Sadutettu sanasto – puhutun kielen leksikaalinen diversiteetti arviointikohteena

Mari Honko

Jyväskylän yliopisto

Abstract

This study analyses lexical diversity (*sums of probabilities*) in spoken narratives of L1 and L2 school age children (n = 99) and compares the results to the lexical diversity of written narratives of the group of comparison. The key research questions are: a) does the lexical diversity of the spoken narratives systematically differ from the lexical diversity of written narratives and b) does the lexical diversity of spoken narratives systematically differ depending on five individual variables: lexical skills, language proficiency, L1, gender and age of the speaker? All the narratives are produced in Finnish in storytelling events during the spring semester of the 2d and 3rd school year. Sums of probabilities is an index of lexical diversity. It is based on widely used D (Malvern & Richards 1997) and can be used with texts that differ in length and genre.

The results reveal a week but significant difference between the lexical diversity of spoken and written narratives and weak but complex correlations between the lexical diversity of spoken narratives and the proficiency level as well as the lexical skills of the child. However, there is no correlation or other appreciable connection between the lexical diversity and language background (L1/L2), gender or the school grade of the child. In addition, in L2 group there is no connection between the lexical diversity and the length of residence in Finnish speaking environment or, between the lexical diversity and the specific language difficulties observed by their teacher, either. The results are discussed and compared with the individual differences in turns and turn-taking during the storytelling event tentatively.

Keywords: lexical diversity, vocabulary skills, language learning, school age

Asiasanat: leksikaalinen diversiteetti, kielen oppiminen, toinen kieli, kouluikäiset

1 Johdanto

Artikkelissa tarkastellaan puhutun kielen leksikaalista diversiteettiä (*lexical diversity*), josta on suomenkielisessä kirjallisuudessa käytetty myös nimityksiä sanaston *monimuotoisuus* ja *rikkaus* (Malin 2012), *monipuolisuus* (Honko 2013) ja *vaihtelevuus* (Taimisto 2014). Leksikaalisen diversiteetin on havaittu olevan yhteydessä mm. kielenkäyttäjän mentaalileksikon kehittyneisyyteen sekä kokonaiskielitaitoon ja vaikuttavan niin kielen ilmaisuvoimaan kuin kielenkäyttäjästä muodostettuihin tulkintoihinkin (ks. esim. Bradac & Wisegarver 1984; Burroughs 1991; Malvern, Richards & Chipere 2004; Jarvis 2013a; Jarvis 2013b). Leksikaalista diversiteettiä on lukuisten tutkimusten perusteella pidetty lupaavana leksikaalisen tiedon ja taidon indikaattorina, sillä sen on katsottu peilaavan paitsi yksilön sanavaraston laajuutta ja kompleksisuutta myös hänen kykyään käyttää leksikaalisia resursseja tehokkaasti (ks. esim. Malvern & Richards 2002: 85; Malvern ym. 2004; Jarvis 2013a ja 2013b). Lisäksi leksikaalisen diversiteetin mittaamista ja tutkimista on pidetty hyödyllisenä tapana kuvata erilaisten tekstien laatua laajemminkin (McCarthy & Jarvis 2007: 476, 482–484; Yu 2010).

Tämän artikkelin tehtävänä on selvittää, a) poikkeaako kouluikäisiltä lapsilta kerättyjen puhuttujen kertomusten leksikaalinen diversiteetti aiemmin (Honko 2013) tutkittujen kirjoitettujen kertomusten diversiteetistä (luku 4.1) ja b) onko puhuttujen kertomusten leksikaalisessa diversiteetissä systemaattisia ryhmäkohtaisia eroja, jotka selittyvät lapsen kielitaidolla, ensikielellä, sukupuolella tai iällä (tarkemmin luku 4.2). Lisäksi tarkastellaan sitä, vaikuttaako aikuisen sadutustilanteessa tuottama sanasto lapsen kertomuksen leksikaaliseen diversiteettiin (luku 4.3). Tulosten kontekstoimiseksi kussakin analyysiosion alaluvussa esitellään myös aiemman tutkimuksen, erityisesti kirjoitelma-aineistoon pohjautuvan verrokkitutkimuksen (Honko 2013), tuloksia[1]. Havaintojen pohjalta arvioidaan, voisiko leksikaalinen diversiteetti toimia kehityksellisenä mittarina vastaavia aineistoja analysoitaessa. Vaikka leksikaalista

[1] Tutkimus on osa myöhemmän kielenkehityksen tarkasteluun keskittyvää hanketta (ks. esim. Pajunen 2012). Kiitän Tampereen yliopistoa, Tampereen Yliopiston Tukisäätiötä, Jyväskylän yliopistoa, artikkelin kahta nimetöntä arvioijaa sekä aivan erityisesti tutkimusavustaja Minna Bogdanoffia tuesta tämän tutkimusartikkelin syntyprosessin eri vaiheissa aineistonkeruusta viimeistelyyn.

diversiteettiä on aiemmassa tutkimuksessa hyödynnetty jo melko laajalti, sen määritelmä ja mittaustapa eivät ole vakiintuneet, minkä vuoksi tämän tutkimuksen rajaukset perusteluineen esitellään yksityiskohtaisesti luvussa 3.

Tämän tutkimusartikkelin varsinaisena aineistona on sadan alakouluikäisen lapsen suomenkielinen sadutusaineisto, joka on kerätty yhtenevän puolistrukturoidun tehtävänannon avulla. Tutkitun ryhmän lapset ovat iältään 8–11-vuotiaita ja edustavat eri kieli- ja kulttuuriryhmiä. Mukana on sekä suomea ensimmäisenä että toisena kielenä omaksuvia lapsia. Kieliaineistojen ohella käytettävissä on kyselytietoa muun muassa tutkimukseen osallistuneen lapsen kielitaidosta sekä iästä, Suomessa asumisen kestosta ja varhaislapsuudessa käytetyistä kielistä. Lisäksi käytettävissä ovat tulokset sanastonhallintaa erikseen mittaavasta strukturoidusta testistä. Sadutusmenetelmä, aineiston eri osiot ja analyysimetodit esitellään tarkemmin luvussa Aineisto ja metodit. Kirjoitetun kielen aineisto mahdollistaa leksikaalisen diversiteetin vertailun modaliteettien eli puheen ja kirjoituksen välillä ja toisaalta pakottaa metodin kriittiseen arviointiin: esimerkiksi se, että puhekieli harvoin jakautuu siististi ehyiksi ja erillisiksi leksikaalisiksi yksiköiksi, asettaa aineiston analysoinnille kirjoitetun kielen teksteistä poikkeavia haasteita.

Alakouluiässä sanaston määrällinen ja laadullinen kehitys on kiivasta ja yksilölliset erot sanaston hallinnassa suuria (ks. Honko 2013). Erot lasten leksikaalisissa valmiuksissa ovat merkityksellisiä, sillä vahva sanastonhallinta on laaja-alaisesti yhteydessä kielelliseen suoriutumiseen kuten luku- ja kirjoitustaitoon ja siten esimerkiksi kykyyn ottaa haltuun koulun oppisisältöjä (Saarela 1997; Alderson 2005; Tannenbaum, Torgesen & Wagner 2006; Milton 2009; Lervåg & Aukrust 2010). Sanastollinen osaaminen myös ennustaa menestymistä kielellisissä taidoissa myöhemmin: vahva sanasto helpottaa sekä kielellä toimimista että sen edelleen kehittämistä (Cain, Oakhill & Lemmon 2004; Dockrell & Messer 2004; Muter, Hulme, Snowling & Stevenson 2004; Qian & Schedl 2004; Honko 2013). Lapsuusiän kielitaitotutkimuksen avulla voidaan tunnistaa sanastollisten valmiuksien puutteita ja pyrkiä ennaltaehkäisemään erojen kasvua paitsi suoran kielellisen tuen avulla myös esimerkiksi tukemalla lasten sosiaalisia suhteita ja monipuolisia kielenkäyttömahdollisuuksia (Verhoeven 1990: 106–

107). Siksi leksikaalisten taitojen tutkiminen on perusteltua juuri koulunaloitusvaiheessa. Leksikaalisen diversiteetin kiinnostavuus arviointi- ja diagnosointivälineenä perustuu ennen kaikkea sen potentiaaliin hyvin monenlaisten tekstien arvioimisessa sekä selkeisiin, laskennallisia menetelmiä hyödyntäviin analyysimalleihin ja niiden toistettavuuteen.

Eroja leksikaalisessa diversiteetissä on aiemmissa tutkimuksissa havaittu muun muassa lasten ja aikuisten sekä eri-ikäisten lasten ja nuorten kirjoittamissa teksteissä (Berman & Verhoeven 2002; Johansson 2008) sekä jo varhaislapsuudessa eri-ikäisten lasten puheessa niin tyypillisen (Durán, Malvern ym. 2004) kuin epätyypillisen kielenkehityksen yhteydessä (Klee, Stokes, Wong, Fletcher & Gavin 2004). Leksikaalista diversiteettiä on pidetty potentiaalisena kehityksellisenä mittarina, mutta näyttö on toistaiseksi vahvinta kielenkehityksen alkuvaiheessa eli ensikielen osalta varhaislapsuudessa ja toisen kielen osalta alimmilla taitotasoilla.

Aiempi tutkimus on tehty suureksi osaksi englanti toisena ja vieraana kielenä -kontekstissa (ks. kuitenkin esim. Castañeda-Jiménez & Jarvis 2014). Koska leksikaalinen diversiteetti on kytköksissä tarkasteltavan kielen morfologiseen ja syntaktiseen rakenteeseen, eri kieliä koskevat tutkimustulokset eivät ole suoraan vertailukelpoisia (Dewaele & Pavlenko 2003: 132–133; Strömqvist, Johansson, Kriz, Ragnarsdóttir, Aisenman & Radvid 2002). Uutta tietoa tarvitaankin sekä puhutun kielen leksikaalisesta diversiteetistä yleisesti että leksikaalisen diversiteetin soveltamisesta erityisesti suomenkieliseen puhutun kielen aineistoon.

Tämä tutkimus rajataan tyypillisesti kehittyvien lasten puhutun kielen leksikaalisen diversiteetin tarkasteluun, sillä aiempien tutkimusten tulokset etenkin puhutun kielen aineistosta ovat ristiriitaisia (ks. Watkins, Kelly & Harbers 1995; Scott C. M. & Windsor 2000; Vermeer 2000; Wong, Klee, Stokes, Fletcher & Leonard 2010; Ellis, Holt & West 2015; Lai & Schwanenflugel 2016). Tuloksia verrataan lasten muuhun kielelliseen osaamiseen sekä aiemmassa kirjoitettuun kieleen kohdistuneessa tutkimuksessa saatuihin tuloksiin. Lisäksi tuloksia arvioidaan suhteessa niihin yksilöllisiin taustatekijöihin, joiden on aiemmassa tutkimuksessa havaittu vaikuttavan

leksikaaliseen diversiteettiin.

2 Leksikaalinen diversiteetti kielitaidon arvioinnissa

Leksikaalisen diversiteetin hyödyntäminen kielitieteellisessä tutkimuksessa juontaa juurensa 1930-luvun loppupuolelle John Carrollin artikkeliin Diversity of vocabulary and the harmonic series law of word frequency distribution, jossa Carroll (1938: 379) määritteli diversiteetin (*diversity*) sanaston suhteelliseksi toisteisuudeksi tai vaihteluksi tietyssä tekstissä ("the relative amount of repetitiveness or the relative variety in vocabulary"). Myöhempiin määritelmiin on vaihtelevasti sisällytetty myös sanojen sironta eli sijoittuminen tekstiin ja sanavalikoiman laatu, kuten käytettyjen sanojen harvinaisuus kielessä ja yksilöllisyys tarkastellussa useamman tekstin aineistossa (McCarthy and Jarvis 2010; Jarvis 2013b). Vaikka leksikaalista diversiteettiä on varhaislapsuuden jälkeen sovellettu enimmäkseen kirjoitetun kielen tutkimukseen, se koskee jo Carrollin mukaan sekä puhuttua että kirjoitettua kieltä ja on riippuvainen monista tekijöistä kuten tekstin tuottajan iästä, älykkyydestä ja taustasta[2]. Leksikaalisen diversiteetin määrittelytapa, sen tutkimisessa käytetty käsitteistö tai analyysimetodit eivät kuitenkaan ole vakiintuneet, mikä vaikeuttaa sekä tutkimustulosten vertailua että itse tutkimusmetodin arviointia.

Ajatus tietyn tekstin tai tekstikokoelman leksikaalisen diversiteetin yhteydestä tekstintuottajan ominaisuuksiin ja taustaan on tehnyt siitä monien tutkijoiden mielestä kiinnostavan metodin juuri kielitaidon arvioinnin näkökulmasta. Tyypillisesti on verrattu kielenoppijoiden ja äidinkielisten kielenpuhujien tekstejä, tarkasteltu eri-ikäisten ja eri kielitaitotasoa edustavien kielenoppijoiden tekstejä tai analysoitu eriasteisista kielihäiriöistä kärsivien tekstejä suhteessa tyypillisen kielenkehityksen ryhmään. Havaintojen mukaan korkea kielitaitotaso (sekä L1 että L2) ja tyypillinen kielenkehitys (vs. kielihäiriö) ovat tietyin varauksin yhteydessä korkeampaan leksikaaliseen diversiteettiin, ja puheen leksikaalisella diversiteetillä on arvioitu olevan

[2] Carroll itse (1938) pohjasi työnsä Zipfin (1935, 1937) aiempiin julkaisuihin, joissa diversiteetti määriteltiin kapeammin yksittäisten sanojen toiston ja esiintymisvälien kautta ("average rate of repetitiveness"). Toisaalta jo Zipf nosti esille muun muassa sanatoisteisuuden vaikutuksen tekstin vastaanottajan (lukijan tai kuulijan) kokemukseen, mikä ei myöhemmässä tutkimuksessa ole saanut juuri huomiota ennen kuin aivan viime vuosina (ks. myös Jarvis 2013b; Castañeda-Jiménez & Jarvis 2014).

vaikutusta myös siihen, millaisia tulkintoja kuulija tekee puhujasta ja kuinka tähän suhtautuu[3]. Miltonin (2009: 127) esittämän tiivistyksen mukaan kielellisen sujuvuuden lisääntyessä ja taitotason noustessa tuotetun kielen sanaston variaatiokin vähitellen kasvaa. Tuoreimmissa tutkimuksissa leksikaalista diversiteettiä on pidetty lupaavana mittarina myös kielellisen attrition osoittamisessa: attrition myötä leksikaalinen diversiteetti vähenee (Sang & Miseon 2013; Schmid & Jarvis 2014).

Leksikaalisen diversiteetin soveltaminen oppijankielentutkimukseen perustuu kahteen keskeiseen oletukseen: a) kielenoppimisen myötä osattujen sanojen määrä ja sanojen käyttämisessä tarvittava tieto karttuu ja b) karttuva osaaminen heijastuu suoraan henkilön tuottamien tekstien sanastoon: käytetystä mittarista riippuen joko pelkästään määrään ja vaihteluun tai määrään, vaihteluun ja laatuun. Tekstin leksikaalista diversiteettiä voidaan kasvattaa vain, jos siihen lisätään uutta sanastoa, ja siksi korkea diversiteetti jo kohtuullisen pitkissä teksteissä väistämättä edellyttää myös kielen harvinaisen sanaston käyttöä – ja siten laajaa aktiivista sanavarastoa (ks. myös Honko 2013).

3 Aineisto ja metodit

3.1 Sadutusaineiston yleisesittely

Perusaineisto koostuu 103 puhutusta kertomuksesta, jotka on kerätty saduttamalla, litteroitu ja syötetty Excel-tietokannaksi (ks. tarkemmin luku 3.2). Sadutus on osallistava metodi, jossa sadutettavaa pyydetään kertomaan suullisesti satu tai tarina vapaasti haluamastaan aineesta tai joskus myös rajatummin tietystä aiheesta. Tässä tutkimuksessa käytössä on aihesadutus, ja aihe (*kerro satu tai tarina Unelmien päivästä*) sekä ohjeistus ovat samat kuin kirjoitettujen kertomusten leksikaalista diversiteettiä käsittelevässä verrokkitutkimuksessa (Honko 2013). Sadutusmetodia on tyypillisesti käytetty arjen työkaluna lasten parissa, mutta sitä on hyödynnetty myös

[3] Ks. Bradac & Wisegarver 1984; Burroughs 1991; Watkins ym. 1995; Scott & Windsor 2000; Dewaele & Pavlenko 2003; Jarvis 2002; Durán ym. 2004: 234; Malvern ym. 2004; Wright, Silverman & Newhoff 2003; Treffers & Daller 2007; Unsworth 2008: 317–318, 322; Yu 2010; Jarvis 2013a ja 2013b; Honko 2013; Gregori-Signes & Clavel-Arroitia 2015.

tutkimuksessa ja aikuisten kanssa. (Ks. menetelmästä esim. Karlsson 2014 ja tutkimuksesta Riihelä 2013.) Tukea juuri sadutusmetodin käyttämiseen diversiteettitutkimuksessa löytyy epäsuorasti myös aiemmasta lingvistisestä tutkimuksesta, sillä Schmid ja Jarvis (2014) pitävät vapaasti tuotetun puheen leksikaalisen diversiteetin tarkastelua validimpana vaihtoehtona kuin muodollisten tehtävien tai kontrolloitujen kertomusten (*elicited narratives*) avulla kerätyn aineiston tarkastelua. Kaikkien sadutustuokioiden olosuhteet, tehtävänanto ja kerronnan aihe sekä aineiston käsittelytapa ovat olleet yhtenevät, mikä on kertomusten leksikaalisen diversiteetin vertailtavuuden kannalta tärkeää (Durán ym. 2004: 75). Sadutustuokion keskimääräinen kesto on noin 20 minuuttia.

Lyhyimmät, alle 50 saneen kertomukset (n = 4) on poistettu aineistosta ennen analyyseja. 50 saneen rajaa on pidetty tulosten luotettavuuden kannalta turvallisena vaihtoehtona, ja se vastaa kirjoitettujen kertomusten analysoinnissa tehtyä rajausta (ks. Durán ym. 2004: 228; Honko 2013: 363). Taulukossa 1 on eritelty tois- ja kolmasluokkalaisten, tyttöjen ja poikien sekä ensikielisten (S1) ja suomi toisena kielenä -oppijoiden (S2) aineisto. Kokonaismäärä (99) on tilastollisten analyysien kannalta riittävä. Alaryhmät (erityisesti ensikieliset suomenpuhujat, n = 32) ovat kuitenkin pienet ja jokaiselta puhujalta on tutkittu vain yhden sadutuskerran aineisto, minkä vuoksi tulosten yleistämisessä täytyy noudattaa varovaisuutta. S2-ryhmän lapsissa ei ole maassaoloajan (≥ 2 vuotta) ja koulunkäynnin vaiheen (yleisopetuksen 2.–3. lk.) perusteella aivan alkeistason suomenoppijoita.

TAULUKKO 1. Sadutettujen kertomusten määrä. Taulukossa on esitetty kertomusten kokonaismäärät ryhmittäin. Sulkeissa on lisäksi esitetty asetetun pituusehdon täyttävien ja siten analyyseihin nostettujen kertomusten kokonaismäärä.

n	2. lk.		3. lk.		
	S1	S2	S1	S2	**yht.**
pojat	5 (4)	20 (19)	8	16	**49 (47)**
tytöt	4	25 (23)	15	10	**54 (52)**
kaikki	**9**	**45**	**23**	**26**	**103 (99)**

Vaikka tehtävänanto ja saduttaja olivat kaikissa sadutustuokioissa samat, tuokioiden kulussa oli paljon vaihtelua: osa lapsista eteni kerronnassa yksittäisin sanoin ja lausekkein, osa kertoi vuolaasti ja pitkään. Sadutustuokion vuorojäsennystä havainnollistaa liitteen 1 esimerkkilitteraatti ja kertomusten pituuseroja taulukko 2, johon on koottu kertomusten sanemäärän osoittavat tunnusluvut (keskipituus, suurin ja pienin pituus, keskihajonta, mediaani). Lapsen puhe on lähtökohtaisesti sisällytetty kertomukseen kokonaisuudessaan (ks. aineiston käsittely ja poistot tarkemmin luku 3.2). Pisin tietokantaistettu kertomus on 1 380 ja lyhin 57 sanetta, keskipituus 305 ja keskihajonta 245 sanetta, eli hajonta on suuri. Aineiston kokonaissanemäärä on 30 641 ja eri sanojen määrä 2 229. Sadutusaineiston kertomukset ovat huomattavasti pitempiä kuin samanikäisten lasten samalla tehtävänannolla kirjoittamat tekstit (2. vuosiluokan keskipituus 53 ja kolmannen 83 sanetta, kun kirjoitusaika oli n. 45 minuutin, ei tarkkaa aikarajausta).

TAULUKKO 2. Sadutettujen kertomusten kokonaissanemäärät.

	min.	maks.	ka	kh	md
saneita	57	1380	305	245	227
eri sanoja	14	312	107	57,6	93

3.2 Aineiston käsittely: kontekstina puhuttu kieli

Puhe ei lukupuhuntaa lukuun ottamatta yleensä jakaudu siististi peräkkäin lausuttujen kokonaisten sanojen muodostamiksi lauseiksi, vaan keskustelu etenee toisiinsa lomittuvina vuoroina, joihin voi sisältyä kesken jääneitä ilmauksia, epäröintejä, toistoja ja korjauksia. Sanojen rajat voivat kadota tai olla häilyviä, jolloin syntyy kahden tai useamman sanan yhteensulautumia. Käytetyissä ilmauksissa ja niiden muodossa voi olla vaihtelua sekä eri yksilöiden välillä että samankin yksilön puheen eri kohdissa ja eri puhetilanteissa. Lisäksi sisältöjä ja merkityksiä voidaan rakentaa yhdessä tai kierrättää toisten puheesta.

Sadutus toisaalta poikkeaa arkisesta vuorovaikutustilanteesta epäsymmetrisyydellään (ks. esim. Habermas 1984): Sadutuksessa on yleensä läsnä tilannetta ohjaava aikuinen sekä lapsi. Vuorovaikutuksessa ei pyritä tasaiseen vuorotteluun, vaan aikuisen perustehtävä on alkuorientaation ja -ohjeistuksen jälkeen tukea lapsen kertomista mutta antaa ensisijainen puhetila lapselle ja välttää kertomisen ohjailua. (Karlsson 2014.) Saduttaja välttää tuomasta kerrontaan omia aineksiaan kuten juonirakenteita tai valmiita ilmauksia, mikä vähentää sekä vuorojen kerrostuneisuutta (kierrättämistä, yhdessä rakennettuja lausekkeita) että päällekkäispuhuntaa ja kesken jääneitä ilmauksia (liite 1; luku 4.3; myös Honko tulossa).

Eri tekstien diversiteettiarvojen vertailu edellyttää saman mittarin käyttämistä sekä tietokannaksi syötettäessä aineiston samanlaista käsittelytapaa eli tekstin kirjoitusasun yhdenmukaistamista ja mittarissa käytettävän leksikaalisen yksikön määrittelemistä (Durán ym. 2004: 228). Tässä tutkimuksessa litteroidut ja toisen kuuntelijan tarkistamat sadutustuokiot on syötetty Excel-tietokannaksi seuraavien periaatteiden mukaan:

1) Lapsen tuottama puhe on tuotu Excel-tietokantaan (sane per rivi) ja lemmattu kokonaisuudessaan sadutuksen orientaatio- ja lopetusvaihetta lukuun ottamatta.
2) Tietokannan perusyksikkö leksikaalisen diversiteetin tarkastelussa on perusmuotoinen sana (lemma), jota kirjoituksessa yleensä vastaa yksi sanavälein erotettu yksikkö (*pelata*, *nopeasti*, *hipi hiljaa*).
3) Taivutusmuodot ja foneettiset sekä murteelliset varieteetit (*pelata*, *belaa*, *pelattii*; *simmottii*, *semmosii*) kuuluvat samaan sanaan yleiskielisen varieteetin kanssa ja on lemmattu samalla tavalla (**pelata**, **semmoinen**). Menettelytapa mahdollistaa tulosten vertaamisen lasten kirjoitetun kielen korpuksesta havaittuun[4].
4) Epäröintiäännähdyksiä tai sanan toistettuja osia ei ole lemmattu (*s-sitten* > **sitten** *niin et ää-* > **niin**, **että**; *tul- mentiin* > **mennä**, *hää- juhlavaatteet* > **juhlavaate**). Toisinaan lapsen vuorosta voi päätellä, että epäröinnillä on myös sisällön rajoittamisen funktio (*Ara- Irakii*; *sitten se Ant- se poika otti*).
5) Kokonaisten sanojen toistamisella on kertomuksissa usein selkeä sisällöllinen funktio (*Kuusi oli hyvin hyvin iloinen*). Siksi toistetut sanat on lemmattu erikseen, kuten kirjoitetun kielen aineistossakin.
6) Monisanaiset erisnimet ja kiteytyneet ilmaukset on lemmattu yhtenä sanana.
7) Sanojen yhteen sulautumat (f = 436, ks. myös ISK § 140) kasautuvat aineistossa samoille puhujille, ja ne on eroteltu yleiskielen mukaan kuten verrokkikorpuksessakin (*son* > **se**, **olla**, *mostettiin* > **me**, **ostaa**; *sillee*/*sillai*/*sillei*

[4] On kuitenkin syytä huomata, että esimerkiksi perosoonapronominien tapauksessa puhekieliset muodot ovat korpuksessa yleiskielisiä muotoja yleisempiä (esim. *mä* f = 960, *minä* f = 604; *sä* f = 40, *sinä* f = 15).

> **sillä lailla**, *miksei* > **miksei**).

8) Epäröinti-ilmauksia kuten *öö-*, *ä-* ja dialogipartikkelia *mm* ei ole lemmattu.
9) Sanat on raakalemmattu aakkostetusta sanalistasta. Kaikki monitulkintaiset muodot (esim. *et* > **että** tai **ei**; *leikkii* > **leikki** tai **leikkiä**; *ankkuja* > **ankka**, *punkkoja* > **punkka**) on merkitty lemmauksen yhteydessä värikoodilla ja lemmattu lopullisesti juoksevasta tekstistä esiintymisympäristön perusteella.
10) Kokonaan tunnistamattomat sanat (esim. hyvin hiljaisesti äännetyt tai häiriöäänen peittämät) on jätetty analyysin ulkopuolelle. Tällaisia sanahahmoja koko aineistossa on kuitenkin vain noin 20 ja osuus aineistosta marginaalinen.

3.3 Kirjoitetun kielen verrokkiaineisto ja kyselyaineistot

Artikkelin ensimmäisessä analyysiosiossa (luku 4.1) puhutun kielen aineiston leksikaalista diversiteettiä peilataan kirjoitetun kielen leksikaaliseen diversiteettiin. Kirjoitelma-aineisto on kerätty, lemmattu ja analysoitu jo aiemmin tämän tutkimuksen kanssa yhteneviä periaatteita noudattaen (ks. Honko 2013). Aineiston määrä on koottu taulukkoon 3 kirjoittajaryhmittäin.

TAULUKKO 3. Kirjoitelma-aineiston kertomusten määrä. Taulukossa on esitetty pituusehdon täyttävien kertomusten kokonaismäärät ryhmittäin.

n	2. lk.		3. lk.		
	S1	S2	S1	S2	**yht.**
pojat	20	10	37	19	**86**
tytöt	47	13	69	25	**154**
kaikki	**67**	**23**	**106**	**44**	**240**

Yksilöllisinä vertailumuuttujina aineiston tarkastelussa käytetään lapsen kielitaitoa, ensikieltä, sukupuolta tai ikää. Kielitaidon yleistasoa on mitattu summamuuttujalla, joka koostuu kyselytietona kerätyistä kielitaitoarvioista kolmella kielitaidon osa-alueella: puhuminen, kirjoittaminen ja sanasto. Summamuuttujat on koostettu sekä lapsen itsearvioinneista että lapsen opettajalta pyydetyistä kielitaitoarvioista. Opettajan arviot ovat 5-portaisia (mahdolliset arvot 1–5, vaihteluväli summamuuttujan arvoissa 2,67–5,0) lapsen arviot 3-portaisia (mahdolliset arvot 1–3). Sekä opettajat että lapset ovat arvioineet myös lapsen mahdollisia erityisiä kielellisiä vaikeuksia. Sanastollisia taitoja on lisäksi arvioitu erillisellä sanastotestillä, joka mittaa sanaston reseptiivistä ja

produktiivista sanastonhallintaa eri yleisyystasoilla.[5]

3.4 Leksikaalisen diversiteetin mittari

Leksikaalisen diversiteetin arvioinnissa käytetyissä mittareissa on sekä yhteneväisyyksiä että eroja. Lähes kaikki mittarit rakentuvat tavalla tai toisella tarkasteltavan tekstin eri lekseemien ja saneiden suhteen varaan ja yksinkertaisimmillaan vain siihen kuten paljon käytetty TTR eli *type–token-ratio* (eri sanojen määrä jaettuna saneiden määrällä). TTR-pohjaiset mittarit ovat kuitenkin herkkiä verrattavien tekstien tai tekstiaineistojen kokoeroille ja siten epäluotettavia heterogeenisessa aineistossa[6]. Sen takia on pyritty yhä tarkempien ja tekstipituuden vaihtelun paremmin sietävien mittareiden kehittelyyn (esim. D: Malvern & Richards 1997; MTLD: McCarthy 2005, McCarthy & Jarvis 2007) sekä näiden mittarien vertailevaan validiointiin[7].

Tässä tutkimuksessa leksikaalisen diversiteetin mittariksi valittiin sama SOP-indeksi (*sums of probabilities*), jota on käytetty myös aiemmin tehdyssä kirjoitettujen kertomusten leksikaalisen diversiteetin arvioinnissa (Honko 2013). SOP:n Excel-version etuna on pidetty tarkkuutta: aineisto on analyysissa mukana kokonaisuudessaan. SOP pohjautuu Malvernin ja Richardsin kehittämään vocd-instrumenttiin, jonka tuottamaa D-arvoa (D-indeksiä) versioineen on hyödynnetty suurimmassa osassa 2000-luvun taitteessa leksikaalista diversiteettiä tarkastelleista tutkimuksista (ks. tarkemmin Malvern & Richards 1997, 2002; McKee, Malvern & Richards 2000; Malvern, Richards, Chipere & Durán 2004; Durán ym. 2004; McCarthy & Jarvis 2010). SOP, kuten D:kin, mahdollistaa hyvinkin erimittaisten tekstien vertaamisen (mm. McCarthy & Jarvis 2007: 460).[8] SOP-indeksin saamiseksi tekstin kullekin eri sanalle lasketaan esiintymistodennäköisyys valitun teoreettisen otoskoon tai otoskokojen avulla ja

---

[5] Ks. yksityiskohtaisempi selostus sekä taustatietolomakkeessa kysytyistä tiedoista että sanastotestistä Honko 2013.

[6] Pitkissä teksteissä sanatoisteisuus on yleensä luonnostaan lyhyitä suurempi, sillä jo sanatoisteisuuden pysyminen tasaisena edellyttäisi jatkuvasti uusien sanojen lisäämistä tekstiin samassa suhteessa kokonaissanemäärän lisääntymisen kanssa (Malvern ym. 2004: 124).

[7] Jarvis 2002; Malvern & Richards 2002; Koizumi & In'Nami 2012; Deboer 2014; Choi & Jeong 2016; Bonvin & Lambelet 2017.

[8] Lupaavina mittareina on pidetty myös tasapitkien tekstikatkelmien analysointiin perustuvaa MSTTR-indeksiä sekä TTR:n vakiointiin perustuvaa MTLD-indeksiä, joilla kuitenkin on omat rajoituksensa ja joiden käyttäminen ei tässä tutkimuksessa olisi mahdollistanut vertailua kirjoitettujen kertomusten verrokkiaineistoon (ks. tarkemmin esim. Jarvis 2013a: 94).

analyysin tulos ilmoitetaan kaikkien todennäköisyyksien summana.

Sadutusaineiston tarkastelussa teoreettiseksi otoskooksi asetettiin verrokkitutkimuksen kirjoitettujen tekstien aineistoa vastaten 42 tekstisanaa. (Teoreettisen otoskoon määrittymisestä ja SOP:n sekä vocD:n vertailusta ks. myös Honko 2013: 175–178.) Excel-taulukon (liite 2) esimerkki tarkentaa SOP-indeksin laskentaperiaatteen. SOP laskettiin erikseen jokaiselle sadutuskertomukselle SOP-riviarvojen summana. SOP-riviarvo puolestaan on määritelty tekstinäytteen jokaiselle leksikaaliselle yksikölle eli eri sanalle. SOP-riviarvo (≤ 1) kertoo todennäköisyyden, jolla kyseinen leksikaalinen yksikkö esiintyisi vähintään kerran teoreettisen otoskoon mittaisessa kyseisen tekstin näytteessä. Pitkässä tekstissä yksittäiset SOP-riviarvot ovat suhteellisen pieniä mutta yhteenlaskettavia arvoja saadaan enemmän. Liitteessä 2 on esitetty aineiston korkeimman (SOP $\approx$ 34,61) sekä matalimman (SOP $\approx$ 10,60) leksikaalisen diversiteetin tekstien SOP-laskentataulukot. Jälkimmäisen esimerkin leksikaalinen diversiteetti on tutkitussa aineistossa kuitenkin poikkeuksellisen matala (seuraavaksi matalin SOP-arvo on 21,24).

3.5 Tilastolliset analyysit ja hajonta-/sirontakuviot

Muuttujan arvojen normaalijakautuneisuuden testaamiseen on käytetty Shapiro-Wilkin testiä ja vertailtavien otosten varianssien yhtäsuuruuden testaamiseen F-testiä (*F test to compare two variances*). SOP-arvojen jakautuminen aineistossa ei noudata normaalijakaumaa (Shapiro-Wilk, $p<0{,}001$) vaan vinoutuu hieman oikealle, ja lisäksi vertailtavien otosten (kuten puhutun ja kirjoitetun kertomusaineiston) SOP-arvojen varianssien yhtäsuuruus ei toteudu (ratio of variances $\approx$ 0,049, $p<0{,}001$). Siksi analyyseissa on käytetty epäparametrisiä menetelmiä.

Koska parametristen testien edellytykset eivät ole voimassa, muuttujien arvojen (lineaarista) riippuvuutta on tutkittu Spearmanin järjestyskorrelaatiolla. Eri otosten muuttujan arvojen vertailussa, kuten puhutun ja kirjoitetun aineiston SOP-arvojen vertailussa on käytetty Mann Whitneyn U -testiä. Luokiteltujen muuttujan arvojen vertailuun perustuvassa testauksessa (SOP-arvojen jakaumien vertailu) on käytetty Khin

neliö -testiä.

Tilastollisten analyysien merkitsevyystasot on asetettu seuraavasti: $p \leq 0{,}05$ = tilastollisesti melkein merkitsevä, $p \leq 0{,}01$ = tilastollisesti merkitsevä, $p \leq 0{,}001$ tilastollisesti erittäin merkitsevä. (Ts. virheen todennäköisyys nollahypoteesin kumoamisessa on asetetuilla merkitsevyystasoilla korkeintaan 5 %, 1 % ja 0,1 %.) Tilastollisten analyysien ohella muuttujien välisiä suhteita on systemaattisesti arvioitu myös tarkastelemalla hajonta-/sirontakuvioita, jotka voivat paljastaa tarkasteltavien muuttujien epälineaarisen yhteyden ja viitata kompleksisiin, usean muuttujan yhteisvaikutuksesta muodostuviin riippuvuussuhteisiin aineistossa. Vain pieni osa kuvioista on kuitenkin tilan säästämiseksi nostettu artikkeliin.

4 Tulokset

Analyysiosio jakautuu kolmeen alalukuun. Ensimmäisessä luvussa käsitellään modaliteetin yhteyttä kertomuksen leksikaaliseen diversiteettiin ja verrataan toisiinsa lasten puhumalla ja kirjoittamalla tuottamien kertomusten leksikaalista diversiteettiä[9]. Toisessa alaluvussa tarkastellaan yksilöllisten tekijöiden yhteyttä puhuttujen kertomusten leksikaaliseen diversiteettiin. Keskeisiä vertailumuuttujia on viisi: sadutettavan lapsen sanastolliset taidot, yleinen kielitaitotaso, kielitausta (S1/S2), sukupuoli sekä ikä. Aikuisen läsnäolo sadutusvuorovaikutuksessa on lähtökohtaisesti kiinteämpi kuin pelkkään alkuohjeistukseen perustuvassa kirjoittamistilanteessa, mikä saattaa vaikuttaa sadutettavan lapsen käyttämiin kielellisiin ilmauksiin ja sitä kautta leksikaaliseen diversiteettiin. Siksi kolmannessa analyysiluvussa tarkastellaan tutkijan puheesta kierrätetyn sanaston mahdollista vaikutusta puhuttujen kertomusten leksikaaliseen diversiteettiin.

4.1 Modaliteetin vaikutus leksikaaliseen diversiteettiin

Vaikka puheen ja kirjoituksen erot ovat sähköisen viestinnän myötä kaventuneet,

[9] Kirjoitetun kielen aineistoa koskevat tulokset on julkaistu aiemmin osana väitöstutkimustani (Honko 2013), jossa tutkimusasetelma (kertomisen ohjeistus, aineiston käsittely- ja analysointitapa) oli modaliteettia lukuun ottamatta sama kuin nyt analysoitavassa sadutusaineistossa.

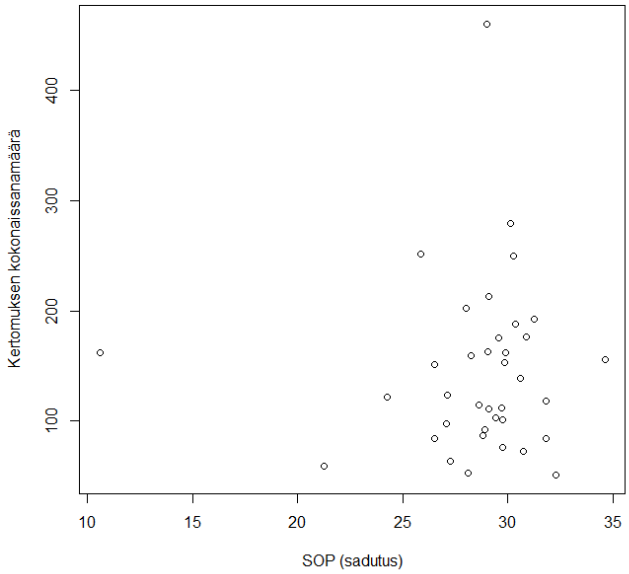
kirjoitettu teksti on usein puhetta prosessoidumpaa ja sen leksikaalinen diversiteetti on todettu puhetta korkeammaksi (ks. Strömqvist ym. 2002; Kuiken & Vedder 2012: 368). Myös sadutusaineistossa puhutun kielen matalin ja korkein SOP-arvo sekä keskimääräinen leksikaalinen diversiteetti ovat jonkin verran kirjoitettujen kertomusten verokkiaineistoa matalammat (taulukko 3 ja liite 3). Ero on tilastollisesti erittäin merkitsevä (Mann Whitneyn U -testi, $W = 8634$, $p<0{,}001$).

Seitsemältä lapselta on käytettävissä sekä sadutusaineisto että saman lukukauden aikana kirjoitettu kertomus. Aineisto on hyvin pieni, mutta suuntaa-antava korrelaatio puhutun ja kirjoitetun kertomustekstin yksilöllisessä leksikaalisessa diversiteetissä on korkea ja tilastollisesti melkein merkitsevä: $r = 0{,}821$, $p = 0{,}023$, $n = 7$ (käytössä Spearmanin järjestyskorrelaatioanalyysi rS). Tämä tarkoittaa, että leksikaalisen diversiteetin yksilöllinen taso näiden puhujien joukossa on suhteellisen pysyvä modaliteetista toiseen siirryttäessä. Jatkossa ilmiön tarkasteluun tarvittaisiin kuitenkin suurempi aineisto.

TAULUKKO 4. Leksikaalisen diversiteetin SOP-jakauma.

	min.	maks.	ka	kh	md
sadutusaineisto (n = 99)	11,00	34,61	28,96	2,94	29,43
ikäverrokkien kirjoitelmat (n = 239)	20,48	38,37	30,22	2,88	30,46
kirjoitelmat, reaaliseuranta (n = 7)	24,06	32,27	28,53	2,74	28,52

Puheaineistossa leksikaalisen diversiteetin ja kokonaissanemäärän välillä on tilastollisesti merkitsevä mutta matala korrelatiivinen yhteys ($rS = 0{,}422$, saneet, $p<0{,}001$). Kirjoitetuissa teksteissä leksikaalisen diversiteetin ja kokonaissanemäärän välillä on alaluokilla havaittu ainoastaan heikko positiivinen yhteys (Honko 2013: 378–380). Puheaineistossa leksikaalisella diversiteetillä on yhteys myös sadutuksen eri sanojen määrään ($rS = 0{,}555$, $p<0{,}001$). Korrelaatioanalyysin tulos ja sirontakuvio (kuvio 1) kertovat, että leksikaalinen diversiteetti ei kuitenkaan ole kokonaissanamäärän funktio: hyvin eripituiset kertomukset ovat saaneet korkeita diversiteettiarvoja.



KUVIO 1. Kertomuksen leksikaalisen diversiteetin suhde sen kokonaissanamäärään.

4.2 Yksilöllisten tekijöiden vaikutus leksikaaliseen diversiteettiin

a) sanastolliset taidot

Sanastollisia taitoja on arvioitu strukturoidulla testillä (ST), joka mittaa sekä produktiivista että reseptiivistä osaamista. Lisäksi lapsen opettajalta (S1/S2) on pyydetty holistinen arvio sanastollisista taidoista.

Sadutusaineistossa leksikaalinen diversiteetti ja sanastotestillä arvioidut sanastolliset taidot eivät korreloi ($rS = 0{,}146$, $p = 0{,}1714$, $n = 89$), eivät myöskään pelkästään produktiivis- tai reseptiivispainotteisten tehtävien avulla arvioituna. Sirontakuvion perusteella muuttujien välillä ei ole muutakaan säännöllistä yhteyttä. Leksikaalisella diversiteetillä on tosin tilastollisesti melkein merkitsevä yhteys opettajan antamaan holistiseen 5-portaiseen arvioon lapsen sanastollisista taidoista ($rS = 0{,}349$, $p = 0{,}025$, $n = 41$). Korrelaatiokertoimen arvo on kuitenkin jälleen niin matala, että yksilötason arvioinnin kannalta yhteyttä ei voi pitää merkityksellisenä.

Kirjoitelmakorpuksesta laskettu kertomuksen leksikaalinen diversiteetti sen sijaan korreloi kaikilla vuosiluokilla kirjoittajan sanastollisiin taitoihin (ST). Korrelatiivinen yhteys on vain kohtalainen ($rS = 0{,}433–0{,}537$) mutta suurehkossa ryhmässä tilastollisesti erittäin merkitsevä ($p<0{,}001$) (Honko 2013: 383). Myös yhteys opettajan arvioimaan kielitaidon yleistasoon (sanastollisen taidon, kirjoitustaidon ja puhetaidon holististen arvioiden indeksi) on tilastollisesti merkitsevä, mutta muuttujien selitysvoima toisiinsa nähden hyvin alhainen ($rS = 0{,}302$, $p=0{,}002$, $n=104$).

Kiinnostavaa on, että sadutetun sanaston leksikaalisella diversiteetillä löytyy yllättävä yhteys sanastolliseen osaamiseen 3 vuotta myöhemmin ($rS = 0{,}419$, $p = 0{,}008$, $n = 39$). Käytetty testi (ST2) on mukautettu uuteen ikätasoon (5.–6. luokka) ja siinä painottuu aiempaa testiä (ST) enemmän sanavaraston laajuuden arviointi. Perusperiaatteet ja erottelukyky ovat kuitenkin samat (ks. Honko 2013). Alaluokilla lasten erot tekstitaidoissa voivat kuitenkin vaikuttaa sanastotestin tulokseen, mikä selittäisi yhteyden puuttumista juuri alaluokilla.

b) yleinen kielitaitotaso

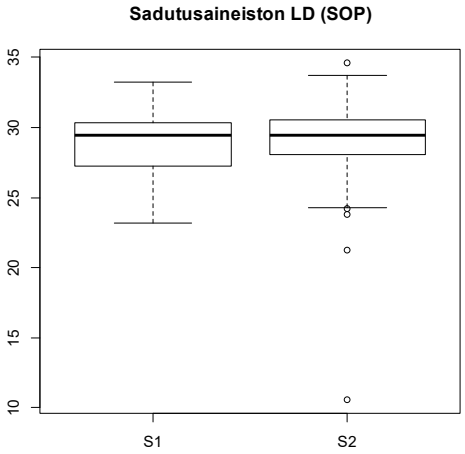
S2-ryhmässä sadutusaineistossa leksikaalisella diversiteetillä on heikosti merkitsevä matala korrelatiivinen yhteys sekä opettajan ($rS = 0{,}32$, $p = 0{,}038$, $n = 41$) että lapsen itsensä arvioimaan kielitaidon yleistasoon ($rS = 0{,}34$, $p = 0{,}033$, $n = 41$). Yhteyttä sen sijaan ei löydy muihin pyydettyihin taustatietoihin: edelliseen suomen kielen todistusarvosanaan (numeerinen tai numeeristettu sanallinen arvio, $n = 28$), opettajan tai lapsen arvioimiin kielellisiin erityisvaikeuksiin (summamuuttuja $n = 40$, $n = 41$) tai Suomessa asumisen kestoon ($n = 69$).

Kielitaidon yleistason yhteys kirjoitettujen kertomusten leksikaaliseen diversiteettiin on opettajan arvioimana merkitsevä mutta matalampi kuin yhteys erillisellä testillä (ST) mitattuun sanastonhallintaan (Honko 2013). Lapsen itsearvioinnin tulos sen sijaan ei ole yhteydessä kirjoitetun kertomuksen leksikaaliseen diversiteettiin. Lisäksi kielitaidon yleisarvioinnin karkeana kuvaajana on käytetty edellistä suomen kielen

todistusarvosanaa (vaihteluväli 6–9) ja opettajan sekä lapsen arvioimien kielellisten erityisvaikeuksien määrää (vaihteluväli 0–9 ja 0–12) sekä maassaolon kestoa (2–11 vuotta). Maassaolon keston yhteys kielitaitoon tosin tiedetään tutkitussa ryhmässä kompleksiseksi (Honko 2013).

c) ensikieli

Sadutusaineistossa leksikaalisessa diversiteetissä ei ilmene eroa kieliryhmien välillä (Mann-Whitney W = 1009, p = 0,8461, ks. kuvio 2). Eroa S1- ja S2-oppilaiden leksikaalisessa diversiteetissä ei ole myöskään erikseen tyttöjen ja poikien tai tois- ja kolmasluokkalaisten sadutuksista arvioituna (W = 276, p = 0,8047; W = 233, p = 0,9088; W = 232, p = 0,1249; W = 210, p = 0,1189). Hajonta on S2-aineistossa hieman suurempi, ja ylin neljännes sijoittuu hieman ensikielisten ryhmää korkeammalle tasolle. Tulosta ei selitä ero tuottamisen runsaudessa, sillä sadutuksen sane- tai sanamäärä eivät eroa kieliryhmittäin.



KUVIO 2. Kielitaustan yhteys sadutetun kertomuksen leksikaaliseen diversiteettiin.

Kirjoitelma-aineistossa S1-oppilaiden keskimääräinen leksikaalinen diversiteetti sen sijaan on ryhmätasolla kaikilla vuosiluokilla S2-oppilaiden tekstien diversiteettiä korkeampi (Honko 2013). Tasavälein ryhmitettyjen SOP-arvojen ristiintaulukointi ja

tarkastelu khiin neliö -testillä osoittavat, että eniten toisistaan poikkeavat jakauman ääripäät: S2-oppijoiden teksteissä on suhteessa enemmän matalan ja vähemmän korkean leksikaalisen diversiteetin kertomuksia. Tähän tulokseen suhteutettuna on yllättävää, että puheaineistossa kieliryhmien leksikaalisessa diversiteetissä ei ilmene eroa.

d) sukupuoli

Sadutusaineistossa poikien leksikaalisen diversiteetin mediaani on hieman tyttöjen leksikaalisen diversiteetin mediaania korkeampi, mutta ero ryhmien välillä ei ole tilastollisesti merkitsevä (Mann-Whitney W = 1494, p = 0,059). Tätä selittää poikien aineiston suurempi hajonta: poikien aineistossa ovat sekä matalimmat että korkeimmat diversiteettiarvot (liite 3).

Alakoululaisten kirjoittamistutkimuksissa tyttöjen kirjoittamien tekstien leksikaalinen diversiteetti on havaittu poikien kirjoittamien tekstien diversiteettiä korkeammaksi (Honko 2013, ks. myös Saarela 1997). Aikuisten puhutun kielen aineistoista tehtyjen tutkimusten tulokset ovat kuitenkin ristiriitaiset: Dewaelen ja Pavlenkon (2003: 134) tutkimuksessa naisten käyttämä sanasto on miesten sanastoa vaihtelevampaa, Singhin (2001: 260–261) ja Härnqvistinin, Christiansonin, Ridingsin ja Tingsellin (2003: 191) tutkimuksessa tulos on päinvastainen.

e) ikä

Tyypillisessä kielenkehityksessä oppijan leksikko karttuu ja syvenee kouluvuosina huimasti, mikä heijastuu spontaanin tuottamisen sanastoon (Berman 2007; Pajunen 2012). Aiemman tutkimuksen perusteella leksikaalinen diversiteetti, myös SOP, erottelee kielitaidoltaan eritasoisia puhujia ja kirjoittajia paremmin varhaisemmassa kielenkehityksen ja kirjoitustaidon vaiheessa (Vermeer 2000; Jarvis 2002; Honko 2013: 370–372).

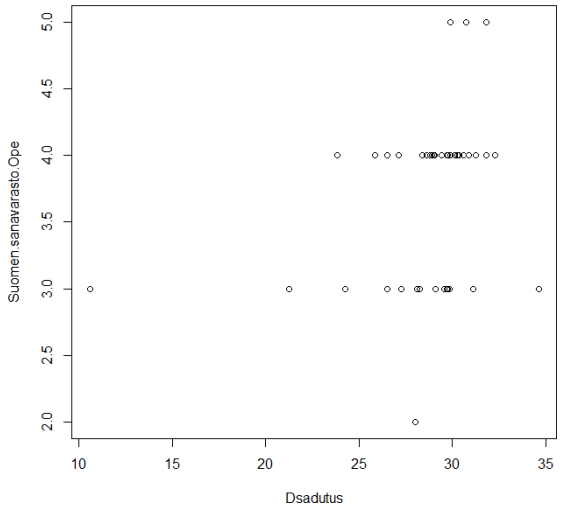
Tois- ja kolmasluokkalaisten sadutusaineistossa SOP-diversiteettiarvot eivät poikkea toisistaan (Mann-Whitney W = 1149, p = 0,6019). Ikäluokittain havaintoja ei ole

riittävästi kaikissa luokissa (7–11 vuotta) Mann Whitneyn U-testin suorittamiseen, mutta Spearmanin järjestyskorrelaatioanalyysi ja sirontakuvio paljastavat, että iän ja leksikaalisen diversiteetin välillä ei tutkitussa aineistossa ole korrelatiivista ($rS = 0{,}088$) tai muutakaan havaittavaa säännönmukaista yhteyttä.

Kirjoitetuissa kertomuksissa leksikaalisen diversiteetin kehitys sen sijaan on kaikissa alaryhmissä vielä alakouluvaiheessa nousujohteista vaikkakaan ei täysin lineaarista luokkatasolta toiselle. Ero vuosiluokkien 2–3 välillä on riippuvien otosten t-testillä mitattuna tilastollisesti erittäin merkitsevä. (Honko 2013: 365–366.)

4.3 Kierrätetyn sanaston vaikutus leksikaaliseen diversiteettiin

Ne lapset, joiden sanastolliset taidot opettaja on arvioinut korkeimmaksi, ovat myös käyttäneet sadutuksessa verrattain vaihtelevaa sanastoa. Toiseen suuntaan yhteys ei kuitenkaan päde (kuvio 3). Tulos saattaa tarkoittaa, että joidenkin oppilaiden sanastollinen osaaminen jää arjessa opettajalta piiloon ja todentamatta myös sanastotestissä (ks. luku 4.1 kohta b). On kuitenkin mahdollista, että sadutustilanteet eivät ole keskenään täysin vertailukelpoisia. Yksi selitys voisi piillä siinä, että heikompaa sanastonhallintaa on vuorovaikutustilanteessa mahdollista kielellisesti kompensoida vuorovaikutuskumppanilta kierrätetyllä sanastolla.



KUVIO 3. Opettajan arvioiman sanavaraston yhteys sadutetun kertomuksen leksikaaliseen diversiteettiin.

Toisen kielenkäyttäjän tekstien käyttäminen lähteenä voi kasvattaa leksikaalista diversiteettiä, mikä on aikaisemmassa tutkimuksessa todennettu kirjoitetun kielen kontekstissa (Gebril & Plakans 2016). Sadutusaineistossa tämä tarkoittaa sitä, että sadutettu lapsi voi omassa puheessaan kierrättää paitsi sadutustuokion ulkopuolella oppimiaan sanoja myös sen aikana aikuisen puheesta poimimaansa kielenainesta – kuten sanastoa. Sadutettujen kertomustekstien leksikaalisen diversiteetin ja kielitaustan sekä luokkatason ja muilla menetelmillä arvioidun sanastollisen osaamisen välisen riippuvuuden puuttuminen (luku 4.2) voisikin selittyä sillä, että sanastollisilta taidoiltaan heikoimmat lapset ovat vuorovaikutustilanteessa aktiivisimpia sanaston kierrättäjiä. Sen takia on tarpeen tarkastella erikseen sanaston mahdollista kierrättämistä sadutusaineistossa.

Jo sadutustuokioista kirjoitettujen keskustelulitteraattien alustava laadullinen tarkastelu osoittaa, että saduttajan rooli uuden sanaston tuojana on tutkitussa vuorovaikutusaineistossa hyvin vähäinen: alkuorientaation jälkeen pääosa saduttajan vuoroista koostuu pelkästä dialogipartikkelista kuten *mm*, *joo*, *nii* (ks. myös liite 1) tai lyhyestä kehotuksesta tai kysymyksestä (*kerro lisää*, *mitä sitten tapahtui?*), johon lapsi

reagoi esimerkiksi jatkamalla kertomista. Silloinkin, kun se olisi mahdollista, lapsi kierrättää aikuisen puheen kautta tarjoutuvaa sanastoa omaan puheeseensa vain harvoin. Saduttajalle (S) sen sijaan on tyypillisempää toistaa sanoja ja laajempia ilmauksia lapsen (L) puheesta (esimerkki 1).

(1) L: mä tulisin **kouluun** ja **tekisin** mitä opettaja olis sanonu.
(.)
S: joo.
(.)
S: sä tykkäät käydä **koulussa**.
(.)
S: kiva .hh
(..)
S: mitä **koulussa tehtäis** (.) semmosena päivänä.

Saduttajan rooli kielellisenä osallistujana korostuu tilanteissa, joissa lapsi tarvitsee tukea: aikuinen toistaa lapsen puhetta, toisinaan myös kokoaa tai jatkaa lapsen vuoroja tai esittää lisäkerrontaan rohkaisevia kysymyksiä (esimerkki 2).

(2) S: ºmikäs olis sinulle **ihanin päivä**. =mitä siellä olis. º
(4 s.)
L: ai **ihanin päi[vä**.
S: [**ihanin päivä** maailmassa. =mitä siellä olis.

Kierrättämistä oli tarpeellista tutkia myös tarkemmin: sadutustuokioista kirjoitettujen litteraattien avulla etsittiin kaikki sellaiset sisältösanaluokkien sanat, joita lapsi käyttää sadutustuokiossa ensimmäisen kerran vasta aikuisen käytettyä sanaa aiemmin omassa puheessaan. Tarkastelu rajattiin eri diversiteettitasoilta poimittuun 20 kertomuksen otokseen (20 % koko aineistosta). Analyysissa huomioitiin sanan kaikki esiintymiskontekstit: esiintyminen osana yhdyssanoja ja erilaisia monisanaisia konstruktioita joko välittömästi aikuisen vuoron jälkeen tai vasta myöhemmin sadutustuokion aikana.

Näin laskettuna kierrätettyjä sanoja esiintyy lapsen puhutuissa kertomuksissa keskimäärin vain 0,1 esiintymää sadutusta kohti (vaihteluväli 0–5). Kierrätettyjen eri sanojen määrä jää vielä matalammaksi (0–3). Kahdessatoista eli yli puolessa (60 %) tarkastelluista sadutustuokioista ei esiinny lainkaan aikuisen vuoroista kierrätettyä sanastoa. Havaintojen perusteella sanojen kierrättäminen aikuisen puheesta ei selitä

aiemmissa luvuissa esitettyjä tuloksia.

Tuloksen voi olettaa olevan vahvasti sidoksissa sadutusvuorovaikutuksen luonteeseen ja heijastavan sadutusmetodin ohjeistusta: saduttajan ei kuulu tarjota lapselle kerronnan sisältöjä (Karlsson 2014). Myös tutkitussa aineistossa saduttajan puhe on karsittua; vuorot ovat lyhyitä ja rakentuvat niukan, pitkälti tehtävänantoa ja lapsen omissa ilmauksissaan käyttämän sanaston varaan. Vuorotterurakenteen tarkempi tarkastelu kuitenkin paljastaa, että sillä saattaa olla muunlaisia – leksikaaliseen diversiteettiin heijastuvia – vaikutuksia sadutettavan tuottamaan puheeseen. Sadutustuokion aikana eniten tukea tarvitsevien lasten vuorot ovat yleensä hyvin lyhyitä lausekkeita, joista puuttuu vapaalle kerronnalle tyypillinen sisällöllinen ja rakenteellinen yhtenäisyys (liite 4). Pidempiin yhtenäisiin vuoroihin perustuvassa kerronnassa sen sijaan yhtenäisyyttä luodaan muun muassa sisäisiä viittaussuhteita rakentavalla leksikaalisella toistolla, jolla voi olla myös puheen prosessointiin ja sadustustuokion vuorojäsennyksen muokkaamiseen liittyviä tehtäviä. Vuorojäsennyksen ja vuorojen rakenteen vaikutusta leksikaaliseen diversiteettiin ei ole aiemmin tutkittu, mutta systemaattinen tarkastelu on kiistatta tarpeen, mikäli puhutun kielen leksikaalista diversiteettiä jatkossakin tutkitaan keskusteluvuorovaikutusaineistoista.

5 Tulosten koonti ja pohdinta

Tutkimuksen tehtävänä oli selvittää, onko a) lasten sadutettujen kertomusten leksikaalisessa diversiteetissä systemaattisia eroja eri modaliteettien tai puhujaryhmien välillä ja b) voisiko leksikaalinen diversiteetti toimia kehityksellisenä mittarina vastaavia aineistoja analysoitaessa. Kirjoitetun kielen leksikaalinen diversiteetti osoittautui puheen diversiteettiä korkeammaksi, mikä tukee aiempien tutkimusten tuloksia. Myös yksilötasolla modaliteettien välinen yhteys on tarkastellussa pienessä ryhmässä havaittavissa. Kieliryhmien (L1/L2), tyttöjen ja poikien tai tois- ja kolmasluokkalaisten välillä leksikaalisessa diversiteetissä ei kuitenkaan ilmennyt eroja, ja sekä kielitaidon yleistason että sanaston hallinnan yleistaso (testisuoritus tai opettajan arvio) yhteys leksikaaliseen diversiteettiin osoittautui epälineaariseksi ja heikoksi. Sadutettujen kertomusten leksikaalisella diversiteetillä ei tarkastellussa aineistossa ollut

yhteyttä myöskään lapsen edelliseen suomen kielen kouluarvosanaan, opettajan listaamien kielellisten erityisvaikeuksien määrään tai Suomessa asumisen kestoon (S2-ryhmä). Tulosten perusteella leksikaalista diversiteettiä (SOP) ei voi pitää riittävän tarkkana menetelmänä yksilötason kielitaidon tai tarkemmin sanastollisten taitojen arvioimiseen.

Myös aiemman puhevuorovaikutuksesta tehdyn tutkimuksen tulokset ovat jättäneet leksikaalisen diversiteetin arviointikäyttöön varauksia. Sadutustuokioiden vuorojäsennyksen laadullinen tarkastelu osoittaa, että eri diversiteettitasojen tekstit saattavat poiketa toistaan laadullisesti muutoinkin kuin sanaston tasolla. Vaikka ryhmätasolla korkea leksikaalinen diversiteetti yhdistyykin tuottamisen runsauteen, yhteys ei ole lineaarinen: Korkean leksikaalisen diversiteetin kertomukset näyttävät usein koostuvan lyhyistä ja rakenteeltaan yksinkertaisista vuoroista, jotka on tuotettu ikään kuin reaktiona saduttajan rohkaisevaan viestintään, matalan diversiteetin kertomuksissa puolestaan on paljon pitkiä, spontaanisti tuotettuja vuoroja. Yksilötasolla kertomisen niukkuus voi siksi selittää korkeaa leksikaalista diversiteettiä ja yhtenäisen, vuolaan kertomuksen leksikaalinen diversiteetti puolestaan olla matala. Vaikka sadun kertomisessa on monologimaisia piirteitä, sadutustuokio on kuitenkin vuorovaikutustilanne ja lapselle mahdollisesti myös uudenlainen ja jännittävä tilanne. Puhevuorovaikutukseen osallistumiseen sadutuksessa tarvitaan sanastollisen osaamisen lisäksi paljon muutakin – tehtävänantoon reagoimisen lisäksi esimerkiksi avoimuutta ja uskallusta sekä henkilökohtainen tarve osallisuuteen. Siksi sadutettu puhe väistämättä antaa kapeana kuvan sadutettavan kokonaiskielitaidosta.

Tämän tutkimuksen perusteella näyttää siltä, että aikuisen tuen suora vaikutus leksikaaliseen diversiteettiin sadutustuokioissa on kuitenkin hyvin pieni: lasten kerronnassa esiintyy hyvin vähän aikuiselta kierrätettyjä sanoja, mikä johtuu osittain aikuisen pelkistetyistä vuoroista (Honko, tulossa). Koska saduttajan vuoroista kierrätettyä sanastoa esiintyy vähän, kierrättämisen vaikutus leksikaaliseen diversiteettiin on minimaalinen. Tukea tarvitsevien lasten niukka kerronta ja lyhyet vuorot sen sijaan johtavat usein välillisesti suhteellisen korkeaan diversiteettiin, kun esimerkiksi tekstiä sidostava leksikaalinen toisto ja funktiosanojen käyttö on niukkaa.

Oletuksia siitä, että leksikaalinen diversiteetti voisi toimia kehityksellisenä mittarina ja jopa diagnostisena työkaluna on kritisoitu muun muassa kielenoppimisprosessin yksinkertaistamisesta ja kielenkäytön tilanteisuuden sivuuttamisesta: Kaikki sanastollinen osaaminen ei esimerkiksi ole luonteeltaan määrällistä eikä siis heijastu sanavaraston kasvuna. Kaikissa tilanteissa ja tekstilajeissa ei myöskään tarvitse tai edes kannata käyttää samanlaista kieltä – ja siten myöskään samalla tavalla varioivaa sanastoa. On selvää, että esimerkiksi ääneen prosessointi sanoja toistamalla yksittäistapauksissa laskee leksikaalista diversiteettiä (*mä olin- mä- mä- mä oon aina nii tehny*). Lisäksi puhutussa kielessä leksikaalisen tiiviyden haittapuolet korostuvat, minkä vuoksi hyvin korkea leksikaalinen diversiteetti ei aina ole vuorovaikutuksessa etu (ks. myös Broeder, Extra & van Hout 1993: 149). Kuten tässä artikkelissa aiemmin esillä olleiden tutkimusten perusteella voidaan todeta, pidemmällä edenneet kielenoppijat kuitenkin tyypillisesti toistavat sanastoa vähemmän ja heidän tuottamissaan teksteissä leksikaalinen diversiteetti on suhteessa korkeampi kuin verrokkiryhmissä (ei-äidinkieliset tai alkeistason oppijat). Heterogeenisissa aineistossa leksikaalinen diversiteetti voikin toimia suuntaa-antavana kielitaidon mittarina, mikäli aineistoa on riittävästi ja se on tuotettu yhdenmukaisilla tavoilla. Jatkossa tarkastelua on kuitenkin syvennettävä laajemmalla aineistolla ja tarpeen mukaan myös monimuuttuja-analyysia käyttäen sen selvittämiseksi, millaisia kerrannaisvaikutuksia tai mahdollisesti myös toisensa pois rajaavia vaikutuksia eri muuttujilla on suhteessa leksikaaliseen diversiteettiin.

Leksikaalisen diversiteetin menetelmällisessä tutkimuksessa on tyypillisesti keskitytty siihen, kuinka hyvin käytetty mittari ennustaa kielellistä suoriutumista verrattuna johonkin toisentyyppiseen mittariin. Oletuksena on ollut, että leksikaalisen diversiteetin määritelmään joka tapauksessa kuuluu ainakin sanojen valikoiman laajuus (*range*) ja vaihtelu (*variety*) (McCarthy & Jarvis 2007: 459). Aivan viime vuosina on kuitenkin havahduttu huomaamaan, että edes diversiteettimittarin tilastollisesti merkitsevä erottelukyky ei takaa sen käytettävyyttä. Samalla on ryhdytty tarkemmin perehtymään leksikaaliseen diversiteettiin tutkittavana ilmiönä, toisin sanoen tutkimaan, mistä – ja erityisesti mistä muusta kuin sanatoisteisuudesta tai tarkasteltavan tekstin

kokonaissanamäärästä – leksikaalinen diversiteetti mahdollisesti koostuu (Jarvis 2013b).

Näyttää siltä, että leksikaalisen diversiteetin määritteleminen pelkästään tekstin sisältämien sanojen vaihteluksi on pelkistys, joka jättää huomiotta mm. sanojen ominaislaadun. Tämä ja monet muut seikat saattavat kuitenkin olennaisesti vaikuttaa tekstin vastaanottajan kokemukseen leksikaalisesta diversiteetistä, sanojen määrän ja vaihtelevuuden ohella (Crossley, Salsbury & McNamara 2011; Crossley, Salsbury, McNamara & Jarvis 2011; Jarvis 2013b). Tämän tutkimuksen perusteella leksikaalisen diversiteetin mahdollisessa jatkotutkimuksissa vaaditaan aiemman diversiteettitutkimuksen perinteestä irrottautumista ja kehitystyötä kahdella tavalla: diversiteetin määrittelyssä ja puhevuorovaikutuksen ominaislaadun huomioon ottamisessa.

Kirjallisuus

Alderson, J. C. 2005. *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Berman, R. 2007. Developing language knowledge and language use across adolescence. Teoksessa E. Hoff & M. Shatz (toim.), *Handbook of Language Development*. London: Blackwell.

Berman, R. & L. Verhoeven 2002. Cross-linguistic perspectives on the development of text-production abilities: speech and writing. *Written language and Literacy*, 5 (1), 1–43.

Bonvin, A. & A. Lambelet 2017. Algorithmic and subjective measures of lexical diversity in bilingual written corpora: a discussion. *Corela*, HS-21 http://corela.revues.org/4843 DOI: 10.4000/corela.4843.

Booth, P. 2014. The variance of lexical diversity profiles and its relationship to learning style. *International Review of Applied Linguistics in Language Teaching*, 52 (4), s. 357–

375.

Bradac, J. J. & R. Wisegarver 1984. Ascribed status, lexical diversity and accent: Determinants of perceived status solidarity, and control of speech style. *Journal of Language and Social Psychology*, 3 (4), 239–255.

Broeder, P., G. Extra, R. van Hout, R. 1993. Richness and variety in the developing lexicon. Teoksessa C. Perdue (toim.), *Adult language acquisition: cross-linguistic perspectives*. Vol I: Field methods. Cambridge: Cambridge University Press, 145–232.

Burroughs, E. I. 1991. Lexical diversity in listeners' judgments of children. *Perception and Motor Skills*, 73 (1), 19–22.

Cain, K., J. Oakhill & K. Lemmon 2004. Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge and memory capacity. *Journal of Educational Psychology*, 96 (4), 671–681.

Carroll, J. B. 1938. Diversity of vocabulary and the harmonic series law of word-frequency distribution. *Psychological Record*, 2, 379–386.

Castañeda-Jiménez, G. & S. Jarvis 2014. Exploring lexical diversity in second language Spanish. Teoksessa K. Geeslin (toim.), *The Handbook of Spanish Second Language Acquisition*, 498–513.

Choi, W. & H. Jeong 2016. Finding an appropriate lexical diversity measurement for a small-sized corpus and its application to a comparative study of L2 learners' writings. *Multimedia Tools and Applications*, 75 (21), 13015–13022.

Crossley, S. A., T. Salsbury & D. S. McNamara 2011. Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29 (2), 243–263.

Crossley, S. A., T. Salsbury, D. S. McNamara & S. Jarvis 2011. What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, 45 (1), 182–193.

Deboer, F. 2014. Evaluating the comparability of two measures of lexical diversity. *System*, 47, 139–145.

Dewaele, J.-M. & A. Pavlenko 2003. Productivity and lexical diversity in native and non-native speech: a study of cross-cultural effects. Teoksessa V. J. Cook (toim.), *L2 effects on the L1*. Clevedon: Multilingual Matters, 120–141.

Dockrell, J. E. & Messer, D. 2004. Lexical acquisition in the early school years. Teoksessa R. Berman (toim.), Language development across childhood and adolescence. Psycholinguistic and crosslinguistic perspectives. Amsterdam: John Benjamins.

Durán, P, D. Malvern, B. Richards, N. Chipere 2004. Developmental Trends in Lexical Diversity. *Applied Linguistics*, 25 (2), 220–242.

Ellis, C., Y. F. Holt & T. West 2015. Lexical diversity in Parkinson's disease. *Journal of Clinical Movement Disorders* 2 (5), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4710975/ DOI: 10.1186/s40734-015-0017-4.

Gebril, A. & L. Plakans 2016. Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency. *Journal of English for Academic Purposes*, 24, 78–88.

Geeslin, K. (toim.) 2014. *The Handbook of Spanish Second Language Acquisition*. Malden: Wiley Blackwell.

Gregori-Signes, C. & B. Clavel-Arroitia 2015. Analysing Lexical Density and Lexical Diversity in University Students' Written Discourse. *Procedia - Social and Behavioral Sciences* 24, 198, 546–556.

Habermas, J. 1984. *The theory of communicative action. Reason and the rationalization of society*. Volume 1. London: Heinemann.

Honko, M. 2013. *Alakouluikäisten leksikaalinen tieto ja taito: toisen sukupolven suomi ja S1-verrokit*. Acta Universitatis Tamperensis 1865. Tampere: Tampere University Press 2013. http://tampub.uta.fi/handle/10024/94544 URN:ISBN:978-951-44-9251-8.

Honko, M. (tulossa). Kieli- ja kielitaitokäsitykset tutkivan opettajan

kenttäpäiväkirjamerkinnöissä. *Puhe ja kieli*.

Härnqvist K., U. Christianson, D. Ridings & J.-G. Tingsell 2003. Vocabulary in interviews as related to respondent characteristics. *Computers and the Humanities*, 37 (2), 179–204.

Jarvis, S. 2002. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19 (1), 57–84.

Jarvis, S. 2013a. Capturing the Diversity in Lexical Diversity. *Language Learning*, 63 (1), 87–106.

Jarvis, S. 2013b. Defining and measuring lexical diversity. Teoksessa S. Jarvis & M. Daller (toim.), *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam: John Benjamins Publishing, 13–44.

Johansson, V. 2008. Lexical diversity and lexical density in speech and writing: a developmental perspective. Lund University. – Dept. of Linguistics and Phonetics Working Papers 53 (2008), 61–79. www.sciecom.org/ojs/index.php/LWPL/article/view/2273/1848

Karlsson, L. 2014 [2003]. *Sadutus. Avain osallistavan toimintakulttuuriin*. Kolmas, uudistettu painos. Jyväskylä: PS-kustannus.

Klee, T., S. F. Stokes, A.M.-Y. Wong, P. Fletcher & W. Gavin 2004. Utterance Length and Lexical Diversity in Cantonese-Speaking Children with and without Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, 47 (6), 1396–1410.

Kuiken, F. & I. Vedder 2012. Speaking and writing tasks and their effects on second language performance. Teoksessa S. M. Gass & A. Mackey (toim.), *The Roudledge handbook of second language acquisition*, 364–377.

Koizumi, R. & Y. In'Nami 2012. Effects of text length on lexical diversity measures: using short texts with less than 200 tokens. *System: An International Journal of Educational Technology and Applied Linguistics*, 40 (4), 554–564.

Lai, S. & P. J. Schwanenflugel 2016. Validating the Use of "D" for Measuring Lexical Diversity in Low-Income Kindergarten Children. *Language, Speech, and Hearing Services in Schools*, 47 (3), 225–235.

Lervåg, A. & V. G. Aukrust 2010. Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry*, 51 (5), 612–620.

Malin, E. 2012. S*uomi toisena kielenä -oppijoiden sanaston kehittyminen taitotasolta toiselle siirryttäessä*. Pro gradu -tutkielma. Kielten laitos. Jyväskylän yliopisto.

Malvern, D. & B. Richards 1997. A new measure of lexical diversity. Teoksessa A. Ryan & A. Wray (toim.), *Evolving models of language. Papers from the Annual Meeting of the British Association of Applied Linguists held at the University of Wales, Swansea, September 1996*. Clevedon, UK: Multilingual Matters, 58–71.

Malvern, D. & B. Richards 2002. Investigating accommodation in language proficiency interviews 426 using a new measure of lexical diversity. *Language Testing*, 19 (1), 85–104.

Malvern D., B. Richards, N. Chipere & P. Durán 2004. *Lexical diversity and language development: Quantification and assessment*. Houndmills, Hampshire: Palgrave Macmillan.

McCarthy, P. M. 2005. An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Opinnäyte

McCarthy, P. M. & S. Jarvis 2007. Vocd: A theoretical and empirical evaluation. *Language Testing*, 24 (4), 459–488.

Milton, J. 2009. *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Muter, V., C. Hulme, M. J. Snowling & J. Stevenson 2004. Phonemes, rimes, vocabulary and grammatical skills as foundations of early reading development:

Evidence from a longitudinal study. *Developmental Psychology*, 40 (5), 665–681.

Pajunen, A. 2012. Kirjoittamistaitojen kehitys 8–12-vuotiailla. Alakoululaisten unelmakirjoitelmat. *Virittäjä*, 114 (1), 481–501.

Riihelä, Monika 2013. Suomalaisten pienten lasten ajatuksia heijastava erittäin laaja aineisto kaipaa tutkijoita! Verkkoartikkeli. http://www.edu.helsinki.fi/lapsetkertovat/lapset/Tutkimus/tutkimus_aineisto.htm [Haettu 30.9.2017.]

Qian, D. D. & M. Schedl 2004. Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21 (1), 28–52.

Saarela, L. 1997. *Peruskoululaisten kirjoitelmien kehittyminen sanastotutkimuksen valossa*. Acta Universitatis Ouluensis. B Humaniora 25. Oulun yliopisto.

Sadeghi, K. & S. K. Dilmaghani 2013. The relationship between lexical diversity and genre in Iranian EFL learners' writings. *Journal of Language Teaching and Research*, 4 (2), 328–334.

Sang, G. K. & L. Miseon 2013. Lexical Diversity and Functional Categories in English Attrition of Korean Returnee Children. 언어 2013, 38 (2), 239–258.

Schmid, M. & S. Jarvis 2014. Lexical access and lexical diversity in first language attrition. *Bilingualism: Language and Cognition*, 17 (4), 729–748.

Scott C. M. & J. Windsor 2000. General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language and Hearing Research*, 43 (2), 324–339.

Singh, S. 2001. A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing*, 16 (3), 251–264.

Strömqvist, S., V. Johansson, S. Kriz, H. Ragnarsdóttir, R. Aisenman & D. Radvid. 2002. Toward a cross-linguistic comparison of lexical quanta in speech and writing. *Written Language and Literacy*, 5 (1), 45–67.

Taimisto, H. 2014. *Taitotasolta toiselle: Korpuspohjainen tutkielma vironkielisten suomenoppijoiden verbisanaston kehittymisestä*. Suomen kielen pro gradu -tutkielma. Oulun yliopisto.

Tannenbaum, K. R., J. K. Torgesen & R. K. Wagner 2006. Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, 10 (4), 381–398.

Tidball, F. & Treffers-Daller, J. 2007. Exploring measures of vocabulary richness in semi-spontaneous French Speech. Teoksessa H. Daller, J. Milton & J. Treffers-Daller (toim.) *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press, 133–149.

Unsworth, S. 2008. Comparing child L2 development with adult L2 development: How to measure L2 proficiency. Teoksessa B. Haznedar & E. Gavruseva (toim.), *Current Trends in Child Second Language Acquisition: A generative perspective*. Amsterdam & Philadelphia: John Benjamins, 301–333.

Verhoeven, L. 1009. Acquisition of reading in a second language. *Reading research Quarterly* 25, (2), 90–114.

Vermeer, A. 2000. Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17 (1), 65–83.

Watkins, R., D. Kelly & H. Harbersh 1995. Measuring children's lexical diversity. Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research*, 38 (6), 1349–1355.

Wong, A., T. Klee, S. Stokes, P. Fletcher & L. Leonard 2010. Differentiating Cantonese-speaking preschool children with and without SLI using MLU and lexical diversity. *Journal of Speech, Language, and Hearing Research* 2010, 53 (3), 794–799.

Wright, H., S. Silverman & M. Newhoff 2003. Measures of lexical diversity in aphasia. *Aphasiology*, 17 (5), 443–452.

Zipf, G. K. 1935. The psycho-biology of language. Boston: Houghton-Mifflin.

Zipf, G. K. 1937. Observations of the possible effect of mental age upon the frequency-distribution of words from the viewpoint of dynamic philology. *Journal of Psychology*, 1937 (4), 239–244.

Yu, G. 2010. Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31 (2), 236–259.