita yhden käsitteen alle voidaan yhdistää ja onko samalla käsitteellä myös muita tarkoituksia ja viimeiseksi päästään siihen vaiheeseen, jossa pohditaan, millaisia suhteita tai yhteyksiä eri sanojen välille voidaan muodostaa. (Henriksen 1999, 308.)

Puro (1999) on koonnut yhteen usean tutkijan (Richards 1976; Carter 1989; Nation 1990; Ringbom 1990) määritelmiä sanan osaamisesta. Niille kaikille yhteistä on se, että sanan osaaminen vaatii monipuolisesti erilaisten osa-alueiden hallintaa. Puron tiivistelmän (1999, 5) perusteella voidaan todeta, että oppija osaa sanan, kun:

- hän ymmärtää sekä osaa tuottaa sanan eri muotoja, kuten taivutusmuotoja,
- hän ymmärtää sen syntaktista käyttämistä eli tietää, millaisissa tehtävissä sana esiintyy lauseessa,
- hän tietää sanan semanttiset tehtävät, kuten merkityksen, kollokaatiot, assosiaatiot, sanan käytön rajoitukset sekä sen, mihin muihin sanoihin sana todennäköisesti liittyy,
- hän tunnistaa sanan merkityksen kontekstissa tai ilman kontekstia ja
- hän osaa käyttää sanoja oikeissa yhteyksissä.

Sanan osaamisessa erityisen tärkeää on, että oppija pystyy yhdistämään sanan muodon ja merkityksen ja käyttämään sitä oikeanlaisissa tilanteissa. Tämä auttaa oppijaa myös tuottamaan kieltä, sillä mitä vahvempia sanojen yhteydet ovat eri merkityksiin, sitä helpompi merkityksiä on tulkita tai ilmaista oikein. Oppija voi toki tunnistaa sanan ilman sen kontekstia ja merkitystä, mutta yhdistämällä nämä hän kykenee ymmärtämään sen yhteyksiä myös muihin sanoihin. (Nation 2011, 48.)

Yleisesti ottaen sanaston hallinta voi olla suomenoppijalle erityisesti opintojen alkuvaiheessa hyvinkin haasteellista (Nissilä 2003, 112). Asiaan vaikuttaa osittain yksilölliset erot, sillä toisille sanaston oppiminen on helppoa, kun taas toisille ei (Pietilä 2014, 45), mutta usein haasteellisuuteen voi vaikuttaa myös oppijan lähtökieli. Suomen sanojen johtaminen sekä päätteiden ja liitteiden lisääminen voivat aiheuttaa hämmennystä ja hankaluuksia suomenoppijoille, etenkin, jos sanoja pitää tunnistaa puheesta tai tekstistä. (Nissilä 2003, 112.) Alkuvaiheessa oppija oppiikin peruskäsitteitä, kuten tuoli, kukka ja perhe. Mitä pidemmälle opinnoissa hän etenee, sitä tarkemmiksi käsitteet muuttuvat. Näin myös sanat saavat uusia merkityksiä ja yhteyksiä toisiin sanoihin. Esimerkiksi tuoleiksi voidaan tunnistaa jakkara tai nojatuoli, jotka puolestaan nähdään kuuluvaksi isompaan ryhmään, huonekaluihin. Lisäksi päätteiden ja liitteiden tunnistaminen ja käyttö auttavat johtamaan ja tunnistamaan uusia sanoja, kun kielentuntemus lisääntyy. (Nissilä ym. 2006, 144.) Olennaista onkin siis, että sanasto muokkautuu koko ajan, kun vanhat sanat saavat uusia yhteyksiä ja uudet sanat luovat uusia mahdollisuuksia käyttää entistä tarkempaa kieltä.

Näiden lisäksi kielen oppimisessa on mukana kulttuurinen ulottuvuus, joka on myös oman tutkimukseni kannalta oleellinen. Kulttuurisella ulottuvuudella tarkoitetaan sitä, että monissa yhteyksissä sanoillakin saattaa olla kulttuurisia merkityksiä, jotka vaikuttavat sanan merkitykseen, tulkintaan ja käyttöön. (Nation 2011, 51.) Oppijan kohdalla onkin hyvä ottaa esiin, miten esimerkiksi ruokaan, perhesuhteisiin ja kohteliaisuuteen liittyvät kulttuuriset merkitykset näkyvät opittavassa kielessä verrattuna lähtökieleen. Tämän tutkimuksen aineistona toimivat kuvasarjat ovatkin osiltaan kulttuurisidonnaisia, ja opiskelijat ovat päässeet tuottamaan kieltä, jossa mahdollisesti kulttuurinen ulottuvuus on nähtävissä. Kuvasarjojen avulla opiskelijoiden oli pystyttävä tuottamaan merkityksiä kuvissa oleville asioille ja tapahtumille sekä lähtökielellään että suomeksi. Sitä kautta heidän oli osattava yhdistää sanat vielä järkeviksi kokonaisuuksiksi, jotta tekstin lukijat tai kuulijat myös ymmärtävät kertomuksen kulkua.

Sen lisäksi, että oppija itse työstää ja harjoittelee sanastoamyös opetuksella on vaikutuksia oppijan sanaston kehittymiseen. Kielen ja sanaston opettamiseen on monia erilaisia menetelmiä, jotka pohjautuvat erilaisiin kielenoppimisteorioihin. Aikaisemmin kielenopetuksessa on keskitytty esimerkiksi kielen rakenteiden ja kieliopin opettamiseen eli kieleen itsessään tai vaihtoehtoisesti fokus on ollut kielessä viestinnän välineenä. Nykyään ajatellaan vahvasti, että molemmat, kieli ja sen rakenteet sekä kielen käyttö vuorovaikutuksessa, ovat tärkeitä oppijalle, mutta oppijalla itselläänkin on suuri rooli, sillä jokainen oppija oppii omalla tavallaan. (Järvinen 2014, 91; 98; 110.) Sanaston opettamisessa pelkkä opetus ei takaa oppimista, vaan oppija joutuu itse tekemään työtä löytääkseen parhaat tavat oppia sanastoa. Opettaja pystyy kuitenkin auttamaan ja helpottamaan oppijan työtä

korostamalla tärkeitä asioita sekä ohjaamalla sanaston käyttöä monipuolisilla tehtävillä ja harjoitteilla, joissa korostuvat sekä suullinen että kirjallinen osaaminen. (Niitemaa 2014, 152–153.)

Tässä tutkimuksessa aineisto on kerätty intensiivikurssilla, jossa treenataan erityisesti puhetaitoja, mutta myös kuuntelu, lukeminen, kirjoittaminen ja kielioppi ovat vahvasti mukana opetuksessa. Niitemaan mukaan (2014, 156) oppijan on helpompi huomata sanoja, jos ne liittyvät selkeästi tehtävien suorittamiseen tai niitä on käytetty jo aikaisemmin. Intensiivikurssilla on tapana käydä oppitunneilla läpi esimerkiksi johonkin tiettyyn aiheeseen liittyvää historiaa tai kulttuuripiirteitä tutustumalla aiheeseen liittyviin teksteihin, videoihin tai vastaaviin, joista nousee myös esiin uusia sanoja. Monesti samana tai seuraavana päivänä vapaa-ajan ohjelmassa tehdään jotakin opiskeltuun aihepiiriin liittyvää, jolloin sanasto pääsee autenttiseen käyttöön ja sitä kautta sanastollinen osaaminen voi vahvistua. Pietilän ja Lintusen (2014, 12–13) mukaan kielen oppimisessa voidaan erottaa toisistaan oppiminen ja omaksuminen. Omaksumista tapahtuu, kun oppijaa pääsee kommunikoimaan opittavalla kielellä, kun taas oppiminen on tiedostettua ja tapahtuu esimerkiksi luokkahuoneessa (mt. 13). Voidaan ajatella, että intensiivikurssilla yhdistyvät tietyllä tavalla kielen oppiminen sekä omaksuminen. Oppimista tuetaan opetuksella ja oppitunneilla oppija saa tietoa säännöistä, kieliopista ja rakenteista, jolloin hän oppii ymmärtämään, miksi asioita ilmaistaan tietyllä tavalla. Toisaalta oppija pääsee myös omaksumaan sanoja omaan sanavarastoonsa, kun kieltä kuullaan ja puhutaan monipuolisesti luokkahuoneen ulkopuolella ja erilaisissa vapaa-ajan ohjelmissa.

3. Puhutun ja kirjoitetun kielen variaatio

3.1. Puhuttu ja kirjoitettu kieli yleisesti

On todettu, että ihmiset kirjoittavat eri tavoin kuin puhuvat, mutta siitä huolimatta sekä kirjoitettu että puhuttu kieli kulkevat vahvasti rinnakkain, eikä niitä voi ajatella toisilleen vastakkaisina, vaan kyseessä on vain kielen eri käyttötavat (Chafe–Danielewicz 1987, 83; Viinikka–Voutilainen 2013, luku Poikkeavat prosessit ja vaihtelevat tilanteet), joilla molemmilla on tärkeä merkitys (Dufva 2000, 88). Kirjoitetun ja puhutun kielen erot ovatkin usein kieli- ja kulttuurikohtaisia (Dufva 1995, 65).

Puhuttu kieli ajatellaan usein keskusteluksi arkitilanteissa, joissa puhe on spontaania kahden tai useamman henkilön välistä vuorovaikutusta (Dufva 1995, 63). Usein puhetta tuotetaankin nopeasti, eikä sitä suunnitella ennakkoon (Viinikka–Voutilainen 2013, luku Poikkeavat prosessit ja vaihtelevat tilanteet). Puhuttu kieli on kuitenkin muutakin kuin perinteistä arkikeskustelua: se voi olla etukäteen valmisteltu puhe, oppitunti, luento tai vaikkapa radio-ohjelma (Dufva 1995, 63–64). Puhutun kielen muodoille yhteistä on kuitenkin se, että viesti ja kieli välittyvät äänen avulla, jolloin myös sanojen lisäksi erilaiset tauot, äänenpainot ja muut nonverbaaliset viestit auttavat hahmottamaan merkityksiä (Dufva 2000, 79). Kirjoitettu kieli puolestaan välittyy vastaanottajalle visuaalisten merkkien avulla. Kirjoitetussa kielessä sanojen lisäksi esimerkiksi välimerkit auttavat vastaanottajaa hahmottamaan viestin sisältöä sekä jäsentämään tekstiä. Kirjoitettua kieltä ovat yhtä lailla niin oppikirjojen kieli, romaanit, sanomalehdissä esiintyvät asiatekstit kuin henkilökohtaiset päiväkirjat, kirjeet tai sähköpostiviestit. (Dufva 1995, 63–64.)

Kirjoitetun ja puhutun kielen lisäksi puhutaan monesti yleis- tai kirjakielestä sekä puhekielestä, joista ensimmäisellä tarkoitetaan sanastoltaan yleistajuista, kielenhuollon sääntöjä ja standardeja noudattelevaa kieltä, kun taas puhekielellä viitataan vapaamuotoisempaan kielenkäyttöön, joka ei aina noudata yleiskielen konventioita. Yleiskieli ja puhekieli eivät kuitenkaan viittaa suoraan siihen, miten kieltä käytetään, vaan ne kertovat kielen tyylistä (Viinikka–Voutilainen 2013, luku Kirjakieli, puhekieli ja sekaannuksen siemen) eli siitä, noudattaako kieli määriteltyjä konventioita, kuten ”*Minä haluan*” vai vapaamuotoisempaa tapaa ilmaista asioita, kuten ”*Mä haluun* / *Mie halluun*”. Yleiskieli ja puhekieli voivat siis esiintyä sekä kirjoitettuna että puhuttuna. Puhuttu ja kirjoitettu kieli taas viittaavat juuri siihen, miten kieltä ilmaistaan: äänellä vai kirjoitettuna. (Viinikka–Voutilainen 2013, luku Kirjakieli, puhekieli ja sekaannuksen siemen.) Tässä tutkimuksessa hyödynnän käsitteitä puhuttu ja kirjoitettu kieli, kun viittaan varsinaisesti kertomusten kieleen. Kun puhun kertomusten tuottotavasta, käytän termejä suullinen ja kirjallinen kertomus, sillä ne kuvaavat selvemmin sitä,

miten kertomukset on tuotettu. Sanaston osalta käytän tarvittaessa termiä yleiskieli, sillä sanat on lemmattu nimenomaan yleiskielen konventioiden mukaisesti.

3.2. Puhutun ja kirjoitetun kielen sanasto

Kuten jo aiemmin luvussa 2.3. esittelin, usein puhutaan reseptiivisistä ja produktiivisista kielitaidon osa-alueista. Koska kirjoittaminen ja puhuminen ovat vahvasti produktiivisia, ja omassa tutkimuksessani tarkastelen kirjoittaen ja suullisesti tuotettuja tekstejä, on olennaista hieman myös pohtia, vaikuttaako kielen tuottamisen tapa sanastoon ja sen käyttöön.

Puhe- ja kirjoitustaidon kehittyminen ovat osittain erilaisia prosesseja. Puhe onkin paljon vanhempi viestintäkeino kuin kirjoitus, ja se opitaan lapsuudessa ensin. Puhe kehittyy dialogissa muiden ihmisten kanssa, kun lapsi ensin havainnoi ympäristöään ja siinä puhuttavaa kieltä sekä alkaa harjoitella erilaisia ilmaisutaitoja syntymästään alkaen. Kirjoitustaito sekä samalla lukutaito kehittyvät puolestaan hitaammin, kun aletaan – tietoisesti ja tiedostamatta – opetella kirjaimia ja niistä koostuvia merkkijonoja, sanoja, jotka välittävät kuitenkin pääosin samoja merkityksiä kuin puhutun kielen sanat ja ilmaukset. (Dufva 1995, 67–68.)

Koska puhuttua kieltä ilmaistaan pääasiassa äänellä ja kirjoitettua kieltä visuaalisilla merkeillä, on ymmärrettävää, että niitä voidaan myös käyttää eri tavoilla (Viinikka–Voutilainen 2013, luku Poikkeavat prosessit ja vaihtelevat tilanteet). Oman tutkimukseni kannalta on tärkeä pitää mielessä, että monesti kirjoitetussa kielessä käytetään vaihtelevampaa ja monipuolisempaa sanastoa kuin puhutussa kielessä (ks. tarkemmin esim. Chafe–Danielewicz 1987), ja tällöin sisältösanojen määrä saattaa olla runsaampi kuin puheessa (Dufva 1995, 65). Syitä tähän on useita, sillä aina tuottaessaan kieltä, puhuja tai kirjoittaja tekee leksikaalisia valintoja, joiden avulla hän ilmaisee haluamansa asian tilanteessa parhaaksi katsomallaan tavalla (Chafe–Danielewicz 1987, 87–88).

Koska puhuttua kieltä tuotetaan yleensä suhteellisen nopeasti hetkessä, jossa kuulija on läsnä, viesti on tarkoitus välittää heti selkeästi. Tästä syystä puhuja usein valitseekin ensimmäisen vaihtoehdon, joka sanasta tai ilmaisusta tulee mieleen sen sijaan, että pohtisi ja muokkaisi viestiä runsaasti. Korjaus ja erilaisten suunnitteluilmausten käyttö on puheessa aina mahdollista ja varsinkin spontaanissa puheessa yleistä, mutta liiallinen korjailu voi vaikuttaa viestin ymmärrettävyyteen ja siihen, miten kuulija puheen tulkitsee. (Chafe–Danielewicz 1987, 87–88; Dufva 1995, 69–70; Viinikka–Voutilainen 2013, luku Poikkeavat prosessit ja vaihtelevat tilanteet.) Tarvittaessa puheen vastaanottaja voi myös kysyä tarkennuksia, jos viestiä on vaikea ymmärtää. Kirjoitettua tekstiä tuottaessaan taas kirjoittajalla on usein aikaa miettiä ja prosessoida sanoja ja ilmauksia sekä pohtia

useampaa eri vaihtoehtoa parhaaseen ilmaisuun ja siksi sanasto on vaihtelevampaa kuin puheessa. Usein kirjoitetussa kielessä ei myöskään ole nähtävissä korjauksia tai suunnitteluilmauksia, sillä tekstiä on ollut mahdollisuus rauhassa muotoilla ja korjata toisin kuin puhetta. (Chafe–Danielewicz 1987, 88; Dufva 1995, 69–70.)

Sanaston vaihtelevuuden lisäksi kirjoitettu ja puhuttu kieli eroavat toisistaan joidenkin kieliopillisten ja rakenteellisten piirteiden osalta. Esimerkiksi suomen kielen yhteydessä kirjoitetun kielen sanat ja lauseet ovat tyypillisesti pidempiä ja monimutkaisempia kuin puhutussa kielessä. (Dufva 1995, 65).

Sanaston osalta on kuitenkin muistettava, että suurin osa käytettävistä sanoista on neutraaleja, jotka esiintyvät sekä puhutussa että kirjoitetussa kielessä, vaikka on myös sellaisia sanoja, jotka esiintyvät tyypillisemmin joko puhutussa tai kirjoitetussa kielessä (Chafe–Danielewicz 1987, 92). Esimerkiksi suomen kielessä *minä* esiintyy yleisemmin kirjoitetussa kielessä, kun taas *mä / mää* puhutussa kielessä. Toisaalta tällaiset kielenpiirteet ovat hyvin vaihtelevia, sillä nykyisin esimerkiksi erilaiset sosiaalisen median kanavat (Facebook, Twitter, Instagram) sekä pikaviestimet (tekstiviestit, WhatsApp, Snapchat) tarjoavat mahdollisuuden nopeaan kommunikointiin, jolloin puheessa tyypillisimmin esiintyvät muodot on otettu ainakin tietyissä yhteyksissä käyttöön myös kirjoitetussa kielessä (ks. esim. Helasvuo 2014; Kuusinen 2018). Variaatiota on paljon ja ihmiset käyttävät kieltä monipuolisesti yhdistelemällä sekä yleis- että puhekieltä (Chafe–Danielewicz 1987, 84). Esimerkiksi puhutussa kielessä suhteellisen yleinen sanojen lyhentäminen (*koulussa* > *koulus*) näkyy myös kirjoitetussa kielessä. Lisäksi yleisistä kirjoitussäännöistä poikkeaminen (ei isoja alkukirjaimia tai välimerkkejä) sekä erilaisten emojien hyödyntäminen tunteiden ilmaisuun tai sanojen korvaamiseen ovat tulleet mukaan kirjoitettuun kieleen ainakin joissain yhteyksissä. (ks. esim. Kuusinen 2018).

Suomenoppijan kohdalla tällaiset erot voivat olla haastavia hahmottaa eikä puheessa välttämättä käytetä näitä tyypillisiä puhutun kielen ilmauksia, vaan pysytellään opituissa yleiskielen ilmauksissa. Suomen kielen oppikirjoissa käsitellään nykyään myös puhekieltä, mutta yleiskieli on edelleen se, mistä lähdetään liikkeelle ja mitä pitkälti tarkastellaan. Oman tutkimukseni kannalta on kuitenkin huomattava, että osa opiskelijoista osaa hyödyntää puhekielenomaisia sanastollisia piirteitä omassa kielenkäytössään. Joissain tilanteissa näitä piirteitä on nähtävissä sekä kirjoitetussa että puhutussa kielessä. Aineiston analyysin kannalta tällaiset muodot on lemmauksen yhteydessä muokattu vastaamaan yleiskielistä vastinettaan. Ratkaisua perustelen sillä, että samat opiskelijat ovat osoittaneet osaavansa myös tämän yleiskielisen vastineen käytön omissa teksteissään eli he tuntevat molemmat variantit.

4. Leksikaalinen diversiteetti tutkimuskohteena

4.1. Leksikaalisen diversiteetin määrittely

Leksikaalinen diversiteetti voidaan määritellä yksinkertaisesti sanastolliseksi monimuotoisuudeksi. Mitä monipuolisempi ja laajempi eri sanojen määrä on, sitä suurempi on myös leksikaalinen diversiteetti. (McCarthy–Jarvis 2010, 381.) Tarkemmin määriteltäessä leksikaalisella diversiteetillä tarkoitetaan kielenoppijan kykyä ilmaista tarkkoja merkityssisältöjä sekä tuottaa laadukasta tekstiä tai puhetta. Tämä tarkoittaa sitä, että kielenoppijan on osattava sanastoa monipuolisesti, mutta myös ymmärrettävä sanojen käyttöä ja niiden merkityksiä. Leksikaalinen diversiteetti voidaan nähdä myös olennaisena osana vuorovaikutuksessa, sillä kyetäkseen ilmaisemaan itseään kattavasti, oppijan on osattava sanoja laajasti ja ymmärrettävä miten sanoja käytetään missäkin yhteydessä. (Honko–Jarvis–Vainio 2019, 44.)

Leksikaalista diversiteettiä on tutkittu aina 1930-luvulta asti, jolloin myös käsite sanaston monimuotoisuudesta (*diversity of vocabulary*, Carrol 1938) otettiin käyttöön ensimmäistä kertaa. Käsitteen määrittely on kuitenkin ollut hyvin vaihtelevaa, sillä monet termeistä ovat olleet synonyymisiä tai päällekkäisiä. (ks. Jarvis 2013b, 15.) Käytettyjä termejä ovat olleet muun muassa leksikaalinen diversiteetti (*lexical diversity*), leksikaalinen tai sanastollinen variaatio (le*xical variation*), sanastollinen vaihtelu (*lexical variety*), sanastollinen joustavuus (*lexical flexibility*) ja leksikaalinen rikkaus (*lexical richness*) (Malvern–Richards–Chipere–Durán 2004, 5; Jarvis 2013a, 15; Honko ym. 2019, 47).

Lingvistiseen diversiteettitutkimukseen on otettu mukaan myös ekologian näkökulmia, metodeja ja termejä (ks. esim. Jarvis 2013a; Honko ym. 2019, 45). Diversiteetin tarkastelussa voidaan lähteä liikkeelle ajatuksesta, että diversiteettiin liittyy useita erilaisia ominaisuuksia, jotka ovat toisiinsa yhteydessä, mutta jotka eivät yksinään ole osoitus diversiteetistä (Jarvis 2013a, 96). Kun leksikaalista diversiteettiä on tarkasteltu, huomio on usein kiinnittynyt vain sanaston määrään ja toisteisuuteen, jonka vuoksi on pyritty löytämään lisää näkökulmia leksikaalisen diversiteetin määrittelyyn ja arviointiin. Leksikaalisen diversiteetin tutkimuksessa on tarkasteltu sanojen määrän (*volume*) lisäksi muun muassa sanojen harvinaisuutta (*rarity*), erityisyyttä (*disparity*), tasaisuutta (*evenness*), sirontaa (*dispersion*) sekä vaihtelevuutta (*variability*). (Jarvis 2013a, 101–102; 2013b, 22–26.) Näiden eri osa-alueiden mittaamiseen ja tarkastelemiseen on pyritty kehittämään mittareita (ks. luvut 4.3. ja 5.3.), joiden avulla saataisiin mahdollisimman kattava kuva leksikaalisesta diversiteetistä. Vaikka ekologisessa diversiteetin tutkimuksessa on kehitetty erilaisia mittareita diversiteetin mittaamiseen,

kuten Shannonin indeksi ja Gini-Simpsonin indeksi, määrällisessä lingvistiikassa niitä ei ole varsinaisesti hyödynnetty (ks. kuitenkin esim. Malin 2012).

Leksikaalista diversiteettiä voidaan kokonaisuutena pohtia monesta eri näkökulmasta. Esimerkiksi Honko, Jarvis & Vainio (2019) kuvaavat kahta tapaa tarkastella leksikaalista diversiteettiä: joko siten, että tarkastelu rajataan tekstiin liittyviin kriteereihin, kuten tekstilajiin tai kohderyhmään, tai vaihtoehtoisesti siten, että tarkastellaan eri yksilöiden tapaa käyttää kieltä. Tekstinäkökulmalle oleellista on se, että sanastollista monimuotoisuutta tarkastellaan itse teksteistä ottaen huomioon tekstien sisältämät sanat tai tarkastelemalla yleisesti kielen sisältämää sanastoa. Yksilönäkökulmassa keskeisenä on kielenoppijan, yksilön, tuottama puhuttu tai kirjoitettu kieli, jonka sanastoa tarkastelemalla pyritään saamaan käsitys oppijan kielitaidosta. Tutkimalla yksilön kieltä on myös saatu välineitä kielitaidon kuvaamiseen, ja sitä kautta on pystytty vertailemaan muun muassa yksilöiden välistä kielitaitoa sekä yksilön kielitaidossa tapahtuvia muutoksia. (Honko ym. 2019, 47–48.)

Omassa tutkimuksessani tarkastelen leksikaalista diversiteettiä nimenomaan yksilönäkökulmasta selvittääkseni sen, miten suomenoppijoiden leksikaalinen diversiteetti kehittyy intensiivikurssilla. Samalla vertaan myös yksilöitä toisiinsa ja pohdin, onko tuloksissa havaittavissa samansuuntaista kehitystä kaikkien oppijoiden kesken.

4.2. Leksikaalisen diversiteetin tutkimus

Leksikaalista diversiteettiä on tarkasteltu erityisesti englanninkielisistä aineistosta, mutta muistakin kielistä tutkimusta on tehty. David Malvern, Brian Richards, Ngoni Chipere ja Pilar Durán tarkastelevat teoksessaan (2004) leksikaalista diversiteettiä ja sen suhdetta kielitaidon kehittymiseen. He selvittävät erilaisia tapoja tarkastella leksikaalista diversiteettiä, ja tutkivat sitä erilaisia aineistoja apuna käyttäen. Teoksessa tarkastellaan muun muassa lähtökielen ja vieraan kielen puhuttua materiaalia, lasten kielen morfologista kehitystä niin lemmatun kuin lemmaamattoman korpuksen avulla, eri mittareiden käyttöä leksikaalisen diversiteetin tutkimuksessa sekä analysoidaan laajaa kirjoitetun kielen korpusta hyödyntäen erilaisia muuttujia. Tutkimusten perusteella he ovat löytäneet ratkaisuja, joiden avulla leksikaalisen diversiteetin mittaaminen helpottuu ja yhdenmukaistuu. Erityisesti antamalla numeerisen arvon D leksikaaliselle diversiteetille, pystytään tarkastelemaan leksikaalista diversiteettiä monenlaisista lähtökohdista.

Scott Jarvis on tutkinut erityisen paljon sekä leksikaalista diversiteettiä, sen määrittelyä ja mittaamista sekä erilaisten inhimillisten arvioijien käyttöä leksikaalisen diversiteetin arvioinnissa. Hän on

pohtinut myös soveltuvien menetelmien ja muuttujien käyttöä leksikaalisen diversiteetin laskemiselle. Jarvis on tutkinut muun muassa sitä, millaisiin asioihin ulkopuoliset arvioijat kiinnittävät huomiota, kun heitä pyydetään arvioimaan leksikaalista diversiteettiä. Tutkimuksessaan (2013b) hän määrittelee ensin kuusi leksikaalisen diversiteetin muuttujaa: määrä, vaihtelevuus, harvinaisuus, tasaisuus, sironta ja erityisyys. Tutkimuksen informantteina toimii yksitoista arvioijaa, jotka tarkastelevat yhteensä viittäkymmentä tekstiä. Jokainen arvioija tarkastelee kymmenestä kahteenkymmeneen tekstiä, eli vähintään kaksi arvioijaa arvioi saman tekstin. Tekstit on korjattu ennen arvioijille lähettämistä, jotta mahdolliset virheet eivät vaikuttaisi arviointeihin. Arvioijien ohjeistus tehtävään on suppea, sillä heitä ei pyritä ohjaamaan tutkimuksessa minkään erityisen piirteen tarkasteluun. Tutkimuksessa osoittautuu, että arvioijien tulokset leksikaalisesta diversiteetistä korreloivat yleisellä tasolla määriteltyjen muuttujien kanssa, mutta Jarvis toteaa, että on vielä tarpeen hienosäätää ja pohtia tarkemmin sitä, miten kuutta eri osa-aluetta mitataan, jotta tuloksista saadaan vielä luotettavampia.

Guoxing Yu (2009) on tarkastellut leksikaalista diversiteettiä kirjoitetuissa ja puhutuissa tehtävissä, joissa arviointi tapahtuu automaattisten systeemien avulla. Hän tarkastelee muun muassa sitä, millaiset osatekijät, kuten sukupuoli, lähtökieli sekä tavoitteet ja tarkoitus testin tekemiselle, vaikuttavat leksikaaliseen diversiteettiin ja sen arviointiin. Lisäksi hän pohtii, miten kirjoitustehtävien aiheet vaikuttavat leksikaaliseen diversiteettiin. Hän tarkastelee myös sitä, miten kirjoitetut ja puhutut diskurssit vaikuttavat yleisesti kielen sujuvuuteen, sekä millaisia yhteyksiä puhuttujen ja kirjoitettujen diskurssien leksikaalisella diversiteetillä on. Tulosten perusteella osoittautuu, että erityisesti leksikaalisella diversiteetillä on yhteyksiä yleiseen kielitaidon tasoon sekä puhuttujen ja kirjallisten tehtävien suorittamiseen. Mielenkiintoista on, että yhteys leksikaalisen diversiteetin ja sukupuolen, lähtökielen ja testin tekemisen tarkoituksen välillä vaihtelee suuresti. Yu nostaa esiin myös sen, että kirjoitustehtävien tehtävänannoilla on suuri merkitys leksikaalisen diversiteetin vaihtelulle teksteissä. Kun kirjoitetaan yleisistä aiheista, leksikaalinen diversiteetti on huomattavasti suurempi kuin kirjoitettaessa henkilökohtaisista aiheista. Lisäksi aiheen tuttuus vaikuttaa selkeästi diversiteettiin: mitä tutumpi aihe, sitä korkeampi leksikaalinen diversiteetti näyttäisi olevan.

Victoria Johansson (2008) puolestaan tarkastelee tutkimuksessaan leksikaalista diversiteettiä ja leksikaalista taajuutta (*lexical density*). Hän vertailee näitä kahta keskenään ja pohtii sitä, miten ne näyttäytyvät teksteissä silloin, kun arvioidaan sanastollista kehitystä. Tarkastellakseen leksikaalisen diversiteetin ja taajuuden välisiä eroja kehityksellisestä näkökulmasta, hän hyödyntää tutkimuksessaan sekä kirjoitettuja että puhuttuja aineistoja. Aineistot on koottu eri ikäisten informanttien kirjoituksista ja suullisista kertomuksista; kokonaisuudessaan jokainen informantti on tuottanut neljä tekstiä. Tutkimuksessa osoittautuu, että leksikaalinen diversiteetti ja leksikaalinen

taajuus korreloivat keskenään ($r = 0.7333$, $p < 0.01$), ja molemmat osoittavat leksikaalista kehittymistä. Tutkimuksesta käy ilmi myös se, että sekä diversiteetti että taajuus kasvavat iän karttuessa. Erityisesti suullisten tekstien osalta on huomattavissa, että sanasto on vaihtelevampaa kertovissa teksteissä 17-vuotiaiden ikäryhmässä kuin nuoremmilla. Aikuisten ikäryhmässä sanaston käyttö on puhutuissa kertomuksissa kaikista laajinta. Taajuuden osalta osoittautuu, että eroja on nähtävillä esimerkiksi kirjoitetuissa teksteissä siten, että 17-vuotiaiden ja aikuisten tekstit ovat huomattavasti erilaisia kuin 10-vuotiaiden, ja puhuttujen tekstien osalta aikuisten tekstit ovat taajuudeltaan kaikkein kehittyneimpiä. Johansson toteaa (2008, 77), että diversiteettiä ja taajuutta ei voida korvata toisillaan, vaikka molempia voidaan hyödyntää laadullisten ja kehityksellisten erojen arvioinnissa. Yhtenä tuloksen Johanssonin mukaan voidaan pitää myös sitä, että leksikaalisen diversiteetin tarkastelu ja hyödyntäminen taajuuden sijaan toimii paremmin ainakin arvioidessa ikäryhmien sanastollista kehitystä, sillä diversiteetti osoittaa taajuutta vahvemmin kehityksellisiä muutoksia.

Suomen kielen osalta esimerkiksi Mari Honko on tutkinut leksikaalista diversiteettiä. Väitöstutkimuksessaan (2013) hän tarkastelee alakouluikäisten sanastollisen tiedon ja taidon kehittymistä verraten toisen sukupolven suomen puhujia S1-puhujiin. Tutkimuksessa osoittautuu, että suomen kielen sanaston hallinnassa on useammin puutteita maahanmuuttajataustaisilla lapsilla kuin S1-puhujilla. Tutkimuksen mukaan erityisesti kielitaidolla sekä kielellisellä itsetunnolla on vaikutusta kielen käytön aktiivisuuteen ja sitä kautta leksikaaliseen diversiteettiin. Mitä aktiivisempaa käyttö on, sen enemmän taito kasvaa ja toisaalta myös toisin päin. Honko (2017) on tarkastellut myös puhutun kielen leksikaalista diversiteettiä alakouluikäisten sadutetuissa kertomuksissa. Hän vertaa puhuttuja tekstejä väitöstutkimuksensa aineiston kirjoitettuihin teksteihin ja pohtii, onko puhuttujen ja kirjoitettujen kertomusten leksikaalisessa diversiteetissä nähtävissä eroja. Tulosten perusteella osoittautuu, että kirjoitetun kielen leksikaalinen diversiteetti on runsaampaa kuin puhutun kielen.

Malin (2012) on tarkastellut pro gradu -tutkielmassaan suomi toisena kielenä -oppijoiden sanaston kehittymistä taitotasolta toiselle siirryttäessä. Oleellinen tulos on, että oppijan sanasto kehittyy taitotasolta toiselle siirryttäessä, mutta erityisesti edistyneemmillä tasoilla ja varsinkin siirryttäessä taitotasolta B2 taitotasolle C1.

Yhtenä uusimpana suomen kielen leksikaaliseen diversiteettiin kohdistuvana tutkimuksena on Hongon, Jarvisin ja Vainion (2019) tutkimus, jossa he tarkastelevat lukijoiden käsityksiä oppijoiden tekstien leksikaalisesta diversiteetistä ja pohtivat millaisiin sanaston piirteisiin lukijat keskittyvät leksikaalista diversiteettiä arvioidessaan. Tutkimuksessa aikuiset lukijat arvioivat koululaisten kirjoittamia kertomuksia ja kertovat käsityksiään teksteistä. Tavoitteena on verrata näitä arvioita sellaisiin sanaston piirteisiin, joihin käsitysten leksikaalisesta diversiteetistä oletetaan perustuvan.

Vertailuun hyödynnetään erilaisia tilastollisia malleja, kuten Cronbachin alfaa ja regressioanalyysia. Tulosten perusteella osoittautuu, että eri arvioijien tekemät arviot leksikaalisesta diversiteetistä ovat keskenään melko samanlaisia, sillä Cronbachin alfa on 0,959, ja se ylittää reliabiliteettitason huomattavasti. Vaikka yksittäisten tekstien kohdalla osoittautuukin olevan hajontaa arvioinneissa, näidenkin tekstien arviot korreloivat yleisen näkemyksen kanssa tyydyttävästi. Tilastollisten analyysien ja tulosten perusteella voidaan todeta, että lukijoiden käsitykset ja arviot leksikaalisesta diversiteetistä ovat samansuuntaisia neljän määritellyn leksikaalisen piirteen kanssa: runsaus, sironta, erityisyys ja vaihtelevuus. Arvioijien huomio keskittyy tekstin eri sanojen tai lekseemien määrään, sanojen sijoittumiseen toisiinsa nähden eli kuinka kaukana saman sanan esiintymät ovat toisistaan, sanojen erityisyyteen eli siihen, millainen merkitys tai käyttötapa niillä on sekä miten paljon uusia sanoja tekstissä käytetään. Se, kuinka paljon tekstissä on saneita ja miten tasaisesti frekvenssijakauma toteutuu, eivät näytä vaikuttavan arvioijien käsityksiin. Kokonaisuudessaan siis osoittautuu, että lukijoiden näkemykset leksikaalisesta diversiteetistä kiinnittyvät sellaisiin piirteisiin, joiden on aiemmin ajateltu olevan osa leksikaalista diversiteettiä. Honko ym. kuitenkin toteavat, että jatkossa olisi tärkeää vielä edistää tutkimusta pidemmälle ja tarkastella asiaa laajemmin.

Kuten aiempia tutkimuksia tarkastellessa on voitu huomata, leksikaalinen diversiteetti on moniulotteinen asia, eikä ole vain yhtä mittaria tai muuttujaa, joka kuvailisi tätä kaikkea. Tästä syystä leksikaalisen diversiteetin arviointiin ja tutkimukseen on kehitetty useita eri mittareita, joista yleisimpänä on pitkään ollut TTR (*type–token ratio*), joka mittaa lekseemien ja saneiden suhdetta eli sanatoisteisuutta. TTR-arvon saavuttamiseksi tekstin erilaisten sanojen eli lemmojen tai lekseemien määrä jaetaan tekstisanojen eli saneiden kokonaismäärällä. TTR on kuitenkin todettu tutkimuksissa haasteelliseksi mittariksi leksikaaliselle diversiteetille, sillä arvo vaihtelee paljon riippuen tekstin otoskoosta eli pituudesta. Onkin esitetty, että mitä pidempi teksti on, sen enemmän sanatoisteisuus kasvaa, mikä vaikuttaa näin TTR-arvoon. (McCarthy–Jarvis 2010; Jarvis 2013b, 18; Honko 2010, 119.)

TTR on saanut rinnalleen useita versioita, joissa taustalla on ajatus sana-sane -suhteen laskemisesta, mutta joissa otoskoon vaikutus arvoihin on pyritty kiertämään. Erityisesti Johnsonin (1939) kehittelemä MSTTR (*mean segmental type/token ratio*) on pyrkinyt vakioimaan tekstikoot, jolloin tekstiosuudet olisivat aina samanmittaisia. Jotta teksteistä on saatu tasakokoiset otokset, niistä on valittu esimerkiksi sadan sanan tekstiosuudet ja ylimenevät sanat on jätetty pois analyysista. (Jarvis 2013b, 16; Honko ym. 2019, 51–52.) Verrattuna muihin kehiteltyihin menetelmiin viimeaikaisten tutkimusten valossa MSTTR:ta voidaan pitää yhtenä onnistuneimmista mittareista (Jarvis 2013b).

Näiden lisäksi kehitelty MTLD (*the measure of textual lexical diversity*) on osoittautunut luotettavaksi diversiteetin mittariksi. MTLD ottaa huomioon tekstin vaihtelevuuden ja se lasketaan

peräkkäisistä sanajonoista, joissa jokaiselle sanalle laskentaan sen oma TTR-arvo. Kun yksittäisen sanan TTR-arvo laskee alle 0,72, se ja sitä edeltävät sanat katsotaan yhdeksi faktoriksi. Toisin kuin MSTTR, MTLD ei jätä huomiotta ylijääviä sanoja, vaan niistä lasketaan niin sanottu häntäfaktori. Tekstin lopullinen faktori muodostuu tekstin kokonaisfaktorien ja häntäfaktorin summasta, ja saadun arvon avulla lasketaan MTLD-arvo. MTLD-arvo saadaan jakamalla tekstin saneiden kokonaismäärä lopullisella faktorilla, jolloin saadaan riittävä tarkkuus ja yhtenäisyys tekstin arvolle. (ks. tarkemmin McCarthy–Jarvis 2010; Malin 2012.)

5. Aineisto ja tutkimusmenetelmät

5.1. Aineiston esittely

Tässä tutkimuksessa hyödynnetty aineisto on kerätty Kansainvälisen liikkuvuuden ja yhteistyön keskuksen CIMO:n (nyk. Opetushallitus) järjestämien intensiivikurssien aikana. Aineistosta on aiemmin tutkittu erityisesti laulamisen, laulujen kuuntelemisen sekä rytmisen lausumisen vaikutuksia oppijan kielitaidon ja kirjoittamisen sujuvuuden kehittymiseen sekä ylipäänsä laulamisen vaikutuksia oppijan kielitaidon sujuvoittamiseen (ks. Alisaari 2016; Alisaari–Heikkola 2016a ja 2016b; Heikkola –Alisaari 2017; Heikkola–Alisaari 2019).

Intensiivikursseille osallistuneet opiskelijat olivat 18–33-vuotiaita yliopisto-opiskelijoita ja suurin osa oli kotoisin Euroopasta tai Pohjois-Amerikasta. Opiskelijoita oli kokonaisuudessaan 67, ja heidät jaettiin kahdelle kurssille kielitaidon taitotason mukaan. CIMO ja kurssien paikalliset järjestäjät jakoivat opiskelijat kursseille I ja IIA sen perusteella, miten opiskelijoiden kotiyliopistojen opettajat olivat oppijoiden kielitaidon arvioineet. Myöhemmin intensiivikurssien opettajat arvioivat opiskelijoiden kielitaidon tason uudelleen. (Alisaari–Heikkola 2016b, 275.) Arvioinnissa hyödynnettiin Eurooppalaisen viitekehyksen kielitaidon taitotasoja, missä tasot A1–A2 kuvaavat peruskielitaitoa, B1–B2 itsenäisen kielenkäyttäjän tasoja sekä C1–C2 taitavan kielenkäyttäjän tasoja (Euroopan neuvosto 2012). Kurssille I osallistuneet opiskelijat olivat opiskelleet suomea puolesta vuodesta vuoteen, ja heidän taitotasonsa vastasi Eurooppalaisen viitekehyksen tasoja A1-A2. Kurssille IIA osallistuneet opiskelijat olivat puolestaan opiskelleet suomea vuodesta kahteen vuoteen, ja heidän taitotasonsa vaihteli alussa tasolta A1 tasolle B1, joskin suurin osa oli taitotasolla A2. (Alisaari 2016, 35; Alisaari–Heikkola 2016a, 317.) Opiskelijat jaettiin lisäksi kurssien sisällä kolmeen opetusryhmään eli kokonaisuudessaan ryhmiä oli kuusi (lauluryhmä I, kuunteluryhmä I, rytmiryhmä I, lauluryhmä IIA, kuunteluryhmä IIA ja rytmiryhmä IIA). Intensiivikurssin normaaliopetuksen lisäksi opiskelijoille järjestettiin seitsemänä päivänä 15 minuuttia kestävä opetushetki, jossa suomen opetukseen liitettiin laulamista, laulujen kuuntelemista tai laulujen tekstien rytmistä lausumista. Ryhmien opetus oli muuten samankaltaista, paitsi näissä lyhyissä 15 minuutin opetustuokioissa, joissa lauluja hyödynnettiin eri tavoin sen mukaan, mikä ryhmä oli kyseessä. (Heikkola–Alisaari 2017, 25.) Tässä tutkimuksessa tarkasteltavan ryhmän oppijat ovat kurssilta IIA, ja heidän kielitaitonsa on määritelty pääsääntöisesti tasoille A2–B1. Lisäksi he kuuluivat rytmiryhmään IIA eli he pääsivät opettelemaan suomea lausumalla laulujen tekstejä, mutta jätän rytmisen lausumisen vaikutusten tarkastelun tämän tutkimuksen ulkopuolelle.

Opiskelijoille teetettiin kurssilla alku- ja lopputesti, jossa he pääsivät tuottamaan kertomuksia kirjallisesti ja suullisesti kuvasarjoja apuna käyttäen. Kuvasarjat olivat alku- ja lopputestauksessa samat, mutta kirjallisissa kertomuksissa kerrottiin eri kuvasarjoista kuin suullisissa kertomuksissa. Kirjallisten kertomusten kuvasarjoissa toisessa oli kuusi kuvaa ja aiheena oli tilanne, jossa kissan omistaja istuu vahingossa kissansa päälle. Toisessa kuvasarjassa oli puolestaan viisi kuvaa ja aiheena oli tilanne, jossa lintu karkaa häkistään. Kirjoitustehtäviin oli aikaa yksi tunti ja kuvasarjat olivat jokaisella opiskelijalla samat. (Alisaari–Heikkola 2016b, 278.) Suullisen kertomuksen kuvasarjoissa puolestaan ensimmäisessä sarjassa oli kuusi kuvaa ja aiheena leipomossa käynti ja leivän ostaminen. Toisessa kuvasarjassa oli viisi kuvaa ja aiheena oli tilanne, jossa kissa leikkii lankakerällä. Suullisissa testeissä opiskelijat kertoivat yksitellen haastattelijalle kuvasarjoista ja keskustelut nauhoitettiin sekä litteroitiin (Heikkola–Alisaari 2017, 25). Kuvasarjat löytyvät liitteistä 1–4.

Ennen kertomusten tuottamista oppijoita pyydettiin täyttämään taustatietolomake, jossa selvitettiin henkilötietoja ja tarkempia tietoja heidän kielitaidostaan ja taustoistaan. Tiedot kerättiin aikaisempia tutkimuksia varten, mutta koska ne eivät ole oman tutkimukseni kannalta relevantteja, en ole missään kohtaa tutkimusprosessin aikana tarkastellut näitä tietoja itse. Oppijoilta pyydettiin lupa vastausten käyttämiseen tutkimustarkoituksessa. Oppijoille myös ilmaistiin selkeästi, että heidän on mahdollista vetäytyä tutkimuksesta missä tahansa vaiheessa niin halutessaan. Tutkimuslupa kerättiin suostumuslomakkeella, jossa tiedot olivat sekä suomeksi että englanniksi, ja tällä pyrittiin varmistamaan se, että oppijat ymmärtävät, mikä aineistonkeruun tarkoitus on. Omassa tutkimuksessani oppijat on järjestetty sattumanvaraiseen järjestykseen eikä heitä tai heidän tekstejään voi mitenkään liittää henkilöön itseensä. Anonymiteetti säilyy siis koko tutkimuksen ajan. Leksikaalinen diversiteetti on lähtökohtaisesti vain yksi osa-alue oppijan kielitaidosta, eivätkä tulokset näin ollen kerro oppijasta itsestään tai hänen kokonaisosaamisestaan. Aineistonani käyttämät kertomukset on kerätty opiskelijoilta luvallisesti eikä yksikään heistä ole vetäytynyt tutkimuksesta. Tutkimuksessani en myöskään viittaa kertomusten sisältöön suoraan, joten aineistojen sisältö on nähtävissä vain numeerisessa muodossa eikä näin ollen paljasta mitään kertomusten tuottajasta.

Omassa tutkimuksessani tarkastelen leksikaalista diversiteettiä ja sen kehittymistä kahdentoista opiskelijan kirjoitetuissa ja suullisissa kertomuksissa. Varsinaisena tutkimusaineistonani on siis vain osa kaikista kerätyistä kertomuksista. Tarkastelen tutkimuksessani edistyneimmän oppijaryhmän tekstejä, joita on yhteensä 48. Jokainen oppija on osallistunut sekä kirjalliseen että suulliseen alku- ja lopputestaukseen, eli jokaisesta testistä on kertynyt 12 tekstiä. Vaikka kukin opiskelija kertoi ja kirjoitti kahdesta eri kuvasarjasta kerrallaan, olen tutkimuksessani luokitellut ne aina yhdeksi tekstiksi, sillä kuvasarjoista kertominen tai kirjoittaminen ovat tapahtuneet samalla testikerralla.

Opiskelijat kirjoittivat vastauksensa käsin paperille, mutta aiemman tutkimuksen yhteydessä kertomukset on kirjattu tietokoneella Word-tiedostoihin, joissa on pyritty noudattamaan opiskelijoiden omia asemointeja paperilla. Suulliset kertomukset on puolestaan nauhoitettu ja litteroitu tutkimusta varten. Litteraateissa on huomioitu myös haastattelijan repliikit, puheen tauot sekä miettimisäänteet ja sanojen toistot. Aiemmissa tutkimuksissa samasta aineistosta on hyödynnetty litteraatteja, joissa ei ole ylimääräisiä (miettimis)äänteitä tai haastattelijan puheenvuoroja, vaan pelkästään opiskelijan käyttämät sanat ja ilmaisut. Hyödynnän omassa tutkimuksessani näitä litteraatteja, sillä tutkimuksen kohteena on nimenomaan opiskelijoiden sanaston monimuotoisuus, ei sujuvuus tai tarkkuus. Omaa tutkimustani varten kuuntelin kaikki analyysin kohteena olevat äänitetyt kertomukset läpi, jotta sain paremman kuvan siitä, miten kertomukset todellisuudessa rakentuvat ja millaista sanastoa niissä käytetään. Samalla kävin läpi litteraattien muotoilua ja pyrin järjestämään tekstit niin, että lauseet ja virkkeet erottuisivat niistä mahdollisimman hyvin.

Aineiston analyysiä varten jokainen teksti lemmattiin, jotta tutkimuksessa pystyttiin keskittymään nimenomaan sanastolliseen vaihteluun, ei muotovaihteluun. Lemmauksella tarkoitetaan sitä, että kullekin sanalle/lekseemille määritellään lemma, joka edustaa sanan perus- tai hakumuotoa. Lemmojen avulla aineistoa on myös helpompi tarkastella, sillä eri lekseemit erottuvat paremmin toisistaan, ja koska tarkastelun kohteen on nimenomaan leksikaalinen diversiteetti, tarkoituksena on pohtia kokonaisten sanojen edustumista, ei niinkään sitä, miten opiskelija osaa käyttää taivutusmuotoja tai yhdistellä erilaisia kielellisiä elementtejä. Lemmatisointia varten kävin läpi kaikki tekstit ja korjasin tarvittaessa sanoja vastaamaan yleiskielen muotoja (esim. **poikan* > *pojan;* **ratiosta* > *radiosta*) sekä tarkensin sanoja (esim. *pela* > *pelaa*), jos niiden merkitys oli selvästi pääteltävistä tekstiyhteydestä. Jos sanan merkitystä ei ollut mahdollista päätellä edes tekstiyhteyden avulla, jätin sen siihen muotoon, jossa se oli alkuperäisessä tekstissä.

Tekstien lemmauksen suoritti pro gradu -tutkielmani ohjaaja, Turun yliopiston Suomen kielen ja suomalais-ugrilaisen kielentutkimuksen apulaisprofessori Ilmari Ivaska hyödyntämällä The TurkuNLP Groupin tutkijoiden luomaa parseria (ks. Kanerva ym. 2018, 133–142). Parserin avulla sanat muutettiin niiden perusmuotoon: nomineilla nominatiivi, verbeillä A-infinitiivi, taipumattomilla sanoilla perusmuoto, johdoksilla johdoksen nominatiivi ja yhdyssanoilla yhdyssanan nominatiivi. Samalla sanoille määriteltiin niiden sanaluokka, taivutustyyppi, persoonamuoto ja aikamuoto. Omaa tutkimustani varten hyödynnän kuitenkin vain perusmuotoisia lemmoja. Varmistaakseni, että lemmat ovat kohdillaan, kävin lemmauksen jälkeen kaikki tekstit ja sanojen lemmat läpi varmistaakseni, että sanat on lemmattu oikein eikä niiden kohdalla ollut tullut

epäselvyyksiä. Muutamassa kohdassa ohjelma ei ollut tunnistanut sanan perusmuotoa, joten ne kohdat muokkasin lopuksi itse oikeaan muotoon (esim. *poja* > *poika*).

5.2. Puhutut ja kirjoitetut tekstit aineistona

Suomenoppijoiden kohdalla keskitytään usein siihen, että heille karttuu tarpeeksi sanastoa ja he oppivat yleisimpiä sääntöjä, joiden avulla ilmausten tuottaminen ja ymmärtäminen on helpompaa. Monesti oppijat käyttävätkin yleiskieltä sekä kirjoittaessaan että puhuessaan (Masonen 2003, 93). Opetuksessa ja oppimateriaaleissa keskitytään monesti yleiskielen sanastoon ja sääntöihin, mutta moniin oppimateriaaleihin on tuotu mukaan tyypillisiä puhekielen piirteitä, joiden avulla oppijoiden on helpompi hahmottaa ja ymmärtää kieltä esimerkiksi autenttisissa kielenkäyttötilanteissa. Monesti yleiskielen opettelu riittää oppijalle alussa, mutta mitä pidemmälle oppija opinnoissaan etenee, sitä enemmän hän luultavasti haluaa oppia käyttämään molempia variantteja omassa kielenkäytössään (Martin 1999, 170). Omassa tutkimuksessani on tärkeää huomata, että monella informantilla on hallussaan jonkin verran myös puhekielen piirteitä ja tästä syystä esimerkiksi suullisissa kertomuksissa esiintyy muotoja *mä, sä* jne. Informantit ovat kuitenkin kirjoitetuissa teksteissään osoittaneet hallitsevansa myös yleiskieliset muodot *minä, sinä,* joten lemmatuissa versioissa puhekieliset muodot on muutettu vastaamaan yleiskielisiä vastineitaan.

Koska sekä puhuttu että kirjoitettu kieli ovat hyvin heterogeenisiä (Dufva 1995, 63), niiden vertailu saattaa olla tietyissä tilanteissa hyvinkin hankalaa. Omassa tutkimuksessani keskityn kuitenkin tekstien sanastoon ja leksikaaliseen diversiteettiin, jolloin huomio ei kiinnity niin vahvasti puhuttua ja kirjoitettua kieltä erottaviin tekijöihin. Koska aineisto on lemmattu suullisten ja kirjallisten tekstien osalta samoin, mahdolliset puhuttua tai kirjoitettua kieltä erottavat tekijät on pystytty yhdenmukaistamaan ja määrittelemään yleiskielisiksi vertailtavuuden parantamiseksi. Tärkeää on kuitenkin huomioida se, että vaikka vertailtavuus on näin toimivaa, puhutun ja kirjoitetun kielen tuottamisen taustalla on osin erilainen psykolingvistinen prosessi, mikä saattaa vaikuttaa leksikaaliseen diversiteettiin esimerkiksi sanojen toiston tai sanavalintojen osalta. Toisaalta tämä tarjoaa kuitenkin hyvän mahdollisuuden tarkastella sitä, miten mahdolliset erot kirjoitetuissa ja puhutuissa teksteissä näyttäytyvät, jolloin voidaan pohtia kokonaisuutena leksikaalisen diversiteetin kehittymistä ja muotoutumista.

Vaikka aineisto on suppea, se antaa hyvän lähtökohdan puhutun ja kirjoitetun kielen leksikaalisen diversiteetin tarkastelulle. Koska aiemmin on keskitytty vahvasti kirjoitetun kielen leksikaalisen diversiteetin tutkimukseen, on tärkeää perehtyä myös siihen, miten leksikaalinen diversiteetti näyttäytyy puhutussa kielessä. On toki mahdollista, että leksikaalinen diversiteetti kehittyy

puhutuissa ja kirjoitetuissa kertomuksissa samoin, jolloin eroja ei ole nähtävissä, mutta toisaalta tämäkin antaisi uutta tietoa siitä, että vaikka kielen tuottamisen prosessit olisivat osin erilaisia, sanasto ja sen tuottaminen voivat kuitenkin molemmissa tapauksissa kehittyä samalla tavalla.

5.3. Analyysimenetelmät

Logistisen regressioanalyysin tekemiseen hyödynnän SPSS-ohjelmaa (IBM SPSS Statistics 27). Logistinen regressioanalyysi valikoitui menetelmäksi, koska se on erittäin käyttökelpoinen silloin, kun selitettävä muuttuja on dikotominen eli kaksiluokkainen ja binäärinen. Lisäksi selittäviksi muuttujiksi on mahdollista valita sekä jatkuvia että luokitteluasteikollisia muuttujia, mikä tarjoaa mahdollisuuden käyttää laajasti erilaisia selittäviä muuttujia. (Jokivuori–Hietala 2007, 58–59.) Tutkimuksessani hyödynnän viittä eri selittävää muuttujaa ja tarkastelen niiden suhdetta leksikaalisen diversiteetin kehittymiseen. Jokaiselle aineiston tekstille lasketaan leksikaaliseen diversiteettiin liittyvät selittävien muuttujien arvot, joita hyödynnetään logistisessa regressioanalyysissa.

Tutkimuksessani selitettävä muuttuja (alku vs. loppu) on koodattu kaksiluokkaiseksi siten, että kurssin alussa tuotettuja kertomuksia edustaa arvo 0 ja kurssin lopussa tuotettuja kertomuksia arvo 1. Koska logistinen regressio ennustaa suuremman numeerisen arvon saavaa luokkaa, on järkevintä asettaa kurssin loppu ennustettavaksi kategoriaksi (Nummenmaa 2009, 337; Jokivuori–Hietala 2007, 58–59). Tällöin kurssin alku toimii referenssikategoriana ja logistinen regressio kertoo, millä todennäköisyydellä havainnot kuuluvat ennustettavaan kategoriaan (loppu) eivätkä referenssikategoriaan (alku). Samalla saadaan käsitystä siitä, miten leksikaalinen diversiteetti kurssin aikana kehittyy.

Tutkimuksessani selittävinä muuttujina ovat harvinaisuus, erityisyys, tasaisuus, sironta sekä vaihtelevuus. Muuttujat on operationalisoitu käyttämällä samaa scriptiä kuin Kajzer-Wietrznyn ja Ivaskan (2020) tutkimuksessa, mutta scripti on muokattu englanninkielisen aineiston sijaan suomenkieliseen aineistoon sopivaksi.

Harvinaisuutta (*rarity*) tarkastelemalla saadaan selville, kuinka yleisiä ja frekventtejä kertomuksissa käytetyt sanat ovat (Kajzer-Wietrzny–Ivaska 2020, 179). Harvinaisuutta voidaan mitata vertaamalla kertomuksissa käytettyjä sanoja suhteessa siihen, kuinka yleisesti käytettyjä sanat ovat kohdekielessä (ks. esim. Jarvis 2013b). Vertailua tehdään usein jonkin korpuksen avulla, jossa yleisin sana saa arvon 1, toiseksi yleisin arvon 2 ja niin edelleen. Tämän jälkeen lasketaan kunkin tekstin sanojen yleisyyden keskiarvo siten, että sanojen saamat arvot lasketaan yhteen ja jaetaan koko tekstin sanamäärällä. Omassa tutkimuksessani käytän frekvenssisanalistaa, joka koostuu yli 38 000 eri sanasta. Sanalista

on koostettu suomenkielisten kaunokirjallisten teosten pohjalta. Harvinaisuusaste määrittyy sen mukaan, kuinka mones sana sanastossa on, ja jos sanaa ei vertailukorpuksesta löydy, se saa korpuksen viimeisen sanan jälkeisen arvon. Tuloksia tarkastellessa voidaan todeta, että mitä korkeampi arvo muodostuu, sen suurempi leksikaalinen diversiteetti on (Kajzer-Wietrzny–Ivaska 2020, 179).

Vaihtelevuus (*variability*) liittyy koko sanastoon, ja se kuvaa eri sanojen kokonaismäärää ja sanojen sijoittumista tekstiin. Vaihtelevuutta voidaan mitata tarkimmin MTLD-arvolla eli sanojen monipuolisuusluvulla, sillä se ei ota vaikutteita tekstin pituudesta tai tasaisuudesta. Kun MTLD-arvo tekstissä kasvaa, sanasto on tiheämpää, jolloin myös leksikaalinen diversiteetti kasvaa. (Honko ym. 2019; Jarvis 2013b.)

Erityisyyden (*disparity*) kohdalla lasketaan, kuinka monta kertaa merkitykset aktivoituvat tekstissä eli kuinka monta eri merkitystä tekstin sanat saavat. Jos sanan merkitys aktivoituu useasti, siihen viitataan monta kertaa joko samalla sanalla tai sanan synonyymeillä ja tällöin arvo kasvaa. Mitä pienempi arvo on, sen suurempaa on leksikaalinen diversiteetti, koska saman sanan merkitykset eivät toistu jatkuvasti (Kajzer-Wietrzny–Ivaska 2020, 179.) Erityisyyden arvojen laskemiseen hyödynnetään WordNetin (versio 2.0) suomenkielistä mallia, joka on leksikaalinen tietokanta. FinnWordNetissä sanat on luokiteltu synonyymijoukoiksi, jotka edustavat tiettyjä käsitteitä. Synonyymijoukot linkittyvät toisiinsa monin eri tavoin ja muodostavat semanttisen verkon. Tätä verkkoa hyödyntämällä on kustakin tekstistä pyritty laskemaan se, miten monia eri merkityksiä tekstin sanat aktivoivat. (Kielipankki 2019.)

Tasaisuus (*evenness*) kertoo siitä, kuinka tasaisesti tekstin eri sanamuodot esiintyvät tekstissä (Kajzer-Wietrzny–Ivaska 2020, 178–179). Mittaamalla tasaisuutta voidaan tarkastella sitä, esiintykö sana tekstissä kerran, kaksi kertaa vai useamman kertaa. Mittari siis pystyy arvioimaan tarkan arvon sanojen tasaisuudelle, ei pelkästään sitä, esiintyykö sana kerran vain monta kertaa. Mitä useammin sana tekstissä esiintyy, sitä toisteisempaa teksti on, ja samalla se vähentää tekstin leksikaalista diversiteettiä. Toisin sanoen, jos tekstissä esiintyvät sanat esiintyvät tasaisesti ja yksittäin, tekstin leksikaalinen diversiteetti on suurempi kuin jos sama sana esiintyy tekstissä monta kertaa. Arvoja tarkastellessa siis korkeammat arvot kertovat suuremmasta leksikaalisesta diversiteetistä.

Sironta (*dispersion*) on samankaltainen tasaisuuden kanssa, mutta se osoittaa sanojen sijoittumisen tekstissä laskemalla tietyn sanan eri esiintymien etäisyyden tai kasaantumisen tarkasteltavassa tekstissä. Oleellista on, että mitä säännöllisemmin samaa lekseemiä edustavat saneet sijoittuvat tekstissä, sen monimuotoisempaa teksti on. Sironnassa tarkastellaan sanakimppuja, joissa sama sana esiintyy kahdesti kahdenkymmenen sanan otoksen sisällä. Mittaaminen tapahtuu niin, että katsotaan,

kuinka monta sanakimppua tekstissä esiintyy sadan sanan otoksen sisällä. (Honko ym. 2019, 58–59.; Jarvis 2013b, 25.)

5.4. Logistisen regressiomallin muodostaminen

Koska tutkimuksessa on tarkoitus tarkastella kirjallisia ja suullisia kertomuksia sekä erikseen että yhdessä, tehdään ensin logistinen regressiomalli, jossa mukana on kaikki kertomukset. Tarkoituksena on tarkastella sitä, toimiiko malli ylipäänsä aineistoon. Sen jälkeen mallia hyödynnetään sekä kirjallisiin että suullisiin kertomuksiin erikseen ja analysoidaan näistä saatuja tuloksia. Kuten jo edeltä kävi ilmi, logistisessa regressioanalyysissa pyritään selvittämään todennäköisyys sille, että jotakin tapahtuu. Hyödyntämällä selittäviä muuttujia pyritään tarkastelemaan sitä, millainen vaikutus niillä on selitettävään muuttujaan ja miten ne ovat muuttuneet kurssin kuluessa. Tutkimuksessani käytetyt muuttujat selviävät tiivistetysti taulukosta 1 tarkemmin.

TAULUKKO 1: Logistisen regressioanalyysin muuttujat

	Muuttuja	**Selitys**	**Luokittelu**	**Muutoksen suunta**
Selitettävä muuttuja	Ajankohta	Tekstin kirjoitusajankohta	0=alussa, 1=lopussa	
Selittävät muuttujat	Vaihtelevuus	Sanojen kokonaismäärä ja sijoittuminen tekstissä	jatkuva 0–X vaihteluväli 13,52–69,45	Suurempi arvo viittaa leksikaalisen diversiteetin kasvuun.
	Tasaisuus	Kuinka tasaisesti sanat esiintyvät tekstissä	jatkuva 0–X vaihteluväli: 0,90–0,96	Suurempi arvo viittaa leksikaalisen diversiteetin kasvuun.
	Harvinaisuus	Sanojen yleisyys ja frekventtiys teksteissä	jatkuva 0–X vaihteluväli 23504–32236	Suurempi arvo viittaa leksikaalisen diversiteetin kasvuun.
	Sironta	Sanojen eri esiintymien etäisyys ja kasautuminen tekstissä	jatkuva 0–99 vaihteluväli 20,18–53,02	Suurempi arvo viittaa leksikaalisen diversiteetin vähenemiseen.
	Erityisyys	Kuinka monta eri merkitystä yhden tekstin sanat aktivoivat	jatkuva 0–X vaihteluväli 1,15–1,45	Suurempi arvo viittaa leksikaalisen diversiteetin vähenemiseen.

Logistisen regressiomallin muodostamista varten leksikaalisesta diversiteetistä kertovat arvot on syötetty ensin Excel-taulukkoon, jonka jälkeen arvot on haettu sieltä suoraan SPSS-ohjelmaan. SPSS:ssä jokainen muuttuja on koodattu tarpeen mukaan, jonka jälkeen on suoritettu logistinen regressioanalyysi. Aiempien tutkimusten perusteella on osoittautunut, että omaan tutkimukseeni valikoidut selittävät muuttujat kertovat oppijan leksikaalisesta diversiteetistä (ks. esim. Jarvis 2013a, 2013b; Honko ym. 2019; Kajzer-Wietrzny–Ivaska 2020), minkä vuoksi tutkimuksessa on käytetty ”Enter”-menetelmää, joka ottaa analyysissa heti mukaan kaikki selittävät muuttujat ja malli muodostetaan niitä hyödyntämällä. Tavoitteena on muodostaa kuva siitä, että miten kertomukset sijoittuvat kurssilla ja miten muuttujat vaikuttavat siihen.

Ensimmäisen logistisen regressiomallin muodostamiseen on käytetty sekä kirjallisia että suullisia kertomuksia, jotta otoskoosta on saatu laajempi ja on pystytty tarkastelemaan, kuinka hyvin malli toimii selittävien muuttujien kanssa. Aineistossa on mukana yhteensä 48 kertomusta, joista 24 kirjallisia ja 24 suullisia. SPSS-ohjelmistossa on otettu mukaan kaikkien näiden kertomusten arvot eli selittävät muuttujat, joiden perusteella logistinen regressiomalli on muodostettu. Sen jälkeen on tehty vielä omat mallit siten, että toisessa mallissa on huomioitu vain kirjalliset kertomukset ja toisessa suulliset kertomukset.

Seuraavassa luvussa 6 tarkastelen ensin logistista regressiomallia, jossa on mukana kaikki kertomukset (N=48) ja sen jälkeen tarkastelen erikseen regressiomalleja kirjallisista ja suullisista kertomuksista, sekä sitä, miten malli selittää tekstin sijoittumista kurssin alkuun tai loppuun. Sen jälkeen luvussa 7 vertailen saatuja tuloksia ja pohdin, onko leksikaalisen diversiteetin kehitys samansuuntaista molemmissa kielenkäyttötilanteissa vai poikkeavatko muutokset jollain tavoin toistaan. Lopuksi luvussa 8 tarkastelen vielä saatuja tuloksia ja pohdin jatkotutkimusmahdollisuuksia.

6. Logistiset regressioanalyysit kertomuksista

6.1. Logistisen regressiomallin tarkastelua

Logistisessa regressioanalyysissa tulee ottaa huomioon se, miten hyvin malli toimii annettujen muuttujien kanssa ja kuinka paljon kokonaisuudessaan malli selittää selitettävän muuttujan vaihtelua. SPSS tulostaakin tätä varten useita erilaisia taulukoita, joiden avulla voidaan arvioida mallin sopivuutta, selitysastetta ja ennustustarkkuutta. (Nummenmaa 2009, 337–339.)

Beginning Block 0 -osuus sisältää ainoastaan malliin sisältyvän vakion eli tässä tapauksessa kertomuksen tuottamisen ajankohdan (alussa vs. lopussa). Siinä ei siis ole mukana ollenkaan selittäviä muuttujia. Taulukkoon 2 on koottu vakion sisältämän mallin luokittelu, joka ilmaisee, kuinka paljon malli luokittelee tekstejä alkuun ja loppuun sekä kuinka suuri osuus prosentuaalisesti teksteistä luokitellaan oikein.

TAULUKKO 2: Vain vakion sisältämän mallin luokittelu

Classification Table[a,b]

Observed			Predicted: Tekstin tuottoajankohta		
			Teksti tuotettu alussa	Teksti tuotettu lopussa	Percentage Correct
Step 0	Tekstin tuottoajankohta	Teksti tuotettu alussa	0	24	,0
		Teksti tuotettu lopussa	0	24	100,0
	Overall Percentage				50,0

a. Constant is included in the model.

b. The cut value is ,500

Kuten taulukosta 2 voidaan huomata, vakion sisältämä malli luokittelee kaikista kertomuksista oikein 50 prosenttia. Lopussa tuotetuista kertomuksista malli luokittelee oikein kaikki 24, mutta alussa tuotetuista kertomuksista malli ei luokittele oikein yhtäkään tekstiä. Ilman selittäviä muuttujia mallin luokittelu on sattumanvaraista, ja koska kokonaisselitysaste on 50 prosenttia, mallin selitysaste ei ole merkityksellinen.

Kun mukaan otetaan kaikki selittävät muuttujat, voidaan tarkastella sitä, onko malli tilastollisesti merkitsevä ja millainen selitysaste mallilla on (Metsämuuronen 2008, 130). Mallin hyvyyttä voidaan

arvioida muun muassa selitysasteiden perusteella, josta kertovat *Cox & Snell- ja Nagelkerke R-square* -arvot. Mitä lähempänä arvo on yhtä (1), sitä paremmin selittävät muuttujat selittävät vaihtelua selitettävässä muuttujassa. (Jokivuori–Hietala 2007, 68; Metsämuuronen 2008, 130.) Nämä arvot löytyvät taulukosta 3.

TAULUKKO 3: Mallin selitysasteet

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	58,334[a]	,157	,210

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Taulukossa 3 esitettyjen arvojen perusteella muuttujien avulla pystytään selittämään jonkin verran ennustettavaa ilmiötä, mutta ei kaikkea siitä. Prosentuaalisesti malli selittää selitettävän muuttujan vaihtelua välillä 15,6–21,0 %. Usein huomio kiinnitetään nimenomaan Nagelkerken muunnettuun neliöarvoon, sillä sen on mahdollista saavuttaa arvo 1 (Metsämuuronen 2008, 130). Tutkittavan mallin tapauksessa Nagelkerke R^2 on 0,210, joten malliin lisättyjen muuttujien avulla pystytään arvioimaan ainakin jossain määrin tekstin sijoittuminen kurssin alkuun tai loppuun.

Mallin hyvyyttä arvioidessa voidaan tarkastella lisäksi *Hosmer & Lemeshowin testiä*, joka testaa sitä, miten hyvin arvot jakautuvat oikeisiin kategorioihin (Metsämuuronen 2008, 131). Se siis ennustaa selitysvoimaisesti sitä, milloin teksti on kirjoitettu. Hosmer & Lemeshowin testissä p-arvon tulee olla suurempi kuin 0.05, jotta malli soveltuu hyvin aineistoon. (Jokivuori–Hietala 2007, 68; Metsämuuronen 2008, 131.) Testin perusteella merkitsevyys on 0,388. Arvo on reilusti yli 0,05, joten malli selittää hyvin kertomusten esiintymistä kurssin lopussa.

Mallin sopivuutta aineistoon voidaan tarkastella myös sitä, miten hyvin selittävät muuttujat sisältävä malli luokittelee kertomukset alkuun tai loppuun. Selittäviä muuttujia apuna käyttäen malli laskee kertomukselle todennäköisyyden siitä, kumpaan ryhmään kertomus kuuluu. Jos $P < 0.5$ kertomus kuuluu alkuryhmään, mutta jos $P > 0.5$ kertomus kuuluu loppuryhmään. Jos P olisi 0,5, niin kumpikin tapahtuma olisi yhtä todennäköinen. (Jokivuori–Hietala 2007, 69–70.) Taulukkoon 4 on koostettu

kokonaisluokittelun tulos eli se, miten kertomukset jakautuvat todennäköisyyksien mukaan alkuun tai loppuun. Tarkat todennäköisyydet löytyvät taulukosta 17.

TAULUKKO 4: Selittävät muuttujat sisältävä malli

Classification Table[a]

Observed			Predicted: Tekstin tuottoajankohta: Teksti tuotettu alussa	Predicted: Tekstin tuottoajankohta: Teksti tuotettu lopussa	Percentage Correct
Step 1	Tekstin tuottoajankohta	Teksti tuotettu alussa	15	9	62,5
		Teksti tuotettu lopussa	6	18	75,0
	Overall Percentage				68,8

a. The cut value is ,500

Tarkastelemalla taulukkoa 4, voidaan huomata, että selittävät muuttujat sisältävässä mallissa kokonaisselitysaste on 68,8 prosenttia, joten luokitus onnistuu suhteellisen hyvin. Tuloksista voidaan havaita, että malli onnistuu selittämään huomattavasti paremmin tekstien sijoittumisen alkuun ja loppuun tässä toisin kuin pelkän vakion sisältämässä mallissa, jossa luokittelu oli satunnaista, sillä tässä mallissa 62,5 % alussa tuotetuista teksteistä ja 75 % lopussa tuotetuista on luokiteltu oikein. Koska kokonaisselitysaste on 68,8 % selittävät muuttujat sisältävässä mallissa, voidaan todeta, että malli toimii ja sen avulla pystytään luokittelemaan reilusti yli puolet kertomuksista oikein.

Tärkein tuloste logistisessa regressioanalyysissa on *Variables in the Equation*, joka kertoo siitä, kuinka todennäköistä on, että tapahtuma toteutuu (Jokivuori–Hietala 2007, 70). Kertomusten osalta tarkoituksena on tarkastella sitä, miten kukin selittävä muuttuja muuttuu riippuen siitä, onko kertomus tuotettu intensiivikurssin alussa vai lopussa. Taulukkoon 5 on koottu logistisen regressiomallin kertoimet ja niiden merkitsevyys kirjallisissa ja suullisissa kertomuksissa.

TAULUKKO 5: Logistisen regressiomallin kertoimet ja niiden merkitsevyys kirjallisissa ja suullisissa kertomuksissa

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Vaihtelevuus	,0382	,077	,244	1	,621	1,0389	,893	1,209
	Tasaisuus	-2,4261	48,108	,003	1	,960	,0884	,000	7,878E+39
	Harvinaisuus	-,0002	,000	,685	1	,408	,9998	,999	1,000
	Sironta	-,0232	,122	,036	1	,850	,9771	,769	1,241
	Erityisyys	-19,2102	10,342	3,450	1	,063	,0000	,000	2,887
	Constant	31,8438	52,751	,364	1	,546	6.7543E+13		

a. Variable(s) entered on step 1: Vaihtelevuus, Tasaisuus, Harvinaisuus, Sironta, Erityisyys.

Taulukossa 5 on nähtävissä jokaisen selittävän muuttujan regressiokerroin (B), regressiokertoimen keskivirhe (S.E.), Waldin testisuure (testaa nollahypoteesia), Waldin testisuureen merkitsevyystaso (Sig. = p-arvo) sekä odds ratio eli todennäköisyyksien suhde (Exp(B)), joka ilmaisee suhteellista riskiä kuulua ennustettavaan ryhmään (Jokivuori–Hietala 2007, 71; Rasi–Kanniainen 2007, 19).

Tapahtuman todennäköisyyttä eli kertomuksen esiintymistä kurssin lopussa, voidaan laskea kaavalla

$$P(y = 1|x) = e^z / 1 + e^z = 1 / 1 + e^{-z}$$

Kaavassa *e* on luonnollinen logaritmi eli Neperin luku (noin 2,718). *Z* puolestaan ilmaisee yhtälöä $B + B_1*X_1 + B_2+*X_2 + \ldots + B_p*X_p$, jossa *B* on regressiokerroin ja *X* selittävä muuttuja. (Metsämuuronen 2008, 117–118; Rasi–Kanniainen 2007, 20–21.) Tarkastelen kertomusten todennäköisyyksiä vielä tarkemmin kirjallisten ja suullisten kertomusten osalta luvuissa 6.2. ja 6.3.

Regressiokerroin (B) ilmaisee sitä, miten paljon selitettävän muuttujan arvot kasvavat silloin, kun selittävän muuttujan arvot kasvavat yhdellä yksiköllä (Metsämuuronen 2008, 91) sekä sitä, mihin suuntaan selittävän muuttujan arvon odotetaan loppukertomuksissa muuttuvan. (Jokivuori–Hietala 2007, 70–71.) Jos arvo on positiivinen, odotuksenmukaista on, että selittävän muuttujan arvo kasvaa kurssin lopussa. Jos arvo on negatiivinen, selitettävän muuttujan arvon voidaan olettaa vähenevän kurssin lopussa. Kun tarkastellaan mallin regressiokertoimia, voidaan huomata, että vain selittävän muuttujan *vaihtelevuus* arvo on positiivinen, mikä tarkoittaa sitä, että todennäköisyys kuulua lopussa tuotettujen kertomusten joukkoon kasvaa, jos arvo kasvaa. Muiden selittävien muuttujien, *tasaisuus, harvinaisuus, sironta* ja *erityisyys*, arvo on negatiivinen, mikä puolestaan kertoo siitä, että arvon voidaan olettaa pienenevän, kun leksikaalinen diversiteetti kasvaa. Tässä pitää kuitenkin huomioida se, että selitettävän ja selittävien muuttujien yhteydet eivät ole lineaarisia, joten selittävän muuttujan arvojen kasvaminen vaihtelee muuttujan eri arvoilla. Pienimmillä ja suurimmilla arvoilla

todennäköisyys muuttuu siis vähän, mutta keskivaiheella arvojen muutokset vaikuttavat ennustetun lopputuloksen todennäköisyyteen huomattavasti. (Nummenmaa 2009, 340.) Oman tutkimukseni osalta on muistettava se, että selittävien muuttujien arvot ovat keskenään hyvin erilaisia, joten toisille muuttujille yhden yksikön muutos tarkoittaa todella suurta muutosta, kun taas toisille muuttujille yhden yksikön muutoksella ei ole juurikaan merkitystä.

Vaihtelevuuden (B=0,0382 / Exp(B)=1,0389) osalta todennäköisyys kuulua lopussa kirjoitettujen kertomusten ryhmään kasvaa 3,89 %, kun muuttujan arvo kasvaa yhdellä yksilöllä. Harvinaisuuden (B=-0,0002 / Exp(B)=0,9998) osalta puolestaan merkitys on suhteellisen vähäinen, sillä jos arvo kasvaa, niin todennäköisyys kuulua loppuryhmään on 0,02 % matalampi. Todennäköisyys kuulua loppuryhmään on matalampi myös sironnan (B=-0,0232 / Exp(B)=0,9771) ja erityisyyden (B=-0,0002 / Exp(B)=0,0000000045) kohdalla. Jos siis muuttujan sironta arvo nousee yhdellä, niin todennäköisyys kuulua loppuryhmään on matalampi 2,29 % ja erityisyyden tapauksessa todennäköisyys on 99,99 % matalampi. Syy sille, että erityisyyden muutos näyttäytyy suurena, liittyy todennäköisesti siihen, että arvojen vaihteluväli on todella pieni (1,25–1,45), joten arvon nousu yhdellä tekisi huomattavan muutoksen todennäköisyyteen. Tasaisuuden ((B=-2,4261 / Exp(B)=0,0884) osalta malli näyttäisi osoittavan, että arvon kasvaminen yhdellä vähentäisi todennäköisyyttä kuulua loppuryhmään 91,16 %, mutta tässäkin on otettava huomioon se, että vaihteluväli tasaisuuden arvoilla on hyvin suppea (0,9–0,97), joten todennäköisyys muutokselle on pieni.

Kaikista selittävistä muuttujista lähimpänä tilastollista merkitsevyyttä on *erityisyys* (p = 0,063). Mallin mukaan se siis lisää parhaiten mallin kykyä lajitella tai ennustaa havainnot oikeisiin selitettävän muuttujan luokkiin. Muiden selittävien muuttujien merkitsevyys on mallissa reilusti yli 0.05, joten ne eivät ole tilastollisesti merkitseviä. Kun tarkastelen kirjallisten ja suullisten kertomusten logistisia regressiomalleja (luvut 6.2. ja 6.3.), jätän p-arvojen tarkastelun pois, sillä yksittäisen muuttujan vaikutus muiden pysyessä vakiona ei tässä tutkimuksessa ole keskeistä.

6.2. Kirjallisten kertomusten tarkastelua logistisella regressiomallilla

Kirjallisia kertomuksia on mukana 24, joista 12 on kirjoitettu kurssin alussa ja 12 kurssin lopussa. Kun kirjallisia kertomuksia tarkastellaan logistisen regressioanalyysin avulla, voidaan todeta, että logistinen regressiomalli luokittelee kaikista kertomuksista oikein 79,2 %. Luokitus onnistuu tässä paremmin kuin silloin, jos mukana olisi myös suullisten kertomusten arvot. Alussa kirjoitetuista kertomuksista malli luokittelee oikein 9 eli 75 %, kun taas lopussa kirjoitetuista kertomuksista malli luokittelee oikein 10 eli 83,3 %. Käytettyjen muuttujien avulla malli selittää kirjallisten kertomusten

sijoittumisen (alku vs. loppu) paremmin kuin silloin, kun mukana on myös suulliset kertomukset, sillä kirjallisten kertomusten logistinen regressiomalli antaa Nagelkerke R^2 -arvoksi on 0,287. Taulukkoon 6 on koostettu kirjallisten kertomusten regressiomallin kertoimet.

TAULUKKO 6: Logistisen regressiomallin kertoimet

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Vaihtelevuus	,2207	,150	2,175	1	,140	1,2470	,930	1,672
	Tasaisuus	-137,4452	117,331	1,372	1	,241	,0000	,000	1,517E+40
	Harvinaisuus	-,0002	,000	,466	1	,495	,9998	,999	1,000
	Sironta	-,0541	,188	,083	1	,774	,9474	,655	1,370
	Erityisyys	,6929	14,909	,002	1	,963	1,9995	,000	9.801E+12
	Constant	128,2202	108,335	1,401	1	,237	4,8455E+55		

a. Variable(s) entered on step 1: Vaihtelevuus, Tasaisuus, Harvinaisuus, Sironta, Erityisyys.

Kuten taulukosta 6 nähdään, minkään selittävän muuttujan merkitsevyystaso ei ole tilastollisesti merkitsevä, mikä selittyy osittain aineiston pienuudella. Malli kuitenkin antaa hieman suuntaa siitä, miten muuttujat selittävät kirjoitettujen kertomusten sijoittumista kurssilla ja sitä kautta myös leksikaalisen diversiteetin kehittymistä kurssin aikana. Kirjallisissa kertomuksissa selittävän muuttujan vaihtelevuus regressiokerroin on positiivinen (B=0,2207), mikä tarkoittaa sitä, että todennäköisyys selitettävän muuttujan arvolle 1 (loppu) kasvaa, kun selittävän muuttujan arvo kasvaa. Lopussa vaihtelevuuden arvo on korkeampi kuin alussa, sillä Exp(B) = 1,2470. Myös erityisyyden (B=0,6929) arvo näyttäytyy siten, että lopussa tuotetuissa kertomuksissa arvo kasvaa melkein kaksinkertaiseksi (Exp(B) = 1,9995). Täytyy kuitenkin huomioida se, että kirjallisissa kertomuksissa erityisyyden arvojen vaihtelu on hyvin suppeaa ja monen kirjoittajan kohdalla arvot ovat alussa ja lopussa hyvin lähellä toisiaan. Muutamalla kirjoittajalla kuitenkin lopussa kirjoitetun tekstin arvo on suurempi kuin alussa, mikä osittain selittää sitä, että riskiluku on positiivinen, vaikka oletuksena on, että erityisyyden arvon tulisi pienentyä loppua kohden, jos leksikaalinen diversiteetti näiltä osin kasvaa.

Muiden selittävien muuttujien osalta voidaan todeta, että regressiokertoimet ovat negatiivisia, mikä viittaa siihen, että selittävän muuttujan arvon tulisi pienentyä, kun ennustetaan kertomuksen sijoittamista kurssin loppuun. Selittävän muuttujan arvon kasvu yhdellä vaikuttaa siis siten, että suhteellinen riski kuulua loppuryhmään on matalampi, ei korkeampi. Erityisesti sironnan (B=-0,0541) kohdalla tämä on nähtävissä, sillä riski (Exp(B)=0,9474) on 5,26 % matalampi, jos muuttujan arvo kasvaa. Harvinaisuuden (B=-0,0002) kohdalla muuttujan arvon kasvu yhdellä yksiköllä

vaikuttaa siten, että riski kuulua loppuryhmään on 0,02 % matalampi (Exp(B)=0,9998). Tasaisuuden (B=-137,4452) kohdalla näyttää siltä, että koska regressiokerroin on negatiivinen, selittävän muuttujan arvot ovat pienempiä lopussa kirjoitetuissa kertomuksissa. Tämä on kuitenkin vastakkain sen kanssa, että arvon tulisi kasvaa silloin, kun leksikaalinen diversiteetti kehittyy.

Taulukosta 7 nähdään, millaisella todennäköisyydellä kukin teksti sijoittuu alussa tai lopussa kirjoitettujen kertomusten ryhmään. Ennustetut arvot on luokiteltu kaksiluokkaisiksi (0 ja 1), joten kun todennäköisyys on suurempi kuin 0,5, niin kertomuksen kirjoitusajankohdan ennusteeksi tulee 1 eli kertomus on kirjoitettu kurssin lopussa. Jos todennäköisyys on pienempi kuin 0,5, niin ennusteeksi tulee 0 eli kertomus on kirjoitettu kurssin alussa.

TAULUKKO 7: Kirjoitettujen kertomusten ennusteet

Informantti	Kirjoitus-ajankohta (alku)	Todennäköisyys	Ennuste	Kirjoitus-ajankohta (loppu)	Todennäköisyys	Ennuste
Inf1	0	0,14452	0	1	0,52496	1
Inf2	0	0,30654	0	1	0,69595	1
Inf3	0	0,30262	0	1	0,21696	0
Inf4	0	0,55731	1	1	0,50149	1
Inf5	0	0,35153	0	1	0,46026	0
Inf6	0	0,31225	0	1	0,50711	1
Inf7	0	0,16482	0	1	0,7986	1
Inf8	0	0,25301	0	1	0,99781	1
Inf9	0	0,55274	1	1	0,7644	1
Inf10	0	0,38699	0	1	0,54647	1
Inf11	0	0,37728	0	1	0,52609	1
Inf12	0	0,90781	1	1	0,84247	1

Taulukosta 7 on korostettu ne viisi kertomusta, jotka mallin ennusteen mukaan poikkeaisivat oikeasta kirjoitusajankohdastaan. Huomionarvoista on se, että teksteistä yhdeksäntoista luokitellaan oikein ja vain viisi väärin. Tärkeä huomio on myös se, että mallin viidestä väärin ennustamasta todennäköisyydestä kolme on hyvin lähellä merkitsevää 0,5, joka erottaa sen, kumpaan luokkaan teksti sijoitetaan. Vain kaksi kertomuksista (12 ja 15) on ennustettu reippaasti väärään kategoriaan todennäköisyyksien ollessa 0,90781 ja 0,21696. Voidaankin siis todeta, että kirjallisten kertomusten osalta malli onnistuu selittävien muuttujien avulla luokittelemaan kertomukset kattavasti oikeisiin ryhmiin, joten selittävät muuttujat kuvaavat luotettavasti alussa ja lopussa kirjoitettujen tekstien eroja.

6.3. Suullisten kertomusten tarkastelua logistisella regressiomallilla

Suullisia kertomuksia on mukana 24, joista 12 on tuotettu kurssin alussa ja 12 kurssin lopussa. Logistinen regressioanalyysi, samoilla selittävillä muuttujilla kuin aikaisemminkin, onnistuu luokittelemaan kaikista tuotetuista kertomuksista oikein 83,3 %. Kokonaisluokittelu on siis parempi suullisten kertomusten kohdalla kuin kirjallisten kertomusten osalta. Malli luokittelee alussa tuotetuista kertomuksista oikein 10 ja lopussa kirjoitetuista myös 10 eli molemmissa tapauksissa 2 kertomusta luokitellaan väärään ryhmään kuuluvaksi. Suullisten kertomusten logistisessa regressiomallissa näyttää siltä, että malli onnistuu tekstien luokittelussa hyvin, sillä Nagelkerke R^2 -arvo on 0,648. Taulukossa 8 on nähtävillä suullisten kertomusten logistisen regressiomallin kertoimet.

TAULUKKO 8: Logistisen regressiomallin kertoimet ja niiden merkitsevyys kirjallisissa kertomuksissa

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Vaihtelevuus	,1634	,335	,237	1	,626	1,1775	,610	2,272
	Tasaisuus	-112,2596	125,203	,804	1	,370	,0000	,000	6,593E+57
	Harvinaisuus	-,0010	,001	2,814	1	,093	,9990	,998	1,000
	Sironta	-,1995	,290	,472	1	,492	,8192	,464	1,447
	Erityisyys	-50,6163	31,597	2,566	1	,109	,0000	,000	81865,226
	Constant	205,4664	160,188	1,645	1	,200	1,7097E+89		

a. Variable(s) entered on step 1: Vaihtelevuus, Tasaisuus, Harvinaisuus, Sironta, Erityisyys.

Taulukkoa 8 tarkastelemalla huomataan, että myöskään suullisten kertomusten kohdalla mikään selittävä muuttuja ei ole tilastollisesti merkitsevä. Näyttää kuitenkin siltä, että kunkin muuttujan regressiokertoimet ja riskiluvut ovat samansuuntaisia kuin kirjoitetuissa kertomuksissa.

Lopussa tuotetuissa kertomuksissa vaihtelevuuden regressiokerroin (B) on 0,1634 ja riskiluku 1,1775. Voidaan siis todeta, että myös suullisissa kertomuksissa vaihtelevuuden arvo kasvaa lopussa tuotetuissa kertomuksissa, jolloin myös leksikaalinen diversiteetti kasvaa. Sekä erityisyyden että tasaisuuden osalta suhteellinen riski Exp(B) näyttäytyy taulukossa 0. Regressiokertoimien perusteella voidaan kuitenkin päätellä, että molempien selittävien muuttujien, erityisyyden (B=-50,6163) ja tasaisuuden (B=-112,2596), arvon pitäisi pienentyä, kun leksikaalinen diversiteetti kehittyy. Erityisyyden osalta tämä pitää paikkansa, sillä leksikaalisen diversiteetin kasvaessa erityisyyden arvo pienenee. Tasaisuuden osalta pitäisi kuitenkin olla niin, että arvon kasvu osoittaisi diversiteetin

kehittymistä. Tasaisuuden regressiokerroin on negatiivinen myös kirjallisten kertomusten mallissa, joten tämä näyttäytyy kuitenkin hyvin samankaltaisena sen kanssa.

Harvinaisuuden regressiokerroin on -0,0010. Negatiivinen luku osoittaa sen, että selittävän muuttujan arvo pienenee, kun ennustetaan tekstin sijoittumista kurssin loppuun. Riskiluku harvinaisuudelle on 0,9990 eli riski kuulua loppuryhmään on 0,1 % matalampi, jos harvinaisuuden arvo kasvaa. Harvinaisuuden osalta muutos on samankaltainen myös kirjallisissa kertomuksissa, sillä niissä riski kuulua loppuryhmään on 0,2 % matalampi, kun arvo kasvaa.

Kuten kirjallisissa kertomuksissa, niin myös suullisissa kertomuksissa sironnan (B=-0,1995) merkitys on nähtävissä niin, että jos sironnan arvo kasvaa yhdellä yksiköllä, suhteellinen riski kuulua loppuryhmään madaltuu 18,08 %, sillä riskiluku on 0,8192. Taulukkoon 9 on koostettu suullisten kertomusten todennäköisyydet ja ennusteet.

TAULUKKO 9. Suullisten kertomusten ennusteet

Informantti	Tuotto-ajankohta (alku)	Todennäköisyys	Ennuste	Tuotto-ajankohta (loppu)	Todennäköisyys	Ennuste
Inf1	0	0,31193	0	1	0,29719	0
Inf2	0	0,06576	0	1	0,95378	1
Inf3	0	0,00547	0	1	0,92771	1
Inf4	0	0,17095	0	1	0,80384	1
Inf5	0	0,10095	0	1	0,66783	1
Inf6	0	0,06432	0	1	0,31664	0
Inf7	0	0,00594	0	1	0,76653	1
Inf8	0	0,00631	0	1	0,87413	1
Inf9	0	0,36888	0	1	0,81062	1
Inf10	0	0,03526	0	1	0,88947	1
Inf11	0	0,8266	1	1	0,94787	1
Inf12	0	0,81107	1	1	0,97095	1

Suullisten kertomusten ennusteissa neljä kertomusta asettuu väärään luokkaan. Kirjallisissa kertomuksissa viidestä väärin luokitellusta kertomuksesta kolme oli kuitenkin aika lähellä oikeaa ennustettaan, kun taas suullisten kertomusten väärin luokitelluista kertomuksista kaikkien neljän todennäköisyydet ovat suuresti alle tai yli 0,5, joten kertomusten arvot ovat olleet muihin kertomuksiin verrattuina poikkeavia. Kaikkiaan kuitenkin 20 kertomusta koko suullisten kertomusten joukosta luokitellaan oikein, joten malli toimii myös tässä selkeyttämään sitä, että selittävillä muuttujilla on vaikutusta siihen, miten kertomus sijoittuu kurssilla.

7. Leksikaalinen diversiteetti kirjoitetuissa ja puhutuissa kertomuksissa

Tässä luvussa tarkastelen vielä tarkemmin kutakin leksikaalisen diversiteetin osa-aluetta kirjallisten ja suullisten kertomusten osalta. Hyödynnän apuna sekä logististen regressiomallin tuloksia että koosteita kullekin kertomukselle lasketuista leksikaalisen diversiteetin osa-alueiden arvoista. Otan tässä huomioon myös sanemäärän, joskin kuten aikaisemmin on todettu, sen merkitys leksikaalisen diversiteetin osana ei ole yksinään merkittävä. Sanemäärää kuitenkin hyödynnetään muissa mittareissa, joten tarkastelen sitä tässä ensin. Sanemäärät kertomustyypeittäin on koottu taulukkoon 10.

TAULUKKO 10: Sanemäärät tuotetuissa kertomuksissa

	Kirjallinen		**Suullinen**	
Informantti	**Alku**	**Loppu**	**Alku**	**Loppu**
Inf1	131	136	185	220
Inf2	183	197	179	145
Inf3	121	140	170	180
Inf4	172	167	213	201
Inf5	146	167	247	237
Inf6	184	219	170	203
Inf7	110	140	186	186
Inf8	190	228	407	472
Inf9	248	217	277	264
Inf10	154	195	125	187
Inf11	122	193	165	205
Inf12	101	132	177	183

Sanemääriltään kirjallisten kertomusten pituudet vaihtelevat välillä 100–250. Suullisten kertomusten sanemäärät vaihtelevat puolestaan pääsääntöisesti välillä 140–280, mutta yksi oppijoista käyttää omissa kertomuksissaan runsaasti enemmän saneita ja hänellä on alkutestissä saneita 407 ja lopputestissä 472. Kahdeksalla oppijalla sanemäärät kasvoivat sekä kirjoitetuissa että puhutuissa kertomuksissa, mutta kahden oppijan kohdalla sanemäärä oli pienempi sekä kirjallisissa että suullisissa loppumittauksissa verrattuna alkumittausten sanemäärään. Kahdella oppijalla puolestaan kasvua sanemäärissä tapahtui kirjallisten kertomusten osalta, mutta suullisissa mittauksissa tapahtui pientä laskua. Lisäksi yhdellä oppijalla sanemäärä kasvoi kirjallisissa kertomuksissa, mutta suullisissa kertomuksissa sanemäärä oli sama sekä alussa että lopussa.

Sanemäärä ei kuitenkaan paljasta, onko oppijan sanasto monimuotoistunut kurssin aikana vai ei. On todennäköistä, että kurssin aikana oppijat ovat oppineet uusia sanoja, mutta sanemäärä ei kuitenkaan kerro, käytetäänkö sanoja monipuolisemmin vai toistuvatko esimerkiksi tietyt ilmaukset uudelleen tekstissä, mikä kasvattaa sanemäärä. Voidaan kuitenkin todeta, että sanemäärien kasvua tapahtuu sekä kirjallisissa että suullisissa kertomuksissa, mikä puolestaan osoittaa, että jonkinlaista sanastollista kehitystä kurssin aikana on tapahtunut.

Taulukkoon 11 on koottu vielä selkeyden vuoksi kirjallisten ja suullisten kertomusten regressiokertoimet sekä riskiluvut.

TAULUKKO 11: Selittävien muuttujien regressiokerroin ja riskiluku kokoavasti

	Kirjalliset kertomukset		**Suulliset kertomukset**	
Tekstin sijoittuminen kurssilla	**B**	**Exp(B)**	**B**	**Exp(B)**
Vaihtelevuus	0,2207	1,2470	0,1634	1,1775
Tasaisuus	-137,4452	0,0000	-112,2596	0,0000
Harvinaisuus	-0,0002	0,9998	-0,0010	0,9990
Sironta	-0,0541	0,9474	-0,1995	0,8192
Erityisyys	0,6929	1,9995	-50,6163	0,0000

Kun kertomuksia tarkastellaan leksikaalisen diversiteetin eri osa-alueiden osalta, voidaan huomata, että kehitys kirjallisissa ja suullisissa kertomuksissa on tietyiltä osin samanlaista molemmissa kertomustyypeissä, mutta joitain erojakin löytyy. Tarkastelen kutakin leksikaalisen diversiteetin osa-aluetta seuraavaksi hieman vertaillen kirjallisten ja suullisten kertomusten arvoja toisiinsa. Taulukoissa 12–16 esitetyt arvot on pyöristetty selkeyden vuoksi, mutta logistiseen regressioanalyysiin käytetyt arvot ovat olleet useamman desimaalin tarkkuudella.

Vaihtelevuus kertoo siitä, millainen eri sanojen kokonaismäärä on tekstissä ja miten sanat sijoittuvat siinä. Mitä suurempi vaihtelevuuden arvo tekstille muodostuu, sen tiheämpi sanasto on. Jos arvo kasvaa kurssin kuluessa, se osoittaa, että leksikaalinen diversiteetti kehittyy. (Kajzer-Wietrzny–Ivaska, 2020, 178.) Kun tarkastellaan regressiokertoimia ja riskilukuja vaihtelevuuden osalta, huomataan, että vaihtelevuuden arvon kasvu vaikuttaa siihen, että suhteellinen riski kuulua loppuryhmään myös kasvaa (kirjallisissa Exp(B)=1,2470, suullisissa Exp(B)=1,1775). Saman suuntaisia tuloksia havaitaan myös, kun tarkastellaan kunkin kertomuksen arvoja vaihtelevuuden osalta (ks. taulukko 12).

TAULUKKO 12: Vaihtelevuuden arvo tuotetuissa kertomuksissa

	Kirjallinen		**Suullinen**	
Informantti	**Alku**	**Loppu**	**Alku**	**Loppu**
Inf1	24,13	34,00	17,27	19,00
Inf2	33,73	45,94	16,49	21,74
Inf3	22,68	30,01	13,96	18,37
Inf4	37,70	36,32	19,60	23,88
Inf5	24,42	30,96	16,20	15,80
Inf6	31,35	34,85	22,94	22,93
Inf7	27,59	41,48	18,39	19,94
Inf8	43,74	69,45	13,52	28,03
Inf9	44,49	46,69	18,91	24,88
Inf10	38,92	39,15	18,02	22,16
Inf11	36,05	41,14	27,32	28,61
Inf12	45,28	48,78	18,92	21,40

Taulukosta 12 voidaan havaita, että arvo kasvaa yhdeksällä oppijalla sekä kirjallisissa että suullisissa kertomuksissa. Lisäksi kahdella oppijalla arvo kasvaa kirjallisten kertomusten osalta, mutta suullisten kertomusten osalta arvo laskee. Yhdellä oppijalla on päinvastoin eli kirjallisten kertomusten arvo laskee, mutta suullisten kertomusten arvo kasvaa.

Arvoja tarkastellessa voidaan kiinnittää huomiota siihen, että lähtökohtaisesti kirjallisten kertomusten arvot ovat suurempia kuin suullisten kertomusten. Kirjallisten kertomusten arvot vaihtelevat alkutestin osalta välillä 22–46 ja lopputestin välillä 30–70, kun taas suullisissa kertomuksissa vaihteluväli on alkutestissä 13–28 ja lopputestissä 18–29. Tästä voidaankin päätellä, että kirjallisissa kertomuksissa arvo on jo alkujaan pääsääntöisesti korkeampi kuin suullisissa kertomuksissa. Lisäksi kirjallisten kertomusten osalta arvo kasvaa lopputestissä suhteessa enemmän kuin suullisissa kertomuksissa, joten voidaan olettaa, että intensiivikurssin aikana osaaminen kehittyy merkittävästi kirjallisen tuottamisen osalta siinä, miten sanojen kokonaismäärä kasvaa ja miten niitä sijoitetaan tekstiin. Suullisessa tuottamisessa on aina omat haasteensa (ks. luku 3.2), mikä voi osaltaan selittää sitä, että kehitys ei ole niin voimakasta kuin kirjallisissa kertomuksissa. Rajallinen aika sekä tietystä asiasta kertominen voivat vaikuttaa siihen, että suullisessa tuottamisessa ei ehdi pohtia niin tarkasti sitä, miten omat sanansa järjestelee puheessaan, kun taas kirjallisissa kertomuksissa omaa tuotosta on mahdollista muokata ja vaihtaa sanajärjestystä sekä pohtia tarkemmin sisältöä ja ilmaisutapoja.

Tasaisuus kertoo siitä, kuinka tasaisesti tekstin eri sanamuodot esiintyvät tekstissä. Voidaan siis tarkastella esiintykö jokin sana tekstissä kerran, kaksi vai useamman kerran. Mitä useammin sana esiintyy, sitä toisteisempaa teksti on. Tuloksia tarkastellessa kiinnitetään huomiota siihen, että mitä

suurempi arvo saadaan, sitä korkeampi tekstin leksikaalinen diversiteetti on. (Kajzer-Wietrzny–Ivaska, 2020, 178–179.) Kun tasaisuutta tarkastellaan logistisen regressiomallin avulla, näyttää siltä, että todennäköisyys kuulua loppuryhmään madaltuu, jos tasaisuuden arvo kasvaa yhdellä yksiköllä, sillä regressiokerroin on kirjallisissa kertomuksissa -137,4452 ja suullisissa kertomuksissa -112,2596. Mallin mukaan lopussa olevien arvojen pitäisi siis olla matalampia kuin alussa, vaikka odotuksenmukaista olisi, että tasaisuuden arvo kasvaa, kun leksikaalinen diversiteetti kehittyy. Taulukkoa 13 tarkastelemalla voidaankin huomata, että arvo on pääsääntöisesti korkeampi lopussa kuin alussa, joten leksikaalinen diversiteetti kasvaa näiltä osin. Syynä sille, että regressiomallien mukaan odotuksenmukaista olisi, että arvo pienenee leksikaalisen diversiteetin kehittyessä, on todennäköisesti se, että arvojen vaihteluväli on hyvin pieni (0,91–0,97). Malli vertaa alku- ja loppukertomusten vaihtelua toisiinsa, mutta koska molemmissa kertomustyypeissä alussa ja lopussa tuotettujen kertomusten arvot sijoittuvat koko vaihteluvälille eivätkä selvästi vaihteluvälin alkuun tai loppuun, malli esittää, että odotuksenmukainen suunta olisi arvon väheneminen, kun leksikaalinen diversiteetti kehittyy eikä päinvastoin.

TAULUKKO 13: Tasaisuuden arvo tuotetuissa kertomuksissa

	Kirjallinen		**Suullinen**	
Informantti	**Alku**	**Loppu**	**Alku**	**Loppu**
Inf1	0,934	0,943	0,911	0,923
Inf2	0,948	0,957	0,912	0,924
Inf3	0,931	0,941	0,909	0,936
Inf4	0,945	0,945	0,908	0,933
Inf5	0,927	0,936	0,900	0,913
Inf6	0,945	0,947	0,928	0,928
Inf7	0,941	0,941	0,916	0,945
Inf8	0,962	0,961	0,928	0,946
Inf9	0,959	0,958	0,924	0,931
Inf10	0,953	0,954	0,916	0,933
Inf11	0,952	0,952	0,940	0,944
Inf12	0,947	0,954	0,934	0,929

Kun tarkastellaan taulukkoa 13, nähdään, että tasaisuuden arvo kasvaa sekä kirjoitetuissa että suullisissa kertomuksissa kahdeksalla oppijalla. Yhdellä oppijalla kasvua tapahtuu kirjallisten kertomusten osalta, mutta ei suullisten, kun taas kolmella muulla oppijalla tapahtuu laskua kirjallisten kertomusten osalta, mutta kasvua suullisten kertomusten osalta. Kirjallisissa kertomuksissa tasaisuuden arvojen vaihteluväli on 0,93–0,96 ja suullisissa 0,91–0,95. Lähtökohtaisesti siis kirjallisten kertomusten alkutesteissä tasaisuus on hieman korkeampi kuin suullisten kertomusten

alkutestissä. Jos taulukkoa 13 tarkastellaan tarkemmin, voidaan kuitenkin huomata, että suullisten kertomusten tasaisuus kasvaa suhteessa enemmän kuin kirjallisten kertomusten. Voidaankin siis pohtia sitä, että kehittyykö suullinen taito intensiivikurssin aikana hieman vahvemmin kuin kirjallinen taito sen suhteen, miten sanoja sijoitetaan omassa tuottamisessa. Voi olla, että intensiivikurssin aikana puhetaito vahvistuu, minkä vuoksi myös hieman sanojen toistoa jää pois.

Harvinaisuus on yksi osa-alueista, joiden osalta oppijoiden välillä oli runsaasti erilaista vaihtelua. Harvinaisuutta tarkastellaan siten, että mitä lähempänä nollaa arvo on, sen yleisempiä sanoja käytetään. Mitä suurempi luku, sitä harvinaisempia sanoja on käytössä eli leksikaalinen diversiteetti on korkeampi. (Kajzer-Wietrzny–Ivaska, 2020, 179.) Logistisia regressiomalleja tarkastellessa voidaan todeta, että vaihtelevuuden kohdalla suhteellinen riski kuulua loppuryhmään pienenee, jos arvo nousee (kirjallisissa Exp(B) = 0,9998, suullisissa Exp(B) = 0,9990). Tämä on täysin linjassa sen kanssa, miten harvinaisuuden arvot on määritelty (taulukko 14), mutta vastakkainen sen ajatuksen kanssa, että mitä harvinaisempia sanoja käytetään, sen monimuotoisempi tekstin leksikaalinen diversiteetti on.

TAULUKKO 14: Harvinaisuuden arvo tuotetuissa kertomuksissa

	Kirjallinen		**Suullinen**	
Informantti	**Alku**	**Loppu**	**Alku**	**Loppu**
Inf1	27534	25995	32236	31530
Inf2	25942	26648	31526	29634
Inf3	23982	28484	30627	25578
Inf4	26919	25743	31435	27220
Inf5	28118	27442	30437	29266
Inf6	25297	23504	28859	30137
Inf7	27594	29860	32200	29153
Inf8	29934	26442	32200	28223
Inf9	26699	25411	277401	29769
Inf10	28434	24790	30945	28097
Inf11	26183	28492	30164	28154
Inf12	25811	27798	28539	27325

Taulukkoa 14 tarkastelemalla voidaan todeta, että oppijat käyttivät harvinaisempia sanoja alkumittauksissa kuin loppumittauksissa. Kaikista oppijoista viidellä kehitystä harvinaisempien sanojen käyttöön ei ole nähtävissä ollenkaan, vaan he käyttävät molemmissa loppumittauksissa yleisempiä sanoja kuin alkumittauksissa. Kokonaisuudessaan viisi oppijaa käyttää harvinaisempia sanoja kirjallisten kertomusten loppumittauksissa, joten seitsemällä muulla harvinaisempia sanoja on käytössään alkumittauksissa. Suullisissa kertomuksissa puolestaan vain kaksi oppijaa käyttää

harvinaisempia sanoja loppumittauksissa kuin alkumittauksissa, joten muilla kymmenellä alkumittauksessa on harvinaisempia sanoja kuin loppumittauksissa.

Saatujen tuloksien perusteella suullisissa kertomuksissa käytettään harvinaisempia sanoja kuin kirjallisissa kertomuksissa. Kirjallisten kertomusten harvinaisuuden vaihteluväli on 23 982–29 860, kun suullisten kertomusten vaihteluväli on 27 741–31 530. On kuitenkin syytä pohtia sitä, että mistä erot voivat johtua. Yhtenä syynä tässä voi olla se, että alkumittausten ja loppumittausten tehtävänannot ovat olleet erilaiset, joten myös sanasto on näissä erilaista. Tekstit ovat lyhyitä, joten yksittäiset, aihekohtaiset sanat voivat vaikuttaa harvinaisuuteen paljonkin. Suullisissa kertomuksissa sanastoon kuuluu esimerkiksi lankakerä, joka saattaa toistua tekstissä monta kertaa, kun taas kirjallisissa kertomuksissa toistuvana sanana voi olla esimerkiksi tuoli, joka ei ole yhtä harvinainen. Suullisissa kertomuksissa siis käytetään lähtökohtaisesti harvinaisempia sanoja, jotka nousevat selvästi kuvasarjan sisällöistä. Toisaalta tulokseen voi vaikuttaa se, että alkumittauksessa on tuotettu sellaisia sanoja, jotka eivät välttämättä sisällöllisesti aivan vastaa tarkoitustaan, jonka vuoksi sanavalinnat näyttäytyvät harvinaisempina kuin loppumittauksissa, joissa sanat ehkä vastaavat paremmin tehtävänantoa ja kuvasarjojen sisältöjä.

Sironta on hyvin samankaltainen tasaisuuden kanssa, mutta se osoittaa sanojen sijoittumisen tekstissä laskemalla tietyn sana eri esiintymien etäisyyden tai kasaantumisen tarkasteltavassa tekstissä. Mitä säännöllisemmin samaa lekseemiä edustavat saneet sijoittuvat tekstissä, sen monimuotoisempaa teksti on. Pienemmät arvot osoittavat, että sanakimppuja on vähemmän, jolloin tekstin saman sanan eri esiintymät ovat sijoittuneet tekstissä tasaisemmin ja tekstin leksikaalinen diversiteetti on silloin korkeampi. (Honko ym. 2019, 58–59.) Logistisissa regressioanalyyseissa sironnan regressiokertoimet ovat negatiivisia (kirjallisissa B=-0,0541; suullisissa B=-0,1995), joten arvojen pieneminen osoittaa leksikaalisen diversiteetin kehittymistä. Tämä on nähtävissä myös, kun tarkastellaan informanttien kertomusten sironta-arvoja (ks. taulukko 15). Suurimmalla osalla oppijoista arvo on pienempi lopussa, kun sitä verrataan alussa tuotettuun kertomukseen.

TAULUKKO 15: Sironnan arvo kertomuksissa

	Kirjallinen		**Suullinen**	
Informantti	**Alku**	**Loppu**	**Alku**	**Loppu**
Inf1	44,27	33,09	48,11	50,00
Inf2	34,97	29,44	51,40	40,69
Inf3	42,98	36,43	47,06	43,89
Inf4	35,47	39,52	46,48	44,28
Inf5	39,73	37,72	53,04	51,48
Inf6	34,24	34,70	47,06	44,83
Inf7	38,18	30,71	43,01	41,40
Inf8	31,58	20,18	51,60	36,65
Inf9	31,05	28,57	50,18	41,29
Inf10	28,57	27,18	46,40	43,32
Inf11	28,69	29,02	41,82	35,61
Inf12	28,71	28,79	44,63	38,25

Kun tutkitaan taulukkoa 15, voidaan huomata, että seitsemällä oppijalla sironnan arvo pienenee alkutestauksesta lopputestaukseen, jolloin tekstien leksikaalinen diversiteetti siis kasvaa. Yhdellä oppijalla arvo kasvaa suullisessa kertomuksessa alkutestiin verrattuna, mutta kirjallisessa kertomuksessa arvo pienenee lopputestissä. Neljällä oppijalla on päinvastoin eli kirjallisissa kertomuksissa arvo kasvaa alkuun verrattuna, mutta suullisissa puolestaan arvo pienenee. Suullisten kertomusten sironta-arvojen pieneneminen on siis nähtävillä yhdellätoista oppijalla kaikista, joten voidaan päätellä, että etenkin näissä tapauksissa leksikaalinen diversiteetti on kehittynyt samansuuntaisesti. Kirjallisissa kertomuksissa vaihteluväli on alussa 28,5–44,3 ja lopussa 20,2–39,5, kun taas suullisissa kertomuksissa vaihteluväli on alussa 41,8–53,0 ja lopussa 35,6–51,5. Suullisissa kertomuksissa saman sanan eri esiintymät toistuvat siis huomattavasti useammin kuin kirjallisissa teksteissä. Tämä liittynee siihen, että pohtimisaika on lyhyempi ja puheessa on helpompi toistaa ja käyttää tuttuja ilmauksia uudelleen, etenkin, jos ei ole aivan varma siitä, miten asian ilmaisi.

Viimeisenä tarkastelussa on erityisyys, joka tarkoittaa sitä, miten monia eri merkityksiä tekstin sanat aktivoivat eli kuinka paljon synonyymisiä ilmauksia tekstissä esiintyy. Pienemmät arvot osoittavat, että synonyyminen redundanssi on pienempi, jolloin tekstin leksikaalinen diversiteetti on suurempi. (Kajzer-Wietrzny–Ivaska 2020, 179.) Logistisia regressiomalleja tarkastellessa voidaan nähdä, että kirjallisten ja suullisten kertomusten regressiokerroin erityisyyden osalta on erilainen. Kirjallisissa kertomuksissa regressiokerroin on 0,6929 ja suullisissa kertomuksissa -50,6163. Mallien mukaan kirjallisissa kertomuksissa erityisyyden arvon pitäisi siis kasvaa lopussa, kun taas suullisissa kertomuksissa arvon pitäisi vähentyä lopussa. Syynä siihen, että kirjallisten kertomusten logistinen regressiomalli arvioi, että arvon kasvu kasvattaa riskiä kuulua loppuryhmään on todennäköisesti se,

että kahdella informantilla arvo kasvaa reilusti loppumittauksessa. Muilla informanteilla vaihtelu on hyvin tasaista, joten korkeat arvot vaikuttavat mallin luokitteluun. Voidaan kuitenkin todeta, että molemmissa kertomustyypeissä arvot pääsääntöisesti pienenevät loppuvaiheessa (ks. taulukko 16) ja näin osoittavat leksikaalisen diversiteetin kehittymistä kurssin aikana.

TAULUKKO 16: Erityisyyden arvo tuotetuissa kertomuksissa

	Kirjallinen		**Suullinen**	
Informantti	**Alku**	**Loppu**	**Alku**	**Loppu**
Inf1	1,325	1,317	1,274	1,261
Inf2	1,346	1,340	1,306	1,266
Inf3	1,358	1,360	1,392	1,306
Inf4	1,308	1,300	1,325	1,319
Inf5	1,301	1,302	1,339	1,280
Inf6	1,303	1,285	1,364	1,309
Inf7	1,317	1,313	1,372	1,255
Inf8	1,373	1,385	1,295	1,302
Inf9	1,328	1,432	1,327	1,287
Inf10	1,295	1,330	1,346	1,285
Inf11	1,308	1,279	1,260	1,294
Inf12	1,317	1,280	1,270	1,300

Taulukon 16 perusteella näyttää siltä, että alkumittauksiin verrattuna viidellä oppijalla arvot pienenevät loppumittauksessa sekä kirjallisissa että suullisissa kertomuksissa, joten näiden oppijoiden kohdalla leksikaalinen diversiteetti on siis kasvanut molemmissa tapauksissa loppua kohden. Yhdellä oppijalla on niin, että molempien kertomustyyppien alkumittauksissa arvo on pienempi kuin loppumittauksessa, joten hänen kohdallaan kehitystä leksikaalisessa diversiteetissä ei suoraan erityisyyden osalta ole nähtävissä. Viidellä oppijalla arvot pienenevät suullisissa kertomuksissa, mutta kasvavat kirjallisissa kertomuksissa. Lisäksi yhdellä oppijalla arvot pienenevät kirjallisessa mittauksessa, mutta kasvavat suullisessa kertomuksessa. Tämän perusteella vaihtelevuutta on siis paljon tutkittavan ryhmän sisällä. Kirjallisissa mittauksissa arvot vaihtelevat alussa välillä 1,30–1,37 ja lopussa 1,28–1,43, kun taas suullisissa mittauksissa arvojen vaihtelu on alussa välillä 1,26–1,39 ja lopussa 1,25–1,32. Voidaan siis ajatella, että jonkinlaista kehitystä intensiivikurssin aikana tapahtuu, sillä yhdellätoista oppijalla kehitystä tapahtuu erityisyyden osalta ainakin toisessa kertomustyypissä. Lisäksi viidellä oppijalla on nähtävissä kehitystä molemmissa kertomustyypeissä.

Logistisen regressioanalyysin ja kertomusten vertailun perusteella on selvää, että muutoksia oppijoiden leksikaalisessa diversiteetissä tapahtuu kurssin aikana. Taulukossa 17 on koottuna vielä kaikkien kertomusten todennäköisyydet sekä ennusteet kuulua alku- tai loppuryhmään.

TAULUKKO 17: Kirjallisten ja suullisten kertomusten todennäköisyydet ja ennusteet

		Kirjalliset kertomukset		**Suulliset kertomukset**	
Informantti	**Ajankohta**	**Todennäköisyys**	**Ennuste**	**Todennäköisyys**	**Ennuste**
Inf1	0	0,14452	0	0,31193	0
Inf2	0	0,30654	0	0,06576	0
Inf3	0	0,30262	0	0,00547	0
Inf4	0	0,55731	1	0,17095	0
Inf5	0	0,35153	0	0,10095	0
Inf6	0	0,31225	0	0,06432	0
Inf7	0	0,16482	0	0,00594	0
Inf8	0	0,25301	0	0,00631	0
Inf9	0	0,55274	1	0,36888	0
Inf10	0	0,38699	0	0,03526	0
Inf11	0	0,37728	0	0,8266	1
Inf12	0	0,90781	1	0,81107	1
Inf1	1	0,52496	1	0,29719	0
Inf2	1	0,69595	1	0,95378	1
Inf3	1	0,21696	0	0,92771	1
Inf4	1	0,50149	1	0,80384	1
Inf5	1	0,46026	0	0,66783	1
Inf6	1	0,50711	1	0,31664	0
Inf7	1	0,7986	1	0,76653	1
Inf8	1	0,99781	1	0,87413	1
Inf9	1	0,7644	1	0,81062	1
Inf10	1	0,54647	1	0,88947	1
Inf11	1	0,52609	1	0,94787	1
Inf12	1	0,84247	1	0,97095	1

Kun mukaan otetaan kaikki selittävät muuttujat ja kirjallisten ja suullisten kertomusten ennusteita tarkastellaan rinnakkain, huomataan myös se, että vain yhden yksittäisen oppijan kohdalla molemmat alussa tuotetut kertomukset luokitellaan lopussa tuotetuiksi. Hänen tapauksessaan ennusteet ovat menneet voimakkaasti väärin, joten voidaan pohtia, onko alussa tuotetuissa kertomuksissa ollut jo vahvasti nähtävissä leksikaalinen diversiteetti ja tästä syystä malli ennustaa tekstien kirjoitusajankohdaksi lopun. Kaikkien muiden oppijoiden kohdalla malli puolestaan on luokitellut vähintään toisen kertomuksen oikeaan luokkaan kuuluvaksi. Erot voivat ainakin osittain selittyä yksilökohtaisella vaihtelulla ja sillä, että kertomusten tuottamiseen liittyvät prosessit ovat vaihtelevia. Toisille kirjoittaminen saattaa olla helpompaa, jolloin myös leksikaalisen diversiteetin kehitys näkyy

kirjallisissa kertomuksissa, kun taas toisille suullinen tuottaminen on helpompaa ja kehitys näkyy suullisissa kertomuksissa. Kaikista 48 kertomuksesta oikein luokitellaan 39 ja luokittelussa menee väärin vain 9 kertomusta. Huomataan myös se, että kirjallisista kertomuksista väärin luokitellaan viisi ja suullisista kertomuksista neljä, ja molemmissa tapauksissa väärin luokiteltu kertomuksia on sekä alussa että lopussa.

8. Yhteenveto ja päätelmät

8.1. Tutkimustulosten tarkastelua ja pohdintaa

Tutkimuksessani olen pyrkinyt selvittämään leksikaalisen diversiteetin kehittymistä intensiivikurssin aikana tarkastelemalla yhden suomenoppijaryhmän tuottamia kirjallisia ja suullisia kertomuksia. Tavoitteena on ollut selvittää, miten leksikaalinen diversiteetti kehittyy niin kirjallisissa kuin suullisissa kertomuksissa kun tarkastellaan intensiivikurssin alussa ja lopussa tuotettuja kertomuksia. Tarkastelun kohteen on ollut myös se, ovatko muutokset oppijoiden kirjallisissa ja suullisissa kertomuksissa samansuuntaisia vai kehittyykö leksikaalinen diversiteetti näissä eri tavoin.

Monet leksikaalista diversiteettiä koskevat tutkimukset ovat keskittyneet joko kirjalliseen tai suulliseen tuottamiseen (ks. esim. Honko ym. 2019; Honko 2017; Jarvis 2013a ja 2013b) eikä vertailua kahden tuottotavan välillä ole tehty laajasti (ks. kuitenkin Yu 2009; Johansson 2008). Tutkimuksessani olen pyrkinyt selvittämään sitä, miten kirjallisten ja suullisten kertomusten leksikaalinen diversiteetti kehittyy lyhyen opiskelujakson aikana oppimisympäristössä, joka on kokonaan kohdekielinen.

Olen tarkastellut oppijoiden leksikaalisen diversiteetin kehittymistä hyödyntämällä ensisijaisena tutkimusmenetelmänäni logistista regressioanalyysia. Tarkasteltavina muuttujina ovat olleet kertomuksen kirjoitusajankohta sekä vaihtelevuus, erityisyys, harvinaisuus, tasaisuus ja sironta. Menetelmää hyödyntämällä olen saanut selville, että leksikaalinen diversiteetti kehittyy kaikilla oppijoilla kurssin aikana, ainakin joiltain osin.

Logististen regressiomallien ennusteiden (ks. taulukko 17) perusteella leksikaalisen diversiteetin kehitys on nähtävissä kahdeksalla oppijalla niin kirjallisten kuin suullistenkin kertomusten osalta ja neljällä oppijalla kehitystä tapahtuu ainakin toisen tuottotavan teksteissä. Huomionarvoista on myös se, että vaikka kehitys on suurella osalla oppijoista samansuuntaista, yksilökohtaisia eroja on myös nähtävissä alku- ja lopputilanteiden välillä. Yksilöllisiä eroja on useimmiten nähtävissä vieraan kielen oppimisessa (Pietilä 2014, 45), joten oletettavaa on, että myös tässä yhteydessä yksilölliset erot vaikuttavat kertomusten tuottamiseen. Tämä näkyy muun muassa siinä, että logistisen regressiomallin ennusteista yhdeksän menee väärin, mutta niistä vain kuudessa ennuste on voimakkaasti pielessä. Ennusteiden perusteella voidaan päätellä, että muutamalla oppijalla leksikaalisen diversiteetin taso on lähtökohtaisesti joko muita oppijoita matalammalla tai huomattavasti korkeammalla, minkä seurauksena ennusteet poikkeavat todellisesta kertomuksen tuottamisajankohdasta.

Sen lisäksi, että kertomusten leksikaalista diversiteettiä on tarkasteltu kertomustyypeittäin ja ennusteittain, olen tarkastellut leksikaalisen diversiteetin eri osa-alueita oppija- ja kertomuskohtaisesti pystyäkseni muodostamaan kuvan siitä, miten leksikaalinen diversiteetti muutuu intensiivikurssin aikana. Tulosten perusteella voidaan todeta, että kirjallisten ja suullisten kertomusten osalta vaihtelevuus, erityisyys, sironta ja tasaisuus kehittyvät odotuksenmukaisesti eli näiden osalta leksikaalinen diversiteetti kasvaa kurssin aikana.

Vaihtelevuus eli se, millainen kertomuksessa käytettyjen sanojen kokonaismäärä on ja miten sanat sijoittuvat tekstissä, kehittyy lähes jokaisella oppijalla intensiivikurssin aikana. Voidaan siis todeta, että loppuvaiheessa oppijat käyttävät sanoja vaihtelevammin kuin alussa, joten heidän sanastollinen osaamisensa on näiltä osin kehittynyt. Kirjallisissa kertomuksissa vaihtelevuuden arvo kasvaa yhdellätoista oppijalla, joten vain yhdellä arvo pienenee. Suullisissa kertomuksissa vaihtelevuuden arvo kasvaa kymmenellä oppijalla, joten vain kahdella arvo pienenee. Molemmissa tuottotavoissa kertomusten leksikaalinen diversiteetti näiltä osin siis kehittyy vahvasti, joskin on huomioitava se, että suullisissa kertomuksissa arvot ovat lähtökohtaisesti matalammat kuin kirjallisissa kertomuksissa. Syynä tähän on mahdollisesti se, että puhuessa teksti tuotetaan kerralla ja sen muotoilemiseen ja korjaamiseen ei käytetä niin paljoa aikaa kuin kirjallisissa kertomuksissa. Koska kehitystä tapahtuu molemmissa tuottotavoissa, on todennäköistä, että intensiivikurssin aikana oppijan sanavarasto kasvaa ja hän oppii myös enemmän lause- ja virkerakenteista sekä siitä, miten tekstiä voi jäsennellä, minkä seurauksena loppumittauksessa sisältöjä on osattu muotoilla tarkemmin ja asioita ilmaista monipuolisemmin kuin alussa.

Nissilä (2003, 112) on todennut, että edetessään opinnoissaan, oppijan käsitys eri sanoista laajenee, kun sanat saavat uusia merkityksiä ja niitä pystytään yhdistämään paremmin toisiin sanoihin. Tämä on nähtävissä muun muassa siinä, että kirjallisissa ja suullisissa kertomuksissa erityisyys ja sironta muuttuvat odotuksenmukaisesti eli niiden arvot pienenevät loppuvaiheessa, jolloin leksikaalinen diversiteetti kertomuksissa kehittyy. Erityisyys eli se, kuinka paljon synonyymisiä ilmauksia tekstissä esiintyy, muuttuu oppijoiden kertomuksissa intensiivikurssin aikana. Erityisyyden arvot pienenevät lopussa seitsemällä oppijalla kirjallisissa kertomuksissa ja yhdeksällä oppijalla suullisissa kertomuksissa. Heidän kertomuksissaan synonyymisten ilmausten käyttö on siis lopussa vähäisempää kuin alussa, joten leksikaalinen diversiteetti kehittyy näiltä osin. Voidaankin todeta, että oppijat ovat kurssin aikana oppineet uusia sanoja tai sanojen yhdistelmiä, ja he pystyvät ilmaisemaan asioita yksilöidymmin ja tarkemmin, eikä samoja sanoja tai niiden synonyymien toistamista tarvita kertomuksissa niin paljon kuin alkuvaiheessa. Sironnan osalta puolestaan on tarkasteltu sitä, miten säännöllisesti samaa lekseemiä edustavat saneet tekstissä esiintyvät. Kun saneet sijoittuvat säännöllisesti, teksti on monimuotoista. Kirjallisissa kertomuksissa sironnan arvo pienenee lopussa

kahdeksalla oppijalla, kun taas suullisissa kertomuksissa arvo pienenee yhdellätoista oppijalla, joten samaa lekseemiä edustavat saneet sijoittuvat kertomuksissa säännöllisemmin loppuvaiheessa kuin alkuvaiheessa. Oppijat pystyvät siis lopussa tuotetuissa kertomuksissaan käyttämään lekseemejä monimuotoisemmin kuin alussa eli heidän kertomuksissaan samat lekseemit eivät toistu enää jatkuvasti peräkkäin, vaan lekseemit sijoittuvat kauemmaksi toisistaan, jolloin kerronta muuttuu sujuvammaksi ja selkeämmäksi.

Sironnan kanssa samankaltainen leksikaalisen diversiteetin osa-alue on tasaisuus, joka kertoo siitä, miten tasaisesti tekstissä käytetyt eri sanamuodot esiintyvät, toisin sanoen kuinka toisteista teksti on. Tasaisuuden osalta odotuksenmukaista on, että sanojen toisteisuus vähenee lopussa, jolloin tasaisuuden arvo kasvaa. Kirjallisten ja suullisten kertomusten tasaisuuden arvoja tarkastelemalla, on huomattu, että pääsääntöisesti arvot kasvavat eli oppijat käyttävät laajemmin eri lekseemejä kurssin lopussa ja pystyvät soveltamaan sanoja sekä kirjoitettuihin että puhuttuihin kertomuksiin. Kirjallisissa kertomuksissa tasaisuuden arvot kasvavat yhdeksällä oppijalla loppumittauksessa, joten vain kolmella arvot pienevät. Suullisissa kertomuksissa tasaisuuden arvot kasvavat yhdellätoista oppijalla, joten vain yhdellä arvo pienenee. Suurilta osin lekseemien käyttö siis tasoittuu oppijoilla intensiivikurssin aikana eli toisteisuus kertomuksissa vähenee.

Tutkimusta tehdessä olen pohtinut myös sitä, millaisia yhteyksiä tai eroja kirjoitettujen ja suullisten kertomusten leksikaalisen diversiteetin kehittymisessä on nähtävissä. Kuten jo edeltä kävi ilmi, molempien tuottotapojen kertomuksissa vaihtelevuus, tasaisuus, sironta ja erityisyys kehittyvät odotuksenmukaisesti, joskin jonkin verran vaihtelua on riippuen kertomuksen tuottotavasta sekä yksilöstä.

Erityisen mielenkiintoisen yhtäläisyyden kirjoitettujen ja suullisten kertomusten leksikaalisessa diversiteetissä muodostaa harvinaisuus eli se, miten yleisiä ja frekventtejä sanoja oppija käyttää. Odotuksenmukaista olisi, että oppija käyttäisi kurssin lopussa harvinaisempia sanoja kuin alussa, mikä osoittaisi leksikaalisen diversiteetin kehittymistä. Sekä kirjallisissa että suullisissa kertomuksissa on kuitenkin niin, että pääasiassa oppijat käyttävät lopussa yleisempiä sanoja kuin alussa. Myös logistisen regressioanalyysin mallit osoittavat, että muutos arvoissa on selkeästi siihen suuntaan, että arvojen pieneneminen osoittaisi leksikaalisen diversiteetin kehittymistä, vaikka aiemmin on osoitettu, että leksikaalisen diversiteetin kehittyessä myös sanaston harvinaisuus kasvaa. Kirjallisissa kertomuksissa harvinaisempia sanoja käyttää vain neljä oppijaa kaikista eli kahdeksalla oppijalla harvinaisuuden arvot pienevät lopussa. Suullisissa kertomuksissa harvinaisempia sanoja käyttää lopussa vain kaksi oppijaa eli kymmenellä oppijalla harvinaisuuden arvot pienevät. Vaikka oletettavaa on, että oppija oppii kurssin aikana harvinaisempia sanoja ja hänen sanavarastonsa laajenee, se ei näissä tuloksissa ole nähtävillä. Yksi selitys tälle on siinä, että alkuvaiheessa oppija on

joutunut etsimään sanoja omasta sanavarastostaan, eikä hän välttämättä ole onnistunut yhdistämään sanan oikeaa muotoa ja merkitystä (Nation 2011, 48,), vaan hän on hyödyntänyt sellaisia ilmauksia, jotka voisivat mahdollisesti merkitä kuvissa olevia asioita eli hän on käyttänyt harvinaisempia sanoja. Kurssin aikana oppija on voinut oppia hahmottamaan sanojen merkityksiä paremmin, minkä vuoksi hän myös osaa ilmaista asioita niiden oikeilla nimityksillä lopputestissä, jolloin tulos näyttäytyy siten, että lopussa käytetään yleisempiä sanoja kuin alussa. Asiaan voi vaikuttaa myös se, että kuvasarjat ovat alku- ja loppumittauksissa samat, joten oppija saattaa muistaa ja tietää jo osan sanoista, joten hän pystyy tuottamaan loppuvaiheessa oikean sanan nopeammin ja selkeämmin eikä hänen tarvitse haparoida sanavalintojen välillä. Lisäksi kielenkäyttö saattaa idiomaattistua intensiivikurssin aikana eli kieleen vakiintuu tiettyjä ilmauksia, joita oppija käyttää lopputestissä sujuvasti. On kuitenkin tärkeää huomata se, että vaikka oppijat käyttävät lopussa yleisempiä sanoja kuin alussa eikä leksikaalinen diversiteetti näin ollen kehity harvinaisuuden osalta, voidaan kuitenkin todeta, että oppijan sanasto on siinä mielessä kehittynyt, että hän pystyy ilmaisemaan asioita ja niiden merkityksiä oikeilla nimityksillä, mikä ehkä tässä vaiheessa on tärkeämpää kuin se, että osattaisiin käyttää todella harvinaisia sanoja.

Saamani tutkimustulokset ja niistä tehdyt päätelmät ovat samansuuntaisia aiempien tutkimusten kanssa. Tutkimuksissa on osoitettu, että leksikaalinen diversiteetti kehittyy muun muassa aktiivisen kielenkäytön seurauksena (ks. Honko 2013), joten voidaan päätellä, että intensiivikurssilla opiskelulla on vaikutuksia leksikaaliseen diversiteettiin osin juuri siksi, että opiskelu on päivittäistä, aktiivista ja vuorovaikutteista. Opetus intensiivikurssilla keskittyy aiheisiin, jotka ovat osittain kulttuurisidonnaisia ja oppituntien lisäksi oppijat pääsevät soveltamaan näitä opetettuja asioita myös vapaa-ajan ohjelmissa. Tällöin myös sanaston käyttö tulee tutummaksi ja mahdollisesti auttaa oppijoita hahmottamaan paremmin sanojen merkityksiä ja kytköksiä toisiinsa, mikä näyttäytyy leksikaalisen diversiteetin kehittymisenä. Lisäksi on myös osoitettu, että aiheen tuttuus voi vaikuttaa siihen, että leksikaalinen diversiteetti on korkeampi tai kasvaa (ks. esim. Yu 2009), joten voi olla, että samoista kuvasarjoista lopussa kertominen on auttanut oppijoita jäsentelemään asioita paremmin ja käyttämään sujuvammin sanoja, joihin on tutustunut jo alkutestissä.

Honko (2013) on myös osoittanut, että ainakin alakouluikäisillä leksikaalinen diversiteetti on kirjoitetuissa kertomuksissa suurempi kuin puhutuissa. Tämä näkyy myös omassa tutkimuksessani, sillä leksikaalisen diversiteetin osa-alueista vaihtelevuus, sironta, erityisyys ja tasaisuus ovat lähtökohtaisesti korkeammalla tasolla kirjallisissa kuin suullisissa kertomuksissa. Toisaalta tulosten perusteella voidaan osoittaa, että suullisissa kertomuksissa leksikaalinen diversiteetti kehittyy tietyillä osa-alueilla enemmän kuin kirjallisissa kertomuksissa (ks. luku 8.2.). Siksi koenkin, että oma

tutkimukseni tarjoaa lisätietoa aiheesta ja mahdollisesti myös laajentaa käsitystä siitä, miten leksikaalinen diversiteetti kehittyy oppijoilla.

8.2. Hypoteesien toteutuminen

Tutkimustulokseni tukevat pääosin hypoteeseja, jotka esitin johdantoluvussa (ks. luku 1). Ensimmäisenä olettamuksenani oli, että leksikaalinen diversiteetti kehittyy yhtä lailla muun kielitaidon kanssa, kun oppija pääsee tutustumaan kieleen ja kulttuuriin kohdekielisessä ympäristössä, jossa oppimisympäristö ja opittavat asiat tukevat toisiaan. Tulosten mukaan oppijoiden leksikaalinen diversiteetti kehittyy sekä kirjallisten että suullisten kertomusten osalta suurella osalla oppijoista. Yksilökohtaista vaihtelua on toki nähtävissä, mutta yleinen suunta on se, että ainakin osa leksikaalisen diversiteetin osa-alueista kehittyy intensiivikurssin aikaan kaikilla oppijoilla. Tulosten perusteella ei voida suoraan päätellä, miten oppijan muu kielitaito on kurssin aikana kehittynyt, mutta voidaan todeta, että sanaston vahvistuessa oppija pystyy ilmaisemaan itseään monipuolisemmin ja laajemmin.

Koska oppijoiden leksikaalinen diversiteetti kehittyy kurssin aikana, voidaan todeta, että oppijan sanasto muuttuu myös kompleksisemmaksi intensiivikurssin loppuvaiheessa. Kielen kompleksisuus on tarkkuuden ja sujuvuuden rinnalla yksi tärkeistä kielellisen osaamisen osa-alueista, sillä kompleksisessa kielenkäytössä oppija osaa muun muassa käyttää eri sanoja ja kieliopillisia rakenteita monipuolisesti (Bulte–Housen 2012, 23–25). Koska intensiivikurssin aikana oppijat pääsevät käyttämään kieltä aktiivisesti, he ovat todennäköisesti myös oppineet kurssin aikana uusia sanoja ja yhteyksiä jo osaamilleen sanoille (Niitemaa 2014, 144–145) sekä erilaisia kielellisiä rakenteita, kuten lauseiden ja virkkeiden muodostamista sekä sanajärjestystä. Tällaiset systeemiseen kompleksisuuteen kuuluvat asiat ovat nähtävissä aineistossa etenkin siinä, että oppijoiden käyttämä sanamäärä kasvaa, sanat esiintyvät teksteissä pääasiassa tasaisemmin ja selkeämmin sekä vaihtelevuutta on enemmän. Leksikaalisen diversiteetin kehittyminen kertookin myös siitä, että oppijat oppivat käyttämään kieltä monipuolisemmin, mikä puolestaan tekee heidän kommunikoinnistaan sujuvampaa ja ymmärrettävämpää, mutta myös mielekkäämpää heidän itsensä kannalta, kun asioita pystyy ilmaisemaan kattavasti (Martin 2003, 85).

Toisena oletuksenani esitin, että suullisten kertomusten leksikaalinen diversiteetti kehittyisi voimakkaammin kuin kirjallisten kertomusten, kun oppija pääsee käyttämään suomen kieltä vuorovaikutteisesti suomenkielisessä ympäristössä ja kielenkäyttöä aktivoidaan päivittäin useiden tuntien ajan.

Niin kirjallisissa kuin suullisissakin kertomuksissa on nähtävissä selkeää kehittymistä leksikaalisessa diversiteetissä, mutta tulosten perusteella voidaan todeta, että suullisten kertomusten leksikaalinen diversiteetti kehittyy tietyiltä osin laajemmin kuin kirjallisissa kertomuksissa. Tällä tarkoitan nyt erityisesti sitä, että suullisia kertomuksia tarkastelemalla on voitu osoittaa, että leksikaalinen diversiteetti kehittyy useammalla oppijalla tiettyjen osa-alueiden osalta enemmän kuin kirjallisissa kertomuksissa.

Suullisissa kertomuksissa tasaisuuden, sironnan ja erityisyyden osalta muutoksia leksikaalisessa diversiteetissä on enemmän nähtävissä, kun tarkastellaan oppijoiden tuottamia kertomuksia ja verrataan suullisia ja kirjallisia kertomuksia toisiinsa. Erityisesti tasaisuuden arvot ovat tässä hyvin mielenkiintoisia, sillä kirjallisissa kertomuksissa tasaisuuden arvot ovat hyvin lähellä toisiaan kummassakin mittauksessa, mutta suullisissa kertomuksissa arvojen vaihtelu alun ja lopun välillä on suurempaa, toisin sanoen suullisissa kertomuksissa oppijoiden leksikaalinen diversiteetti kehittyy voimakkaammin intensiivikurssin aikana. Oppijat siis kykenevät suullisissa kertomuksissaan käyttämään ilmauksia tasaisemmin loppuvaiheessa kuin alkuvaiheessa. Erityisen kiinnostavaa tämä on siksi, että monesti suullinen tuottaminen on toisteista korjausten ja miettimisen takia, joten oppijoiden puheen tuottaminen tasoittuu ja näiltä osin myös kehittyy kurssin aikana.

Sironnan ja erityisyyden osalta leksikaalisen diversiteetin kehitys suullisissa kertomuksissa näkyy nimenomaan siinä, että kertomuksien sironnalle ja erityisyydelle muodostetut arvot kehittyvät useammalla oppijalla kuin kirjallisissa kertomuksissa. Tässä täytyy kuitenkin muistaa se, että kirjallisten kertomusten arvot ovat lähtökohtaisesti korkeampia kuin suullisissa kertomuksissa, mikä selittynee ainakin osin sillä, että puhe- ja kirjoitusprosessit ovat erilaisia. Kirjoituksessa on enemmän aikaa miettiä ja tehdä korjauksia omiin teksteihinsä, kun taas suullisessa tuottamista pohdinta-aikaa on vähemmän (Viinikka–Voutilainen 2013, luku Poikkeavat prosessit ja vaihtelevat tilanteet). Voidaan kuitenkin todeta, että puheen kuuleminen sekä suullisen tuottamisen aktiivinen harjoitteleminen intensiivikurssilla ovat auttaneet siihen, että useamman oppijan leksikaalinen diversiteetti kehittyy sen suhteen, miten he sanoja sijoittavat suullisissa kertomuksissaan ja millaisia sanavalintoja he tekevät kertoessaan kuvasarjoista.

8.3. Tutkimuksen toteuttaminen ja jatkotutkimusmahdollisuudet

Tarkasteltava aineisto on ollut pieni, joten tulokset eivät ole yleistettävissä laajempaan joukkoon. Vaikka minulla olisi ollut mahdollisuus käyttää laajempaa aineistoa, niin koen, että valintani rajata tarkastelu vain pienen oppijaryhmän kertomuksiin on perusteltu, sillä heidän taitotasonsa oli lähes

sama ja he osallistuivat samoille oppitunneille, joten saadut tulokset ovat luotettavia ja keskenään vertailukelpoisia.

Koen tutkimukseni onnistuneen siinä mielessä hyvin, että olen pystynyt tarkastelemaan kertomuksia monipuolisesti ottamalla huomioon sekä kirjoitusajankohdan että osa-alueet, jotka katsotaan kuuluvaksi leksikaaliseen diversiteettiin. Alustavan suunnitelmani mukaan minun oli tarkoitus huomioida logistista regressiomallia luodessa satunnaismuuttujana kertomuksen tuottanut informantti, mutta jätin sen lopulta pois mallista. Jälkikäteen ajateltuna koen, että se olisi ollut kuitenkin erittäin hedelmällinen lisä tutkimukseeni, sillä oppijoiden kesken on selvästi nähtävissä vaihtelua etenkin kirjallisten kertomusten tuottamisessa, joten informantin lisääminen malliin olisi voinut antaa lisää viitteitä siitä, miten eri osatekijät vaikuttavat diversiteetin kehittymiseen yksilötasolla. Oman tutkimukseni kannalta ei ollut välttämätöntä tarkastella selittävien muuttujien tilastollista merkitsevyyttä, mutta sekin on asia, joka olisi tarjonnut vielä lisäarvoa tutkimukseeni. Jokainen osatekijä on varmasti omalta osaltaan merkityksellinen, mutta tarkemmat tilastolliset merkitsevyydet olisin voinut vielä selvittää hyödyntämällä esimerkiksi ANOVA-testiä.

Jatkossa olisi kiinnostavaa tutkia laajemminkin sitä, miten leksikaalinen diversiteetti kehittyy intensiivikurssilla muilla oppijolla, joiden kielitaito on eri taitotasoilla. Tarkastelemalla eri taitotasoilla olevien ryhmien kehittymistä voitaisiin myös päätellä, vaikuttaa taitotaso jollakin tavalla kehitykseen ja onko esimerkiksi edistyneemmillä opiskelijoilla nähtävissä voimakkaampaa kehitystä kuin aloittelevilla oppijoilla. Tällaisen tutkimuksen tekeminen olisi erityisen mielenkiintoista siksi, että saatuja tuloksia voitaisiin pohtia myös oppimisen ja opettamisen kannalta. Jos osoittautuisi, että esimerkiksi vain kirjallisten kertomusten leksikaalinen diversiteetti kehittyisi aloittelevilla opiskelijoilla, voitaisiin pohtia, miten sanastollista monimuotoisuutta voitaisiin kehittää myös suullisten taitojen osalta. Intensiivikurssilla koottujen kertomusten tarkastelun lisäksi olisi mielenkiintoista laajentaa tutkimusta siihen suuntaan, miten esimerkiksi suomea toisena tai vieraana kielenä opiskelevien oppijoiden leksikaalinen diversiteetti mahdollisesti kehittyy pidemmällä aikavälillä, kuten vaikkapa kokonaisen vuoden aikana. Olisi myös kiinnostavaa pohtia sitä, että vaikuttaako pidempiaikaisen opiskelun aikana jatkuva kohdekielisessä ympäristössä opiskelu jollakin tavoin leksikaalisen diversiteetin kehittymiseen. Koska leksikaalista diversiteettiä on tutkittu suomen kielen osalta verrattain vähän, toivon, että aihe herättää mielenkiintoa ja jatkotutkimusta tehdään etenkin suullisten taitojen osalta.

Lähteet

Alisaari, Jenni 2016: Songs and poems in the second language classroom. The hidden potential of singing for developing writing fluency. Väitöskirja. Turun yliopisto. [Viitattu 2.12.2019.] Saatavissa: https://www.utupub.fi/bitstream/handle/10024/128287/AnnalesB426Alisaari.pdf?sequence=2&isAllowed=y

Alisaari, Jenni – Heikkola, Leena Maria 2016a: Laulamalla sujuvuutta suomenoppijoiden kirjoittamiseen. Kasvatus 47 (4). S. 313–326.

Alisaari, Jenni – Heikkola, Leena Maria 2016b: Increasing fluency in L2 writing with singing. *Studies in Second Language Learning 6/2*. S. 271–292. [Verkkojulkaisu. Viitattu 2.12.2019.] Saatavissa: https://www.ceeol.com/search/article-detail?id=411481

Aitchison, Jean 1994 [1987]: Words in the mind – An introduction to the mental lexicon. Oxford: Blackwell Publishers Ltd.

Bulté, Bram – Housen, Alex 2012: Defining and operationalising L2 complexity. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. S. 21–46.

Chafe, Wallace – Danielewicz, Jane 1987: Properties of Spoken and Written Language. *Comprehending oral and Written Language*. S. 83–112. Toim. Rosalind Horowitz ja S. Jay Samuels. Academic Press Inc.

Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. Toim. Alex Housen, Folkert Kuiken ja Ineke Vedder. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Dufva, Hannele 1995: Puhutun ja kirjoitetun kielen eroista. *Kirjoitetun kielen prosessointi*. S. 63–72. Toim. Jukka Hyönä, Heikki Lang ja Marja Vauras. Oppimistutkimuksen keskus, TY, Turku.

---------- 2000: Puheen ja kirjoituksen maailmat: eräs näkökulma lukemaan oppimiseen. *Kielikoulussa – Kieli koulussa*. AFinLAn vuosikirja 2000. S. 71–93. Toim. Paula Kalaja ja Lea Nieminen. Jyväskylä: Suomen soveltavan kielitieteen yhdistyksen julkaisuja no. 58. Jyväskylä.

Euroopan neuvosto 2012*: Eurooppalainen viitekehys. Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys*. (Alkuteos Common European Framework of Reference for Languages: Learning, teaching, assessment [2001].) Suom. Irma Huttunen & Hanna Jaakkola. SanomaPro, Helsinki.

Heikkola, Leena Maria – Alisaari Jenni 2017: Laulun sanoja lausumalla taitavaksi ääntäjäksi*? Näkökulmia toisen kielen puheeseen. Insights into second language speech*. S. 18–44. [Verkkojulkaisu. Viitattu 2.12.2019.] Saatavissa: https://journal.fi/afinla/article/view/73122

Heikkola, Leena Maria – Alisaari Jenni 2019: Increasing fluency in L2 speaking with singing. *Fluency in SLA.* Multilingual Matters. S. 166–185. Toim. Pekka Lintunen, Maarit Mutta ja Pauliina Peltonen.

Helasvuo, Marja-Liisa 2014: ”Jotta suomalaiset voisivat puhua enemmän” – Puhetilanteet osallistujat tekstiviestikeskustelussa. *Kieli verkossa. Näkökulmia digitaaliseen vuorovaikutukseen.* Toim. Marja-Liisa Helasvuo, Marjut Johansson ja Sanna-Kaisa Tanskanen. SKS, Helsinki.

Henriksen, Birgit 1999: Three dimensions of vocabulary development. *Studies in second language acquisition.* Cambridge University Press (CUP) 21(2). S. 303–317. [Verkkojulkaisu. Viitattu 24.4.2021.] Saatavissa: https://doi.org/10.1017/s0272263199002089

Honko, Mari 2013: Alakouluikäisten leksikaalinen tieto ja taito – Toisen sukupolven suomi ja S2-verrokit. Väitöskirja, Tampereen yliopisto. [Viitattu 24.4.2021.] Saatavissa: https://trepo.tuni.fi/bitstream/handle/10024/94544/978-951-44-9251-8.pdf?sequence=1&isAllowed=y

---------- 2017: Sadutettu sanasto: puhutun kielen leksikaalinen diversiteetti arviointikohteena. *Näkökulmia toisen kielen puheeseen. Insights into second language speech.* S. 163–192. [Verkkojulkaisu. Viitattu 22.4.2021.] Saatavissa: https://doi.org/10.30660/afinla.73136

Honko, Mari – Jarvis, Scott – Vainio, Seppo 2019: Lukijat sanaston monimuotoisuutta määrittämässä: Leksikaalisen diversiteetin tarkastelua määrällisen ja laadullisen tutkimuksen rajapinnassa. Virittäjä, 123 (1). [Viitattu 24.4.2021] Saatavilla: https://journal.fi/virittaja/article/view/59025 DOI: 10.23982/vir.59025

Housen, Alex – Kuiken, Folkert – Vedder, Ineke 2012: Complexity, accuracy and fluency: Definitions, measurement and research. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA.* S. 1–20.

Jarvis, Scott 2013a: Capturing the Diversity in Lexical Diversity. *Language Learning* 63 (1). S. 87–106. University of Michigan. [Verkkojulkaisu. Viitattu 24.4.2021] Saatavissa: https://onlinelibrary-wiley-com.ezproxy.utu.fi/doi/pdfdirect/10.1111/j.1467-9922.2012.00739.x

---------- 2013b: Defining and measuring lexical diversity. *Vocabulary Knowledge: Human ratings and automated measures.* S. 13–44. Toim. Scott Jarvis – Michael Daller. John Benjamins Publishing Company, Amsterdam / Philadelphia.

---------- 2017: Grounding lexical diversity in human judgments. *Language Testing*, 34 (4). S. 537–553. [Verkkojulkaisu. Viitattu 22.4.2021] Saatavissa: https://doi.org/10.1177%2F0265532217710632

Johansson, Victoria 2008: Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers* 53, Lund University, Dept. of Linguistics and Phonetics. S. 61–79. [Viitattu 18.4.2021] Saatavilla: https://journals.lub.lu.se/LWPL/article/view/2273

Jokivuori, Pertti – Hietala, Risto 2007: *Määrällisiä tarinoita. Monimuuttujamenetelmien käyttö ja tulkinta.* WSOY, Helsinki.

Järvinen, Heini-Marja 2014: *Kielen opettamisen menetelmiä. Kuinka kieltä opitaan*. *Opas vieraan kielen opettajalle ja opiskelijalle.* S. 89–113.

Kajzer-Wietrzny, Marta – Ivaska, Ilmari 2020: A Multivariate Approach to Lexical Diversity in Constrained Language. *Across Languages and Cultures*, 21 (2). S. 169–194. [Verkkojulkaisu. Viitattu 22.4.2021] Saatavissa: https://doi.org/10.1556/084.2020.00011

Kanerva, Jenna – Ginter, Filip – Miekka, Niko – Leino, Akseli – Salakoski, Tapio 2018: Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 133–142. Brussels, Belgium: Association for Computational Linguistics. [Verkkoartikkeli. Viitattu 17.4.2021] Saatavissa: https://doi.org/10.18653/v1/K18-2013

Kielenoppimisen kysymyksiä. Toim. Kari Sajavaara ja Arja Piirainen-Marsh. Soveltavan kielentutkimuksen teoriaa ja käytäntöä. Jyväskylän yliopisto 1999.

Kielipankki 2019: FinnWordNet. [Viitattu 13.5.2021.] Saatavissa: https://www.kielipankki.fi/corpora/finnwordnet/

Koivisto, Vesa 2013*: Suomen sanojen rakenne.* SKS, Helsinki.

KS = Kielitoimiston sanakirja. 2020. Helsinki: Kotimaisten kielten keskuksen verkkojulkaisuja 35. URN:NBN:fi:kotus-201433. [Viitattu 25.4.2021.] Saatavissa: https://www.kielitoimistonsanakirja.fi. Päivitettävä julkaisu. Päivitetty 11.11.2020.

Kuinka kieltä opitaan. *Opas vieraan kielen opettajalle ja opiskelijalle.* Toim. Päivi Pietilä ja Pekka Lintunen. Gaudeamus, Helsinki.

Kuusinen, Karoliina 2018: "😂 Käytäx sä useinki x-ää -ks tilalla? 😂 " – Kahden eri ikäryhmän kielenkäytön erot WhatsApp-viestipalvelun perhekeskusteluissa. Kandidaatintutkielma, Turun yliopisto,

Malin, Essi 2012: Suomi toisena kielenä -oppijoiden sanaston kehittyminen taitosalta toiselle siirryttäessä. Pro gradu -tutkielma. Jyväskylän yliopisto. Saatavissa: https://jyx.jyu.fi/handle/123456789/40415

Malvern, David – Richards, Brian – Chipere, Ngoni – Durán, Pilar 2004: Lexical Diversity and Language Development: Quantification and Assessment. Palgrave MacMillan, New York.

McCarthy, Philip – Jarvis, Scott 2010: MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42 (2), 381–392.

[Verkkolehti.] The Psychonomic Society, Inc. [Viitattu 16.4.2021] Saatavissa: https://search-proquest-com.ezproxy.utu.fi/docview/347860708?accountid=14774. DOI: 10.3758/BRM.42.2.381

Martin, Maisa 1999: Suomi toisena ja vieraana kielenä. *Kielenoppimisen kysymyksiä.* S.157–178.

Martin, Maisa 2003: Kieli on kuin lammikko – johdatusta toisen kielen oppimiseen. *Suolla suomea. Perustietoa maahanmuuttajien suomen kielen opettajille.* S. 75–90.

Masonen, Virpi 2003: Sanasto, opettaja ja kielenoppija. *Suomi kakkonen. Opas opettajille*. Toim. Merja Mela ja Pirjo Mikkonen. S. 93–105. SKS, Helsinki.

Metsämuuronen, Jari 2001: Monimuuttujamenetelmien perusteet. *Metodologia-sarja 7*. Gummerus, Jyväskylä.

Nation, I. S. P. 2011 [2001]: Learning Vocabulary in Another Language. *Cambridge Applied Linguistics*. Cambridge University Press.

Niitemaa, Marja-Leena 2014: Kuinka vieraan kielen sanoja opitaan ja opetetaan. *Kuinka kieltä opitaan*. *Opas vieraan kielen opettajalle ja opiskelijalle.* S. 138–164.

Nissilä, Leena 2003: S2-opetuksen didaktiikkaa. *Suolla suomea. Perustietoa maahanmuuttajien suomen kielen opettajille.* S.103–114.

Nissilä, Leena – Martin, Maisa – Vaarala, Heidi – Kuukka, Ilona 2006: *Saako olla suomea? Opas suomi toisena kielenä -opetukseen.* Opetushallitus.

Nummenmaa, Lauri 2009: *Käyttäytymistieteiden tilastolliset menetelmät*. Tammi, Helsinki.

Näkökulmia toisen kielen puheeseen. Insights into Second Language Speech. Toim. Mikko Kuronen – Pekka Lintunen – Tommi Nieminen. AFinLA-e. Soveltavan kielitieteen tutkimuksia 10. Suomen Soveltavan Kielitieteen yhdistys AFinLA ry.

Pietilä, Päivi 2014: Yksilölliset erot kielenoppimisessa. *Kuinka kieltä opitaan*. *Opas vieraan kielen opettajalle ja opiskelijalle.* S. 45–67.

Pietilä, Päivi – Lintunen, Pekka 2014: Kielen oppiminen ja opettaminen. *Kuinka kieltä opitaan. Opas vieraan kielen opettajalle ja opiskelijalle.* S. 11–25.

Puro, Tarja 1999: Sanastollinen tieto ja suomen kielen oppikirjojen sanasto. Virittäjä, 103 (1). S. 2–26. [Verkkoartikkeli. Viitattu 24.4.2021] Saatavissa: https://journal.fi/virittaja/article/view/39126

---------- 2002: Suomi toisena kielenä -aikuisoppijan verbien kehittyminen alkeiskurssilla. Lisensiaatintyö. Jyväskylän yliopisto. [Viitattu 23.4.2021] Saatavissa: http://urn.fi/URN:NBN:fi:jyu-2002893574

Rasi, Ilkka – Kanniainen, Aila 2007: *SPSS for Windows – Menetelmiä*. Oulun yliopisto.

Read, John 2005 [2000]: *Assessing Vocabulary*. Cambridge University Press, Cambridge.

Sajavaara, Kari 1999: Toisen kielen oppiminen. *Kielenoppimisen kysymyksiä.* S. 75–102.

Singleton, David 1999: *Exploring the Second Language Mental Lexicon*. Cambridge Applied Linguistics. Cambridge University Press.

Suolla suomea. Perustietoa maahanmuuttajien suomen kielen opettajille. Toim. Leena Nissilä, Heidi Vaarala ja Maisa Martin. Äidinkielen opettajain liiton vuosikirja XLVII, 2003.

Schwartz, Mila – Katzir, Tami 2012: Depth of lexical knowledge among bilingual children: the impact of schooling. *Reading and Writing*, Vol. 25 (8). S. 1947–1971. [Verkkojulkaisu. Viitattu 22.4.2021] Saatavissa: https://link.springer.com/article/10.1007%2Fs11145-011-9308-9. DOI: 10.1007/s11145-011-9308-9

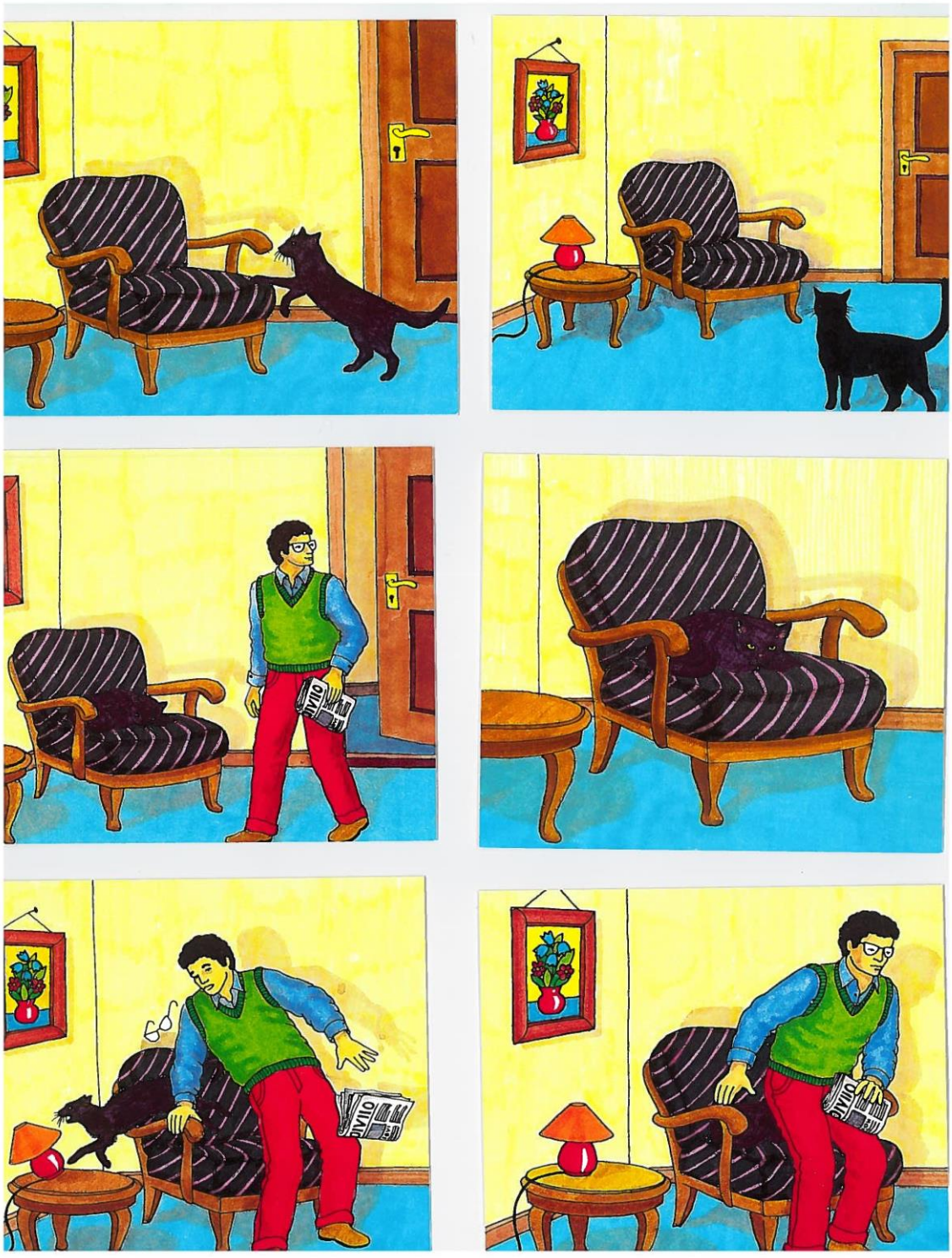
Viinikka, Jenni – Voutilainen, Eero 2013: Ääniä ilmassa, merkkejä paperilla – puhutun ja kirjoitetun kielen suhteesta. *Kielikello 3.* [Verkkoartikkeli. Viitattu 24.4.2021] Saatavissa: https://www.kielikello.fi/-/aania-ilmassa-merkkeja-paperilla-puhutun-ja-kirjoitetun-kielen-suhteesta#wrapper

Yu, Guoxing 2009: Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics*: 31/2. S. 236–259. Oxford University Press. [Verkkojulkaisu. Viitattu 25.4.2021] Saatavissa: http://web.a.ebscohost.com.ezproxy.utu.fi/ehost/detail/detail?vid=0&sid=68b3ad39-c756-4f78-908f-

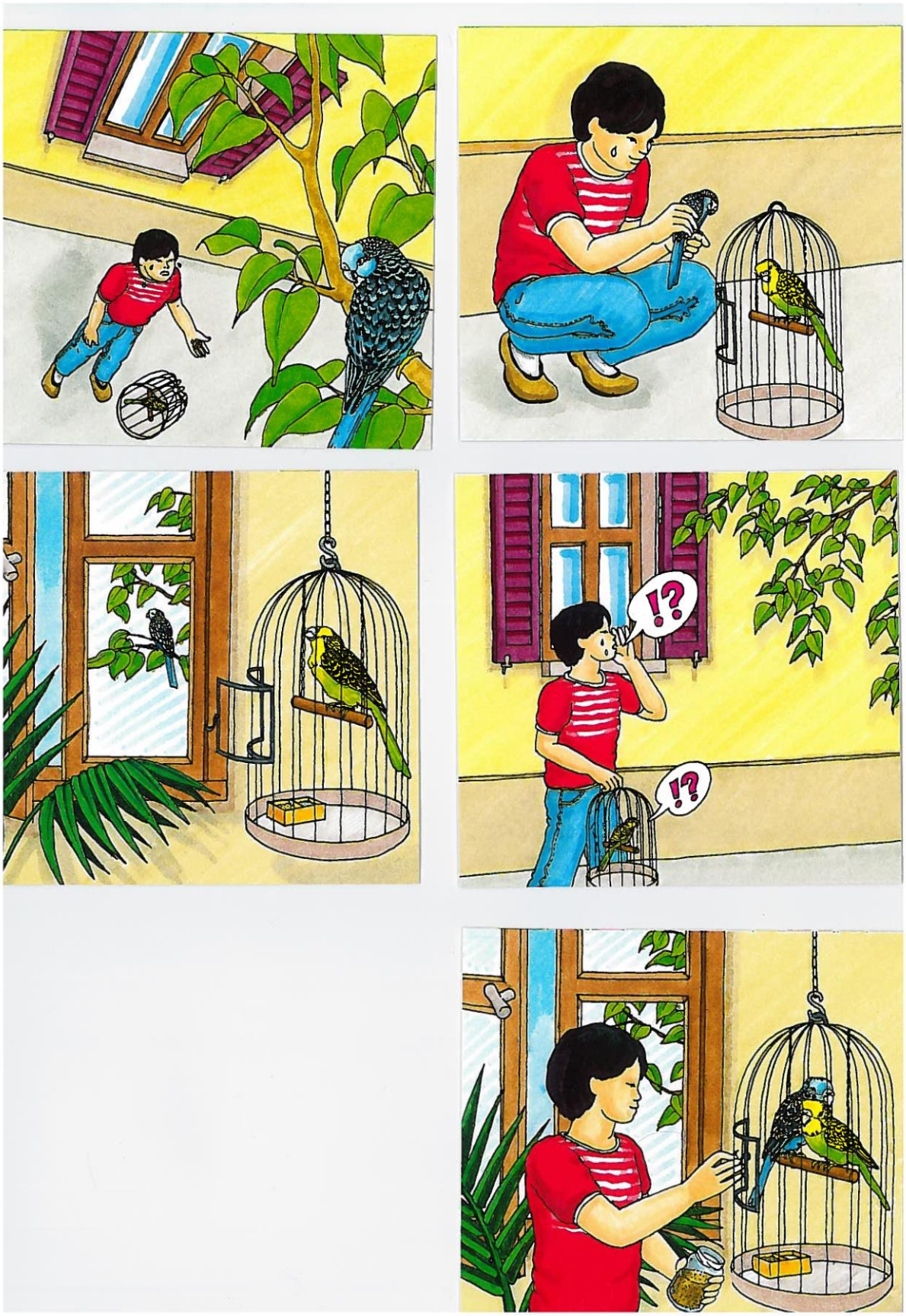
8623fb5532bc%40sessionmgr4006&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#AN=2010933035&db=mlf. DOI:10.1093/applin/amp024

Liitteet

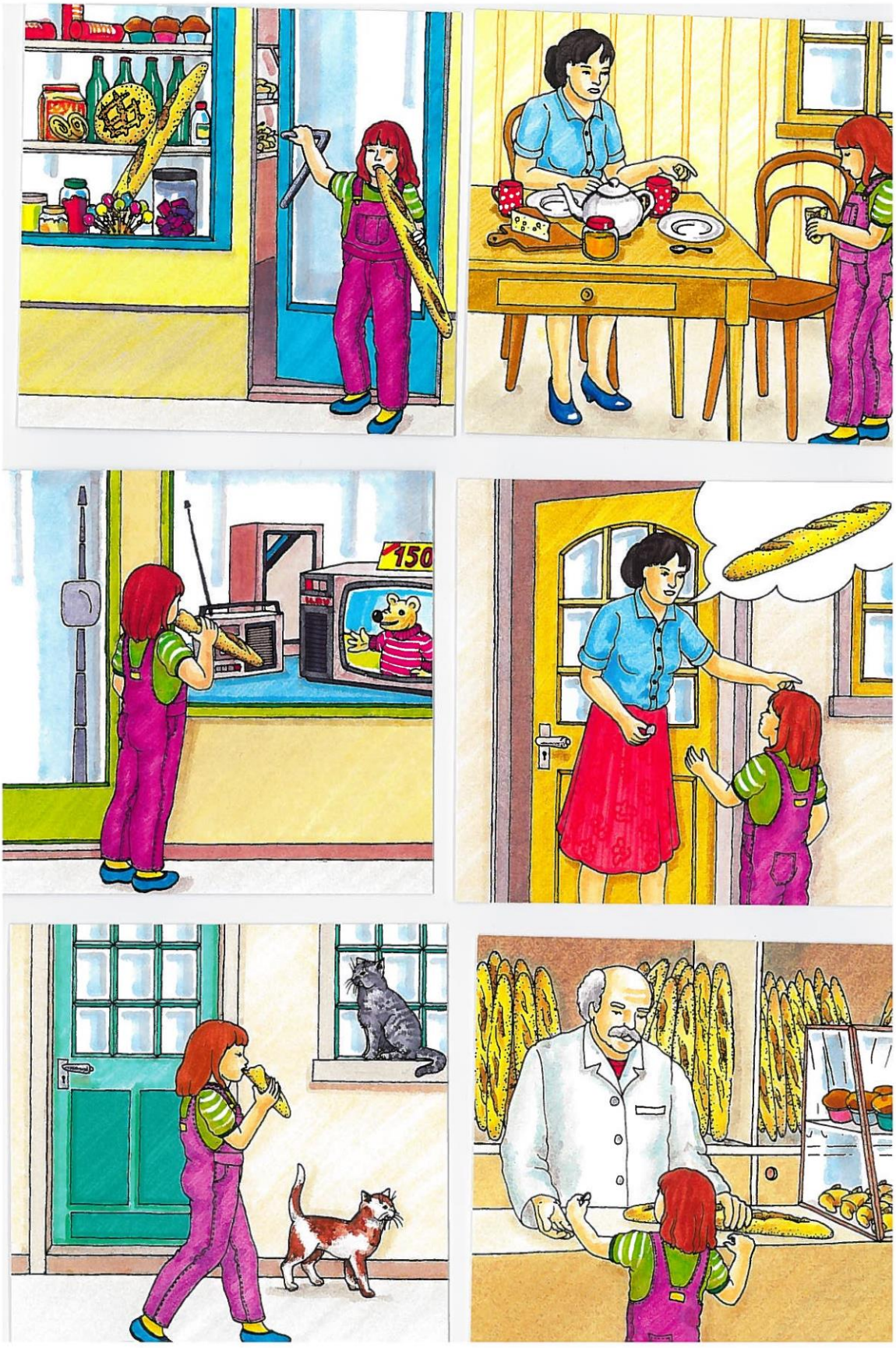
Liite 1: Kirjoitettujen kertomusten kuvasarja 1



Liite 2: Kirjoitettujen kertomusten kuvasarja 2



Liite 3: Suullisten kertomusten kuvasarja 1



Liite 4: Suullisen kertomuksen kuvasarja 2