Suomi toisena kielenä -oppijoiden sanaston kehittyminen taitotasolta toiselle siirryttäessä

Pro gradu

Essi Malin

Jyväskylän yliopisto
Kielten laitos
Suomen kieli
Marraskuu 2012

JYVÄSKYLÄN YLIOPISTO

Tiedekunta – Faculty Humanistinen tiedekunta	Laitos – Department Kielten laitos
Tekijä – Author Malin, Essi Pauliina	
Työn nimi – Title Suomi toisena kielenä -oppijoiden sanaston kehittyminen taitotasolta toiselle siirryttäessä	
Oppiaine – Subject Suomen kieli	Työn laji – Level Pro gradu -tutkielma
Aika – Month and year Marraskuu 2012	Sivumäärä – Number of pages 71

Tiivistelmä – Abstract

Tutkimuksen tavoitteena on selvittää, kuinka sanaston käyttö kehittyy suomi toisena kielenä -oppijoiden teksteissä. Tarkastelun kohteena on sanavaraston rikkaus eli leksikaalinen diversiteetti. Aineistona on 1197 aikuisten ja koululaisten eri tekstilajeja edustavaa kirjoitusta, jotka on arvioitu valmiiksi Eurooppalaisen viitekehyksen mukaisille taitotasoille. Työ on osa Suomen Akatemian ja Jyväskylän yliopiston vuosina 2007-2009 rahoittamaa Cefling-hanketta, jonka tehtävänä on selvittää, millaista kielitaito on kullakin Eurooppalaisessa viitekehyksessä kuvatulla kuudella taitotasolla. Tavoitteena on myös selvittää, onko aikuisten ja koululaisten kielenoppimisessa eroja. Keskeisimmät teoreettiset taustaajatukset tulevat Cefling-hankkeen lisäksi sanaston tutkimuksen perinteestä ja erityisesti sanaston määrällisen tutkimuksen lähtökohdista.

Perinteisten menetelmien ja mittareiden lisäksi työ esittelee nykyaikaisia sanastollisen diversiteetin mittaamiseen kehitettyjä, entistä parempia tunnuslukuja, joita ei ole aikaisemmin sovellettu suomenkieliseen aineistoon. Tutkimuksessa käytetyistä uusista tunnusluvuista erityisesti MTLD, sanojen monipuolisuusluku, on osoittautunut luotettavaksi ja otoskoosta riippumattomaksi keinoksi mitata sanaston monimuotoisuutta. Toinen yhtä luotettava tunnusluku on Shannonin indeksi. Aineiston sanastoa on verrattu tunnuslukujen avulla myös Suomen sanomalehtikielen taajuussanastoon, sillä aikaisempien tutkimusten mukaan oppijan kielessä esiintyvien sanojen yleisyys tai harvinaisuus kertoo osaltaan kielenkäyttäjän sanastollisesta osaamisesta.

Tutkimus osoittaa, että suomi toisena kielenä -oppijan sanasto kehittyy taitotasojen myötä. Erityisen selkeä harppaus sanaston kehittymisessä tapahtuu MTLD-tunnusluvun ja Shannonin indeksillä laskettujen tulosten mukaan taitotasolta B2 taitotasolle C1 siirryttäessä. Alemmilla tasoilla (A1 ja A2) sanaston kehittymisessä ei tapahdu ratkaisevia askelia, vaan sanaston rikkaus ja monipuolisuus alkavat kehittyä vasta myöhemmin. Johtopäätöstä tukee myös aineistosta laskettujen TTR-arvojen vertailu. Sen sijaan tutkimuksessa mukana olevien tekstilajien välille ei synny niin suuria eroja, että niistä olisi mahdollista tehdä päätelmiä eri tekstilajien osaamisesta.

Aineiston yleisimpien lekseemien tarkastelu paljastaa, että mitä ylempää taitotasoa tekstit edustavat, sitä pienemmäksi yleisimpien lekseemien kattavuus teksteissä laskee. Korkeimmilla taitotasoilla yleisimmätkään lekseemit eivät kata enää prosentuaalisesti yhtä suurta osaa kaikista tekstien saneista kuin alemmilla tasoilla. Tämä kertoo osaltaan suomi toisena kielenä -oppijoiden sanaston monipuolistumisesta kielitaidon kehittyessä.

Asiasanat – Keywords suomi toisena kielenä, kielen oppiminen, leksikologia, kirjoittaminen, eurooppalainen viitekehys
Säilytyspaikka – Depository Fennicum
Muita tietoja – Additional information

ESIPUHE

Suunnitellessani aihetta lopputyöhöni oli sanastontutkimus päällimmäisenä mielessäni. Jo kandidaatintutkielmassa olin uponnut syvälle sanojen merkilliseen maailmaan. Kun graduohjaajani sitten ehdotti oppijan sanaston tarkastelua osana Jyväskylän yliopiston Cefling-hanketta, aihe tuntui heti omalta. Suomi toisena kielenä -oppijoiden sanaston tutkiminen määrällisillä menetelmillä muodostui koko työni kantavaksi ajatukseksi. Gradun kirjoittamisen ohella minulla oli mahdollisuus tutustua suomena vieraana kielenä -oppijoiden sanaston kehittymiseen myös käytännössä, kun opetin vaihtovuotenani suomen kieltä Kaarlen yliopistossa Prahassa. Opetustyö lisäsi motivaatiota tutkimuksen kirjoittamiseen ja asioiden teoreettiseen pohdiskeluun.

Haluan kiittää kaikkia minua projektin varrella auttaneita. Kiitän Mari Honkoa, Sanna Ravia ja Ari Huhtaa tutkimukseni lukemisesta ja arvokkaan palautteen antamisesta työn eri vaiheissa. Erityiskiitoksen haluan lähettää Ohion yliopistoon professori Scott Jarvisille, joka analysoi lemmaamani aineiston ohjelmallaan ja antoi muutenkin paljon apua käyttämieni tunnuslukujen kanssa. Olen kiitollinen myös Cefling-hankkeen toimikunnalle: lemmaustyöhön osoitettu apuraha helpotti keskittymistä tutkimustyöhön. Suurin kiitos kuuluu ohjaajalleni professori Maisa Martinille tärkeästä palautteesta ja kannustuksesta.

Sisällys

1 JOHDANTO

Sanastollinen osaaminen on merkittävä osa kielitaitoa. Ilman sanoja ihmisen toiminta käy hankalaksi: pyytäminen, kysyminen ja selittäminen onnistuvat parhaiten sanojen avulla. Sanoja tarvitaan muodostamaan lauseita, kappaleita ja tekstiä, joten myös kielen rakennetta tai kielioppia on mahdotonta opettaa ilman sanastoa. Oikeastaan kaikki kielenoppiminen lähtee liikkeelle sanoista ja niiden välisten suhteiden hahmottamisesta. Mitä rikkaampi kielenkäyttäjän sanavarasto on, sitä tarkemmin hän pystyy itseään ilmaisemaan.

Jyväskylän yliopistossa vuonna 2007 aloitettu Cefling-hanke (Linguistic Basis of the Common European Framework for L2 English and L2 Finnish) tarkastelee toisen ja vieraan kielen oppimisprosessia. Tutkimuksen taustalla on olettamus, että kielenoppiminen etenee vaiheittain ja noudattaa tiettyä oppimisjärjestystä. Hanke pyrkii tutkimaan, kuinka toisen ja vieraan kielen oppijan kielitaito kehittyy taitotasolta toiselle, sekä kartoittamaan tietyllä taitotasolla olevien oppijoiden yhteneviä kielellisiä ja kielitaidon kehitykseen liittyviä piirteitä. Hankkeessa tarkastellaan tietyn taitotason kirjallisia taitoja, ei yksittäisen oppijoiden kielitaidon tasoa. Tavoitteena on myös selvittää, onko aikuisten ja nuorten kielenoppimisessa eroja ja minkälaisia mahdolliset eroavaisuudet ovat.

Oman tutkimukseni tavoitteena on selvittää, kuinka sanaston osaaminen kehittyy suomi toisena kielenä -oppijoiden teksteissä. Tarkastelun kohteena on sanaston rikastuminen taitotasolta toiselle siirryttäessä. Aineistona ovat Cefling-hankkeessa kerätyt ja valmiiksi arvioidut aikuisten ja koululaisten tekstit. Vertailun kohteena on toisaalta lasten ja aikuisten sanasto, toisaalta eri tekstilajien vaikutus sanaston rikkauteen. Tekstilajeina on sekä muodollisia että epämuodollisia tekstejä.

Sanastollisesta osaamisesta puhuttaessa käytetään usein termiä sanavarasto. Sanavarasto ei kuitenkaan ole yksiselitteinen termi. Pelkkien yksittäisten sanojen osaamisen lisäksi sanastolliseen osaamiseen kuuluu tietoa siitä, kuinka sanoja käytetään tarkoituksenmukaisesti. Suurenkaan sanavaraston hallitseminen ei vielä takaa sujuvaa viestintää: kielenkäyttäjällä on oltava tietoa siitä, kuinka sanoja käytetään ja liitetään suuremmiksi kokonaisuuksiksi. Sanavaraston laajuutta on myös mahdotonta mitata Cefling-hankkeessa kerättyjen lyhyiden tekstien perusteella. Koska jonkinlaisia tuloksia sanastollisesta osaamisesta kuitenkin kaivataan, päädyin omassa tutkimuksessani keskittymään teksteissä ilmenevään sanaston rikkauteen. Tässä tutkimuksessa sanaston rikkaus on yhtä kuin leksikaalinen diversiteetti.

Leksikaalista diversiteettiä mitataan tutkittavasta aineistosta toisaalta eri sanojen ja sanaesiintymien välisestä suhteesta, toisaalta käytettyjen sanojen taajuuden perusteella. Monet aikaisemmista sanaston osaamisen kvantitatiivisista tutkimuksista perustuvat TTR-mittauksiin (type/token ration) eli tutkittavan aineiston eri sanojen ja sanaesiintymien väliseen suhteeseen. Sanoilla (eng. type) tarkoitetaan eri sanojen määrää. Saneilla eli sanaesiintymillä (eng. token) tarkoitetaan juoksevien sanojen määrää. Sanojen ja esiintymien väliseen suhteeseen perustuvat mittaukset eivät kuitenkaan ole kovin luotettavia, eikä niillä esimerkiksi voida vertailla eripituisia tekstejä keskenään. Siksi tutkittavien tekstien sanastoa analysoidaan myös suhteessa suomenkieliseen taajuussanastoon.

Cefling-hankkeesta on jo julkaistu useita eri tutkimuksia (ks. Cefling-hankkeen verkkosivut) ja oma työni liittyy hankkeesta tehtyjen aikaisempien pro gradu -tutkielmien kasvavaan sisarusparveen. Hanketta varten kerätystä suomi toisena kielenä -aineistosta on aikaisemmin tutkittu muun muassa olla-verbirakenteita (Kynsijärvi 2007), verbiketjujen kehittymistä (Paavola 2008), sananmuodostamista (Penttinen 2010), mennä ja tulla verbejä (Puhakka 2010) ja ajanilmauksia (Varis 2010). Sanaston kehittymistä taitotasolta toiselle on tutkittu kuitenkin vasta englantia vieraana kielenä -oppijoiden teksteistä. Suomi toisena kielenä -oppijoiden osalta hankkeessa oli sopiva gradun mentävä aukko. Siihen aukkoon sovitin omat tutkimuskysymykseni.

Tutkielmani toisen luvun aloitan Cefling-hankkeen esittelyllä ja jatkan Eurooppalaisen viitekehyksen taitotasojen kuvauksella. Kolmannessa luvussa esittelen sanastontutkimuksen teoreettisia lähtökohtia ja aikaisempia tuloksia, joille oma tutkimukseni perustuu. Teoreettiseen taustaan kuuluvat myös keskeisimpien käsitteiden tarkempi määrittely. Neljännessä luvussa esittelen oman tutkimukseni tutkimuskysymykset, aineiston ja menetelmät. Samalla esittelen käyttämäni sanastollista monimuotoisuutta mittaavat tunnusluvut. Tutkimukseni tulokset on esitelty viidennessä ja kuudennessa luvussa. Seitsemäs luku on päätäntöluku, jossa kokoan yhteen tutkimukseni keskeiset tulokset ja arvioin tekemääni tutkimusta.

2 CEFLING-HANKE

Pro gradu -tutkielmani liittyy Suomen Akatemian ja Jyväskylän yliopiston vuosina 2007-2009 rahoittamaan Cefling-hankkeeseen (Linguistic Basis of the Common European Framework for L2 English and L2 Finnish) sekä sitä seuranneeseen ja samoihin lähtökohtiin perustuvaan Topling-hankkeeseen. Cefling-hankkeen tavoitteena on tutkia, kuinka kielen oppijan kielitaito kehittyy taitotasolta toiselle siirryttäessä (Cefling-hankkeen verkkosivut). Hankkeen pohjana on Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys (EVK, eng. CEFR) ja viitekehyksessä kuvatut kielenoppijan taitotasot. Taitotasot luotiin, jotta kielitaitoa voitaisiin arvioida aikaisempaa yhtenäisemmin eri puolilla Eurooppaa. Cefling-hanketta seuranneen Topling-hankkeen päätavoitteena on selvittää, miten suomen (suomi toisena kielenä), englannin ja ruotsin kielen oppijoiden kirjoittamistaidot kehittyvät suomalaisessa koulutusjärjestelmässä (Topling-hankkeen verkkosivut). Oppimista seurataan pitkittäisaineiston avulla, ja tuloksia verrataan Cefling-hankkeen poikittaisaineistosta saatuihin tuloksiin. Molempien hankkeiden syntyyn on osaltaan vaikuttanut alan eurooppalainen tutkimusverkosto Second Language Acquisition and Testing in Europe (SLATE-verkoston verkkosivut). Verkoston avainkysymyksiä on, miten kielenoppiminen etenee ja mitkä tekijät vaikuttavat toisen tai vieraan kielen oppimiseen. Tavoitteena on myös yhdistää kielenoppimisen tutkimustuloksia ja kielitaidon arviointiin liittyvää tietoutta toisiaan tukevaksi kokonaisuudeksi.

2.1 Yhteinen eurooppalainen viitekehys ja funktionaalinen kielikäsitys

Yhteinen eurooppalainen viitekehys on Euroopan neuvoston julkaisema järjestelmällinen kuvaus vieraiden kielten oppimisesta, opettamisesta ja arvioinnista. Viitekehyksen ja siinä määriteltyjen taitotasojen tarkoituksena on tarjota kielenopetukselle, kielikoulutuksen suunnittelulle ja kielitaidon arvioinnille yhtenäinen perusta (Eurooppalainen viitekehys 2003: 19, myöhemmin EVK). Suomessa Yhteistä eurooppalaista viitekehystä käytetään laajasti esimerkiksi kieltenopetuksen ja arvioinnin perustana sekä perus- että aikuisopetuksessa. Myös muualla Euroopassa viitekehys on saavuttamassa vakiintuneen aseman. (Cefling-hankkeen verkkosivut.) Viitekehys perustuu funktionaaliseen kielikäsitykseen, jonka mukaan oppiminen sosiaalista toimintaa: kieltä opitaan ja käytetään todellisissa kielenkäyttötilanteissa. Tällöin oppija on toimija, joka käyttää kieltä tarkoituksenmukaisesti erilaisissa sosiaalisissa konteksteissa. Pelkkien kieltä koskevien sääntöjen ohella oppijan huomion tulisi kohdistua myös erilaisiin kielellisiin konstruktioihin ja niiden oppimiseen. Kaiken kaikkiaan

opetus ja oppiminen tulisi viitekehyksen mukaan nähdä aikaisempaa laajemmin yhteiskunnallisessa ja kulttuurisessa kontekstissa. (Martin, Mustonen, Reiman, Seilonen 2010: 57-59.)

2.2 Viitekehyksen taitotasot ja tutkimusaineisto

Kielen oppiminen kuvataan viitekehyksessä kuusiportaisena taitotasoasteikkona. Kielenkäyttäjät on määritelty kolmeen pääluokkaan: A, B ja C. A-tason oppijat ovat perustason kielenkäyttäjiä, B-tason oppijat itsenäisiä kielenkäyttäjiä ja C-tason taitavia kielenkäyttäjiä. Nämä tasot jakautuvat kukin vielä kahdeksi, jolloin asteikko jakautuu yhteensä kuuteen tasoon (suluissa englanniksi): A1 alkeistaso (breakthrough), A2 selviytyjän taso (waystage), B1 kynnystaso (threshold), B2 osaajan taso (vantage), C1 taitajan taso (effective operational profiency) ja C2 mestarin taso (mastery). (EVK 2003: 46-47.)

Hanketta varten kerättiin toisen kielen oppijoiden kirjoitussuorituksista koostuva aineisto. Toisen kielen oppijoiden tekstejä edustavat Yleisten kielitutkintojen kirjoitustehtävät (YKI-aineisto) ja erityisesti Cefling-hanketta varten kerätyt yläkoululaisten, 7.-9. luokkalaisten tekstit. Koko aineisto on arvioitu eurooppalaisen viitekehyksen mukaan. Aikuisten tekstit jakautuvat A-, B- ja C-tasoille, koululaisaineisto A- ja B-tasoille.

2.3 Cefling-hankkeen yleiset tavoitteet ja tutkimuskysymykset

Cefling-hankkeen tavoitteena on yhdistää toisen kielen oppimisen ja kielitaidon arvioinnin tutkimuksesta saatavaa tietoa. Tutkimuksen taustalla on olettamus, että toisen kielen oppiminen etenee vaiheittain ja että nämä vaiheet voidaan kuvata taitotasoina. Oletuksen mukaan tietyllä taitotasolla olevien oppijoiden väliltä on löydettävissä yhteneviä kielellisiä ja kielitaidon kehitykseen liittyviä piirteitä. Analyysin avulla pyritään selvittämään, minkälaisia nämä tyypilliset piirteet tai niiden yhdistelmät ovat. Hankkeen tavoitteena on myös selvittää, eroaako koululaisten osaaminen aikuisten suomen kielen oppijoiden tavasta käyttää kieltä ja minkälaisia eroja aikuisten ja nuorten suorituksissa ilmenee samalla taitotasolla. Aikuisten ja nuorten tekstit ovat keskenään vertailukelpoisia, koska molemmat perustuivat samantasoisiin ja samankaltaisiin viestinnällisiin tehtävänantoihin. Vaikka Celfling-aineistossa on vain kirjoitetuttuja tekstejä, tutkimuksessa käytettyjen metodien ajatellaan soveltuvan myös puhuttuun kieleen. (Cefling-hankkeen verkkosivut.)

3 SANASTON TUTKIMUS

3.1 Tärkeimmät käsitteet

Tämän tutkimuksen tärkeimmät käsitteet ovat **lekseemi** ja **sane**. Lekseemillä tarkoitan tietyn sanan kaikkia taivutusmuotoja käsittävää abstraktiota, jota englanninkielisessä kirjallisuudessa vastaa termi *type*. Sane eli sanaesiintymä tarkoittaa puolestaan aineistossa itsenäisenä esiintyvää ja realistuvaa konkreettista tekstiyksikköä, jonka englanninkielinen vastine on *token*. Saneiden määrä on siis yhtä kuin ”juoksevien sanojen” määrä. Lekseemin ja saneen käsitteitä on helppo havainnollistaa laskemalla ne tietystä virkkeestä. Esimerkiksi virkkeessä ”Herra antoi, herra otti.” on neljä sanetta, mutta vain kolme lekseemiä. (Penttilä 1963: 115-118, Niemikorpi 1991: 21-22, Puro 1999: 9.)

Koska primitiivikäsite **sana** esiintyy arkikielessä sekä saneen että lekseemin synonyyminä, monet sanastontutkijat ovat luopuneet sen käytöstä tieteellisessä kielessä. Myös omassa tutkimuksessani pyrin selkeyden vuoksi käyttämään vain termiä lekseemi. Toisinaan tutkimuksessani kuitenkin esiintyy sanan käsite lekseemin rinnalla sellaisissa paikoissa, joissa väärinkäsityksen vaaraa ei ole. Tämä tuntuu luontevimmalta esimerkiksi silloin, kun aiheena ovat sana-loppuiset yhdyssanat kuten epäsanat, HL-sanat tai yhdyssanat.

Vaikka edellä esitetyt määritelmät saneesta ja lekseemistä riittävät hyvin tämän työn perustaksi, aikaisemmat sanastontutkijat ovat pyrkineet määrittelemään tärkeimpiä käsitteitä tarkemmin ja syvemmin. Varsinkin sanan määritteleminen on tuottanut ongelmia. Esimerkiksi Singletonin (1995) mukaan sana voidaan määritellä ortografisesti, foneettisesti, fonologisesti, semanttisesti tai kieliopillisesti, mutta mikään niistä ei yksistään riitä. Ortografisesti määriteltynä sana on joukko kirjaimia, joiden molemmin puolin on tyhjä tila. Foneettisen määritelmän mukaan sana on foneettinen kokonaisuus, yhtäjaksoinen ääniryöppy, jolla on tietyt akustiset ominaisuutensa. Fonologinen sanan määritelmä kuvaa sanan sarjaksi yksiköitä, jotka sopivat tietyn kielen äännejärjestelmään. Tietyn kielen äännejärjestelmän mukaan sanassa voi olla esimerkiksi vain yksi pääpainollinen tavu tai sen tulee noudattaa vokaaliharmoniaa. Semanttisesti määriteltynä sana on kielen pienin merkityksellinen yksikkö esimerkiksi morfeemi. Kieliopillisen määritelmän mukaan sanat ovat puolestaan lauseen yksikköjä, jotka ovat vapaasti liikkuvia, mutta sisäisesti stabiileja. Jokaisella edellä mainituista määritelmistä on kuitenkin ongelmakohtansa, joissa määritelmän todenmukaisuus horjuu. Siksi mikään niistä ei voi yksiselitteisesti määritellä sanan koko käsitettä. (Singleton 1995: 2, 10-14.)

Sanastontutkimuksessa voi lekseemin, saneen ja sanan lisäksi törmätä myös **lemman** käsitteeseen. Lemma on lähinnä tutkimustekninen käsite, jota käytetään sanan perusmuodosta silloin, kun lekseemien välillä saattaa ilmetä homonymiaa tai polysemiaa esimerkiksi lemmauksen yhteydessä (Jaakola 2004: 9-10). Esimerkiksi lemma *kuusi* voi jakautua vielä kahteen lekseemiin *kuusi* (substantiivi) ja *kuusi* (numeraali), jolloin kyseessä on homonyminen tapaus. Polysemiasta on kyse esimerkiksi lemman *myös* kohdalla, joka voi olla joko partikkelin tai konjuktion lekseemi. Tässä tutkimuksessa joudun kuitenkin käyttämään lemman käsitettä vain harvoin, sillä raakalemmauksen jälkeen kävin kaikki potentiaaliset homonyymitapaukset erikseen läpi juoksevasta tekstistä ja määritin jokaiselle lemmalle oikean lekseemin. Tarkemmin lemmauksesta kerron luvussa 4.3.

Sanaston käsite vaikuttaa olevan niin yksiselitteinen, ettei sitä ole monissa yhteyksissä vaivauduttu määrittelemään ollenkaan tai määritelmä on hyvin yksinkertainen. Esimerkiksi Voionmaa (1993: 12) määrittelee väitöskirjassaan sanaston tarkoittavan sanojen kokoelmaa ja monissa tutkimuksissa termillä tarkoitetaan jonkinlaista sanalistaa (Read 2006: 36). Termi sanasto liitetäänkin usein opetukseen liittyvään materiaaliin ja sanakirjoihin. Kielen omaksumista teoreettisemmin käsittelevässä kirjallisuudessa sanaston rinnalla käytetään termiä **leksikko**. Termiä käytetään muun muassa psykolingivistiikan piirissä, jolloin leksikko määritellään mentaaliseksi leksikoksi, joka pyrkii kuvaamaan, miten sanoja vastaanotetaan, miten sanat ovat varastoituneet mieleen ja miten niitä haetaan käyttöön (Puro 2002: 3). Sanastoon verrattuna leksikon merkitys on nimenomaan mielen rakenteessa, dynaamisessa organisaatiossa, joka mukautuu jatkuvasti uusia sanoja ja merkityksiä opittaessa ja toisia unohdettaessa. Puron (2002: 3) mukaan termiin leksikko kuuluu vahvasti oletus siitä, että kieli on varastoitunut mieleen ja kielitaito on erilainen kognitiivinen taito kuin muut kognitiiviset taidot. Toisaalta toisen kielen tutkimuskirjallisuudessa sanastoa ja leksikkoa käytetään myös synonyymisesti tai toinen termeistä on valittu kattamaan molemmat termit.

Sanastoa ja leksikkoa määriteltäessä tärkeintä onkin tehdä ero kielen kaikkia sanoja tarkoittavan määritelmän ja yksilön mentaalisen leksikon välille. Tarkoitettaessa yksilön mieleen varastoituneiden äidinkielisten tai muunkielisten sanojen joukkoa puhutaan toisinaan myös sanavarastosta. Sanavaraston määritelmää on kuitenkin kritisoitu, koska todellisuudessa yksilön sanasto ei ole yksityisten sanojen ja merkitysten kokoelma, eikä sanastoa ja kielioppia ole mahdollista erottaa toisistaan (esim. Cook 1991: 11, Voionmaa 1993: 12, Meara 1993: 69). Sekä yksilön sanavarastoa että yleisesti käsitettyä tietyn kielen sanojen joukkoa on lisäksi mahdoton kuvata tai edes mitata. Yksilön sanavaraston koosta on toisinaan tehty erilaisia arvauksia ja joitakin suuntaa-antavia tuloksia on pyritty antamaan erilaisten mittausmenetelmien avulla. Yksilön sanavaraston tarkan koon laskemista pidetään kuitenkin mahdottomana tehtävänä sanastontutkijoiden keskuudessa.

3.2 Sanaston oppimiseen vaikuttavia tekijöitä

Sanaston oppimisen näkökulmasta suomen kieltä on pidetty sekä helppona että vaikeana (Puro 2002: 2). Helpoksi suomen sanaston tekee perussanojen suhteellisen vähäinen määrä, sillä suuri osa sanastosta muodostuu perussanoista rakennetuista johdoksista ja yhdyssanoista. Toisaalta suomen sanastoa pidetään vaikeana, koska kielessä on vain vähän muista kielistä tuttuja sanoja ja koska sanahahmot vaihtelevat paljon. Indoeurooppalaista kieltä äidinkielenään puhuvalle suomen sanat näyttäytyvät usein pitkinä ja outoina. Sanojen opittavuuteen vaikuttaa myös suomen taivutusjärjestelmän monimutkaisuus (Martin 1999: 169).

Sanastolla on erityisen tärkeä rooli kielen oppimisen alkuvaiheessa. Alusta asti oppija tarvitsee sanoja, eikä rakenteita tai kielioppia voida opettaa ilman sanastoa (Aalto 1993: 34). Kaikki kielen oppiminen lähtee liikkeelle sanoista ja niiden välisten suhteiden hahmottamisesta. Sanaston kehittymisen prosessiin vaikuttavat useat seikat, jotka voidaan jakaa Lauferin (1997) mukaan sananSisäisiin ja sananulkoisiin tekijöihin.

Sanojen omaksumiseen vaikuttavia sanansisäisiä tekijöitä eli sanojen muotoon ja merkitykseen liittyviä tekijöitä ovat sanan ääntäminen ja siihen liittyvä äänne- ja kirjoitusasun vastaavuus sekä sanojen läpinäkyvyys, säännönmukaisuus, pituus, sanaluokka ja abstraktiotaso (Laufer 1997: 142). Sanojen läpinäkyvyydellä ja säännönmukaisuudella tarkoitetaan sitä, kuinka helposti kompleksisten sanojen merkitys on pääteltävissä kantasanasta ja siihen liitetyistä suffikseista kuten johtimista. Esimerkiksi sana *savuton* on läpinäkyvä, koska sen merkityksen voi suoraan päätellä sanan osien *savu* ja *-ton* merkityksistä. Päinvastainen esimerkki on sana *merkillinen*, koska kantasanalla *merkki* ei ole paljon tekemistä *erikoisen* tai *omituisen* kanssa, jotka ovat sanalle vakiintuneet varsinaiset merkitykset (Penttinen 2010: 12).

Sanan pituus voi vaikuttaa sanan oppimiseen negatiivisesti: mitä pidempi sana, sitä vaikeampi se voi olla oppia. Toisaalta kaikki lyhyet sanat eivät ole kaikkia pitempiä sanoja helpompia. Jos pitkä sana muuten koostuu tutuista äänteistä ja osista, se voi olla helpommin hahmotettavissa kuin vieraampi lyhyempi sana. Myöskään sanaluokan vaikuttavuudesta sanan opittavuuteen ei ole yksiselitteistä tietoa. Joidenkin tutkimusten mukaan substantiivit ovat helpompia oppia kuin adjektiivit, verbit ja adverbit (esim. Aalto 1994: 96). Substantiiveja opitaan helpommin kuin verbejä, koska niiden semanttisen sisällön havaitseminen ja ymmärtäminen on helpompaa kuin verbien (Puro 1999: 12). Substantiiveissa ei myöskään ole verbeihin verrattuna ilmaisukyvyltään yhtä pal-

jon monimerkityksisiä ja moneen tilanteeseen ja lauseyhteyteen sopivia sanoja. Verbien merkitykset ovat puolestaan usein helpompi arvata tekstistä kuin adjektiivien ja adverbien (Nation 1990: 48).

Sanansisäisten tekijöiden lisäksi sanan opittavuuteen vaikuttavat myös kontekstisidonnaiset tekijät, kuten sanan frekvenssi ja opittavan kielen ja äidinkielen välinen suhde. Esimerkiksi yhtenevät ääntämis-, kirjoittamis- ja kirjainjärjestelmät sekä kohdekielisten sanojen ja äidinkielisten vastineiden samankaltaisuus voivat helpottaa sanaston omaksumista. Toisaalta opittavan kielen eri sanojen samankaltaisuus ääntämisessä tai kirjoitusasussa voi hankaloittaa oppimista, jos oppilas sekoittaa samalta kuulostavat tai näyttävät sanat keskenään (Penttinen 2010: 12).

3.3 Sanastollisen osaamisen mittaaminen ja arviointi

Oppijoiden kielitaitoa arvioitaessa on totuttu kohdistamaan arviointi toisaalta kielen ymmärtämiseen eli luetun ja kuullun ymmärtämiseen ja toisaalta kielen tuottamiseen eli kirjoittamiseen ja puhumiseen. Mittaamisen kohteena on harvoin ollut suoraan sanasto. Kun lukemisen, kuuntelemisen, kirjoittamisen ja puhumisen lisäksi omaksi alueekseen jaotellaan sanasto ja rakenteet, kyseessä on perinteinen, niin sanottu kämmenmalli. Nykyisten käsitysten mukaan kielitaitoa ei kuitenkaan tulisi arvioida ja käsitellä toisistaan erillisinä osa-alueina. Vaihtoehdoksi on esitetty kielitaidon käsittelemistä kehityksellisien piirteiden sujuvuuden, tarkkuuden ja kompleksisuuden kautta. Sujuvuus on nopeutta, tuotteliaisuutta sekä omien resurssien soveltamista tarkoituksenmukaisella tavalla. Tarkkuudella tarkoitetaan oikeakielisyyttä, kohdekielen mukaisia sanavalintoja ja kielen käytön konventioiden hallitsemista. Kompleksisuus puolestaan näyttäytyy sidosteisuutena, abstraktiotason nousuna ja kykynä tuottaa tyylillisesti ja kielellisesti erilaisia tekstilajeja. (Nissilä ym. 2006: 66, 86, 120, 159.)

Oma sanaston laajuuden ja rikkauden mittaamiseen keskittyvä tutkimukseni sivuaa sujuvuuden, tarkkuuden ja kompleksisuuden käsitteistä lähinnä vain viimeistä. Mitä laajempaa ja rikkaampaa oppijan sanasto on, sitä kompleksisempana voidaan pitää sanastoa ja siten myös kieltä. Tarkkuus on jätettävä tässä tutkimuksessa huomiotta, sillä sanaston mittaukset perustuivat lemmattuun aineistoon, jolloin alkuperäiset virheelliset ilmaisut ovat saaneet lemmausvaiheessa oikeakielisen vastineen. Myös sujuvuutta on hankala lähteä arvioimaan tutkimusaineistoni pohjalta.

Nykyisten tutkimusten valossa sanaston osaamiseen kuuluu pelkkien yksittäisten sanojen lisäksi myös tietoa sanojen tarkoituksenmukaisesta käytöstä. Sanavaraston laajuuden lisäksi

sanastolliseen osaamiseen kuuluu tehokas ja tarkoituksenmukainen sanavalinta, tyylivalinnat ja sanojen oikeakielinen ja rakenteellisesti oikeakielinen käyttö. Nykyisen käsityksen mukaan kieliopin ja sanaston osaaminen kehittyvät siis samanaikaisesti ja tosiaan tukien. (Puro 1999: 5-8.)

Toinen kielitaidon arvioinnissa yleisesti käytetty jako on reseptiivinen ja produktiivinen sanasto. Reseptiivisellä sanasto-osaamisella tarkoitetaan muiden tuottaman kielisyötteen vastaanottamisen taitoja eli lukemista, kuuntelemista ja syötteen ymmärtämistä. Produktiiviset taidot tarkoittavat kielen tuottamisen taitoja eli merkitysten välittämistä puhumalla ja kirjoittamalla. Reseptiivinen sanasto aktivoituu silloin, kun kielenkäyttäjä havaitsee syötteen ja produktiivinen silloin, kun täytyy tuottaa merkitys ja löytää sille parhaiten sopivat sanat ja ilmaisut. (Nation 2001: 24-26.) Luonnollisesti reseptiivinen sanasto on yleensä paljon produktiivista sanastoa laajempi – joidenkin arvioiden mukaan produktiivisen sanaston oppiminen on jopa 50-100 prosenttia reseptiivistä vaikeampaa (Nation 1990:48). Kuitenkin produktiivisen ja reseptiivisen sanaston hallitseminen limittyvät kielenoppimisessa ja erillisten taitojen sijasta ne tulisikin nähdä jatkumona. Omassa tutkimuksessani aineistona on suomi toisena kielenä -oppijoiden kirjoittamat tekstit, jolloin kyseessä on enimmäkseen produktiivisen sanaston tarkastelu.

Produktiivisen ja reseptiivisen sanaston laajuutta on vaikea arvioida, kuten ylipäätään sanaston laajuutta. Varsinkaan lyhyen tekstin perusteella on mahdoton tehdä päätelmiä kirjoittajan sanavaraston laajuudesta. Tutkimuksen kannalta on kuitenkin tarpeellista pystyä kohdistamaan mittaaminen erityisesti sanastoon. Siksi menetelmiä ja mittareita sanaston arvioimiseen on pyritty kehittämään niiden monista haasteista huolimatta. Seuraavassa luvussa esittelen tunnetuimpia sanaston määrälliseen mittaamiseen perustuvia menetelmiä.

3.4 Sanastontutkimuksen kvantitatiiviset menetelmät

Sanastontutkimuksessa ja erityisesti määrällisessä mittaamisessa on tärkeää tehdä selväksi, mitä mitataan ja mitä mittaamisen kohteena olevilla käsitteillä tarkoitetaan. Arkikielessä sanastoa saatetaan kuvailla rikkaaksi, laajaksi, monipuoliseksi, runsaaksi tai vaihtelevaksi ilman tarvetta tarkemmalle määrittelylle. Tällöin kyseessä on subjektiivinen näkemys ja intuitio kyseessä olevan tekstin sanastosta, mikä usein riittääkin ja toimii esimerkiksi oppilaalle annettavassa palautteessa tarkoituksenmukaisesti. Kvantitatiivisessa tutkimuksessa käytettävät käsitteet on kuitenkin syytä määritellä objektiivisen tarkasti.

Sanaston määrällisessä mittaamisessa tarkastelun kohteena voi olla sanaston rikkaus, syvyys tai laajuus. Toisen kielen sanaston kehittymistä tutkineen Lauferin (1994) mukaan sanavaraston rikkaudella voidaan tarkoittaa sitä, miten paljon eri sanoja oppija osaa käyttää, kuinka yleisiä, harvinaisia, monimutkaisia tai ainutkertaisia oppijan käyttämät sanat ovat ja kuinka suuri osa niistä on leksikaalisia sanoja. Suomessa ruotsinkielisten suomen oppimista tutkineen Grönholmin (1993) käsitys sanaston rikkaudesta on hyvin samansuuntainen: sanasto on sitä rikkaampaa, mitä enemmän se sisältää kompleksisempia ja vähätaajuisempia sanoja. Sanan kompleksisuudella Grönholm (1993: 42) tarkoittaa yksinkertaisesti sanan pituutta. Sanaston rikkautta voidaan pitää tavoiteltavana esimerkiksi kaunokirjallisuudessa ja oppilaan kirjoittamissa teksteissä, mutta syötöksessä se voi olla myös oppimista hidastava tekijä. Varsinkin kielenopetuksen alkuvaiheessa sanaston tulisi toistua sekä oppikirjoissa että opetuspuheessa, jotta sanojen oppiminen olisi mahdollisimman helppoa (Nation 1990: 7).

Tarkasteltaessa oppilaan sanastoa syvyyden näkökulmasta, tarkoitetaan usein sitä, kuinka hyvin oppilas hallitsee sanastollisen tiedon eri osa-alueita eli kuinka paljon hänellä on tietoa esimerkiksi tietyn sanan fonologiasta, morfologiasta, syntaksista, semantiikasta, pragmatiikasta ja esiintymistodennäköisyydestä (ks. esim. Schmitt 2008: 333-335, Puro 1999: 7-8). Toisen kielen sananmuodostustaitoja Cefling-aineistosta tutkineen Penttisen pro gradu -työ (2010) käy myös esimerkiksi sanaston syvyyden tutkimuksesta.

Sanaston laajuudella tarkoitetaan puolestaan sitä, kuinka paljon sanoja oppilas osaa eli kuinka laaja oppilaan sanavarasto on. Sanavaraston laajuuden mittaamista pidetään kuitenkin hankalana ja vaikeampana kuin sanaston syvyyden ja rikkauden tutkimusta. Viime aikoina sanaston laajuutta on kuitenkin tutkittu niin sanotulla sana-assosiaatio menetelmillä (esim. Meara & Fitzpatrick 2000). Yksinkertaisimmillaan sana-assosiaatio tehtävässä henkilöä pyydetään kirjoittamaan ylös mahdollisimman monta sanaa, joita hän assosioi eli joita hänelle tulee mieleen ennakkoon annetusta termistä. Assosiaatiotutkimusten on todettu antavan realistisia tuloksia tutkittavien sanavaraston laajuudesta, mutta ne eivät sovellu sanaston mittaamiseen valmiista teksteistä.

Oma tutkimukseni painottuu sanaston rikkauden mittaamiseen. Nykyään sanaston rikkaudesta saatetaan käyttää myös termiä diversiteetti eli monimuotoisuus. Esimerkiksi nykytutkijoista Jarvis kirjoittaa diversiteetistä (lexical diversity) ja Vermeer rikkaudesta (lexical richness), mutta lähempi tarkastelu paljastaa molempien tarkoittavan samaa ilmiötä. Sekä rikkaus että diversiteetti viittaavat tekstin sanojen erilaisuuden asteeseen, jolloin korkeampi aste osoittaa korkeampaa erilaisuutta. Koska diversiteettiä voidaan tarkastella lähes kaiken tyyppisistä teksteistä, sen mittaamiseen

on kehitetty monenlaisia menetelmiä. Menetelmiä on käytetty hyvin monenlaisissa yhteyksissä: Jarvis (2010) viittaa diversiteetin mittaamista käsittelevään artikkeliin (Malvern, Richards, Chipere & Duran, 2004), jonka mukaan diversiteettimittareita on käytetty muun muassa osoittamassa kirjoittamisen laatua, sanastollista tietoutta ja puheen kompetenssia sekä tutkittaessa tyylin- ja kielen omaksumista. Diversiteetin mittaamiseen kehitettyjä tunnuslukuja on hyödynnetty myös kuullun ymmärtämisessä, sosioekonomisen aseman indikaattorina ja jopa neuropatologiassa ennustamassa Alzhaimerin puhkeamista (McCarthy & Jarvis 2010: 381). Suomessa rikkaustunnuslukuja on laskettu kaunokirjallisista teksteistä, iskelmäteksteistä, puolueohjelmista ja Uuden testamentin teksteistä (Särkkä 1987, Räsänen 1975, Vehmaskoski 1976) sekä oppikirjoista (Voionmaa 1993: 130-132; Grönholm 1993: 97, Puro 1999: 15, Jaakola 2004) esimerkiksi oppikirjojen luettavuutta tarkasteltaessa. Tunnuslukuja on käyttänyt myös Mäkinen (1997) tutkiessaan opetuspuheen sanastoa suomen kielen alkeiskurssilla, Saarela (1997) arvioidessaan peruskoululaisten kirjoitelmien sanastollista kehittymistä ja Niemikorpi (1991) kuvatessaan suomen kielen sanaston yleisiä piirteitä väitöskirjassaan.

Suurimpana ongelmana sanaston diversiteetin mittaamisessa ovat olleet rikkaustunnusluvut, jotka vaihtelevat herkästi tekstin pituuden mukaan. Monien tunnuslukujen kohdalla tekstin pituus vaikuttaa tunnusluvuista saataviin tuloksiin, joten keskenään eripituiset tekstit eivät ole vertailukelpoisia. Toisaalta tekstien diversiteetistä on siten julkaistu harhaanjohtavia tuloksia, toisaalta tutkijat ovat tämän välttääkseen joutuneet rajaamaan aineistoa ja valitsemaan vertailtavaksi vain samanpituisia tekstejä. Erityisen hyvin tämä ongelma on tunnettu käytettäessä TTR-tunnuslukua, joka on tunnetuin ja samalla yksinkertaisimpia sanaston diversiteetin tunnuslukuja.

3.5 Rikkausluvut

Vaikka Särkkä (1987: 129) varoittaakin sanaston kvantitatiivisten tutkimusmenetelmien liiallisesta ihannoinnista, on esimerkiksi Yhdysvalloissa kvantitatiivista tutkimusta ja kvantitatiivisia menetelmiä kehitetty paljon eteenpäin. Uusien sanaston rikkauden mittaamiseen ja arvioimiseen tarkoitettujen indeksien ja kaavojen on todettu antavan aikaisempaa luotettavampia tuloksia.

Särkkä on jaotellut sanaston rikkautta mittaavat matemaattiset kaavat kahteen ryhmään: sanojen ja saneiden suhteeseen perustuviin kertoimiin sekä hajontaan perustuviin indekseihin (Särkkä 1987: 131). Uusimmissa sanaston rikkautta käsittelevissä tutkimuksissa on mukana myös sanojen harvinaisuuteen perustuvia menetelmiä, jotka ottavat huomioon sanojen frekvenssin. Tässä

tutkimuksessani en ota kantaa jaotteluun, sillä varsinkaan uudet tunnusluvut eivät asetu kovin luontevasti vanhojen ylälukujen alle, vaan sisältävät useampien luokkien piirteitä.

Tunnetuin esimerkki rikkausluvuista lienee **TTR-arvo** (type/token ration) eli tekstissä esiintyvien lekseemien ja saneiden osamäärä. TTR-arvo ilmoitetaan usein prosentuaalisesti: mitä pienempi prosenttiosuus, sitä toistuvampaa eli ”köyhempää” sanasto on. Sanasto on siis sitä rikkaampaa ja monimuotoisempaa, mitä enemmän eri lekseemejä tekstissä on ja mitä vähemmän ne toistuvat. Lekseemien ja esiintymien väliseen suhteeseen perustuva TTR-arvo ei kuitenkaan ole kovin luotettava, sillä TTR ei ota huomioon otoksen suuruutta. TTR-arvo on yleensä suhteellisesti sitä pienempi, mitä suurempi otos on (Särkkä 1974: 104) ja liian pienet otokset (joidenkin arvioiden mukaan alle 5000 sanaa) antavat aineistolle liian suuria rikkauslukuja, koska yksittäisten sanojen toistoa on vähän. TTR-arvon suurin ongelma onkin sen riippuvuus otoksen koosta: sen avulla ei voida vertailla eripituisia tekstejä keskenään.

TTR-arvon käänteisluku on **M-kerroin**, joka saadaan jakamalla saneiden määrä lekseemien määrällä. M-kerroin on toistuvuusluku, joka ilmoittaa kuinka monta kertaa lekseemi keskimäärin esiintyy otoksessa. Mitä suurempi M-kerroin, sitä toistuvampaa eli köyhempää sanasto on. (Särkilahti 1977: 49.) M-kertoimen käyttöä koskevat kuitenkin samat ongelmat otoskoon vaihtelun aiheuttamasta vinoumasta kuin TTR-arvoakin.

TTR-arvon puutteista johtuen tutkijat ovat pyrkineet kehittämään luotettavampia tunnuslukuja, jotka olisivat riippumattomia aineiston koosta. Näitä ovat mm. Carrollin TTR, Guiraud’n rikkausindeksi (1954), Brunet’n W-indeksi (1973) ja Honorén R-indeksi, TTR_{log}-indeksi (Richards – Malvern 1999), Yulen K-indeksi sekä jälkimmäisestä kehitetty hajonnan huomioiva Herdanin V_m-indeksi (1960) (Tarkemmin kyseisistä tunnusluvuista suomeksi esim. Jaakola 2004: 100-103). Näistä indekseistä Guiraud’n indeksi lienee saanut eniten kannatusta sanastontutkimuksessa. Yhteistä kaikille mainituille indekseille on kuitenkin se, että diversiteettiä mitattaessaan ne pyrkivät hyödyntämään saneiden ja lekseemien välistä suhdetta. Indeksit eivät myöskään ota huomioon sanojen frekvenssiä. Vaikka edellä mainitut indeksit antavatkin TTR-arvoa luotettavampia tuloksia, ne ovat saaneet osakseen myös kritiikkiä eikä niiden luotettavuudesta olla yksimielisiä tutkijoiden keskuudessa.

Yksi ehdotus luotettavammaksi mittariksi, joka ottaa sanojen taajuuden huomioon, on saksalaisten lasten sanaston diversiteettiä tutkineen Vermeerin (2004) kehittämä MLR-menetelmä (Measure of Lexical Richness). Vermeerin käyttämä sanastollisen rikkauden mittari MLR perustuu

sanojen vaikeusasteille. Sanojen vaikeusaste (the degree of difficulty of the words) on Vermeerin mukaan yhtä kuin sanojen taajuus, koska päivittäisessä kielenkäytössä yleisimmin esiintyvät sanat opitaan ensimmäisinä ja ovat täten helppoja sanoja. Vastaavasti harvinaiset sanat ovat vaikeita. Vaikka sanan oppimiseen ja muistamiseen vaikuttavat myös muut seikat, kuten äänne- ja muotorakenne, Vermeer pitää sanan taajuutta merkittävimpänä sanan osaamiseen liittyvänä tekijänä. Vermeerin tutkimus keskittyy kirjoitetun kielen sijasta puhuttuun kieleen, mutta se perustuu muuten hyvin samanlaisille lähtökohdille kuin oma tutkimukseni. Siksi esittelen seuraavaksi MLR:n tutkimusprosessin hieman tarkemmin.

Vermeerin tutkimuksessa sanan yleisyys määriteltiin sen esiintymistaajuutena noin kahden miljoonan sanan korpuksessa, joka oli kerätty esi- ja alakoulussa käytetystä kieliaineistosta. Aineiston sanat olivat peräisin opettajien suullisista ja kirjallisista ohjeista, eri oppiaineiden tehtävä- ja lukukirjoista sekä kirjoissa esiintyvistä kuvista. Tästä aineistosta syntyi tutkimuksessa käytetty korpus. Tutkittavana oli 16 natiivia saksanpuhujaa ja 16 saksaa toisena kielenä oppivaa syntyperäistä saksalaista lasta. Suullisessa haastattelussa jokaiselta lapselta kerättiin noin 200 lausetta, joista leksikaalista diversiteettiä mitattiin eri menetelmillä ja joiden antamia tuloksia sitten vertailtiin keskenään. Tutkimukseen osallistuneilla lapsilla teetettiin myös reseptiivinen sanastotehtävä ja määrittelytehtävä, joiden tuloksia verrattiin lekseemin ja saneen suhteeseen perustuvien mittareiden sekä MLR:n antamiin tuloksiin.

Vermeerin tutkimukset osoittivat, että MLR pystyy tekemään eron L1 ja L2 puhujien välille. MLR-tulokset korreloivat myös lapsen reseptiivisestä sanastotehtävästä ja määrittelytehtävästä saamien tulosten kanssa. Tulosten perusteella MLR vaikuttaa luotettavammalta mittarilta spontaanin puheen sanaston rikkauden analysoimisessa kuin lekseemi-sane-suhteeseen perustuvat mittarit. Sanojen taajuuteen perustuvia sanaston diversiteetin mittareita kannattaisi siis hyödyntää entistä enemmän sanastontutkimuksessa. Tässä tutkimuksessa käyttämäni rikkausluvut olen esitellyt tarkemmin luvussa 4.5.

4 AINEISTO JA MENETELMÄT

4.1 Tutkimuksen tavoitteet ja tutkimuskysymykset

Oma tutkimukseni tarkastelee sanaston kehittymistä Eurooppalaisessa viitekehyksessä kuvattujen taitotasojen näkökulmasta. Tutkimukseni tavoitteena on selvittää, kuinka sanaston osaaminen kehittyy suomi toisena kielenä -oppijoiden teksteissä taitotasolta toiselle siirryttäessä. Tarkastelun kohteena on sanavaraston rikkaus eli leksikaalinen diversiteetti. Aineistona ovat Yleisen kielitutkinnon kirjoittamisen osakokeen vastaustekstit ja peruskoululaisten kirjoitelmat.

Työni lähtökohta noudattaa Cefling-hankkeen päätavoitteita eli tarkoituksena on selvittää, millaista kielitaito on kullakin taitotasolla. Vertailun kohteena on toisaalta lasten ja aikuisten sanavarasto ja sanastollinen osaaminen, toisaalta eri tekstilajien vaikutus sanaston runsauteen. Tekstilajeina on sekä muodollisia että epämuodollisia tekstejä. Tutkimuskysymykseni muotoilin seuraavasti.

Kuinka sanaston osaaminen kehittyy suomi toisena kielenä -oppijoiden teksteissä?

– Millaista sanaston osaaminen on kullakin taitotasolla?

– Kuinka koululaisten ja aikuisten sanavarastot eroavat toisistaan?

– Miten tekstilaji (muodollinen vr. epämuodollinen tekstilaji) vaikuttaa sanaston rikkauteen?

4.2 Aineisto

Tutkimuksessani olen käyttänyt Cefling-hankkeessa käytettyä kirjallista aineistoa. Hanketta varten kerättiin toisena kielenä -oppijoiden kirjoitussuorituksista koostuva aineisto, joka arvioitiin Yhteiseen Eurooppalaiseen viitekehykseen pohjautuvan taitotasoasteikon mukaisesti. Tekstejä arvioimassa olivat tehtävään erityisen koulutuksen saaneet arvioijat.

Aineisto muodostuu Yleisten kielitutkintojen aikuisten kirjallisista suorituksista (myöhemmin Yki-aineisto) ja erityisesti Cefling-hanketta varten kerätyistä yläkoululaisten kirjoitelmista (koululaisaineisto). Yki-aineisto ja koululaisaineisto ovat vertailukelpoisia keskenään, koska kumpikin perustuu tavoitteiltaan ja tehtäviltään Euroopan neuvoston kehittelemään Eurooppalaiseen viitekehykseen (YKI-verkkosivut, Cefling-verkkosivut). Aikuisten tekstit jakautuvat A-, B- ja

C-tasoille, koululaisaineisto A- ja B-tasoille. A, B ja C. A-tason oppijat ovat perustason kielenkäyttäjiä, B-tason oppijat itsenäisiä kielenkäyttäjiä ja C-tason taitavia kielenkäyttäjiä. Nämä tasot jakautuvat kukin vielä kahdeksi, jolloin asteikko jakautuu yhteensä kuuteen tasoon (suluissa englanniksi): A1 alkeistaso (breakthrough), A2 selviytyjän taso (waystage), B1 kynnystaso (threshold), B2 osaajan taso (vantage), C1 taitajan taso (effective operational profiency) ja C2 mestarin taso (mastery). (EVK 2003: 46-47.)

Koululaisaineistossa on viisi tehtävätyyppiä: viesti ystävälle, viesti opettajalle, sähköpostiviesti verkkokauppaan, mielipide sekä kertomus. Kaksi ensimmäistä ovat tekstilajeiltaan epämuodollisia ja sähköpostiviesti verkkokauppaan edustaa muodollista tekstilajia. Aikuisten aineisto sisältää kolme eri tehtävätyyppiä: epämuodollinen viesti, muodollinen viesti ja mielipide. Aineisto koostuu yhteensä 1197 tekstistä, joista 527 on koululaisten ja 670 aikuisten kirjoittamia.

4.2.1 Koululaisten ja aikuisten kirjoitustehtävät

Aineiston keräämisessä käytetyt tehtävänannot oli suunniteltu simuloimaan arjen kommunikatiivisia tilanteita. Jotta tehtävät osoittaisivat oppilaan osaamista mahdollisimman hyvin, tehtävistä vastaava työryhmä pyrki suunnittelemaan tehtävät niin, että ne rajaisivat oppilaan antaman vastauksen tiettyyn tekstilajiin, mutta antaisivat silti oppilaalle vapauden tuottaa itsenäinen teksti. Työryhmä joutui pohtimaan myös tehtävänantotekstiä, jotta oppilas ei voisi liikaa hyödyntää niissä annettuja sanoja. Ennen varsinaisten kirjoitussuoritusten keräämistä tehtävien toimivuutta myös kokeiltiin pilottiryhmällä. (ks. Alanen, Huhta & Tarnanen 2010.)

Kolmessa koululaisten tehtävänannossa pyydettiin kirjoittamaan sähköposti. Ystävälle ja opettajalle suunnatut viestit edustavat epämuodollista tekstilajia, verkkokauppaan lähetettävä reklamaatioviesti on puolestaan muodollinen. Kaikissa kolmessa tehtävänannossa ohjeistettiin kertomaan yhteydenoton syy sekä muistutettiin erikseen sopivasta aloituksesta ja lopetuksesta. Tehtävälomake jäljitteli tavanomaista sähköpostiformaattia, jossa oli valmiiksi annetut paikat *lähettäjä, vastaanottaja* ja *aihe.* Esimerkiksi Sähköposti verkkokauppaan -tehtävänanto oli seuraavanlainen:

Nimi:
Koulu ja luokka:

Sähköposti verkkokauppaan

Isoveli on tilannut sinulle verkkokaupasta tietokonepelin. Peli toimii huonosti.
Kirjoita sähköpostiviesti verkkokauppaan ja kerro

- kuka olet
- miksi kirjoitat (kerro kaksi ongelmaa pelistä)
- mitä haluat, että asialle tehdään
- omat yhteystietosi.

Kirjoita selvällä käsialalla **suomeksi**. Muista sopiva aloitus ja lopetus.

Vastaanottaja: **Lähettäjä:** **Aihe:**

Sähköpostiviestien lisäksi koululaisten tehtävänantoihin kuuluivat myös mielipiteen ja kertovan tekstin tehtävätyypit. Tehtävänannot kuuluivat:

Mielipide
Valitse aihe 1 tai 2 ja kirjoita koulun lehteen, mitä mieltä olet. Perustele mielipiteesi.
1. Kännykät pois koulusta!
2. Vanhemmat saavat päättää, miten lapset käyttävät Internetiä.
Kirjoita suomeksi selvällä käsialalla alla olevaan tilaan. Kirjoita vähintään viisi lausetta.

Kerro!
Kerro jokin pelottava tai hauska asia, joka sinulle on tapahtunut.
- Mitä tapahtui.
- Miksi tapahtuma oli pelottava tai hauska.

Kirjoita selvällä käsialalla **suomeksi** alla olevaan tilaan.

Vaikka sekä mielipiteen että kertovan tekstin tehtävänannot jättävät oppilaan omalle ajattelulle runsaasti tilaa, mielipiteen tehtävänanto on rajattu kahteen aiheeseen, kun taas kertovan tekstin aihe on täysin vapaa. Kaikista koululaisille suunnatuista tehtävistä juuri kertomuksen kirjoittamista on ohjattu vähiten. On oletettavaa, että vähiten ohjatuista tehtävänannoista syntyy keskenään vaihtelevampaa sanastoa sisältäviä tekstejä.

Koululaisten tehtävänannot olivat samat kaikille. Sen sijaan Yki-tutkintoa suorittaneet aikuiset oppijat oli jo valmiiksi jaettu tasoille A, B ja C ja jokaiselle tasolle oli laadittu omat tehtävänannot. Aikuisten epämuodollisessa tehtävätyypissä A ja B tasoilla kirjoittajaa pyydetään kirjoittamaan ystävälle viesti, jossa pitää esimerkiksi vastata kutsuun tai perua tai suunnitella tapaamista. C-tasolla epämuodollista tehtävänantoa edustaa puolestaan talkookutsun kirjoittaminen. Aikuisten tehtävänannot tuntuvat olevan koululaisille suunnattuja tehtävänantoja pitempiä ja ne muuttuvat yhä yksityiskohtaisemmiksi ja pidemmiksi taitotason noustessa. Esimerkiksi talkookutsutehtävänannossa ohjeistetaan tarkasti, mitä asioita kutsun on sisällettävä (motivointi, ajankohta, tarvittavat välineet, tarjoilu jne.).

Aikuisille suunnatuissa muodolliseen tekstin kirjoittamiseen tähtäävissä tehtävänannoissa oli kyse muun muassa palautteen antamisesta, reklamaatioviestistä, lisäajan pyytämisestä työprojektiin tai kuvitteellisen esitelmän tiivistelmän laatimisesta. Palautteenantotehtäviä oli muodollisen tehtävätyypin lisäksi myös A-tason mielipide-tekstilajin tehtävänannoissa. Tehtävänlaatijoiden mukaan palautteen kirjoittaminen voi siis edustaa sekä mielipidettä että muodollista viestiä. Tietyn tekstilajin määritteleminen itsenäiseksi lajiksi on lopulta aina tutkijan valinta ja kuten Saukkonen (2001, 165-166) huomauttaa, rajat tekstilajien välillä ovat häilyviä. Tehtävänannoissa palaute on luultavasti päädytty määrittelemään mielipiteeksi, koska varsinaisen mielipidetekstin teettäminen A-tason oppivilta olisi paljon vaadittu. Toisaalta koululaisille laadituissa tehtävänannoissa myös A-tason oppijoita pyydettiin kirjoittamaan lyhyt mielipideteksti. Koululaisten aiheet olivat kuitenkin yksinkertaisempia ja tehtävänannoissa ohjattiin esimerkiksi kirjoittamaan ”vähintään viisi lausetta”. B- ja C-tason mielipidetehtävänannoissa pyydettiin valitsemaan valmiista otsikoista yksi ja laatimaan sen pohjalta mielipidekirjoitus. Otsikoiden aiheet liittyivät muun muassa politiikkaan, urheiluun, hyvinvointiin, matkusteluun ja

työntekoon. Tässä tutkimuksessani en voi esittää suoria esimerkkejä aikuisten tehtävänannoista, sillä samoja tehtävänantoja käytetään Yki-tutkintojen testeissä edelleen.

4.2.2 Esimerkkejä aineistosta

Tutkimuksessani en tarkastele yksittäisiä kirjoitussuorituksia, vaan tietyn taitotason kirjoituksia tehtävätyypeittäin. Seuraavassa on kuitenkin muutama esimerkkiteksti eri tasoilta sekä koululais- että Yki-aineistosta. Koulaisaineiston esimerkkitekstit ovat sähköposteja verkkokauppaan, aikuisten tekstit edustavat epämuodollisia viestejä.

Koululaisaineisto: Sähköposti verkkokauppaan

A1
Terve Minä olen Matti mun isoveli osti mulla tietokonepeli se toi kotiin ja annoi mulle sitten kun mä avasi n sitä peli sitten kun laitoin sita tietokoneesen ja sitten kun halusin pelata se ei toiminut ja sitten se on käytetty (I) mä haluun että annat mulle uus peli kiitos.

A2
Moi, Mun isoveli osti eilen teidän kaupasta yhen pelin, mutta silloin kun mä pistiin koneeseen peli oli ilman ääntä ja en pystyy pelaa netissä sitä peliä (kannessa luki että voidan pelaa netissä). Tuunks mä vahtaa tän pelin tai tiedätte että mistä se johtuu.
Vastakaa heti kun te saitte tämän sähköpostiin!
Kiitos
T. Matti Solki

B1
Hei olen Maija. Eilen Isovelini kävi, ostaa Teiltä tietokonen peli. ja mulla olisi pikku Ongelma, Ja se olisi että siinä tietokone pelissä vähän ongelmia. Kun se ei toimi oike hyvin, kun käynistän sitä peliä CD sanoo että levy on tyhja välillä mutta toisalta se toimi välillä ja sanoo yhessä vaiheessa että levy on tyhjä ja haluaisin jotenki palauta jos on Mahdollista? odottelen vastaustasi
Terveisin: Maija.

B2
Moi!
Minä olen Maija Solki.
Isovelini on tilannut minulle verkkokaupasta tietokonepelin, mutta se toimii huonosti. Siellä on ääni-virhe, siis ääni ei kuulostu ollenkaan. Ja sitten pelissä on vielä joitakin muita häiriöitä, kuten värit, pelini on mustavalkoisena minun tietokoneella ja joskus käy niin että tietokone menee kiinni keskipelissä, mutta kaikki muut pelit toimivat ihan normaalisti.
Voisitteko korjata sen tietokonepelin jotenkin tai lähettää uuden? Tai onko mahdollista saada rahaa siitä takaisin?

Yki-aineisto: Epämuodollinen viesti

A2
Hei Kalle! Pyydan anteksi, koska en voi menna saunaan sinun kanssa illalla. Minulla on pieni ongelma. Kerron sitten. Ehkä menemme huomenna? Soitan aamulla, sitten sopimme. Maija.

B2
Moi Kalle,
Kiitos hääkutsusta. Olin yllättynyt kuulla, että menet Kaisan kanssa naimisiin. Lämpimät onnittelut, valitsit hyvin. Valitettavasti en pääse osallistumaan häihin, koska olen juuri silloin Ruotsissa työkomennuksella. Lähetän kuitenkin häälahjan jo etukäteen, jos sopii. Onko teillä toivomuksia vai voinko itse valita jotain?
Täällä kotimaassa, kaikki on niin kuin ennenkin. Maija tekee vieläkin kovasti töitä ja on illallakin usein toimistossa. Mikko lähtee ensi kuussa esikouluun. Hän puhuu siitä jo päivittäin. Toivon teille paljon onnea! Pidäkää huolta toisistaan.
Matti

C2
4.9.2005
Hei kaikki asukkaat!
Taloyhtiömme perinteiset kevättalkoot pidetään la 16.4.2005 klo 14-18 00, A ja B talojen pihalla. Talkooissa siivotaan kävelytiet, leikataan pensaat ja siistitään piha kevätkuntoon. Mukaan tarvitaan puutarhahanskat ja hyvä mieli. Työvälineet (harjat, oksasakset yms.) löytyvät B talon pesutuvan eteisestä. Sään mukainen vaatetus päälle! (Lapsille kannattaa laittaa kumisaappaat ja kurahousut, takapihalta löytyy ihania lätäköitä.) Nautitaan yhdessä liikunnasta, seurasta ja siististä pihasta! Huomio! Hyvin tehdyn työn päätteeksi taloyhtiö tarjoaa makkaraa ja olutta. (Lapsille mehua.)
Tervetuloa mukaan!
taloyhtiön puolesta: Maija

4.3 Aineiston lemmaus

Ennen varsinaista analyysiä lemmasin eli sanastin koko aineiston. Lemmatessa määritin jokaiselle saneelle lekseemin ja sanaluokan. Ensimmäisessä vaiheessa lemmasin saneet aakkosjärjestyksessä työn nopeuttamiseksi. Monissa tapauksissa saneella oli kuitenkin useampi eri lekseemivaihtoehto (esimerkiksi sane *asua* voi olla joko verbi *asua* tai partitiivi substantivista *asu*) ja monella lekseemillä useampi sanaluokka vaihtoehto (esimerkiksi sane *myös* voi olla joko adverbi tai konjuktio tilanteesta riippuen). Tällaisia epäselviä tapauksia en voinut lemmata lopullisesti ensimmäisessä vaiheessa.

Ensimmäisen vaiheen raakalemmauksen jälkeen kävin aineiston uudelleen läpi palauttamalla sen aakkosjärjestyksestä luonnolliseen juoksevaan järjestykseen, jolloin pystyin päättele-

mään useimpien epäselvien saneiden lekseemin ja sanaluokan tekstiyhteydestä. Mikäli sane jäi edelleen epäselväksi, merkitsin lekseemiksi ja sanaluokaksi pelkän kysymysmerkin. Merkitsin kysymysmerkillä myös kaikki aineistossa esiintyneet numerot, erikoismerkit ja vieraskieliset saneet. Lopulta näitä kysymysmerkillisiä epäselviä saneita oli yhteensä noin 2500 eli noin kolme prosenttia koko aineistosta.

Lemmatessa jouduin tekemään seuraavia ratkaisuja:

a) Partisiippimuodot

Kaikki aineiston partisiippimuodot määrittelin verbeiksi, vaikka monet muodoista olivatkin mielestäni lähempänä adjektiivia kuin verbiä. Tähän ratkaisuun päädyin, koska aineistoa verrattiin suomen sanomalehtikielen taajuussanastoon, jossa kaikki partisiippimuodot on niin ikään määritelty verbeiksi.

b) Slangisanat

Aineistossa oli jonkin verran slangisanoja kuten *leffa*, *treenata* ja *sori* jotka lemmasin sellaisinaan omiksi lekseemeikseen. Taajuussanastoon kuuluu vain yleiskielisiä lekseemejä, joten omiksi lekseemeikseen lemmatut slangisanat näyttäytyvät ehkä virheellisesti todellisuutta harvinaisempina lekseemeinä. Slangisanojen kääntäminen vastaaviksi yleiskielisiksi lekseemeiksi olisi kuitenkin ollut keinotekoista. Sanoma välittyy perille slangisanoistakin ja osaltaan ne kertovat sanaston hallitsemisesta, vaikka ovatkin usein lainasanoja. Kuitenkin persoonapronominit *mä* ja *sä* päädyin kääntämään yleiskielisiksi pronomineiksi *minä* ja *sinä* niiden yleisyyden vuoksi.

c) Lyhenteet

Yksittäisten kirjainten kohdalla ei aina voinut päätellä onko kyseessä lyhenne vai muuten vaan irrallinen kirjainyksikkö. Osa tapauksista selvisi tekstiyhteydestä, mutta osan jouduin merkitsemään vain kysymysmerkillä. Rajatapauksia olivat myös vieraskieliset lyhenteet kuten *pc, gsm* ja *sos*. Osan lyhennetyistä erisnimistä kuten *vr, sdp* ja *bmw* merkitsin ensin systemaattisesti erisnimiksi, mutta huomattuani, että taajuussanasto käsittelee kaikki nämä lyhenteinä, vaihdoin logiikkaa. Vakiintuneet alkujaan vieraskieliset sanat kuten *cd, dvd* ja *wc* luokittelin lyhenteiksi, vaikka saneiden määritteleminen substantiiveiksi voisi olla yhtä perusteltua. Kaiken kaikkiaan eniten epäjohdonmukaisuuksia lemmauksessa saattaa olla juuri lyhenteiden kohdalla.

d) Epäsanat

Aineistossa tuli vastaan useita kummallisia yhdyssanoja, joita ei varsinaisesti esiinny natiivien suomen puhujien teksteissä kuten *serkkukaveri, ulkotarve* ja *rekanveturi*. Saneet olivat kuitenkin oikeakielisesti muodostettuja ja niitä esiintyi suhteellisen harvoin, joten lemmasin ne sellaisinaan. Nämä harvinaiset lekseemit saattavat nostaa joidenkin tekstien sanastollista rikkautta, vaikka kyseessä on kirjoittajan oma keksimä ilmaisu, jolle saattaisi yleiskielessä löytyä parempikin vaihtoehto.

e) Virheelliset saneet ja ilmaisut

Virheellisesti kirjoitetut, mutta tekstiyhteydessään ymmärrettävät saneet lemmasin kuten mitkä tahansa oikeinkirjoitetut saneet. Useissa tapauksissa ilmaisut olivat hyvin tulkinnanvaraisia, jolloin päädyin lemmaamaan saneet omasta mielestäni todennäköisimpiin lekseemeihin tai mikäli en sellaista pystynyt määrittelemään, lemmasin saneet kysymysmerkillä. Helposti ymmärrettäviä ilmaisuja oli esimerkiksi *mina selka poliisi* (minä pelkään poliisia) ja *olen ollut viikon pikeä* (olen ollut viikon kipeä), mutta hankalampia esimerkiksi ilmaisut *Olen matkalla etäränällä* (itäraja vai eteläranta?) ja *Afganistanin veillä on suotta* (Afganistanin veljellä/veljillä on suota/suoja vai Afganistanin teillä on suota?).

4.4 Lemmauksessa käytetty taajuussanasto

Lekseemien määrittämisessä sanaluokkiin käytin mahdollisimman pitkälle samaa sanaluokkajaottelua kuin on käytetty Suomen sanomalehtikielen taajuussanastossa. Sanasto koottiin vuonna 2004 ja se on vapaasti saatavilla Tieteen tietotekniikan keskuksen (CSC) verkkosivuilla. Sanasto sisältää sanomalehtikielen 9996 yleisintä lemmaa ja lähdeainestossa on ollut 43 999 826 sanetta. Taajuussanasto on olennaisessa osassa tutkimusta laskettaessa aineistosta eri tunnuslukuja, erityisesti harvinaisuustunnuslukuja laskettaessa. Aikaisemmin vastaavaa sanaston harvinaisuuteen perustuvaa tutkimusta on tehty suomessa lähinnä vain englannin kielisestä aineistosta, jolloin taajuussanastona on käytetty joko amerikkalaista tai englantilaista kansalliskorpusta (American National Corpus, British National Corpus). Esimerkiksi englantilainen kansalliskorpus ulottuu kuitenkin vain 6 500:an yleisimpään sanaan, kun suomenkielinen taajuussanasto kattaa 9 996 sanaa. Toisaalta suomenkielisessä sanastossa on kyse vain sanomalehtikielestä, kun englanninkielisissä korpuksissa lähdeaineistossa on mukana muun muassa kaunokirjallisia tekstejä, jolloin korpusta voidaan pitää syvempänä.

4.5 Lemmauksen jälkeinen analysointi ja sanastollisen diversiteetin tunnusluvut

Lemmattuani koko aineiston lähetin sen ja sanomalehtikielen taajuussanaston professori Scott Jarvisille Ohion yliopistoon. Jarvis on tutkinut leksikaalista diversiteettiä englannin kielisestä aineistosta ja kehittänyt tutkimustensa pohjalta ohjelman, joka vertaa aineistoa taajuussanastoon ja kertoo sanaston rikkaudesta erilaisia tunnuslukuja. Sanojen ja lekseemien lisäksi ohjelma laski aineistosta Shannonin indeksin, harvinaisuustunnusluvun, sisältösanojen harvinaisuustunnusluvun, monipuolisuustunnusluvun (MTLD), tasapuolisuustunnusluvun ja hajaannustunnusluvun. Esittelen nämä tunnusluvut seuraavaksi.

4.5.1 Shannonin indeksi

Shannonin moninaisuustunnusluku lasketaan sanojen osuuksien perusteella eli sen perusteella, mitä prosenttiosuutta kukin sane edustaa, ja nämä osuudet kerrotaan logaritmiarvoilla. Shannonin indeksi vaihtelee jonkin verran tekstin pituuden mukaan, mutta se ennustaa silti tehokkaasti oppijoiden kielitaitotasoa. Shannonin indeksi tunnetaan myös nimillä Shannonin diversiteetti-indeksi (moninaisuustunnusluku), Shannonin-Weaverin indeksi ja Shannonin entropia. (Malvern ym. 2004.)

Shannonin indeksiä on jo pitkään käytetty monilla aloilla esimerkiksi ekologiassa laskettaessa eliöyhteisön monimuotoisuutta. Lingvistisessä tutkimuksessa Shannonin tunnusluku on otettu käyttöön myöhemmin. Lingvistisen tutkimuksen kohteena on monimuotoisuus kuten ekologiassakin, mutta indeksi lasketaan eläinlajien ja yksilöiden sijaan tekstin eri lekseemien määrän ja lekseemien esiintymien määrän perusteella. Shannonin indeksi lasketaan kaavalla

$$H' = -\sum_{i=1}^{R} p_i \log p_i$$

missä R on sanemäärä, pi on lekseemin i osuus näytteen koko sanemäärästä. Indeksiarvo on pienin, kun kaikki saneet ovat peräisin samasta lekseemistä (kaikki saneet edustavat samaa lekseemiä eli H = log 1 = 0) ja suurin, kun yhdelläkään saneella ei ole samaa lekseemiä toisen saneen kanssa (kaikkia lekseemejä esiintyy yhtä paljon). Indeksin mukaan teksti on sitä monimuotoisempi, mitä enemmän ja tasaisemmin sanemäärältään jakautuneita lekseemejä tekstissä esiintyy.

Aikaisemmin Shannonin indeksiä on käyttänyt suomenkieliseen aineistoon muun muassa Leena Saarela (1997), joka tutki väitöskirjassaan peruskoululaisten kirjoitelmien sanaston ke-

hittymistä. Saarelan mukaan tekstin rikkaudesta kertoo sanaston monipuolinen käyttö ja informaatiotiheys. Saarelan informaatiotiheyden laskemiseen käyttämiä tietoja ovat sanojen keskipituus, infinitiivi- ja partisiippimuotojen sekä relatiivipronominien määrä ja eksplikatiivisen *että*-konjunktion käyttö. (Saarela 1997: 49, 52.) Näiden lisäksi hän laski teksteistä Shannonin indeksin. Saarela pitää Shannonin indeksin etuna sen kykyä tasoittaa suuritaajuisten sanojen (kuten *olla* ja *ja*) ja suhteellisen pienen otoskoon aiheuttamaa vinoumaa. Otoskoon lisäksi kielen suurifrekvenssiset sanat ovat usein ongelmana hajontaan pohjautuvissa mittareissa. Jos suurifrekvenssien sanojen osuus aineistosta on pieni, keskimääräinen poikkeama keskiarvosta on pieni ja sanasto on mittarin mukaan rikasta (Särkkä 1987: 134).

Shannonin puolesta puhuu myös Cefling-aineiston englanninkielisestä aineistosta tehty tutkimus. Tutkimuksessa Shannonin tunnuslukua käytettiin oman tutkimukseni tapaan. Tutkimuksessa tehdyn analyysin mukaan juuri Shannonin tunnusluku ennustaa hyvin oppijoiden taitotasoa: se korreloi taitotason kanssa vahvemmin kuin mikään muu tutkimuksessa käytetty tunnusluku. Vaikka oma tutkimukseni laskee arvoja suomenkielisestä aineistosta, kyseisen tutkimuksen tuloksia on mielenkiintoista verrata omiini.

4.5.2 Harvinaisuustunnusluku

Harvinaisuustunnusluku on sanojen keskimääräinen järjestysluku suomenkielen taajuussanastossa. Jokaisella taajuussanastossa esiintyvällä sanalla on siis yksi järjestyslukunsa, jolloin harvinaisuustunnusluku on kaikkien otoksessa esiintyvien sanojen järjestyslukujen keskiarvo (Jarvis 2011, diaesitys). Esimerkiksi A1-tason keskimääräiseksi harvinaisuustunnusluvuksi tuli 2064,04, joka tarkoittaa, että keskimääräisen sanan järjestysnumero taajuussanastossa on noin 2064, kun taajuussanastossa oli yhteensä 9996 sanaa. Esimerkkisanat havainnollistavat, millaisista sanoista tässä harvinaisuusluokassa on kysymys: järjestyslukua 2064 edustaa taajuussanastossa sana *lupaus*. Lähimmät tätä harvinaisuustunnuslukua edustavat sanat ovat lisäksi *lentokenttä, sanna, säveltäjä, keskuspankki, sektori, sotilaallinen, johanna, opiskelu*. Sellaiset sanat, joita ei löytynyt taajuussanastosta, jätettiin kokonaan pois laskuista. Tämän takia harvinaisuustunnusluku voi olla hieman todellisuutta pienempi; taajuussanastossa on vain 9996 yleisintä sanaa ja tätä harvinaisempia sanoja ei ole mukana.

4.5.3 Sisältösanojen harvinaisuustunnusluku

Sisältösanojen harvinaisuustunnusluku on harvinaisuustunnusluku liittyen ainoastaan sisältösanoihin eli substantiiveihin, verbeihin ja adjektiiveihin. Sisältösanojen harvinaisuustunnuslukua laskettaessa tarkastelun kohteeksi rajataan sekä aineistosta että taajuussanastosta vain sisältösanat kun taas funktiosanat eli artikkelit, pronominit, post- ja prepositiot sekä konjunktiot jätetään pois laskuista. Näin saadaan tarkempaa tietoa tekstin merkityksellisten ja tekstin sisältöä eteenpäin vievien sanojen harvinaisuudesta. (Jarvis 2011, dia-esitys.)

4.5.4 MTLD, sanojen monipuolisuustunnusluku

Tämän luvun MTLD-tunnusluvun esittely ja kuvaus pohjautuvat McCarthyn ja Jarvisin artikkeeliin *MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment*.

MTLD (the measure of textual diversity) eli sanojen monipuolisuustunnusluku lasketaan peräkkäisistä sanajonoista, joiden tulee ylläpitää annettu TTR-arvo. Menetelmä perustuu sanojen peräkkäiseen järjestykseen ja jokaisen yksittäisen saneen omalle TTR arvolle (McCarthy & Jarvis 2010: 384). Toisin sanoen MTLD lasketaan sen mukaan, kuinka monta peräkkäistä sanaa kohdataan keskimäärin tekstissä, ennen kuin lekseemien ja saneiden välinen suhde laskee tietyn rajan alle. MTLD-tunnusluvun suhteen rajaksi ja oletusarvoksi on määritelty 0,72.

MTLD-tunnuslukua ei voida esittää lyhyellä matemaattisella kaavalla kuten esimerkiksi Shannonin indeksiä, vaan se vaatii paljon pitemmän sanallisen kuvaamisen. Esimerkiksi Abraham Lincolnin lauseessa *of the people by the people for the people* jokaisen yksittäisen saneen TTR on esimerkiksi seuraava *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) *for* (.714) *the* (.625) *people* (.556). Kun TTR laskee ensimmäisen kerran alle 0,72 eli kohdassa *people (0,667)*, annettu TTR-arvo on saavutettu, jolloin teksti saa yhden kokonaisen faktorin. Tämän jälkeen TTR-mittari nollataan ja TTR-arvoja aletaan uudelleen laskea seuraavasta sanasta lähtien. Annetussa esimerkissä tapahtuu siis seuraavasti *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) *for* (1.00) *the* (1.00) *people* (1.00). Esimerkkilause saavuttaa oletusarvon 0,72 yhden kerran, jolloin lauseen kokonaisfaktorien määrä on 1.

Kokonaisten faktoreiden lisäksi MTLD:n laskemiseksi tulee jäljelle jääneistä sanoista (loput sanat, jotka eivät muodosta kokonaista faktoria) laskea osittainen faktori eli häntäfaktori.

Häntäfaktori lasketaan sen mukaan, kuinka suuren osuuden se muodostaa kokonaisesta faktorista eli kuinka kauaksi häntäfaktorin TTR jää sovitusta arvosta 0,720. Esimerkiksi jos tekstin viimeisen yksittäisen sanan TTR on 0,887, sen etäisyys lähtökohdasta on 0,113 (1-0,887), joka on 40,4 prosenttia kokonaisfaktorin vastaavasta arvosta 0,28 (1-0,72).

MTLD ei siis jätä huomiotta kokonaisten faktoreiden jälkeen jäljelle jääneitä sanoja, vaan häntäfaktori lasketaan mukaan lopulliseen faktoriin. Lopullinen faktori saadaan laskemalla yhteen kokonaisfaktorien määrä ja häntäfaktori. Jos teksti koostuu neljästä kokonaisesta faktorista ja jäljelle jääneistä sanoista, joiden TTR on 0,887, lopullinen faktori on siis 4 + 0,404 = 4,404.

Häntäfaktorin laskemisessa on kuitenkin muutamia ongelmia. Ensiksikin jäljelle jääneistä sanoista laskettu häntäfaktori on aina likiarvo ja täten alttiimpi virheille. Toiseksi, mitä lyhyemmän tekstin jäljelle jääneet sanat muodostavat, sitä korkeammaksi häntäfaktorin TTR jää, jolloin sen lisääminen kokonaisten faktoreiden määrään nostaa harhaanjohtavasti lopullista faktoria. Yleisesti ottaen mitä lyhyemmästä tekstistä on kysymys, sitä hankalampaa on laskea MTLD luotettavasti. Tutkimusten mukaan noin sata sanaa sisältävien tekstien MTLD-lukuja voidaan kuitenkin jo pitää luotettavina.

Jotta MTLD:n lopulliseen arvoon ei vaikuttaisi liikaa häntäfaktorin likimääräisyys, arvioitavalle tekstille tehdään niin sanotusti kaksoiskäsittely. Kaksoiskäsittelyssä tekstin faktorit lasketaan sekä vasemmalta oikealle että oikealta vasemmalle (teksti ”luetaan” sekä etu- että takaperin), jolloin häntäfaktori muodostuu erilaiseksi toiseen suuntaan laskettaessa. MTLD:n lopullinen arvo lasketaan vasemmalta oikealle ja oikealta vasemmalle laskettujen lopullisten faktoreiden keskiarvosta. Koska häntäfaktori muodostuu erilaiseksi eri suuntiin laskettaessa, se tasoittaa riittävästi sitä satunnaisen vaihtelun ongelmaa, joka aiheutuu jäljelle jääneiden sanojen määrän vaihtelusta. Näin kaksoiskäsittely takaa riittävän yhtenäisyyden ja tarkkuuden MTLD:n laskemiseksi.

Seuraava esimerkki havainnollistaa MTLD:n laskemista ja kaksoiskäsittelyä faktorien laskemisessa. Taulukossa A faktoreita lasketaan vasemmalta oikealle ja taulukossa B oikealta vasemmalle. Esimerkkiteksti on poimittu tutkimukseni koululaisaineistosta ja edustaa A2 tasoa. Tehtävätyyppinä on viesti ystävälle. Siltä osin esimerkki on kuitenkin huono, että tässä tekstissä sattumalta sekä vasemmalta oikealle että oikealta vasemmalle laskettaessa juuri 31. sane saavuttaa arvon 0,72 ja häntäfaktorin viimeisen sanan TTR on molempiin suuntiin laskettaessa 0,933. Yleensä eri suuntiin laskettaessa faktorit saavat hieman eri arvot.

A) Vasemmalta oikealle laskettu (etuperin)

Juokseva	Lemmattu	Sanaluokka	Tokens	Types	TTR
Moi,	moi	interjektio	1	1	1
Kaisa!	kaisa	erisnimi	2	2	1
Me	me	pronomini	3	3	1
sovitimme	sovittaa	verbi	4	4	1
aiemmin,	aiemmin	adverbi	5	5	1
että	että	konjuktio	6	6	1
me	me	pronomini	7	6	0,857143
tavataan	tavata	verbi	8	7	0,875
kahvilassa	kahvila	substantiivi	9	8	0,888889
huomenna,	huomenna	adverbi	10	9	0,9
mutta	mutta	konjuktio	11	10	0,909091
minulla	minä	pronomini	12	11	0,916667
on	olla	verbi	13	12	0,923077
muuto	muu	pronomini	14	13	0,928571
menoa.	meno	substantiivi	15	14	0,933333
Minä	minä	pronomini	16	14	0,875
en	ei	verbi	17	15	0,882353
voi	voida	verbi	18	16	0,888889
huomenna,	huomenna	adverbi	19	16	0,842105
voisitko	voida	verbi	20	16	0,8
sina	sinä	pronomini	21	17	0,809524
tavata	tavata	verbi	22	17	0,772727
minun	minä	pronomini	23	17	0,73913
kanssa	kanssa	adverbi	24	18	0,75
ensin	ensin	adverbi	25	19	0,76
viikona	viikko	substantiivi	26	20	0,769231
tiistaina?	tiistai	substantiivi	27	21	0,777778
Minä	minä	pronomini	28	21	0,75
haluan	haluta	verbi	29	22	0,758621
tavata,	tavata	verbi	30	22	0,733333
minä	minä	pronomini	31	22	0,709677
en	ei	verbi	1	1	1
näi	nähdä	verbi	2	2	1
sinua	sinä	pronomini	3	3	1
aika	aika	adverbi	4	4	1
paljon	paljon	adverbi	5	5	1
aikaa.	aika	substantiivi	6	6	1
Olisi	olla	verbi	7	7	1
kiva,	kiva	adjektiivi	8	8	1
jos	jos	konjuktio	9	9	1
me	me	pronomini	10	10	1
olemme	olla	verbi	11	10	0,909091

B) Oikealta vasemmalle laskettu (takaperin)

Juokseva	Tokens	Types	TTR
maija	1	1	1
t	2	2	1
usein	3	3	1
tavata	4	4	1
olla	5	5	1
me	6	6	1
jos	7	7	1
kiva	8	8	1
olla	9	8	0,888889
aika	10	9	0,9
paljon	11	10	0,909091
aika	12	11	0,916667
sinä	13	12	0,923077
nähdä	14	13	0,928571
ei	15	14	0,933333
minä	16	15	0,9375
tavata	17	15	0,882353
haluta	18	16	0,888889
minä	19	16	0,842105
tiistai	20	17	0,85
viikko	21	18	0,857143
ensin	22	19	0,863636
kanssa	23	20	0,869565
minä	24	20	0,833333
tavata	25	20	0,8
sinä	26	20	0,769231
voida	27	21	0,777778
huomenna	28	22	0,785714
voida	29	22	0,758621
ei	30	22	0,733333
minä	31	22	0,709677
meno	1	1	1
muu	2	2	1
olla	3	3	1
minä	4	4	1
mutta	5	5	1
huomenna	6	6	1
kahvila	7	7	1
tavata	8	8	1
me	9	9	1
että	10	10	1
aiemmin	11	11	1

tavata	tavata	verbi	12	11	0,916667
useimpi!	usein	pronomini	13	12	0,923077
T.	t	lyhenne	14	13	0,928571
Maija.	maija	erisnimi	15	14	0,933333

sovittaa	12	12	1
me	13	12	0,923077
kaisa	14	13	0,928571
moi	15	14	0,933333

Esimerkkiteksti (tässä tapauksessa molempiin suuntiin luettaessa) saavuttaa kokonaisen faktorin vain kerran, 31. saneen kohdalla. Jäljelle jääneiden sanojen viimeisen sanan TTR on 0,933, jonka etäisyys lähtökohdasta on täten 0,067. Häntäfaktorin arvo on silloin 0,239 (0,067 : 0,28) eli se muodostaa 23,9% kokonaisfaktorin vastaavasta arvosta (1 - 0,72 = 0,28). Esimerkkitekstin lopullinen faktori on siis **1,239** (1 + 0,239).

MTLD lasketaan jakamalla tekstin saneiden kokonaismäärä lopullisella faktorilla. Annetussa esimerkkitekstissä on yhteensä 46 sanetta ja lopullinen faktori on 1,239, jolloin MTLD:n arvoksi saadaan **37,13** (46 : 1,239). Jos toiseen suuntaan laskettaessa lopullinen faktori saisi eri arvon, laskettaisiin sen perusteella myös toinen MTLD arvo. Näiden kahden MTLD arvon keskiarvo olisi lopullinen MTLD arvo.

4.5.5 Tasapuolisuustunnusluku

Tasapuolisuustunnusluku kertoo, kuinka tasaisesti eri lekseemien saneet jakautuvat tekstissä. Esimerkiksi virkkeessä *Minä kävelin, sinä kävelit, Kaisa käveli, Enni käveli, mutta Sini ajoi autoa.* lekseemien saneet eivät jakaudu kovin tasaisesti, koska lekseemi kävellä toistuu huomattavasti useammin kuin muut lekseemit. Toisin on virkkeessä *Jotkut meistä kävelivät ja jotkut meistä menivät autolla,* jossa toistuvia lekseemeitä on useampia. Ensimmäisessä virkkeessä saneiden määrän keskihajonta (SD, standard deviation) on suurempi kuin jälkimmäisessä virkkeessä, jonka keskihajonta on pienempi. Jokaiselle lekseemille voidaan siis määrittää oma tasapuolisuustunnusluku eli balance, joka saadaan saneiden keskihajonta laskemalla. Lopullinen tasapuolisuustunnusluku on kaikkien lekseemien tunnuslukujen keskiarvo. (Jarvis 2011, dia-esitys.)

4.5.6 Hajaannustunnusluku

Hajaannustunnusluku ilmoittaa, kuinka kaukana keskimäärin kunkin lekseemin saneet ovat toisistaan. Hajaannustunnusluku siis kertoo, toistuuko tietty lekseemi tasaisesti saman aineiston alusta loppuun vai vain yhdessä tai muutamassa kohtaa aineistoa. Esimerkiksi suurifrekvenssiset funktiosanat kuten konjunktio *ja* toistuvat usein tasaisesti läpi tekstin, jolloin niiden hajaannustunnusluku on muita sanoja suurempi. Sen sijaan sisältösanat saattavat toistua tiiviisti vain tietyssä kohtaa tekstiä esimerkiksi uutta aihetta käsiteltäessä. Esimerkiksi tässä kappaleessa lekseemi *hajaannustunnusluku* toistuu tiiviisti, vaikka koko tekstiin nähden se ei olekaan kovin yleinen. (Jarvis 2011, dia-esitys.)

5 SANASTON YLEISTÄ TARKASTELUA

5.1 Aineiston sane- ja lekseemimäärät

Koko aineistoni käsittää 1197 tekstiä, joissa on yhteensä 75 485 sanetta. Lekseemejä on luonnollisesti huomattavasti vähemmän, 11 650 lekseemiä. Aineisto koostuu aikuisten Yleisten kielitutkintojen kirjallisista suorituksista, joita on yhteensä 670 ja erityisesti Cefling-hanketta varten kerätyistä yläkoululaisten 527 kirjoitelmasta. Koululaisaineistossa on viisi tehtävätyyppiä: viesti ystävälle, viesti opettajalle, sähköpostiviesti verkkokauppaan, mielipide sekä kertomus. Aikuisten aineisto sisältää kolme eri tehtävätyyppiä: epämuodollinen viesti, muodollinen viesti ja mielipide. Koululaisaineiston ja yki-tutkintojen eri tehtävätyyppien ja taitotasojen kirjoitussuoritusten absoluuttiset sane- ja lekseemimäärät on esitetty seuraavassa kahdessa taulukossa. Taulukko perustuu yhdenmukaisesti arvioituun aineistoon.

Taulukko1: Absoluuttiset sanemäärät tehtävätyypeittäin ja taitotasottain

Saneet	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
Kaikki	8252	14515	21536	11061	9584	10537
Koululaisaineisto						
Viesti ystävälle (epämuod.)	322	1764	1693	192		
Viesti opettajalle (epämuod.)	477	1679	1725	570		
Sähköposti verkkokauppaan (muod.)	1032	1707	2277	496		
Mielipide	796	1656	2306			
Kertomus	799	2088	3025	909		
Yki-aineisto						
Epämuodollinen	1730	1693	2617	2357	1513	1021
Muodollinen	804	1605	2670	2301	3744	5562
Mielipide	2295	2323	5223	4236	4327	3954

Taulukko2: Absoluuttiset lekseemimäärät tehtävätyypeittäin ja taitotasottain

Lekseemit	A1	A2	B1	B2	C1	C2
Kaikki	1189	1547	2274	1886	2235	2528
Koululaisaineisto						
Viesti ystävälle (epämuod.)	122	375	329	111		
Viesti opettajalle (epämuod.)	183	289	316	202		
Sähköposti verkkokauppaan (muod.)	249	312	429	180		
Mielipide	223	336	432			
Kertomus	257	524	759	386		
Yki-aineisto						
Epämuodollinen	425	351	622	598	523	418
Muodollinen	293	469	593	573	1098	1526
Mielipide	497	500	1012	1038	1335	1332

Sane- ja lekseemitaulukot poikkeavat odotusten mukaisesti toisistaan: sanemäärät vaihtelevat sattumanvaraisesti kun taas lekseemimäärät kasvavat kohti korkeampaa taitotasoa. Sanemäärältään laajimmaksi taitotasoksi erottuu selvästi B1-taitotaso ja toiseksi eniten saneita on A2-tasolla. Tämä johtuu näille tasoille arvioitujen tekstien suuresta määrästä. Sanemäärältään suppeimpia ovat A1-tason tekstit, mikä johtuu toisaalta tälle tasolle osuneiden tekstien vähäisestä määrästä, mutta toisaalta myös alinta tasoa edustavien tekstien lyhyydestä. Erot aineiston sane- ja lekseemimäärissä tulee huomioida erityisesti sellaisissa menetelmissä, jossa otoskoon on todettu vaikuttavan lasketuihin tuloksiin.

5.2 Saneiden jakautuminen sanaluokittain

Aineisto oli hyvin verbivoittoinen. Kaikista aineiston saneista 27,62 % eli yli neljännes oli verbejä. Melkein yhtä suuri ryhmä on substantiivit (26,73 %) ja mikäli myös erisnimet (4,38 %) lasketaan mukaan substantiiveihin, substantiivien osuus (31,11 %) ylittää verbien osuuden. Seuraavaksi suurimmat ryhmät ovat pronominit (11,33 %), adverbit (10,77 %) ja konjunktiot (9,30 %). Kaikki aineiston saneet jakautuivat sanaluokkiin seuraavasti:

Taulukko3: Saneiden jakautuminen sanaluokittain

Verbit	20 846	27,62 %
Substantiivit	20 173	26,73 %
Pronominit	8 556	11,33 %
Adverbit	8 135	10,77 %
Konjuktiot	7 021	9,30 %
Adjektiivit	4 668	6,18 %
Erisnimet	3 307	4,38 %
Prepositiot	1 101	1,46 %
Interjektio	667	0,88 %
Lukusanat	506	0,67 %
Lyhenteet	505	0,67 %
Kaikki saneet	**75 485**	**100,00 %**

Vaikka aineiston yleisin sanaluokka on verbit, yleisimpänä sanaluokkana suomen kielessä pidetään tavallisesti substantiiveja. Vanhemman taajuussanaston Suomen kielen taajuussanaston (1979) mukaan yleisin sanaluokka on substantiivit: yleiskielen sanoista 35,8 % eli noin kolmannes on substantiiveja. Myös Karlsson ja Niemikorpi ovat omissa tutkimuksissaan laskeneet yleiskielisen kirjoitetun tekstin ja asiatekstin keskimääräisiksi substantiiviosuuksiksi suunnilleen saman, noin 34-42 % (Karlsson 1983: 220; Niemikorpi 1991: 199). Seuraavaksi suurin ryhmä on verbit: Suomen kielen taajuussanaston (1979) sanoista noin neljännes (24,3 %) on verbejä. Aineiston sanaluokkajakaumassa verbejä oli kuitenkin enemmän kuin substantiiveja. Verbien suureen määrään voi osaltaan vaikuttaa kaikkien partisiippimuotojen määritteleminen verbeiksi.

5.3 TTR-arvo

Vaikka lekseemien ja saneiden suhteeseen perustuva TTR-kerroin ei ole kovin luotettava vertailtaessa eripituisia tekstejä keskenään, olen ottanut sen mukaan tutkimukseeni sen yksinkertaisuuden ja tunnettuuden vuoksi. Aineistoni TTR-arvoja on myös mielenkiintoista verrata tutkimuksessa käytettyihin uudempiin rikkaus- ja diversiteettilukuihin. TTR-arvon suurin puute on sen riippuvuus otoskoosta, mikä aiheuttaa epävarmuutta erisuuruisten otosten tuloksissa (enemmän TTR-kertoimesta

luvussa 3.5). Aineistoni taitotasojen laajuudet vaihtelevat 322 saneesta 5 562 saneeseen, jolloin osa TTR-arvoista on odotettavasti harhaanjohtavia. Mukana on kuitenkin myös samanpituisia tekstejä, joiden välillä TTR-arvojen vertailu on mahdollinen.

Taulukko 4: TTR-arvot

TTR	A1	A2	B1	B2	C1	C2
Kaikki	14,41	10,66	10,56	17,05	23,32	23,99
Koululaisaineisto						
Viesti ystävälle (epämuod.)	37,89	21,26	19,43	57,81		
Viesti opettajalle (epämuod.)	38,36	17,21	18,32	35,44		
Sähköposti verkkokauppaan (muod.)	24,13	18,28	18,84	36,29		
Mielipide	28,02	20,29	18,73			
Kertomus	32,17	25,10	25,09	42,46		
Yki-aineisto						
Epämuodollinen	24,57	20,73	23,77	25,37	34,57	40,94
Muodollinen	36,44	29,22	22,21	24,90	29,33	27,44
Mielipide	21,66	21,52	19,38	24,50	30,85	33,69

TTR-arvoilla on taipumus jäädä suhteettomasti sitä pienemmiksi mitä suuremmasta otoskoosta on kysymys. Esimerkiksi kaikkien tekstien TTR-arvo laskee odotustenvastaisesti A1 tasolta A2 ja B1 tasoille siirryttäessä, sillä A2 ja B1 tasoilla sanemäärät ovat kaikkein korkeimmat. Koska TTR-arvot ovat riippuvaisia otoskoosta, taulukkoa pitää lukea vertaamalla sitä tiiviisti absoluuttisten sanemäärien taulukkoon (Taulukko 1). Taulukkoja vertaamalla voidaan kuitenkin tehdä joitakin havaintoja suomi toisena kielenä -oppijoiden sanaston rikkaudesta. Esimerkiksi yki-aineiston epämuodollisen tekstilajin sanemäärät laskevat B1 tasolta aina C2 tasolle, mutta TTR-arvot kasvavat tästä huolimatta kohti korkeampia arvoja. Näin ollen sanasto selkeästi rikastuu aikuisten epämuodollisessa tekstilajissa korkeammille taitotasoille siirryttäessä. Hyvin samanlainen ilmiö on nähtävissä myös aikuisten mielipiteen tekstilajissa: TTR-arvot kasvavat B1 tasolta lähtien vaikka sanemäärät laskevat. Yllättäen alemmilla tasoilla tapahtuu päinvastainen ilmiö. TTR-arvo ei kasva mielipiteen tekstilajissa A1-tasolta A2-tasolle siirryttäessä, vaikka sanemäärä pysyy lähes samana. Vielä jyrkemmin näillä tasoilla käy epämuodolliselle tekstilajille: TTR-arvo laskee vaikka sanemäärät laskevat.

Tulosten perusteella voisi tehdä varovaisia päätelmiä siitä, että alemmilla tasoilla (A1 ja A2) aikuisten sanaston rikkaus ei vielä kasva, vaikka muu kielitaito kehittyisikin. Vasta ylemmil-

lä tasoilla (B1-tasosta ylöspäin) sanasto alkaa rikastua. Johtopäätöstä ei kuitenkaan voi yleistää koskemaan myös koululaisten sanastoa, sillä koululaisten TTR-arvot eivät ole vertailukelpoisia eri tasoilla epätasaisten sanemäärien vuoksi.

5.4 Kerran esiintyvät lekseemit

Yksi keino tutkia sanaston rikkautta ja vivahteikkuutta on laskea aineiston **hapaks legomenon -sanat** eli **HL-sanat**. Hapaks legomenon -sanoiksi kutsutaan lekseemeitä, jotka esiintyvät otoksessa vain kerran. Mitä rikkaampaa ja vaihtelevampaa otoksen sanasto ja tyyli ovat, sitä enemmän HL-sanoja yleensä esiintyy. HL-sanoja kutsutaankin luonnehtimis- eli karakterisoimissanoiksi. (Mäkinen 1997: 50.)

Sanaston tutkimuksessa HL-sanoja pidetään kiinnostavina, koska ne ovat melkein poikkeuksetta sisältösanoja eli substantiiveja, adjektiiveja tai verbejä. Sisältösanat vievät tekstiä eteenpäin ja esimerkiksi puheesta mitattuna kerran esiintyvien sisältösanojen ja muiden pienifrekvenssisten sanojen on jopa väitetty kuvaavan eri puhujien välisiä maailmankuvien ja arvomaailmojen eroja (Mustonen 1997: 54). Tällaiset tutkimukset vaativat kuitenkin aina kvalitatiivista analyysiä sanojen laskemisen ja tunnistamisen lisäksi. Vaikka HL-sanojen määrää pidetään yhtenä sanastollisen diversiteetin ja vivahteikkuuden mittarina, luettavuuden kannalta niiden runsas esiintyminen esimerkiksi oppikirjateksteissä on tulkittu myös haittaavaksi tekijäksi (Jaakola 2004: 55).

Sanaston kvantitatiivisessa tutkimuksessa lasketaan usein kerran esiintyvien lekseemien prosenttiosuus tekstin kokonaissanamäärästä eli kaikista saneista (Särkkä 1987: 132). HL-sanojen osuus voidaan laskea myös lekseemien määrästä. Molemmissa laskutavoissa HL-sanojen osuus on kuitenkin voimakkaasti riippuvainen otoskoosta ja tämä on pidettävä mielessä myös tämän tutkimuksen tuloksia analysoitaessa. Niemikorven mukaan pienissä alle 2000 saneen otoksissa HL-sanojen osuus voi olla lähes 80 % esiintymistä, mutta otoksen laajetessa osuus saattaa olla vain murto-osa tästä. (Niemikorpi 1991: 78-79.)

Kerran esiintyvien lekseemien prosentuaalisen osuuden voi suomi toisena kielenä -oppijoiden teksteissä olettaa jäävän melko pieneksi, koska tulokset on laskettu kokonaisesta samaa taitotasoa edustavien tekstien ryhmästä. Tällöin toisteisuus on paljon suurempaa, sillä yhdessä tekstissä vain kerran esiintyvä lekseemi voi esiintyä vain kerran myös toisessa tekstissä, jolloin käyttä-

mäni menetelmä ei laske kyseistä lekseemiä kuuluvaksi HL-sanoihin. Seuraavassa taulukossa on esitetty suomi toisena kielenä -oppijoiden HL-sanojen määrät ja osuudet taitotasoittain.

Taulukko 5: HL-sanojen määrät ja osuudet taitotasoittain

Taitotaso	**HL-sanojen määrä**	**HL-sanojen osuus saneista / %**	**HL-sanojen osuus lekseemeista / %**
A1	585	7,05	50,17
A2	710	4,83	46,80
B1	1031	4,72	46,34
B2	971	8,67	52,09
C1	1213	12,62	54,81
C2	1335	12,64	53,40
Kaikki	*2749*	*3,61*	*48,37*

Suomi toisena kielenä -oppijoiden teksteissä kerran esiintyvien lekseemien prosentuaaliset osuudet jäävät odotetusti varsin mataliksi kaikilla taitotasoilla. HL-sanojen osuudet vaihtelevat välillä 4,72-12,64 % kaikista tekstikimpun saneista ja 46,34-54,81 % kaikista lekseemeistä.

Vaikka kyseessä on tietyn taitotason tekstikimpun kerran esiintyvien lekseemien tulokset, HL-sanojen osuus aineistossa on silti pieni. Aikaisempien tutkimusten mukaan esimerkiksi suomen kielen oppikirjoissa HL-sanojen osuudet kaikista saneista ovat 33-43 % (Nissinen-Taivainen 1996: 80) ja suomi toisena kielenä -opettajan opetuspuheessa kerran esiintyviä lekseemejä on keskimäärin noin 33,7 % (Mäkinen 1997: 51). Kaunokirjallisessa tekstissä HL-sanojen osuudet ovat tätäkin suurempia: esimerkiksi Ilmari Kiannon teksteissä HL-sanoja on lähes 72 % ja Väinö Linnalla yli 65 % (Särkkä 1987: 133). Toisen ja kolmannen luokan koulun oppikirjateksteissä kerran esiintyvien HL-sanojen osuudet ovat myös suhteellisen korkeat, 5,8 – 33,0 % saneista ja 40-70 % kaikista lekseemeistä (Jaakola 2004: 56). Jaakolan mukaan erityisen monimuotoista sanasto on HL-sanoilla mitattuna äidinkielen oppikirjoissa.

Mikäli sanaston monimuotoisuus kasvaa taitotasojen myötä, myös HL-sanojen osuuksien tulisi olla sitä suurempia, mitä korkeammasta taitotasosta on kysymys. Taulukon mukaan HL-sanojen osuudet saneista (kolmas sarake) ja osuudet lekseemeistä (neljäs sarake) osoittavat kyllä kasvavaa suuntaa, mutta osuudet eivät kasva tasaisesti. Molempien sarakkeiden epätasainen kasvu

johtuu otoskoon vaihtelun aiheuttamasta vinoumasta: otoskoko vaikuttaa selvästi kerran esiintyvien lekseemien osuuteen. Esimerkiksi B1-tasolla, jolla saneiden määrä on suurin, HL-saneiden osuudet jäävät pienimmiksi. HL-sanojen osuuksilla on taipumus jäädä sitä pienemmiksi mitä suuremmasta otoskoosta on kysymys.

Otoskoon vaihtelun johtuvan vinouman lisäksi tuloksiin vaikuttaa ratkaisevasti se, ettei osuuksia ole laskettu yksittäisistä teksteistä vaan tietylle taitotasolle arvioitujen tekstien ryhmästä. Tuloksia tarkasteltaessa tulee huomioida myös tehtävänantojen vaikutus tekstien vaihtelevuuteen. Tutkimuksessani ei ole eritelty kerran esiintyvien sanojen määrää tehtävätyypeittäin, mutta tehtävänantojen vaihtelevuus näkyy varmasti tuloksissa. Esimerkiksi mielipidetekstin tekstilajin tehtävänannossa B ja C -taitotasoilla oli annettu huomattavasti enemmän vaihtoehtoja kirjoitettavan tekstin aiheeksi kuin A-taitotasolla.

Vertaamalla HL-sanojen taulukkoa absoluuttisten sanemäärien taulukkoon, voidaan kuitenkin tehdä karkea päätelmä siitä, että HL-sanojen osuudet kasvavat suomi toisena kielenä -oppijoiden teksteissä taitotason myötä. Esimerkiksi A1 tason ja C1 tason kaikkien tekstien yhteen lasketut sanemäärät ovat lähellä toisiaan, mutta C1 tasolla HL-sanojen osuudet ovat huomattavasti suuremmat. Lyhyt vilkaisu näiden tasojen HL-sanojen listaan antaa myös vaikutelman, että C1 tason HL-sanat ovat harvinaisempia ja spesifimipiä kuin A1 tasolla. Ensimmäiset HL-sanat C1 tasolla ovat aakkosjärjestyksessä *aamujuna, Adzharian, ahdistella, aikakausi, aikamoinen, aikuisikä, aikuiskasvatustiede, aikuiskoulutus, ainainen, ainakaan, aineopinto, aivot, ajanvaraus, ajatusmaailma* ja *ajautua.* Ensimmäiset HL-sanat A1 tasolla ovat puolestaan *aamu, adjektiivi, ahkera, ai, aikuiskoulutus, aivot, aktiivinen, ala-aste, Ala-solkela, aloittaa, Amerikka, ammatti, ammattikoulu, ananas* ja *Anne.* En ole kuitenkaan tehnyt eri tasojen HL-sanoista tarkempaa frekvenssianalyysiä.

5.5 Yleisimmät lekseemit

Kerran esiintyvien lekseemien vastakohtana ovat teksteissä yleisimmin esiintyvät lekseemit, joita eri tutkimuksissa on kutsuttu kärkisanoiksi, tyyppisanoiksi tai avainsanoiksi. Suuritaajuiset lekseemit ovat usein merkitykseltään yleisiä ja niiden informaatioarvo on vähäinen (Vehmaskoski 1976: 141). Semanttisesti laajamerkityksiset lekseemit toimivatkin usein kieliopillisissa tehtävissä ja niiden sisällöllinen rooli tekstin eteenpäin viemisessä ei ole kovin suuri. Vehmaskosken mukaan tällaisia, ei tarkkarajaisia lekseemejä ovat muun muassa hänen omassa aineistossaan yleisimmät leksee-

mit *hyvä*, *pieni*, *vanha*; *lapsi*, *poika*, *asia*, *aika*; *tulla*, *mennä*, *saada*, *sanoa*, *pitää*, *voida*, *antaa*; *ja*, *ei* ja *minä*.

Yleisimmät eli suurifrekvenssiset lekseemit ovat kiinnostaneet kielenoppimisen ja -opettamisen tutkijoita, koska kielen yleisimmät lekseemit ovat usein kaikkein helpoimpia ja tärkeimpiä kielenoppijalle (esim. McCarthy 1990: 66, Cook 1991: 40-41). Myös Suomen kielen taajuussanaston (1979: 8) mukaan lekseemin yleisyys on suoraan verrannollinen sen tärkeyteen viestinnässä ja ihmismielessä: sanaston yleisimpien käsitteiden on jopa nähty hallitsevan kielenkäyttäjän ajattelua. Kielenoppimisen tutkimuksessa onkin toisinaan päädytty suosittelemaan sanaston opettamisen aloittamista kielen suuritaajuisimmista sanoista tai ainakin opetuksen päähuomion kohdistamista niihin (Sinclair & Renouf 1988: 148). Toisaalta nykytutkimusten valossa suuritaajuisuus ei voi olla ainut peruste sille, mitä sanoja on tärkeintä opettaa kielenoppimisen alkuvaiheessa (esim. Carter & McCarthy 1988: 9, Cook 1991: 41, Little 1994: 116, Nuutinen 1994: 33), sillä sanan oppimiseen vaikuttavat monet muutkin tekijät ja käytännön seikat (ks. luku 3.2). Lisäksi funktionaalisen kielikäsityksen ja kielenopettamisen mukaan sanoja tulisi kaiken kaikkiaan opettaa suhteessa toisiinsa eikä irrallisina yksikköinä (Nuutinen 1994: 33).

Suomen kielen viisi yleisintä lekseemiä ovat taajuussanaston (1979) mukaan taajuusjärjestyksessä *olla, ja, se, ei* ja *joka.* Suomen sanomalehtikielen viisi yleisintä lekseemiä ovat lähes samat: *olla, ja, ei, se* ja *että*. Myös puhekielessä nämä lekseemit ovat yleisimpien joukossa (Niemikorpi 1990: 145, Mäkinen 1997: 55). Suomen kielen kaikkein yleisimmät lekseemit tuntuvatkin olevan samat aineistosta riippumatta. Taulukossa 6 on esitetty oman aineistoni eri taitotasojen kolmekymmentä yleisintä lekseemiä.

Taulukko 6: Yleisimmät lekseemit ja niiden osuus otoksen kaikista saneista.

Kaikki		**A1**		**A2**		**B1**		**B2**		**C1**		**C2**	
olla	7,09	olla	8,49	olla	7,99	olla	6,90	olla	6,64	olla	6,20	olla	6,45
ja	3,74	minä	5,71	minä	4,58	ja	3,98	ja	3,84	ja	3,75	ja	3,67
minä	3,08	ja	3,36	ja	3,54	minä	3,36	minä	2,31	se	1,84	ei	1,48
ei	2,18	ei	2,77	ei	2,76	ei	2,38	se	2,25	että	1,48	että	1,41
se	2,17	se	2,48	se	2,47	se	2,34	ei	1,96	ei	1,33	se	1,35
että	1,40	koska	1,63	voida	1,37	että	1,64	että	1,54	minä	1,00	minä	1,06
voida	1,15	siinä	1,58	koska	1,33	voida	1,34	voida	1,21	voida	0,82	joka	0,78
tulla	0,87	hyvä	1,42	että	1,29	mutta	1,03	Maija	0,96	joka	0,75	voida	0,78
Solki	0,80	haluta	1,08	tulla	1,19	Solki	1,03	Solki	0,95	te	0,70	Solki	0,67

mutta	0,80	peli	1,05	mennä	1,10	tulla	0,93	mutta	0,92	tämä	0,67	saada	0,66
hyvä	0,79	paljon	0,99	koulu	1,02	koulu	0,93	tulla	0,86	hyvä	0,60	tämä	0,66
koska	0,79	voida	0,99	me	1,00	Maija	0,89	lapsi	0,79	pitää	0,59	hyvä	0,63
sinä	0,75	tulla	0,98	sinä	0,96	sinä	0,86	saada	0,79	Solki	0,57	myös	0,56
Maija	0,75	Matti	0,95	haluta	0,87	mitä	0,80	koulu	0,74	tulla	0,57	työ	0,54
me	0,68	mitä	0,87	hyvä	0,86	pitää	0,80	kun	0,71	ihminen	0,56	Maija	0,50
haluta	0,68	mennä	0,77	Matti	0,85	koska	0,79	joka	0,71	me	0,56	mutta	0,49
mennä	0,66	kun	0,76	Maija	0,80	lapsi	0,79	sinä	0,71	juna	0,55	tulla	0,46
koulu	0,65	että	0,73	kun	0,80	mennä	0,76	te	0,64	kaikki	0,54	jos	0,45
kun	0,65	mutta	0,73	mutta	0,80	kun	0,75	hyvä	0,63	mutta	0,54	kaikki	0,45
joka	0,64	me	0,70	paljon	0,78	hyvä	0,74	haluta	0,62	asia	0,53	ihminen	0,44
matti	0,63	hei	0,69	peli	0,78	matti	0,74	kaikki	0,62	saada	0,52	juna	0,44
jos	0,61	Suomi	0,67	känny	0,76	peli	0,72	Matti	0,62	jos	0,51	yhdistys	0,43
saada	0,61	kaikki	0,64	Solki	0,75	saada	0,71	jos	0,62	ottaa	0,50	me	0,42
pitää	0,61	jos	0,63	opettaja	0,70	jos	0,71	mennä	0,61	Maija	0,49	haluta	0,41
mitä	0,56	sitten	0,61	jos	0,64	haluta	0,70	me	0,60	myös	0,48	pitää	0,41
paljon	0,54	kännykkä	0,57	pitää	0,58	me	0,70	pitää	0,59	työ	0,46	te	0,38
peli	0,52	Maija	0,57	hei	0,56	joka	0,65	auto	0,58	aika	0,45	vuosi	0,38
kaikki	0,51	koulu	0,53	joka	0,54	te	0,62	tämä	0,54	auto	0,43	asia	0,35
te	0,51	Solki	0,53	kanssa	0,54	tai	0,58	hän	0,54	oma	0,42	Solkila	0,33
lapsi	0,48	opettaja	0,52	niin	0,54	auto	0,57	mitä	0,54	niin	0,40	klo	0,32

Suomi toisena kielenä -oppijoiden teksteissä kaikkein yleisimmät lekseemit toistuvat kaikilla taitotasoilla melko samoina. Viisi yleisintä lekseemiä kaikilla tasoilla ovat lähes samat: *olla, ja, minä, ei* ja *se.* Vain taitotasoilla C1 ja C2 lekseemi *minä* tulee vasta kuudentena ja viiden yleisimmän joukossa on puolestaan lekseemi *että*. Viiden yleisimmän mainitun lekseemin lisäksi kahdenkymmenen yleisimmän lekseemin joukossa kaikilla tasoilla ovat lekseemit *hyvä, voida, tulla, että* ja *mutta* ja kolmenkymmenen joukossa yleisimpiä yhteisiä lekseemejä ovat myös *me, jos, Maija* ja *Solki*. Näiden lekseemien lisäksi on paljon sanoja, jotka toistuvat lähes jokaisella taitotasolla kolmenkymmenen yleisimmän lekseemin joukossa.

Yleisimmät, kaikilla tasoilla kolmenkymmenen joukkoon kuuluvat verbit ovat *olla, ei, voida* ja *tulla*, joiden lisäksi lähes kaikilla tasoilla esiintyy myös verbit *haluta, mennä, saada* ja *pitää*. Olla-verbin yleisyyteen vaikuttaa Taajuussanaston (1979: 9) mukaan muun muassa sen käyttö aikamuotojen apuverbinä. Olla-verbillä onkin ensimmäinen sija lähes kaikilla suomen kielen frekvenssilistoilla (Taajuussanasto 1979, Suomen sanomalehtikielen taajuussanasto 2004, Mäkinen

1997: 55). Ainoan poikkeuksen tekee Niemikorven (1990: 145) frekvenssianalyysi, jonka frekvenssilista perustuu eri puolelta Suomea saatuihin puhuttuihin murrenäytteisiin. Tällä listalla *olla*-verbi jää vasta toiseksi, kun pronomini *se* esiintyy *olla*-verbiäkin useammin.

Yleisimpien lekseemien taulukko osoittaa, että eri taitotasojen kärkisanat eivät vaihtele paljon. Syitä vaihteluun voi löytyä tehtävänannoista ja siitä, että taitojen karttuessa esimerkiksi tietylle verbille opitaan uusia merkityksiä ja käyttötapoja, jolloin se alkaa esiintyä oppilaan kielessä entistä useammin. Yleisesti ottaen kaikki kärkisanat ovat kuitenkin suhteellisen neutraaleja, epätarkkoja ja jokapäiväisiä sanoja, eikä joukossa ole suuria yllätyksiä verrattuna muihin taajuussanastoihin. Kärkisanat ovatkin usein aineistosta riippumattomia ja monimerkityksisiä yleisiä lekseemejä, jotka eivät vaihtele paljon puheenaiheen mukaan (Mäkinen 1997: 56). Näin näyttäisi olevan myös aineistoni teksteissä.

Suomi toisena kielenä -oppijoiden sanaston 30 yleisimmän lekseemin listasta tasan puolet löytyy myös Suomen sanomalehtikielen 30 yleisimmän sanan joukosta. Tällaisia sanoja ovat verbit *olla, ei, saada, voida, tulla* ja *pitää*; pronominit *se* ja *kaikki*; konjunktiot *ja*, *tai*, *että, mutta* ja *kun* sekä substantiivi *Suomi* ja adjektiivi *hyvä*. Kaikki näistä yhteisistä frekventeistä lekseemeistä kuuluvat sanastoon, jotka ovat hyvin laajamerkityksisinä ja sikäli myös hyvin monikäyttöisiä. Lopuissa suomi toisena kielenä -oppijoiden 30 yleisimmän lekseemin joukossa näkyy muun muassa tehtävänannon vaikutus. Yleisimpien sanojen taulukossa esiintyvät esimerkiksi sanat *peli, kännykkä, opettaja* ja *koulu,* jotka ovat oleellisessa osassa myös tehtävänantojen sanastoa tarkasteltaessa. Erisnimien *Maija* ja *Solki* esiintymiselle jokaisella taitotasolla 30 yleisimmän sanan joukossa löytyy hyvin tekninen syy: näitä nimiä käytettiin peiteniminä teksteissä esiintyneille todellisille nimille, jotta tutkittavien henkilöiden yksityisyyden suoja toteutuisi mahdollisimman hyvin. Aineistot anonymisoitiin eli todelliset nimet vaihdettiin peitenimiin oppilaiden tekstejä sähköiseen muotoon siirrettäessä. Samasta syystä johtuu myös nimien *Matti* ja *Suomi* yleisyys yleisimpien sanojen listalla.

Aineiston viisi yleisintä lekseemiä ovat hyvin yhteneviä taajuussanaston ja Suomen sanomalehtikielen viiden yleisimmän lekseemin kanssa. Poikkeuksen tekee kuitenkin *minä*-pronominini, joka taajuussanastoissa ei ole yhtä yleinen kuin suomi toisena kielenä -oppijoiden teksteissä. Taajuussanastossa (1979) lekseemi *minä* tulee sijalla 27. ja sanomalehtikielessä vasta sijalla 9586. *Minä*-pronominin yleisyys paljastaakin kirjoitustehtävien ja sanomalehden tekstilajien yhden keskeisimmistä eroista: monissa kirjoitustehtävissä yksikön ensimmäisen persoonan käyttö on ehdoton valinta, kun taas sanomalehtikielessä se on selkeästi harvinaisempi. Kaiken kaikkiaan

taajuussanastoa ja sanomalehtikielen taajuussanastoa varten kerätty aineisto on ollut tiukemmin asiapitoista ja vähemmän persoonaan liittyvää.

Minä-pronomini tuntuu olevan selkeästi yleisempi myös puhekielessä verrattuna kirjoitettuun kieleen. Suomen kielen murteiden frekvenssilistalla *minä* on kymmenen yleisimmän lekseemin joukossa (Niemikorpi 1990: 145) ja alkeiskurssin opettajanpuheessa 20 yleisimmän joukossa (Mäkinen 1997: 55). Lähtökohtaisesti kieli onkin usein hyvin minä-keskeistä (Nuutinen 1994: 34), ja siksi esimerkiksi kielen oppikirjojen olisi hyvä sisältää paljon *minä*-muodossa kerrottuja tekstejä ja *minän* suhtautumista ulkomaailmaan olisi jatkuvasti pidettävä esillä (Mäkinen 1997: 28). Suomen kielen oppikirjoissa *minä* onkin tutkittu olevan kymmenen yleisimmän lekseemin joukossa (Nissinen-Taivainen 1996: 47). Myös funktionaalisen kielikäsityksen mukaan puhekieli on kielen ensisijainen muoto, joten puhekielen mukailemista kirjoituksissa voidaan pitää hyvin luontevana vaiheena kielen oppimisen alussa. Aineistoni kaikista *minä*-lekseemeistä osa on todellisuudessa ollut kirjoittajien teksteissä alun perin puhekielisessä muodossa *mä*. Lemmauksen yhteydessä kaikki *mä*-pronominit kuitenkin vaihdettiin ja korjattiin yleiskieliseen muotoon *minä*.

Mielenkiintoisin huomio suomi toisena kielenä -oppijoiden yleisimpien lekseemien listaa tarkasteltaessa liittyy yleisimpien lekseemien kattavuuteen eri taitotasoilla. Mitä ylempää taitotasoa tekstit edustavat, sitä pienemmäksi yleisimpien lekseemien kattavuus koko aineistosta laskee. Tämä ei johdu pelkästään sanemäärien kasvusta, sillä taitotasojen B1 ja C1 välillä sanemäärät itse asiassa laskevat, mutta yleisimpien lekseemien kattavuus laskee silti. Korkeimmilla taitotasoilla yleisimmätkään lekseemit eivät kata enää prosentuaalisesti yhtä suurta osaa kaikista saneista kuin alemmilla tasoilla. Tämä kertoo osaltaan siitä, että kielitaidon kehittyessä suomi toisena kielenä -oppijoiden sanasto monipuolistuu ja teksteissä ilmenee yhä enemmän eri sanoja.

5.6 Kärkisanojen sanaluokkajakauma

Taulukko 7: Oppilaiden sanaston kolmenkymmenen yleisimmän lekseemin sanaluokkajakauma (lyhenteet on sisällytetty mukaan varsinaisiin sanaluokkiin).

Sanaluokka	**Kaikki**	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
Verbit	8	6	7	8	8	7	7
Pronominit	7	5	5	6	9	7	7
Substantiivit	6	8	7	7	6	8	10
(joista erisnimiä)	3	4	3	3	3	2	3
Konjunktiot	6	6	6	7	5	5	5
Adverbit	2	3	3	1	1	1	0
Adjektiivit	1	1	1	1	1	2	1
Interjektiot	-	1	1	-	-	-	-
Numeraalit	-	-	-	-	-	-	-
Yhteensä	30	30	30	30	30	30	30

Suomi toisena kielenä -oppijoiden sanaston 30 yleisimmästä lekseemistä (kattavuus kaikilla taitotasoilla 35,9 %) eniten on verbejä (8 kpl) ja pronomineja (7 kpl). Substantiiveja ja konjunktioita on molempia kuusi, adverbeja kaksi ja adjektiiveja yksi kappale.

Puolet listalta löytyvistä substantiiveista on erisnimiä, joten varsinaisten substantiivien osuus jää hyvin pieneksi siihen nähden, että koko aineistossa substantiiveja (erisnimet mukaan luettuna) on 31 %. Syynä suhteellisen alhaisille frekvensseille lienee lähinnä substantiiveille ominainen merkityksen spesifisyys. Sisältösanoina substantiivit ovat heterogeenisempia kuin esimerkiksi funktiosanoihin kuuluvat pronominit ja konjunktiot, joiden kattavuus suuritaajuisista sanoista on substantiiveja suurempi. Pronominit ovat kärkisanojen toiseksi yleisin sanaluokka, vaikka kaikista saneista niiden osuus jääkin paljon pienemmäksi. Vaikka muutamaa pronominia toistetaankin erittäin paljon, muiden pronominien määrä on kuitenkin niin pieni, etteivät ne pysty kilpailemaan esimerkiksi substantiivien suuren sanaluokan kanssa. Pronominit ovat niin sanotusti suljettu sanaluokka, kun taas substantiiveja avoimena sanaluokkana voi kielenkäyttäjä jopa tarvittaessa keksiä itse.

Yleisimpien lekseemien sanaluokkajakaumassa ei ole havaittavissa suuria eroja eri taitotasojen välillä. Alemmilla taitotasoilla (A1 ja A2) 30 yleisintä sanaa sisältävät kuitenkin enemmän adverbejä ja konjunktioita ylempiin tasoihin nähden. Ylemmillä tasoilla yleisimpien sanojen joukkoon kuuluu puolestaan enemmän substantiiveja ja pronomineja.

5.7 Kumuloituva frekvenssi

Tarkasteltaessa aineiston yleisimpien lekseemien kattavuutta voidaan puhua myös kumuloituvasta frekvenssistä (myös summa- tai kertymäfrekvenssi). Kumuloituva frekvenssi antaa kokonaiskuvan siitä, kuinka suuren osan suuritaajuisimmat lekseemit muodostavat koko aineistosta (Jaakola 2004: 61). Esimerkiksi Suomen kielen taajuussanaston (1979) mukaan suomen kielen sata yleisintä lekseemiä kattaa noin 35,1 % juoksevasta kirjoitetun kielen tekstistä. Edelleen 500 yleisimmän lekseemin osuus on 55,0 %, 1000 yleisimmän lekseemin osuus 65 %, 2200 yleisimmän lekseemin osuus 75 % ja 3500 yleisintä lekseemiä kattaa 80 % tekstistä. Puhutussa kielessä yleisimpien lekseemien kattavuus on vielä suurempi. Esimerkiksi 1000 yleisimmän lekseemin on tutkittu kattavan puheesta jopa 87 % (Jussila 1996: 26-27).

Kumuloituva frekvenssi on kaikista todellisista tekstikorpuksista laskettuna epätasainen ja painottuu alkupään lekseemeihin. Niemikorven Oulun korpuksesta esittämien tulosten mukaan aineistossa vähintään 21 kertaa esiintyvien lekseemien prosentuaalinen osuus kaikista lekseemeistä on vain 5 % mutta kaikista tekstisaneista jopa 75 % (Niemikorpi 1990). Kumuloituvan frekvenssin laskemista voidaan käyttää sanaston koostumusten mittaamisen lisäksi apuna myös tekstin tekstuaalisen rakenteen selvittämiseen. Kerran esiintyvien lekseemien ohella kumuloituva frekvenssi kertoo esimerkiksi sidossanojen määrän, joka voidaan ottaa huomioon otoksen tekstuaalista rakennetta määriteltäessä (Jaakola 1994: 65).

Suomenkielisistä aineistoista laskettuja kumuloituvan frekvenssin arvoja on vertailtu myös muihin kieliin. Vertailussa on kuitenkin oltava varovainen, sillä eri kielien rakenteelliset erot tuovat tuloksiin epävarmuutta. Lisäksi eri kielissä taajuussanastoja on koottu hieman eri periaattein. Kuitenkin ainakin kattavuusasteikon keskivaiheen yleisimpien lekseemien osuudet ovat melko samansuuntaisia useissa eri kielissä. Kun suomenkielisessä tekstissä 500 yleisintä lekseemiä kattaa 55,0 % kaikista saneista, englanninkielisessä tekstissä vastaava luku on 61,9 %, ruotsinkielisessä 57,3 % ja unkarinkielessä 54,7 % (Niemikorpi 1991: 49). Sen sijaan kumuloituvan frekvenssin ääripäissä, esimerkiksi vertailtaessa sadan yleisimmän lekseemin osuuksia, eri kielten kesken on selkei-

tä eroja. Erot lekseemien taajuuksissa ja yleisimpien lekseemien osuuksissa johtuvat kielten rakenteellisista eroista: kieliopillisia suhteita ilmaistaan toisissa kielissä erillisillä lekseemeillä, kun taas toisissa kielissä sama tehtävä on erilaisilla päätteillä ja liitteillä. Niemikorven (1991: 50) mukaan esimerkiksi englannin ja ruotsin taajuusluetteloita pitäisi oikeastaan käsitellä sananmuotoluetteloina, jotta suhteellisia sanamääriä ja kumulatiivista frekvenssiä olisi mielekästä vertailla keskenään.

Lekseemien taajuus ja kumuloituva frekvenssi ovat lisäksi olleet kiinnostuksen kohteena tekstinymmärtämistä koskevissa sanastontutkimuksissa. Tutkijat ovat esittäneet eri arvioita siitä, kuinka monta prosenttia tekstin lekseemeistä tulisi osata, jotta tekstin voisi ymmärtää. Takalan (1989: 7) mukaan yleiskuvan syntymiseen luettavasta tekstistä vaaditaan 70-75 % tekstin lekseemien ymmärtäminen eli noin 1300 eri sanaa. Niemikorven (1991: 50-51) mukaan vasta 2000 lekseemin osaaminen riittää hyvin helpohkon kaunokirjallisen tekstin ymmärtämiseen. Toisaalta Takala esittää, että yksityiskohtaisempaan tekstin ymmärtämiseen vaaditaan osaamista jo n. 90-95 % tekstin lekseemeistä.

Suomi toisena kielenä -oppijoiden kumuloituvan frekvenssin arvot on esitetty seuraavassa taulukossa.

Taulukko 8: Kumuloituva frekvenssi

Yleisintä lekseemiä (kpl)	**Kattavuus taitotason saneista (%)**						
	Kaikki	A1	A2	B1	B2	C1	C2
5	18,26	22,81	21,34	18,96	17,00	14,6	14,36
10	23,28	29,57	27,62	24,93	22,58	18,54	18,31
20	30,32	38,04	36,36	33,04	29,78	24,15	23,49
30	35,9	44,00	42,75	39,74	35,64	28,81	27,36

Suomi toisena kielenä -oppijoiden kahdenkymmenen yleisimmin käytettyjen lekseemien frekvenssien prosentuaalinen osuus taitotason saneista on keskimäärin 30,32 %. Kaksikymmentä yleisintä lekseemiä kattaa siis noin 30 % kaikista otoksen teksteistä. Taajuussanastossa vastaava luku on 27,10 % ja Suomen kielen alkeiskurssin opettajan opetuspuheessa kaksikymmentä yleisintä lekseemiä kattaa tekstistä 43 % (Mäkinen 1997: 56). Toisen ja kolmannen luokan oppikirjoissa 20 yleisintä lekseemiä kattaa 21 % oppikirjateksteistä (Jaakola 2004: 64). Taulukon frekvenssitiedot on kuitenkin laskettu tiettyä taitotasoa edustavien tekstien ryhmästä, joten siinä esitettyjä kattavuuslukuja ei voida suoraan verrata muista aineistoista laskettuihin lukuihin.

5.8 Sane- ja lekseemi pituudet

Yhtenä sanastontutkimuksen kvantitatiivisena menetelmänä on pidetty saneiden keskipituuksien mittaamista. Suomi toisena kielenä -oppijoiden saneiden- ja lekseemien keskipituuksia on aikaisemmin laskenut muun muassa Grönholm (1993), jonka mukaan oppilaat pyrkivät välttämään pitkien lekseemien käyttöä varsinkin kielenoppimisen alkuvaiheessa. Muihin kieliin nähden Suomen kielen taivutusjärjestelmä monine suffikseineen tekeekin erityisesti substantiiveista melko pitkiä verrattuna muunkielisiin vastineisiin, mikä voi osaltaan hankaloittaa suomi toisena kielenä -oppivan sanaston omaksumista (Martin 1999:169). Saneiden- ja lekseemien keskipituuksia ovat laskeneet myös Niemikorpi Oulun korpuksesta (Niemikorpi 1991: 181-187), Leena Saarela peruskoululaisten kirjoitelmista (Saarela 1997: 108) ja Mari Jaakola oppikirjoista (Jaakola 2004: 73).

Saneiden keskipituudeksi saadut tulokset vaihtelevat eri aineistossa. Oulun korpuksen saneiden keskipituus on 7,42 grafeemia (Niemikorpi 1991: 86-89) kun taas oppilaiden kirjoituksissa grafeemipituudet ovat toisen luokan oppilailla 5,74, neljännen luokan oppilailla 6,08, kuudennella luokalla 6,04 ja kahdeksannella 6,43 (Saarela 1997: 108-109). Jaakolan tutkimissa oppikirjoissa keskimääräiset saneen pituudet vaihtelivat 6,06 ja 7,66 grafeemin välillä: pisimpiä tekstisaneet olivat luonnontiedon ja 3. luokan äidinkielen tekstikirjoissa, lyhyimpiä taas matematiikan kirjoissa sekä 2. luokan äidinkielen tekstikirjoissa ja 3. luokan harjoituskirjoissa (Jaakola 2004: 72-73).

Suomen kielen saneiden pituusjakauma on Niemikorven (1991: 89-91.) mukaan kolmihuippuinen. Korkein huippu on 6-grafeemisissa saneissa, joita on 12,28 % kaikista saneista. Seuraavaksi yleisimpiä ovat 2- ja 8-grafeemiset saneet. Kaksigrafeemisten saneiden suurta määrää selittävät suomen kielen monet suuritaajuiset yksitavuiset sanat kuten *ja, on, ei, se, jo* ja *ne*. Jaakolan mukaan oppikirjojen yleisin sanaesiintymä on 5-grafeeminen sane (Jaakola 2004: 74).

Saneiden keskipituus kasvaa yhtä aikaa lauseiden keskipituuden kanssa. Pitkissä lauseissa on keskimäärin enemmän substantiiveja kuin lyhyissä ja substantiivien määrän lisääntyessä myös saneiden keskipituudet kasvavat, sillä substantiiveissa on paljon pitkiä yhdyssanoja (Niemikorpi 1991: 181-187, 200-207). Saarelan (1997: 109) tutkimusten mukaan peruskoululaisten äidinkielen virkkeet pitenevät yläluokilla, jolloin virkkeisiin ilmestyy enemmän lauseenvastikkeita. Virkkeidem kehittyessä monimutkaisemmiksi tekstiin tulee infinitiivi- ja partisiippirakenteita, jotka nostavan saneiden keskipituutta (Saarela1997: 49).

Saneiden keskipituuden kasvuun vaikuttaa myös suffiksien kuten sijamuotojen ilmaantuminen oppijan kieleen ja sanastoon. Natiivin suomen kielen puhujan kieleen lähes kaikki

suffiksit ilmaantuvat jo ennen kouluikää (Toivainen 1980: 160-190). Lapsen kielen kehittymistä tutkimalla on todettu, että esimerkiksi kolmivuotias suomalaislapsi hallitsee ainakin 12 eri suffiksikategoriaa: partitiivin, monikon, lokaalisia adverbeja, paikallissijoja, välineen adessiivin, genetiivin (Toivainen 1980: 188-190). Yläluokilla lapset kykenevät käyttämään ja muistamaan pitempiä sanoja ja osaavat niiden taivutusmuodot.

Grafeemitiedoista ollaan kiinnostuneita myös luettavuustutkimuksessa, sillä grafeemi- tai tavulukuihin perustuvan tekstisanojen pituutta on pidetty yhtenä luettavuutta heikentävänä tekijänä (esim. Björnsson 1968, Karvonen 1970, Jaakola 2004). Vaikka yksittäiset pitemmät lekseemit tekstissä eivät yleensä aiheuta lukijalle ongelmia, runsasmerkkiset sanat kuitenkin hidastavat mekaanista lukemista. Varsinkin vieraiden pitkien lekseemien kohdalla lukija joutuu prosessoimaan sanan eri osat tulkintajaksoihin, jolloin sanahahmon tunnistaminen on työläämpää ja merkityksen ymmärtäminen vaatii enemmän aikaa. (Arajuuri 1980: 151-153.)

Taulukko 9: Keskimääräinen saneen pituus eri taitotasoilla

Taitotaso	**Saneet**	**Lekseemit**	**TTR (%)**	**M-kerroin**	**Saneen keskipituus**
A1	8 300	1 166	14,05	7,11	5,03
A2	14 701	1 517	10,32	9,69	5,02
B1	21 852	2 223	10,18	9,83	5,15
B2	11 201	1 864	16,64	6,01	5,41
C1	9 611	2 213	23,03	4,34	6,04
C2	10 558	2 500	23,69	4,22	6,24

Omassa aineistossani saneen keskipituus vaihtelee 5,03 ja 6,24 grafeemin välillä. Odotusten mukaisesti grafeemimäärät nousevat taitotasolta toiselle siirryttäessä, tosin tasojen A1 ja A2 välillä kasvua ei tapahdu. Vasta B2 tasolta lähtien suomi toisena kielenä -oppijoiden keskimääräiset sanepituudet alkavat olla lähellä natiivien suomenkielisten peruskoululaisten tasoa. Tasoilla C1 ja C2 opiskelijat saavuttavat suunnilleen samat grafeemipituudet kuin peruskoululaisten neljännen (6,08 grafeemia) ja kuudennen (6,04 grafeemia) luokan oppilaat. Kahdeksannella luokalla olevien oppilaiden saneen keskipituus on jo 6,43 grafeemia, johon C2-tason kirjoittajatkaan eivät vielä yllä. Vapaata yleispuhekieltä tutkittaessa on saatu saneen keskipituudeksi noin 5,98 grafeemia, johon yltävät vain C1- ja C2-tasoiset suoritukset.

Suomi toisena kielenä -oppijoiden sanaston keskimääräiset grafeemipituudet näyttävät siis asettuvan vastaamaan lähinnä yleispuhekielestä ja peruskoulun alakoululaisten teksteistä saatuja arvoja. Onkin varsin loogista, että oppivan suomen kieli jää vielä puhekielen tasolle grafeemipituuksissa. Kielen oppimisen alkuvaiheessa kirjoitettu kieli mukailee usein puhuttua kieltä, joten puheessa usein esiintyvät lyhyet sanat saattavat siirtyä myös kirjoitettuun kieleen. Samoin on odotuksenmukaista, että taitotasojen suoritukset jäävät kauaksi jälkeen Oulun korpuksen tuloksista. Pääosin tiedotusvälineiden asia- tai muista melko muodollisista teksteistä kootun Oulun korpuksen sanepituudet ovat peräisin tarkasta ja kompleksisesta kielestä: Sanomalehtikielessä ja muissa muodollisissa teksteissä pyritään yleensä eksakteihin ilmauksiin ja spesifiseen kieleen, jolloin myös keskimääräiset sanepituudet kasvavat.

Saneiden keskipituuksien lisäksi aineistosta olisi mahdollista laskea myös lekseemien keskipituudet. Saneiden ja lekseemien keskimääräiset pituudet poikkeavat yleensä toisistaan merkittävästi. Lekseemien keskipituudet ovat saneiden keskipituuksia huomattavasti suuremmat, sillä suuritaajuisimmat lekseemit ovat tyypillisesti hyvin lyhyitä ja vaikuttavat täten pienentävästi saneiden pituusjakauman arvoihin (esim. Jaakola 2004: 78). Eri sanaluokkien lekseemien keskipituudet vaihtelivat esimerkiksi Niemikorven tutkimuksissa, joissa järjestyksessä pisimmistä lyhyimpiin lekseemeihin sisältävät sanaluokat ovat numeraalit (15,1 g), substantiivit (10,9 g), adjektiivit (10,9 g), adverbit (8,4 g), verbit (8,3 g), partikkelit (6,7 g) ja pronominit (4,8 g) (Niemikorpi 1991: 148). Tässä tutkimuksessa lekseemien keskipituudet on kuitenkin jätetty huomiotta, sillä laskettujen saneiden keskipituuksien rinnalla lekseemitulokset tuskin toisivat lisää merkittävää tietoa suomi toisena kielenä -oppijoiden sanastosta.

6 SANASTON DIVERSITEETIN TARKASTELUA

6.1 Tunnuslukujen tulokset ja eri taitotasojen leksikaalinen diversiteetti

Lemmatusta aineistosta laskettiin taajuussanastoa apuna käyttäen luvussa 4.5 esitellyt sanaston rikkauden tunnusluvut eli Shannonin indeksi, harvinaisuustunnusluku (rarity), sisältösanojen harvinaisuustunnusluku (contentrarity), monipuolisuus-, tasapuolisuus- ja hajaannustunnusluku (MTLD, balance, dispersion). Tunnuslukujen tuloksia vertailemalla syntyy selkeitä eroja eritasoisten tekstien ja tehtävätyyppien välille. Käsittelen seuraavaksi tuloksia ensin pelkän taitotason kannalta ja otan sen jälkeen kantaa siihen, kuinka tehtävätyyppi vaikuttaa sanaston rikkauteen. Taulukossa 10 on esitetty eritasoisista teksteistä lasketut tunnusluvut yleisesti, ilman jaottelua tehtävätyyppeihin.

Taulukko 10: Eri taitotasojen leksikaalisen diversiteetin tunnusluvut

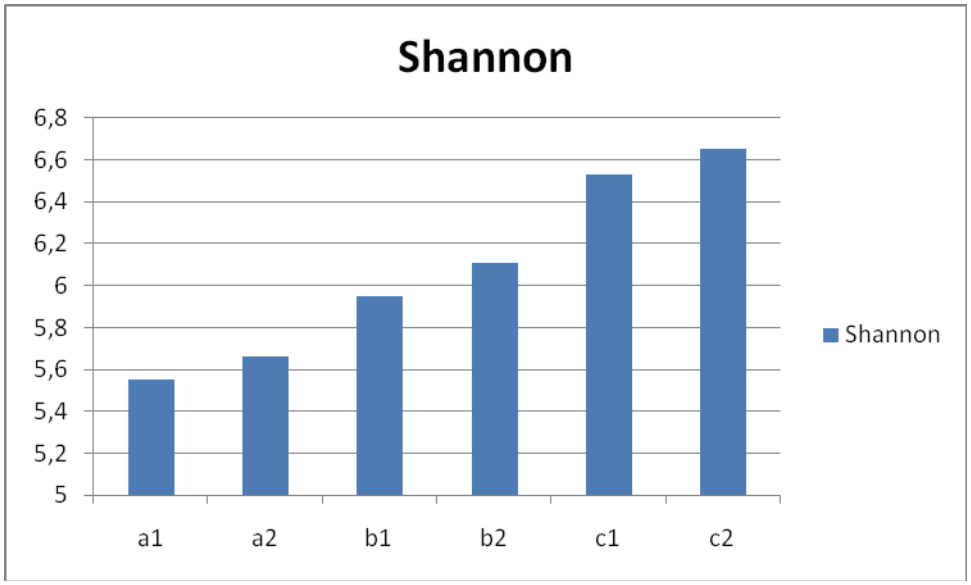
level	tokens	types	shannon	rarity	contentrarity	MTLD	balance	dispersion
a1	8252,00	1189,00	5,55	2064,04	2115,22	38,41	0,78	307,58
a2	14515,00	1547,00	5,66	2147,40	2230,08	41,82	0,77	421,84
b1	21536,00	2274,00	5,95	2421,01	2498,74	53,48	0,77	631,76
b2	11061,00	1886,00	6,11	2278,82	2400,93	73,19	0,81	493,70
c1	9584,00	2235,00	6,53	2343,36	2432,30	119,73	0,85	516,09
c2	10537,00	2528,00	6,65	2402,64	2524,54	125,97	0,85	585,55

Käyttämieni tunnuslukujen mukaan eri taitotasojen sanaston rikkaudessa on selkeitä eroja. Kaikista aineistosta lasketuista tunnusluvuista Shannonin indeksi ja MTLD näyttävät kiinnostavimmilta, koska ne vaikuttavat olevan täysin riippumattomia tekstimäärästä. Toisin sanoen tekstikimpussa esiintyvien saneiden määrä ei vaikuta Shannonin ja MTLD:n antamiin tuloksiin, jolloin tunnuslukuja voidaan pitää luotettavina. Sen sijaan muita, erityisesti harvinaisuustunnuslukuja ja hajaannustunnuslukua ei voida pitää erityisen luotettavina, sillä niiden antavat arvot vaihtelevat voimakkaasti otoskoon mukaan. Mitä enemmän taitotasolla on saneita, sitä korkeammat ovat kyseisten tunnuslukujen tulokset. Tasapuolisuustunnusluvun kohdalla vaikuttaa puolestaan olevan päinvastoin: mitä enemmän saneita sitä matalammat tulokset. Niinpä tasapuolisuustunnusluvunkaan luotettavuus ei ole Shannonin indeksin ja MTLD-tunnusluvun veroinen.

Yleisesti ottaen tunnuslukujen arvot kasvavat sitä enemmän, mitä korkeammasta taitotasosta on kysymys. Näin ollen tulokset vahvistavat tutkimushypoteesin sanaston rikastumisesta taitotason kohoamisen myötä. Mielenkiintoinen siirtymä tapahtuu erityisesti taitotasojen B2 ja C1

välillä, jolloin Shannonin ja MTLD-tunnuslukujen arvot kasvavat voimakkaasti. Tulosten mukaan suomi toisena kielenä -oppijoiden sanasto siis monipuolistuu ja rikastuu erityisesti oppijoiden päästessä kirjoituksissaan C-tasolle. Sen sijaan kehitys A-tasolta B-tasolle on tasaista Shannonin ja MTLD-tunnusluvun mukaan, vaikka harvinaisuustunnusluvut antavatkin harhaanjohtavan vaikutelman voimakkaasta kasvusta.

6.1.1 Shannonin indeksi

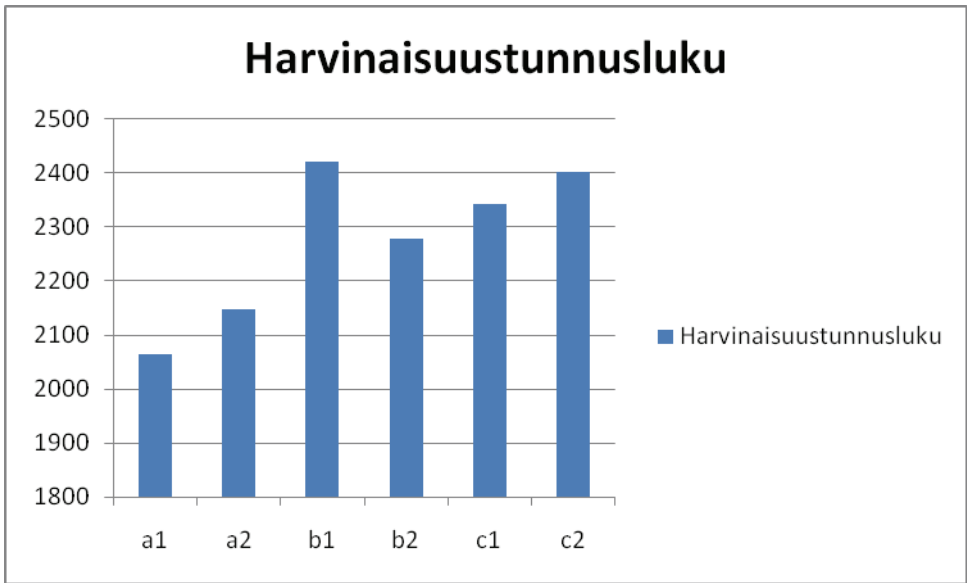


Shannon-tunnusluku kasvaa kauniisti taitotasojen välissä. Kasvu on tasaista, noin 0,22 yksikköä yhtä taitotasosiirtymää kohden. Shannonin antamat tulokset vaikuttavat olevan täysin riippumattomia tekstien pituudesta tai määrästä. Alimmillaan tunnusluku on 5,55 taitotasolla A1 ja korkeimmillaan 6,65 taitotasolla C2. Shannonin tunnusluku vahvistaa näin osaltaan tutkimushypoteesin, jonka mukaan sanaston rikkaus kasvaa taitotasolta toiselle siirryttäessä. Voimakkainta kasvu on B2 ja C1 välillä, jolloin tunnusluku kasvaa 0,42 yksikköä.

Shannonin indeksiä ei ole aikaisemmin tutkittu suomi toisena kielenä -oppijoiden teksteistä, mutta suomenkielisten peruskoululaisten kirjoitelmiin verrattuna arvot ovat epäilyttävän korkeat. Peruskoululaisten äidinkielen sanastoa tutkineen Saarelan (1997: 106) mukaan Shannonin indeksillä mitattuna toisluokkalaisten saama keskiarvo on 3,50, neljäsluokkalaisten 3,81, kuudesluokkalaisten 4,36 ja kahdeksasluokkalaisten 4,24. Niinpä saamieni arvojen mukaan edes natiivit kahdeksasluokkalaiset eivät kirjoittaisi yhtä rikkaasti kuin oman aineistoni suomi toisena kielenä -oppijat jo A1-tasolla. Tulokset eivät kuitenkaan ole keskenään vertailukelpoisia. Heikkoon vertailtavuuteen vaikuttaa käyttämäni menetelmä, jossa olen laskenut tunnuslukuja tiettyä taitotasoa edus-

tavista tekstikimpuista, enkä niinkään yksittäisistä teksteistä kuten Saarela omassa tutkimuksessaan. Siinä mielessä Saarelan ja oman tutkimukseni tulokset ovat kuitenkin linjassa, että molemmissa aineistoissa Shannon-tulokset kasvavat luokkatason tai taitotason mukaisesti.

6.1.2 Harvinaisuustunnusluku

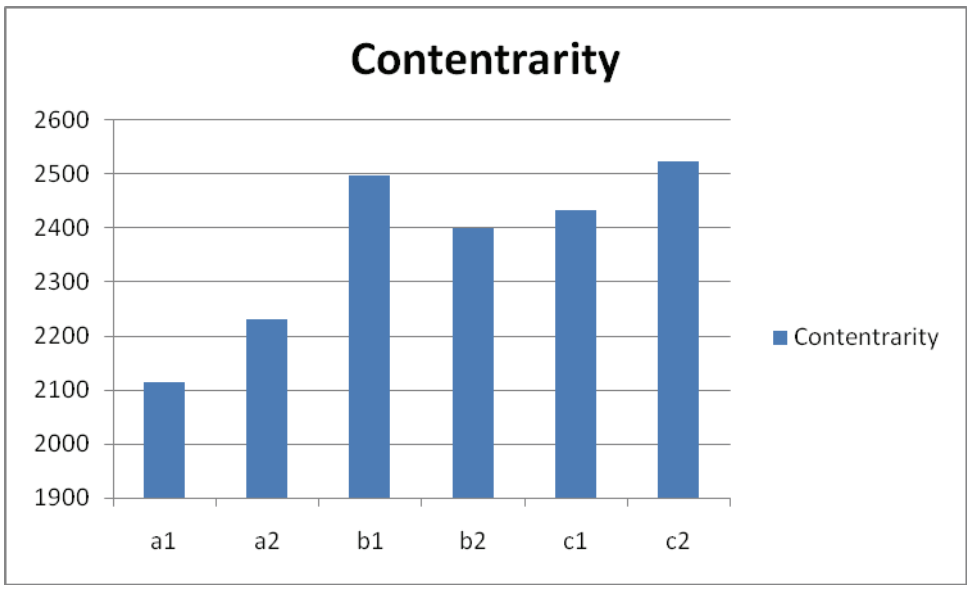


Lukuun ottamatta B1 taitotasoa harvinaisuustunnusluku kasvaa sitä suuremmaksi, mitä korkeammasta taitotasosta on kyse. Taitotasolla B1 harvinaisuustunnusluku tekee poikkeuksellisen piikin eli nousee kaikkia muita tasoja korkeammalle. Toisin sanoen taitotasojen B1 ja B2 välillä tapahtuu jyrkkä notkahdus, jolloin harvinaisuustunnusluku ei vastoin oletuksia kasvakaan, vaan laskee huomattavasti. Tulosten mukaan B2 taitotasolla käytettäisiin siis vähemmän harvinaisia sanoja kuin alemmalla B1 taitotasolla, mikä on vastoin tutkimushypoteesia. Harvinaisuustunnusluvun antamiin tuloksiin ei kuitenkaan voida luottaa, sillä otoskoko vääristää laskettuja arvoja merkittävästi: B1-tasoisia kirjoituksia on aineistossa kaikkein eniten ja ne sisältävät eniten saneita muihin tasoihin verrattuna.

Harvinaisuustunnusluku kasvoi tasolta toiselle siirryttäessä 59,28-273,61 yksikköä, keskimäärin 67,72 yksikköä. Alimmillaan harvinaisuustunnusluku on 2064,04 taitotasolla A1. Taitotason keskimääräisen sanan järjestysnumero taajuussanastossa on siis noin 2064, jota taajuussanastossa edustaa sana *lupaus*. Lähimmät tätä harvinaisuustunnuslukua edustavat sanat ovat lisäksi *lentokenttä, sanna, säveltäjä, keskuspankki, sektori, sotilaallinen, johanna, opiskelu*. Ylimmillään harvinaisuustunnusluku on 2421,01 ja kuten todettu, hieman yllättäen taitotasolla B1. Taajuussanas-

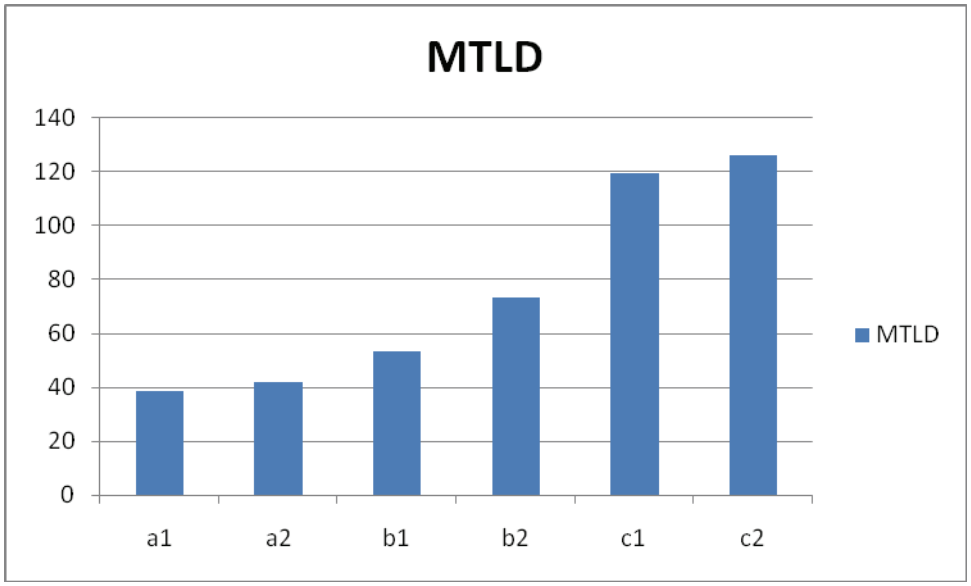
tossa kyseistä keskiarvoa edustaa sana *nuoriso* ja seuraavaksi lähimmät sanat ovat *huumori, kirjallinen, seuraavaksi, tosissaan, tähdentää, nälkäinen, sosiaalidemokraatti* ja *mänttä.* Vaikka suurin osa tekstissä esiintyvistä sanoista olisi hyvinkin yleisiä, muutamak hyvin harvinainen sana nostaa harvinaisuustunnuslukua tuntuvasti.

6.1.3 Sisältösanojen harvinaisuustunnusluku



Sisältösanojen harvinaisuustunnusluku kasvaa yleisesti sitä suuremmaksi, mitä korkeammasta taitotasosta on kyse. Samoin kuin harvinaisuustunnusluvun kohdalla, sisältösanojen harvinaisuustunnusluku nousee kuitenkin poikkeuksellisen korkealle taitotasolla B1. Niinpä arvoissa tapahtuu notkahdus seuraavalle tasolle siirryttäessä. Syynä piikkiin ja sitä seuraavaan notkahdukseen on jälleen kerran otoskoon aiheuttama vinouma. Harvinaisuustunnuslukuun verrattaessa sisältösanojen harvinaisuustunnusluku kasvaa kuitenkin tasaisemmin, keskimäärin noin 80,98 yksikköä tasolta toiselle siirryttäessä.

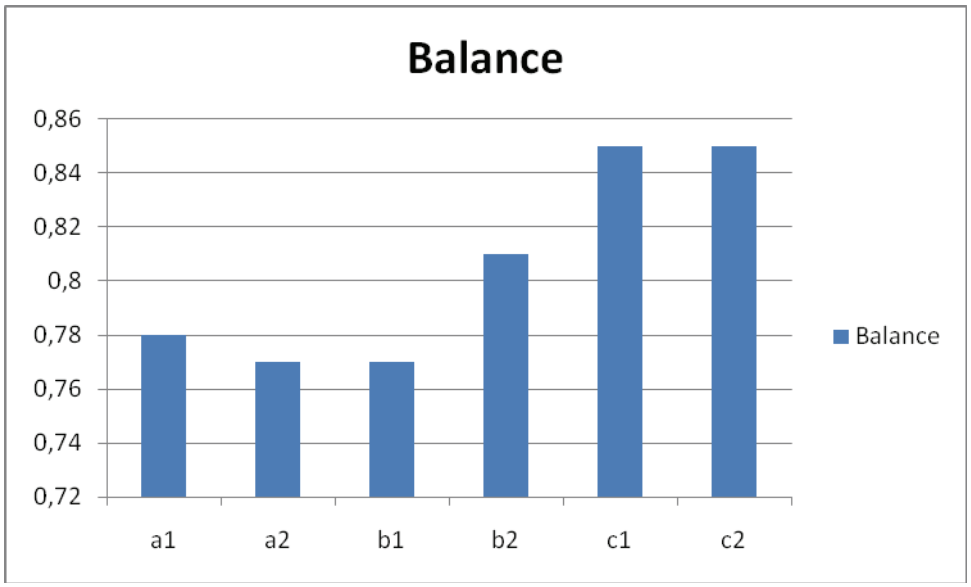
6.1.4 MTLD, sanojen monipuolisuustunnusluku



Sanojen monipuolisuustunnusluku kasvaa taitotasolta toiselle siirryttäessä ja vahvistaa näin tutkimushypoteesin. Kasvu on keskimäärin 17,33 yksikköä yhtä taitotasosiirtymää kohden. Alimmillaan tunnusluku on 38,41 taitotasolla A1 ja korkeimmillaan 125,97 taitotasolla C2. Huimaa kasvua tapahtuu erityisesti taitotasolta B2 taitotasolle C1 siirryttäessä, jolloin tunnusluvun saama arvo kasvaa 46,54 yksikköä. Myös taitotasojen B1 ja B2 välillä kasvu on voi voimakasta: 19,71 yksikköä. MTLD-tunnusluvun tulokset ovat tärkeitä sanaston rikkautta arvioitaessa, sillä Shannonin tavoin sen antamia arvoja voidaan pitää luotettavina.

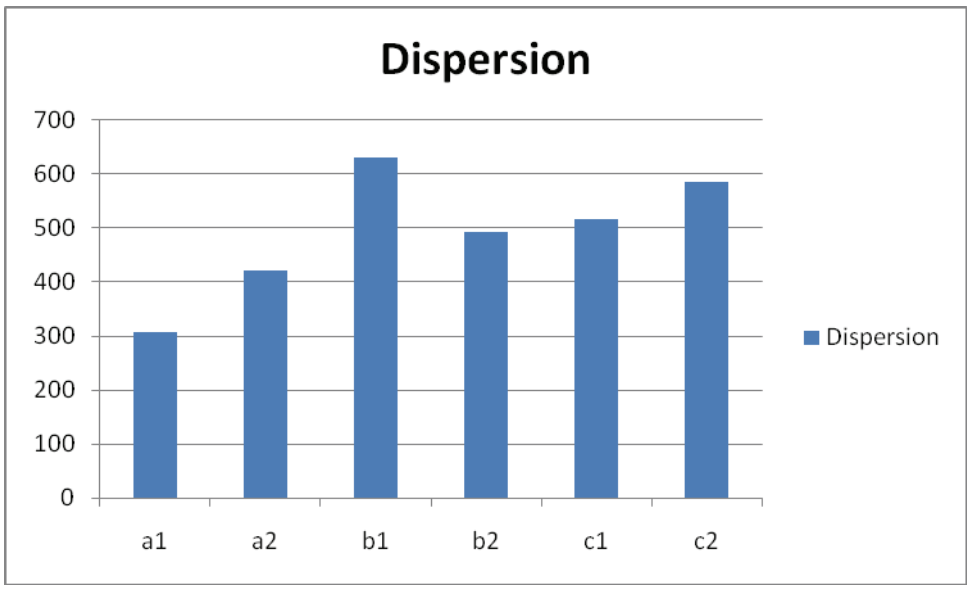
Suomi toisena kielenä -oppijoiden sanastollisen diversiteetin kannalta juuri B2 ja C1 taitotasojen välinen siirtymä vaikuttaa oleelliselta. MTLD-tunnusluvun lisäksi tasolta B2 tasolle C1 siirryttäessä sanaston kehittymisessä tapahtuu selkeä harppaus myös shannonin-indeksillä mitattuna. Sen sijaan molempien tunnuslukujen mukaan tasojen A1 ja A2 välinen kasvu on kaikkein vähäisintä. Tulosten perusteella näyttää siis siltä, että alemmilla tasoilla (A1 ja A2) sanaston kehittymisessä ei tapahdu ratkaisevia askelia vaan sanaston rikkaus ja monipuolisuus alkavat kehittyä vasta myöhemmin ja voimakkainta kasvu on B2 ja C1 taitotasojen välillä. Johtopäätöstä tukee myös aineistosta laskettujen TTR-arvojen vertailu. Vaikka TTR-arvot ovat voimakkaasti riippuvaisia sanemääristä, aineistossani oli mukana myös samankokoisia tekstikimppuja, joiden välillä TTR-arvojen vertailu oli mahdollinen. Joissakin tehtävätyypeissä TTR-arvot kasvavat B1 tasolta lähtien vaikka sanemäärät laskevat. Sen sijaan TTR-arvo ei kasva A1-tasolta A2-tasolle siirryttäessä, vaikka sanemäärä laskee tai pysyy samana (ks. luku 5.3).

6.1.5 Tasapuolisuustunnusluku



Tasapuolisuustunnusluvun antamien arvojen perusteella piirretyt pylväät vaikuttavat laskevan ja nousevan sattumanvaraisesti: pylväiden mukaan sanasto köyhtyisi odotusten vastaisesti esimerkiksi A1 tasolta A2 ja B1 tasoille noustessa. Alimmillaan tasapuolisuusluku on juuri näillä tasoilla eli 0,77 yksikköä ja ylimmillään taitotasoilla C1 ja C2, jolloin tunnusluku pysyy arvossa 0,85. Ilmiötä selittää jälleen otoskoon aiheuttama vinouma, mutta sanemäärän aiheuttamat vääristymät ilmenevät nyt päinvastoin kuin harvinaisuustunnuslukujen kohdalla: tasapuolisuustunnusluku jää sitä alhaisemmaksi, mitä enemmän saneita taitotasoa edustavissa teksteissä yhteensä on.

6.1.6 Hajaannustunnusluku

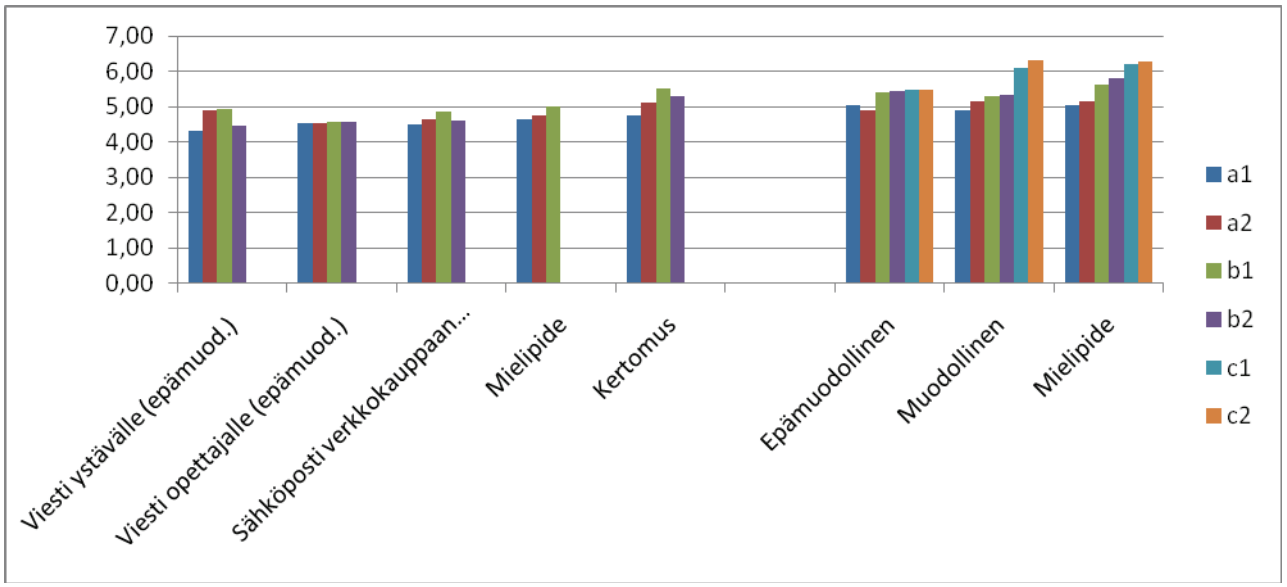


Samoin kuin harvinaisuustunnusluvut myös hajaannustunnusluku vaihtelee harhaanjohtavasti aineiston sanemäärän mukaan. Mitä enemmän taitotasolla on saneita, sitä korkeammat ovat kyseisten tunnuslukujen tulokset. Sanemäärä ei kuitenkaan vaikuta vääristävän tuloksia yhtä paljon kuin harvinaisuustunnusluvun kohdalla. Alimmillaan hajaannustunnusluku on 307,58 taitotasolla A1 ja ylimmillään 631,76 taitotasolla B1.

6.2 Tehtävätyypin vaikutus leksikaaliseen diversiteettiin ja erot koululaisten ja aikuisten sanastoissa

Tehtävätyyppi ja tekstilaji vaikuttivat sanastollista rikkautta mittaavien tunnuslukujen tuloksiin. Seuraavat taulukot esittävät tehtävätyypin vaikutuksen Shannonin indeksiin, harvinaisuustunnuslukuihin sekä tasapuolisuus-, hajaannus- ja monipuolisuustunnuslukuun (MTLD). Seuraava kuvio havainnollistaa tehtävätyypin ja tekstilajin vaikutusta tulosten Shannon-arvoihin. Ensimmäiset viisi diagrammia edustavat koululaisten eri tehtävätyyppejä ja jälkimmäiset kolme yki-aineiston tekstilajeja.

Taulukko 11: Shannon

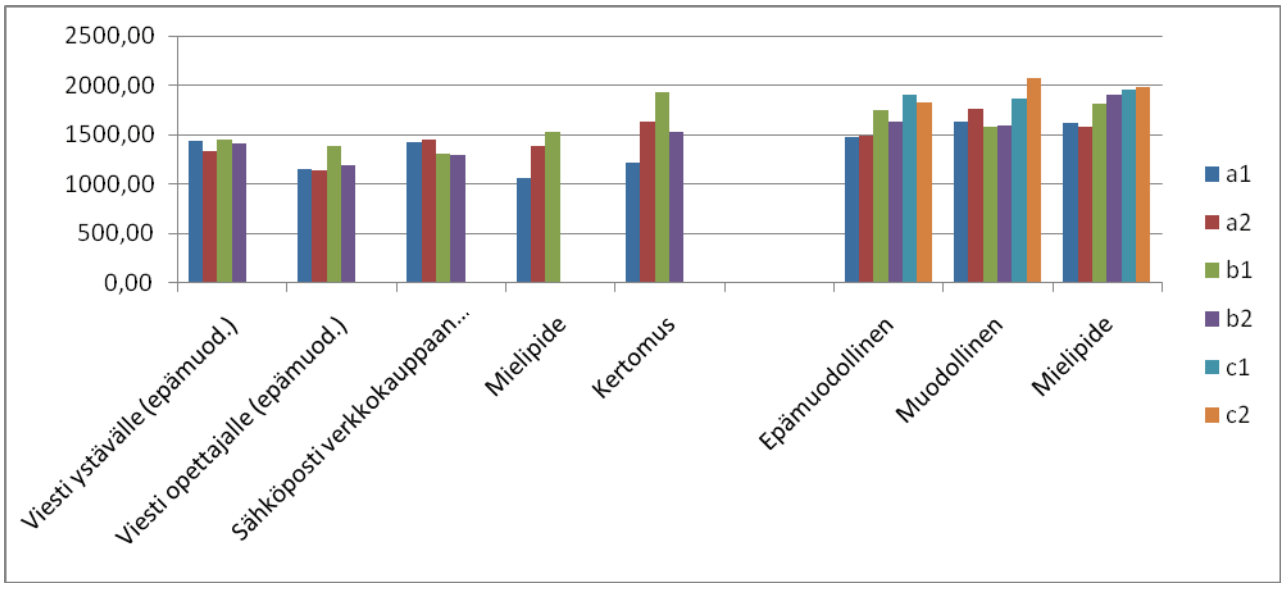


Tekstien saamat Shannon arvot pysyvät melko tasaisina tekstilajista riippumatta. Kuitenkin muita tekstilajeja korkeimpia arvoja koululaisten aineistossa saa erityisesti kertomuksen tekstilaji, jossa taitotasoilla A2-B2 Shannon on enemmän kuin viisi yksikköä. Yki-teksteissä korkeimmat Shannon arvot saa mielipiteen tekstilaji, erityisesti tasoilla A2-C1, joissa Shannon nousee nopeasti yli 5,5.

Vähiten rikasta Shannonin tunnusluvun mukaan sanasto tuntuu olevan epämuodollisissa tekstilajeissa sekä koululaisten (viesti opettajalle) että aikuisten kirjoitelmissa. Luvussa 6.1.1 esille noussut havainto Shannonin indeksin osoittamasta sanastollisen osaamisen harppauksesta B2 ja C1 tasojen välillä on nähtävissä myös aikuisten muodollisen tekstilajin kohdalla.

Hyvin samansuuntaisia, joskaan ei yhtä tasaisia tuloksia, antavat myös harvinaisuustunnusluku ja sisältösanojen harvinaisuustunnusluku vertailtaessa tekstilajien saamia tuloksia keskenään. Seuraava taulukko havainnollistaa, millaisia arvoja eri tekstilajit saavat harvinaisuustunnusluvun osalta.

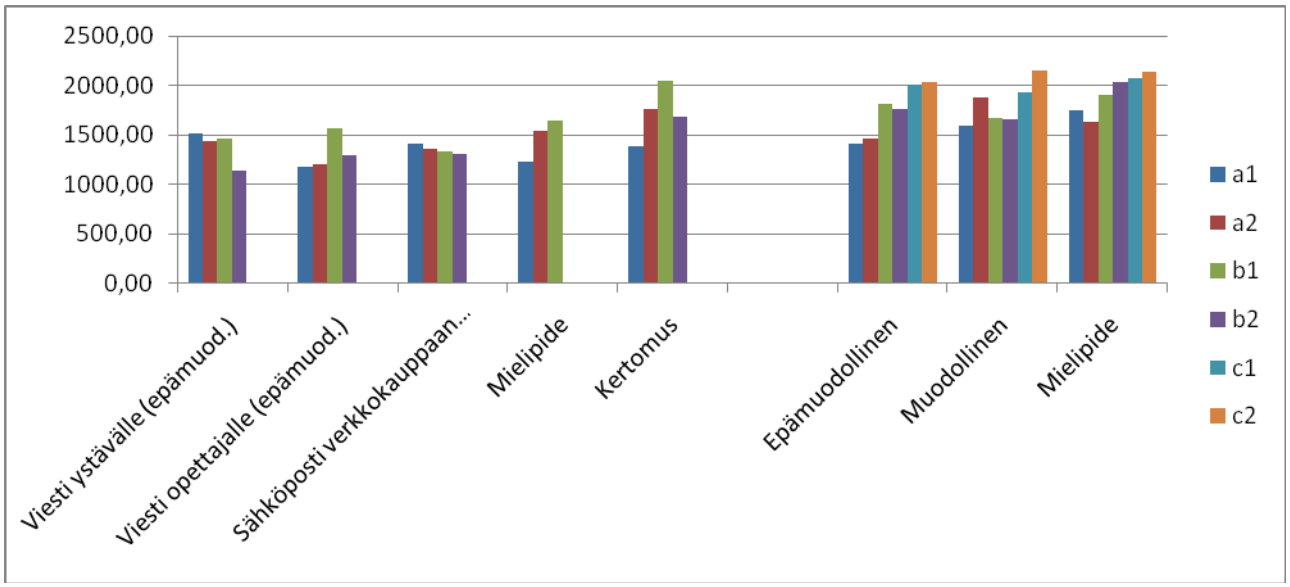
Taulukko 12: Harvinaisuustunnusluku



Harvinaisuustunnusluvun perusteella koululaisten sanasto on harvinaisempaa kertomuksissa kuin muissa kirjoituksissa. Ensimmäistä taitotasoa lukuun ottamatta kaikki kertomuksen taitotasot ylittävät harvinaisuustunnusluvun 1500, mitä ei tapahdu minkään muun tehtävätyypin kohdalla. Sisältösanojen osalta vastaava luku on jopa 1600, mikä myös käy ilmi kuviosta. Epämuodollinen viesti opettajalle saa alhaisimmat arvot. Viesti ystävälle, sähköpostiviesti verkkokauppaan ja mielipide pyörivät arvon 1250 molemmilla puolin.

Aikuisten osalta harvinaisuustunnusluku on keskimäärin korkeimmillaan mielipiteen tekstilajissa, mutta muodollinen tekstilaji yltää melkein samoihin lukemiin ja joidenkin tasojen kohdalla ylittää mielipiteen vastaavat arvot.

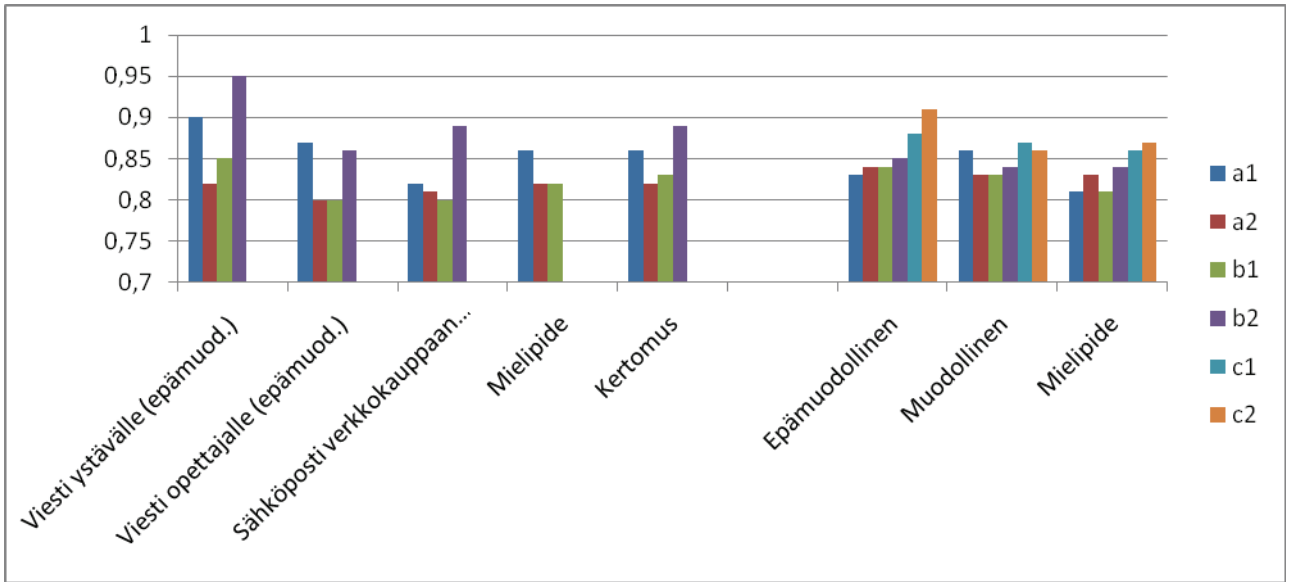
Taulukko 13: Sisältösanojen harvinaisuustunnusluku



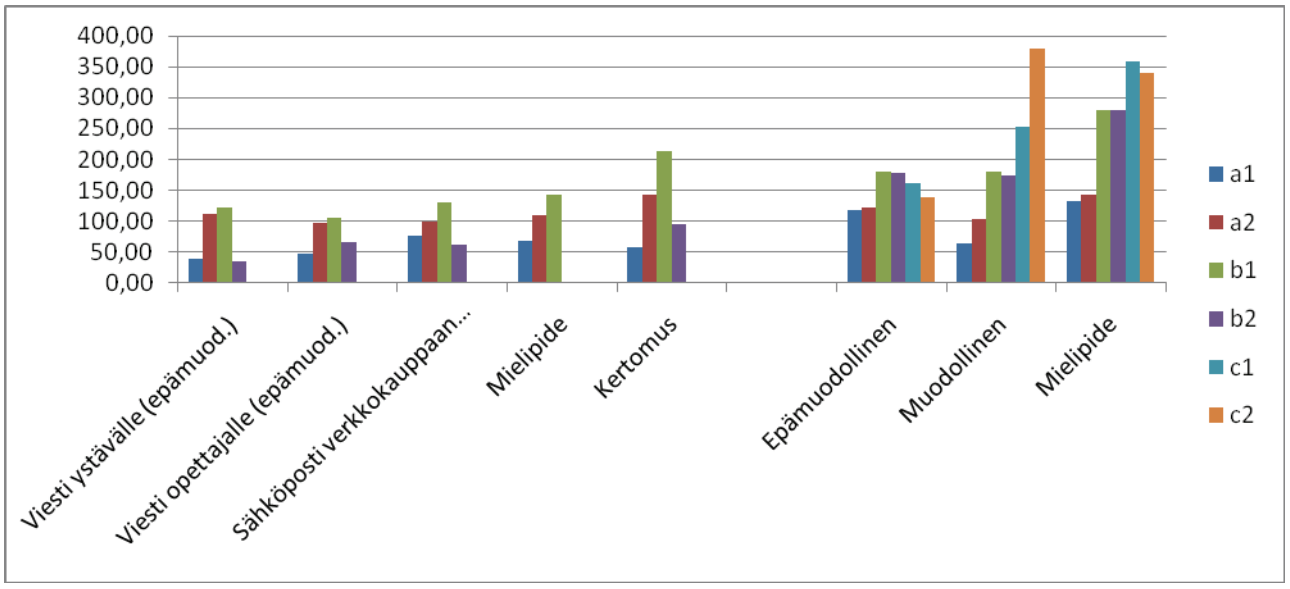
Eri tehtävätyypeistä lasketut sisältösanojen harvinaisuustunnusluvun tulokset ovat lähes identtiset harvinaisuustunnusluvun saamien arvojen kanssa: kertomuksen tekstilaji saa korkeimmat arvot koululaisten teksteissä, mutta aikuisilla ei ole suuria eroja eri tekstilajien välillä. Kertomuksen tekstilaji tuntuu siis koululaisaineistossa hallitsevan tilastoja rikkaus- ja harvinaisuustunnuslukujen osalta.

Toisaalta molempien harvinaisuustunnuslukujen antamiin tuloksiin on suhtauduttava hyvin varauksellisesti, sillä kuten tunnuslukujen vertailu edellisessä luvussa (6.1) osoitti, vain Shannonin ja MTLD-tunnusluvun antamiin tuloksiin voidaan todella luottaa. Seuraavat kaksi taulukkoa tasapuolisuus- ja hajaannustunnuslukujen tuloksista havainnollistavat hyvin otoskoon aiheuttamaa vääristymää myös eri tehtävätyyppejä ja tekstilajeja vertailtaessa. Sanojen tasapuolisuusluku, jolle on ominaista antaa suhteettoman korkeita arvoja sanemäärältään pienille taitotasoille, muodostaa melkein peilikuvan hajaannustunnusluvulle, joka puolestaan antaa sanemäärältään pienille taitotasoille suhteettoman matalia arvoja.

Taulukko 14: Tasapuolisuustunnusluku

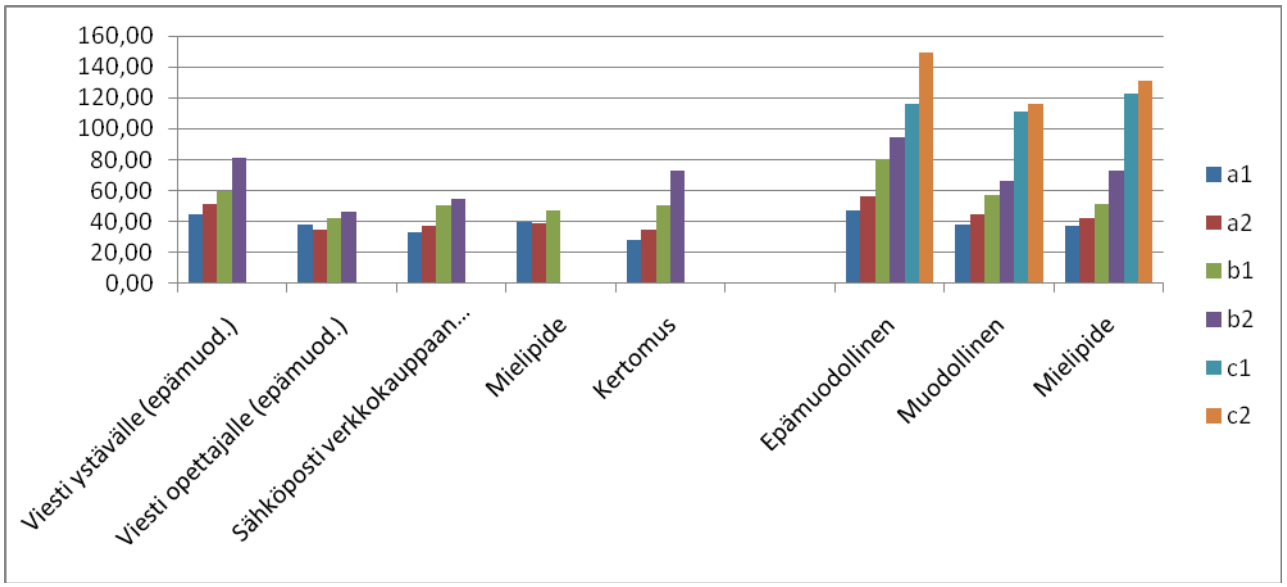


Taulukko 15: Hajaannustunnusluku



Vaikka harvinaisuus-, tasapuolisuus- ja hajaannustunnuslukujen luotettavuudet ovat hyvin kyseenalaisia, niiden tulokset kertomuksen tekstilajista vastaavat luotettavaksi todetun Shannonin arvoja. Lähes kaikkien edellä mainittujen tunnuslukujen mukaan juuri kertomuksen tekstilaji olisi koululaisaineistossa sanastollisesti rikkain. Selkeän poikkeuksen tehtävätyyppien keskinäiseen järjestykseen tekee kuitenkin MTLD, sanojen monipuolisuustunnusluku, jonka kohdalla koululaisten teksteissä viesti ystävälle saa järjestelmällisesti korkeimmat arvot. Seuraavassa kuviossa näkyy, kuinka viesti ystävälle saa kaikilla tasoilla arvon >40. Myös aikuisten aineistossa tapahtuu poikkeus. Muodollisen- ja mielipiteen tekstilajien sijasta korkeimmat arvot saa MTLD:n kohdalla epämuodollinen tekstilaji.

Taulukko 16: MTLD, sanojen monipuolisuustunnusluku



MTLD-tunnusluku antaa korkeimmat arvot koululaisten viesti ystävälle -tehtävätyypille ja aikuisten aineistossa epämuodolliselle tekstilajille. Kaikkien tekstilajien arvot kasvavat taitotasolta toiselle siirryttäessä. Aikaisempi havainto MTLD:n ja Shannonin indeksin osoittamasta sanastollisen osaamisen harppauksesta B2 ja C1 tasojen välillä näkyy mielipiteen ja muodollisen tekstilajin kohdalla tässäkin taulukossa.

Yleisesti ottaen aikuisten tekstit tuntuvat saavan hieman korkeampia rikkaus-, harvinaisuus- ja monipuolisuustunnuslukuja kuin koululaisten tekstit vastaavilla taitotasoilla. Vaikka Yki-aineiston sanemäärät ovat koululaisten aineistoa suuremmat, ilmiö ei johdu tällä kertaa otoskoon aiheuttamasta vinoumasta: aikuiset saavat korkeampia arvoja myös luotettaviksi todettujen Shannonin ja MTLD:n tunnuslukujen kohdalla. Lisäksi esimerkiksi koululaisten mielipiteen ja aikuisten muodollisen tekstilajien otoskoot ovat kaikilla taitotasoilla lähes samat. Joukossa on kuitenkin muutama poikkeus. Esimerkiksi MTLD-tunnusluvun antama tulos koululaisten mielipiteessä tasolla A1 on suurempi kuin Yki-aineiston mielipiteessä vastaavalla tasolla. Tällaiset poikkeukset eivät kuitenkaan muuta sitä tosiasiaa, että tunnusluvut ovat aikuisten aineistossa suurempia kuin koululaisten aineistossa. Tämän pohjalta vaikuttaisi, että Yki-tutkintojen kirjoitelmat on arvioitu taitotasoille kriittisemmin kuin koululaisten kirjoitelmat. Toisin sanoen koululaiset tuntuivat saavan helpommin paremman tasomäärityksen Yki-tutkintoa suorittaviin verrattuna.

Vaikka Yki-aineisto ja koululaisaineisto ovat vertailukelpoiset, Yki-korpus poikkeaa kuitenkin koululaisaineistosta siinä, että korpuksessa olevat tasoarviot eivät välttämättä ole yksittäiselle tekstille annettuja tasoarvioita vaan kyseisen kirjoittajan saama kokonaisarvio, joka perustuu

kaikkiin (2-3 kpl) hänen kirjoittamiinsa teksteihin. Yksittäinen teksti voi siis olla Yki-arvioidenkin mukaan muulla tasolla kuin kirjoittaja, jolta teksti on saatu. Tämä tieto ei kuitenkaan ole mukana korpuksessa. Lisäksi Yki-aineiston arvioissa on tavallaan kattoefekti: jos osallistuu keskitason testiin, niin kirjoittamisen arvosanaksi voi saada korkeintaan B2-tason, vaikka osoittaisi kirjoituksessaan esimerkiksi C1-tason osaamista. Sama pätee perustason testiin, jossa korkein mahdollinen arvio on B1, vaikka olisikin parempi kirjoittaja. Siitä syystä voi osaltaan näyttää siltä, että Yki-kirjoituksissa vaaditaan enemmän kuin nuorilta tai että nuoret saavat helpommin arvosanan.

6.3 Pohdintaa tehtävätyypin ja tekstilajin vaikutuksesta leksikaaliseen diversiteettiin

Tunnuslukujen vertailu osoittaa, että sanaston kehittymisessä on pieniä eroja eri tekstilajien välillä. Shannonin tunnusluvun mukaan rikkainta ja vaihtelevinta koululaisten sanasto on kertomuksissa. Tämä ei ole erityisen yllättävää niiden tutkimusten valossa, joiden mukaan oppilaat kirjoittavat yläkoulussa sekä äidinkielen että vieraiden kielten tunneilla eniten tarinoita ja kertovia tekstejä (Luukka 2008: 152). Narratiivinen teksilaji on oppilaille hyvin tuttu ja saattaa siksi vaikuttaa rikastuttavasti myös sanastoon, jota he käyttävät kirjoittaessaan kertomuksia. Suurempana syynä saattaa kuitenkin olla Cefling-aineiston tekstien heterogeenisyys erityisesti kertomuksen tekstilajissa. Tehtävänannon rajoissa oppilaiden oli mahdollisuus kirjoittaa hyvin monista eri aihealueista, jolloin tietyn tason kaikkien tekstien keskinäinen vaihtelu on luultavasti ollut suurempaa, kuin esimerkiksi viesti opettajalle -tehtävänannossa, jossa on melko tarkasti ohjeistettu kertomaan tietyt asiat.

Seuraavaksi rikkainta ja vaihtelevinta koululaisten sanasto on epämuodollisessa viestissä ystävälle, joka tekstilajien vertailussa osoittautui MTLD-tunnusluvulla mitattuna sanastoltaan myös kaikkein monipuolisimmaksi tekstilajiksi. Ystävälle osoitettu viesti saatetaan kokea epäviralliselmmaksi ja epämuodollisemmaksi kuin opettajalle lähetettävä viesti, vaikka jälkimmäisessäkin olisi kyse epämuodollisesta tekstilajista. Epävirallisuus saattaa rohkaista kirjoittamaan rönsyilevämmin ja kokeilemaan myös sanoja, joiden oikeinkirjoituksesta tai täsmällisistä merkityksistä ei olisikaan varma. Koululaisten aineistossa viesti opettajalle sai jokaisen tunnusluvun kohdalla keskimäärin matalimmat arvot. Viestit opettajalle olivat lyhyitä ja homogeenisiä ja keskenään viestit toistivat itseään.

Aikuisilla sanasto on Shannonin tunnusluvun mukaan rikkainta ja vaihtelevinta mielipiteen tekstilajissa, mutta erot muodolliseen tekstilajiin eivät ole suuret. Samoin kuin koululaisilla

aikuisten sanasto on monipuolisinta MTLD-tunnusluvun mitattuna epämuodollisessa tekstilajissa. Tehtävätyyppien ja tekstilajien välillä havaittavat pienet sanastolliset erot sekä koululaisten että aikuisten aineistossa johtuvat luultavasti juuri tehtävänantojen eroista. Esimerkiksi mielipiteen ja kertomuksen tekstilajeissa tehtävänannot olivat vapaampia tai valmiiksi annettuja aiheita oli useampia. (ks. lisää tehtävänannoista luku 4.2.1).

Sen sijaan yki-tutkintoa suorittaneiden saama opetus ja käytetyt oppikirjat tuskin vaikuttavat kovin paljon eri tekstilajien sanastoon. Oili Jäppinen (2011) on gradussaan tutkinut, millaisia tekstitaitoja aikuisille maahanmuuttajille tarkoitetuissa oppikirjoissa tarjotaan ja millaisia tekstilajeja kirjoissa esiintyy. Sekä oppikirjoissa olevat esimerkkitekstit että tehtävänannoissa vaadittavat tekstilajit edustavat monipuolisesti tekstilajien kirjoa. Varsinkin uudemmissa, 2000-luvun oppikirjoissa funktionaalinen tekstikäsitys on omaksuttu kaiken perustaksi ja tehtävänannot pyrkivät tuottamaan mahdollisimman autenttisia tekstejä. Sen sijaan 2000-lukua aikaisemmin ilmestyneissä oppikirjoissa oli pääasiassa oppikirjatekstejä eli varta vasten oppikirjaan laadittuja opetustekstejä ja -dialogeja sekä kielioppikuvauksia.

Uusissa oppikirjoissa on Jäppisen (2011) mukaan panostettu juuri autenttisiin teksteihin ja kirjoitustehtäviin: todellisen elämän tekstilajit ovat sekä oppimisen kohteena että kielenoppimisen apuna. Lisäksi uusimpien, erityisesti opetussuunnitelmasuositusta noudattavien oppikirjojen kirjoitustehtävät ovat saaneet mallinsa YKI-testien tehtävistä, jolloin niiden tekstit luonnollisesti limittyvät Eurooppalaisen viitekehyksen sisältöihin ja taitotasokuvauksiin sekä YKI-perusteiden funktioihin ja aihealueisiin. Epämuodollisia tekstilajeja kirjoissa edustivat muun muassa tekstiviesti ystävälle, kutsu juhliin ja kirje ja muodollisia esimerkiksi asunto-ilmoitus, työhakemus, työpaikkailmoitus ja reklamaatio jne. Nämä oppikirjat tarjoavat suomi toisena kielenä -oppiville tasapuoliset valmiudet eri tekstilajien omaksumiseen.

Myös Kauppinen (2008) on tutkinut oppikirjojen tekstivalikoimaa. Hänen mukaansa vieraiden kielten tekstivalikoima on oppikirjoissa hieman äidinkielen oppikirjoja laajempi ja niissä esiintyy hieman enemmän esimerkiksi mediatekstejä. Hänen tutkimissaan oppikirjoissa esiintyi toisaalta vielä paljon koulun tekstejä, ei niinkään todellisen elämän tekstilajeja.

Sekä Shannonilla ja MTLD-tunnusluvulla mitattuna eri tekstilajien keskinäiset erot sekä aikuisten että koululaisten aineistoissa ovat kuitenkin niin pieniä, että pitkälle meneviä johtopäätöksiä niistä ei kannata tehdä. Ylipäätään rajat eri tekstilajien välillä eivät ole kovin selvät ja limittymistä niiden välillä tapahtuu paljon. Esimerkiksi muodollisen ja epämuodollisen tekstilajin

eroa on joskus jopa vaikea määritellä (Saukkonen 2001, 165-166). Myös aikaisempi tekstilajien sanaston kvantitatiivinen tutkimus on ollut vähäistä. Vaikka tekstilajeja on tutkittu paljon, tarkastelun kohteena on perinteisesti ollut rakenne eikä niinkään sanasto (esim. Hyland 2004: 7–9). Eri tekstilajeilla on kyllä tutkittu olevan omat tyypilliset sanastolliset ominaisuutensa, mutta tulokset ovat usein laadullisia ja kytkeytyvät sanaston rakenteisiin. Tulosten mukaan esimerkiksi mielipideteksteissä sanasto on affektiivista (Myntti 2012: 77), kertomuksissa käytetään paljon imperfektiä, kuvaavat tekstit ja esimerkiksi työpaikkailmoitukset sisältävät paljon adjektiiveja ja uutiset puolestaan verbejä (esim. Jäppinen 2011: 65). Oman aineistoni vertaaminen tällaisiin tutkimuksiin vaatisi kuitenkin laadullista tarkastelua määrällisen analyysin rinnalle.

7 TUTKIMUKSEN LOPUKSI

7.1 Tutkimuksen keskeisimmät tulokset

Tutkimukseni tavoitteena on ollut selvittää, kuinka sanaston osaaminen kehittyy suomi toisena kielenä -oppijoiden teksteissä. Tarkastelun kohteena on ollut sanavaraston rikkaus eli leksikaalinen diversiteetti ja aineistona aikuisten ja koululaisten eri taitotasoille valmiiksi arvioidut tekstit. Työni on osa Cefling-hanketta, jonka tarkoituksena on selvittää, millaista kielitaito on kullakin Eurooppalaisessa viitekehyksessä kuvatulla kuudella taitotasolla. Keskeisimmät teoreettiset tausta-ajatukset ovat tulleet Cefling-hankkeen lisäksi sanaston tutkimuksen perinteestä ja erityisesti määrällisen tutkimuksen lähtökohdista.

Perinteisten menetelmien ja mittareiden lisäksi olen esitellyt työssäni nykyaikaisia sanastollisen diversiteetin mittaamiseen kehitettyjä, entistä parempia tunnuslukuja, joita ei ole aikaisemmin sovellettu suomenkieliseen aineistoon. Tutkimuksessa käyttämistäni uusista tunnusluvuista erityisesti MTLD, sanojen monipuolisuusluku, on osoittautunut luotettavaksi ja otoskoosta riippumattomaksi keinoksi mitata sanaston monimuotoisuutta. Toinen yhtä luotettava tunnusluku on Shannonin indeksi. Aineistoni sanastoa on tunnuslukujen avulla verrattu myös Suomen sanomalehtikielen taajuussanastoon, sillä aikaisempien tutkimusten mukaan oppivan kielessä esiintyvien sanojen yleisyys tai harvinaisuus kertoo osaltaan kielenkäyttäjän sanastollisesta osaamisesta.

Työtäni ohjaava tutkimuskysymys on ollut, kuinka sanaston osaaminen kehittyy suomi toisena kielenä -oppijoiden teksteissä. Kysymystä olen tarkentanut kolmella alakysymyksellä: millaista sanaston osaaminen on kullakin taitotasolla, miten tekstilaji vaikuttaa sanaston rikkauteen ja kuinka koululaisten ja aikuisten sanavarastot eroavat toisistaan.

Tutkimukseni osoitti, että suomi toisena kielenä -oppivan sanasto kehittyy taitotasojen myötä. Erityisen selkeä harppaus sanaston kehittymisessä tapahtuu MTLD-tunnusluvun ja Shannonin indeksillä laskettujen tulosten mukaan taitotasolta B2 taitotasolle C1 siirryttäessä. Sen sijaan molempien tunnuslukujen mukaan taitotasojen A1 ja A2 välinen kasvu on kaikkein vähäisintä. Tulosten perusteella näyttää siis siltä, että alemmilla tasoilla (A1 ja A2) sanaston kehittymisessä ei tapahdu ratkaisevia askelia, vaan sanaston rikkaus ja monipuolisuus alkavat kehittyä vasta myöhemmin. Johtopäätöstä tukee myös aineistosta laskettujen TTR-arvojen vertailu. Vertailukelpoisten tiettyä taitotasoa edustavien tekstien joukoissa TTR-arvo kasvoi erityisesti taitotasojen B2 ja C1 välillä, kun taas tasojen A1 ja A2 välillä ei tapahtunut kasvua tai se oli hyvin vähäistä. Suomi toise-

na kielenä -oppijoiden sanastollisen diversiteetin kannalta juuri B2 ja C1 taitotasojen välinen siirtymä vaikuttaa siis oleelliselta. Tulosten mukaan suomi toisena kielenä -oppijoiden sanasto monipuolistuu ja rikastuu erityisesti oppijoiden yltäessä kirjoituksissaan C-tasolle.

Aineistossani on ollut mukana eri tekstilajeja edustaneita tekstejä. Koululaisaineiston tehtävätyypit olivat viesti ystävälle, viesti opettajalle, sähköpostiviesti verkkokauppaan, mielipide sekä kertomus ja aikuisten aineisto sisälsi epämuodollinen viestin, muodollinen viestin ja mielipiteen. Tutkimuksessani eri tekstilajien välille ei kuitenkaan syntynyt niin suuria eroja, että niistä olisi mahdollista vetää eri tekstilajien sanastojen osaamista koskevia päätelmiä. Aikuisten aineisto sai korkeimmat Shannon-arvot mielipiteen tekstilajissa ja korkeimmat MTLD-arvot epämuodollinen tekstilajissa ja koululaisten sanasto on arvioitu korkeimmalla kertomuksen ja viesti ystävälle -tehtävätyypin kohdalla. Oletettavasti tehtävätyyppien ja tekstilajien välillä havaittavat pienet sanastolliset erot johtuvat kuitenkin tehtävänantojen eroista. Esimerkiksi mielipiteen ja kertomuksen tekstilajeissa tehtävänannot olivat vapaampia tai valmiiksi annettuja aiheita oli useampia.

Sen sijaan aikuisten ja koululaisten keskinäinen tekstien vertaaminen osoitti, että aikuisten aineisto saa hieman korkeampia rikkaus- ja monipuolisuustunnuslukuja kuin koululaisten aineisto vastaavilla taitotasoilla. Tämän pohjalta vaikuttaisi, että Yki-tutkintojen kirjoitelmat on arvioitu taitotasoille kriittisemmin kuin koululaisten kirjoitelmat. Toisin sanoen koululaiset tuntuvat saavan helpommin paremman arvosanan Yki-tutkintoa suorittaviin verrattuna.

Yksittäinen mielenkiintoinen tutkimustulos suomi toisena kielenä -oppijoiden sanaston kehittymisestä nousi esille myös aineistojen kaikkein yleisimpiä lekseemejä tarkasteltaessa. Kolmenkymmenen yleisimmän lekseemien kattavuus eri taitotasoilla paljasti, että mitä ylempää taitotasoa tekstit edustavat, sitä pienemmäksi yleisimpien lekseemien kattavuus koko aineistosta laskee. Korkeimmilla taitotasoilla yleisimmätkään lekseemit eivät kata enää prosentuaalisesti yhtä suurta osaa kaikista saneista kuin alemmilla tasoilla. Tämä kertoo osaltaan siitä, että kielitaidon kehittyessä suomi toisena kielenä -oppijoiden sanasto monipuolistuu ja teksteissä ilmenee yhä enemmän eri sanoja.

7.2 Tutkimuksen arviointia ja jatkotutkimusideoita

Olen tarkastellut tässä tutkimuksessa suomi toisena kielenä -oppijoiden sanaston kehittymistä eri taitotasoille luokitelluissa teksteissä. Tutkimukseni edustaa kvantitatiivista tutkimusta. Aineistona on ollut valmis Cefling-aineisto, jonka aikaisemmat tutkijat ovat siirtäneet sähköiseen muotoon. Koska valmista tutkimuksessani tarvittavaa lemmattua aineistoa ei ollut, päädyin lemmaamaan aineistoni itse.

Lemmaustyössä olen pyrkinyt tekniseen luotettavuuteen yleisellä huolellisuudella ja käyttämällä mahdollisimman pitkälle samaa sanaluokkajaottelua kuin on käytetty Suomen sanomalehtikielen taajuussanastossa. Sama luokittelu on ollut tärkeää, jotta aineistoni olisi mahdollisimman vertailukelpoinen taajuussanaston kanssa. Sanaluokkajaottelu taajuussanaston ja tietokoneohjelman avulla lisää myös tutkimuksen toistettavuutta ja tarkistettavuutta: olen voinut tarkistaa luokittelut useampaan kertaan ja muilla tutkijoilla on halutessaan mahdollisuus tarkistaa analyysini. Luonnollisesti lemmattu aineisto sisältää myös virheanalyysejä: kaikkien saneiden merkityksiä on ollut mahdotonta tarkastaa tekstiyhteydestä.

Koko työn kannalta tutkimukseni aineistossa on ollut kaksi isoa haastetta. Ensimmäinen on eri taitotasojen ja eri tekstilajien sanemäärien keskinäinen vaihtelu. Tämä tuotti ongelmia erityisesti viidennessä luvussa, jossa tarkastelin sanastoa yleisesti perinteisten kvantitatiiviset menetelmien avulla. Vaikka sanemääristä riippuvaisten menetelmien ongelmat perinteisissä menetelmissä olivat tiedossa alusta alkaen, otoskoonkoon aiheuttama vinouma vei pohjaa myös useiden uusien tunnuslukujen luotettavuudelta. Tutkimukseni pelasti kuitenkin Shannonin ja MTLD-tunnuslukujen luotettavuus, joiden avulla sain aineistostani irti otoskoosta riippumattomia tuloksia. Joillakin tasoilla joidenkin tekstilajien sanemäärät olivat myös niin lähellä toisiaan, että vertailu niiden välillä oli mahdollinen myös muilla tunnusluvuilla.

Toinen, jopa sanemäärien vaihtelua suurempi haaste muodostui siitä, että taitotasojeni aineistot eivät olleet yksittäisiä tekstejä, vaan koko samalle taitotasolle arvioitujen tekstien ryhmiä. Eri tunnuslukuja ja esimerkiksi eri lekseemien osuuksia ei ole siis laskettu yksittäisistä teksteistä vaan tiettyä taitotasoa edustavien tekstien ryhmästä. Näin ollen saamani tulokset eivät ole suoraan vertailukelpoisia aikaisempiin tutkimuksiin, joissa samoja tunnuslukuja ja muita menetelmiä on käytetty yksittäisiin teksteihin. Esimerkiksi verrattaessa aikaisempiin tutkimuksiin (esim. Saarela 1997) saamani Shannon-arvot ovat aivan liian korkeita. Tästä johtuen en ole voinut tehdä tuloksistani kovin monipuolista vertailevaa analyysia aikaisempiin tuloksiin, mikä kavensi jo entuudestaan

kapeaa lähde- ja tutkimuskirjallisuuden joukkoa. Esimerkiksi MTLD-monipuolisuustunnusluku on Suomessa vielä niin uusi, ettei sen käytöstä ole aikaisemmin julkaistu mitään tutkimustuloksia suomenkielisestä aineistosta. Tältä osin oma tutkimukseni sisältääkin uutta perustutkimusta sanaston alueelta.

Jotta tutkimuksesta saatuja tuloksia voitaisiin paremmin verrata muihin vastaaviin tuloksiin, tutkimusta olisi mahdollista jatkaa tarkastelemalla yksittäisiä tekstejä kokonaisten taitotasojen sijaan. Tässä työssä käytettyjä muuttujia voisi laskea samalla tapaa yksittäisistä teksteistä niin, että samat tunnusluvut laskettaisiin jokaisesta tämän tutkimuksen aineiston muodostaneesta noin 1200 tekstistä erikseen. Varsinkin korkeammilla tasoilla, joilla yksittäisten tekstien sanepituus on jo useampia satoja, tunnuslukujen soveltaminen olisi mielekästä.

Tutkimustani voisi lähteä laajentamaan myös moneen muuhun suuntaan. Yksi kiinnostava vaihtoehto olisi lähteä tekemään laadullisempaa tutkimusta tässä tutkimuksessa tehdyn määrällisen avauksen rinnalle. Laadullisin menetelmin olisi mielenkiintoista tarkastella esimerkiksi sanaston kehittymistä B2 ja C1 taitotasojen välillä, jolloin tulosteni mukaan sanaston kehittyminen suomi toisena kielenä -oppijoiden teksteissä on erityisen nopeaa.

LÄHTEET

Painetut lähteet

Aalto, Eija 1994: Alussa on sana – Systemaattisuutta sanaston opettamiseen. Teoksessa Suni, Minna & Aalto, Eija (toim.) *Suuntaa suomenopetukseen – tuntumaa tutkimukseen.* Korkeakoulujen kielikeskuksen selosteita 4. Jyväskylä: Jyväskylän yliopisto, s. 93-117.

Alanen, Riikka – Huhta, Ari – Tarnanen, Mirja 2010: Designing and assessing L2 writing tasks across CEFR proficiency levels. – Inge Bartning, Maisa Martin & Ineke Vedder (toim.), Communicative proficiency and linguistic development: intersections between SLA and language testing research s. 21–56. European Association of Second Language Acquisition. http://eurosla.org/monographs/EM01/21-56Alanen_et_al.pdf.

Arajuuri, Yrjö 1980: Lukutapahtuma, tulkintajakso ja tekstin luettavuus. Psykolingvistinen tutkimus. Suomen kielen lisensiaatintyö. Tampereen yliopiston suomen kielen ja yleisen kielitieteen laitos.

Carter, Ronald – McCarthy, Michael 1988: Vocabulary and Language Teaching. London.

Cook, Vivian 1991: Second Language Learning and Language Teaching. London: Arnold. (New York)

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

Grönholm, Maija 1993: TV on pang pang -verbisanaston kehitys toisen kielen kirjoittamisessa. Vasa: Åbo Akademi.

Hyland, Ken 2004: Genre and second language writing. Michigan series on teaching multilingual writers. Ann Arbor : University of Michigan Press, cop. 2004.

Jaakola, Mari 2004: Peruskoulun 2. ja 3. luokan oppikirjojen sanasto – dynamiikasta luettavuuteen. Pro gradu -tutkielma. Tampere: Tampereen yliopisto.

Jarvis, Scott 2011: Constructs and dimensions of vocabulary deployment. Paper presented at the annual meeting of the American Association for Applied Linguistics (AAAL) 2011, Chicago, March 27. (PowerPoint-esitys).

Jussila, Raimo 1996: Onko muodikkain yleisintä? – Hiidenkivi 6, s. 26-27.

Jäppinen, Oili 2011: Oppikirjat tavoitteiden tulkkeina ja tukijoina: tekstitaidot aikuisille maahanmuuttajille tarkoitetuissa suomen kielen oppikirjoissa. Pro gradu -tutkielma. Jyväskylä: Jyväskylän yliopisto.

Karlsson, Fred 1983: Suomen kielen äänne- ja muotorakenne. Helsinki: WSOY.

Laufer, Batia 1997: What's in a word makes it hard or easy: some intralexical factors that affect the learning of words. Norbert Schmitt & Michael McCarthy (toim.), Vocabulary: Description, Acquisition and Pedagogy s.140-155. Cambridge University Press.

Little, David 1994: Words and their Properties: Argument for a Lexical Approach to Pedagogical Grammar. Teoksessa Odlin, Terence (toim.) Perspectives on Pedagogical Grammar. Cambridge.

Luukka, M.R. ym. 2008: Maailma muuttuu – mitä tekee koulu? Jyväskylä: Jyväskylän yliopisto, Soveltavan kielentutkimuksen keskus.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. 2004: Lexical diversity and language development: quantification and assessment. Basingstoke: Palgrave Macmillan.

Martin, Maisa 1999: Mikä on keskeisintä suomen kielessä. – Seppo Pekkola (toim.), Sadanmiehet. Aarni Penttilän ja Ahti Rytkösen juhlakirja. Jyväskylän yliopiston suomen kielen laitoksen julkaisuja 41, s. 155-181.

McCarthy, Michael 1990: Vocabulary. Oxford.

McCarthy, P.M. – Jarvis, S. 2010: MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. – Behavior Research Methods 42 (2), 381-392. The Psychonomic Society, Inc.

Meara, Paul 1993 [1992]: Network structures and vocabulary acquisition in a foreign language. Pierre J. L. Arnauld – Henri Béjoint (toim.), Vocabulary and applied linguistics, s. 62-70. London: Macmillan.

Meara, Paul & Fitzpatrick, Tess 2000: Lex 30: an improved method of assessing productive vocabulary in an L2. *System*, 28, 19-30.

Myntti, Mari 2012: Sanomalehden tekstiviestipalstojen mielipidekirjoitukset tekstilajina. Pro gradu -tutkielma. Joensuu: Itä-Suomen yliopisto.

Mäkinen, Auli 1997: *Sanasto* suomen kielen alkeiskurssin opettajan opetuspuheessa. Pro gradu -tutkielma. Jyväskylä: Jyväskylän yliopisto, suomen kielen laitos.

Nation, Paul 1990: Teaching and Learning Vocabulary. New York: Newbury House Publishers.

Niemikorpi, Antero 1990: Suomen kielen sanaston frekvenssianalyysia. Vaasan korkeakoulun julkaisuja. Tutkimuksia No 150.

Niemikorpi, Antero 1991: Suomen kielen sanaston dynamiikkaa. Acta Wasaensia. No 26. Kielitiede 2. Vaasa.

Nuutinen, Olli 1994: Hetkisen pituus ja muita kirjoituksia kielestä. Tietolipas 128. Tampere.

Penttilä, Aarni 1963: Suomen kielioppi. 2. painos. Porvoo: WSOY.

Penttinen, Kati 2010: Voisitko apua? Suomi toisena kielenä -oppijoiden sananmuodostustaitojen jäljillä. Pro gradu -tutkielma. Jyväskylä: Jyväskylän yliopisto.

Puro, Tarja 1999: Sanastollinen tieto ja suomen kielen oppikirjojen sanasto. Virittäjä. Kotikielen Seuran aikakauslehti 1/1999, s. 2–26.

Puro, Tarja 2002: Suomi toisena kielenä -aikuisoppijan verbien kehittyminen alkeiskurssilla. Lisensiaatintyö. Jyväskylä: Jyväskylän yliopisto.

Read, John 2006: Assessing Vocabulary. Cambridge: Cambridge University Press.

Räsänen, Seppo 1975: Sanastollis-kvantitatiivinen tutkimus Aleksis Kiven pääteosten tyylistä. Tampereen yliopiston suomen kielen laitos. Monistesarja, moniste 4.

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

Saukkonen, Pauli 2001: Maailman hahmottaminen teksteinä. Tekstirakenteen ja tekstilajien teoriaa ja analyysia. Helsingin yliopistopaino -kustannus.

Sinclair, John – Renouf, Antoinette 1988: A lexical syllabus for language learning. Teoksessa Carter, Ronald – McCarthy, Michael (toim.) Vocabulary and Language Teaching. London.

Singleton, David 1995: First catch your lexicon: the tribulations of the L2 vocabulary acquisition researcer. Plenary paper 10.9. Fifth Eurosla Annual Conference, Dublin.

Suomen kielen taajuussanasto 1979: Toim. Saukkonen, Pauli & Haipus, Marjatta & Niemikorpi, Antero & Sulkala, Helena. Porvoo: WSOY.

Särkilahti, Sirkka-Liisa 1977: Tyylintutkimuksen kvantitatiiviset metodit. Suomen sovelletun kielitieteen yhdistyksen (AFinLA) julkaisuja n:o 19. Turku.

Särkkä, Tauno 1987: Sanaston rikkaudesta ja sen mittaamisesta. – Virittäjä. Kotikielen Seuran aikakauslehti 91, s. 129-136. Helsinki.

Takala, Sauli 1989: Sanaston opettamisen uudet haasteet. – Sauli Takala (toim.), Sanaston opettaminen ja oppiminen. Kasvatustieteiden tutkimuslaitoksen julkaisusarja B. Teoriaa ja käytäntöjä 44, s. 1-11. Jyväskylä: Kasvatustieteiden tutkimuslaitos.

Toivainen, Jorma 1980: Inflectional affixes used by Finnish-speaking children aged 1-3 years. SKS, Helsinki.

Vermeer, Anne 2004: The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. Teoksessa Bogaards, Paul & Laufer, Batia (toim.), Vocabulary in a Second Language: Selection, Acquisition, and Testing, s.173-189. Amsterdam: John Benjamins Publishing Co.

Voionmaa, Kaarlo 1993: Frekventtien verbien asemasta kielenoppimisessa ja -opetuksessa. – Jyrki Kalliokoski – Siitonen, Kirsti (toim.), Suomeksi maailmalla. Kirjoituksia suomen kielen ja kulttuurin opettamisesta, s. 63-75. Castrenianumin toimitteita 44. Helsinki: Yliopistopaino.

Vehmaskoski, Maila 1976: Sanaston frekvensseistä ja laadusta eräiden vuosina 1935 ja 1965 ilmestyneiden romaanien repliikeissä. Artikkeli kokoelmassa Kielitieteellisiä lehtiä. Raija Lehtinen, Tapani Lehtinen, Pirkko Nuolijärvi, Heikki Paunonen (toim.). Kotikielen seuran sadannen toimintavuoden täyttyessä. Suomi 120: 4. Helsinki: Suomalaisen Kirjallisuuden seura.

Sähköiset lähteet

Cefling-hankkeen Jyväskylän yliopiston kielten laitoksen www-sivut:

https://www.jyu.fi/hum/laitokset/kielet/cefling/suom (luettu 1.8.2012)

Second Language and Testing in Europe (SLATE) -verkoston verkkosivut:

http://www.slate.eu.org/ (luettu 1.8.2012)

Suomen sanomalehtikielen taajuussanasto, CSC – Suomen tietotekniikan keskus:

http://www.csc.fi/tutkimus/alat/kielitiede/taajuussanasto-B9996/view (luettu 8.8.2012)

Topling-hankkeen Jyväskylän yliopiston kielten laitoksen www-sivut:

https://www.jyu.fi/hum/laitokset/kielet/topling/suom (luettu 1.8.2012)