Ulkomailta tulleiden lääkäreiden sanasto laillistamiskuulustelussa

Jarkko Hakala
Tampereen yliopisto
Viestintätieteiden tiedekunta
Suomen kielen tutkinto-ohjelma
Pro gradu -tutkielma
Syyskuu 2018

Tampereen yliopisto
Viestintätieteiden tiedekunta
Suomen kielen tutkinto-ohjelma

HAKALA, JARKKO: Ulkomailta tulleiden lääkäreiden sanasto laillistamiskuulustelussa

Pro gradu -tutkielma, 71 sivua
Kesäkuu 2018

Tässä tutkimuksessa tarkastellaan, millaista sanastoa ulkomailta Suomeen saapuneet lääkärit ovat käyttäneet Tampereen yliopiston järjestämässä laillistamiskuulustelussa, joka EU- ja ETA-maiden ulkopuolella kouluttautuneiden lääkäreiden on läpäistävä saadakseen oikeuden harjoittaa ammattiaan Suomessa. Tutkimus liittyy Tampereen yliopistossa toteutettuun tutkimushankkeeseen, joka pyrki selvittämään, millainen yhteys ulkomailta saapuvien lääkäreiden kielitaidolla on kuulustelun heikkoon läpäisyprosenttiin.

Tutkimuksen tarkoitus on hahmottaa yleiskuva kuulusteluun osallistuneiden lääkärien käyttämästä sanastosta ja heidän sanastotaidoistaan. Lisäksi tutkimus pyrkii selvittämään, millainen yhteys sanastotaidoilla ja koemenestyksellä on mahdollisesti havaittavissa. Empiiristä aineistoa tarkastellaan sanatasolla, mutta sanojen kontekstia on hyödynnetty esimerkiksi aineiston luokittelussa. Empiirinen aineisto koostuu laillistamiskokeen vastauksista, jotka 46 kokeeseen osallistunutta lääkäriä on kirjoittanut suomeksi kolmiosaisen kuulustelun ensimmäisessä vaiheessa. Kyseessä on kirjallinen koe, jonka kysymykset koskevat käytännön potilastapauksia. Aineiston tuottaneet lääkärit oli jaettavissa kielitaustan perusteella kahteen ryhmään, sillä heistä kahdeksan on tulkittu natiiveiksi suomenpuhujiksi.

Analyysi hyödyntää kvantitatiivisia menetelmiä ja tilastollista analyysiä. Se on toteutettu kahdessa vaiheessa, joista kumpikin nostaa esiin sekä yksittäisten lääkärien että kieliryhmien välisiä eroja. Ensin tutkimus rakentaa deskriptiivisesti yleiskuvaa käytetystä sanastosta ja erittelee sitä muun muassa sanojen sanaluokkien, kompleksisuuden ja pituuden näkökulmasta. Lisäksi käytetty leksikko on luokiteltu ammatillisen erityisyyden mukaan erikoisammattisanastoon, terveyssanastoon ja yleissanastoon. Toisessa vaiheessa tutkimus keskittyy käytettyyn yleissanastoon, jota tarkastellaan frekvenssiin, käytetyn sanaston laajuuteen ja toiston määrään perustuvilla menetelmillä. Frekventeintä sanastoa esitellään sanaluokittain. Käytettyä leksikkoa on suhteutettu ulkoiseen taajuussanastoon, ja teksteistä on laskettu leksikaalista diversiteettiä kuvaavia tunnuslukuja (TTR, MTLD, Shannonin entropia). Lopuksi tutkimus tarkastelee diversiteetti-indeksien, lekseemien frekvenssien avulla laskettujen kumulatiivisten prosenttien ja laillistamiskokeen pistemäärien tilastollista yhteyttä.

Tutkimuksen mukaan kuulusteluun osallistuneiden lääkärien sanastotaidot vaihtelivat huomattavasti ja korreloivat koemenestyksen kanssa selvästi. Yleissanaston perusteella kokeessa heikoimmin menestyneillä sanasto oli muita köyhempää ja frekventimpää, kun taas hyvin suoriutuneet käyttivät yleensä monimuotoisempaa sanastoa. Suomen kaikkein yleisimmän sanaston käytön määrä oli käänteisessä korrelaatiossa koemenestyksen kanssa. Lisäksi kokeen hyväksytysti läpäisseet käyttivät keskimäärin harvinaisempaa sanastoa kuin kokeessa hylätyt.

Avainsanat: sanasto, sanastotaidot, leksikaalinen diversiteetti, suomi toisena kielenä, lääkärit

Luettelo taulukoista ja kuvaajista

SISÄLLYS

1 JOHDANTO

1.1 Tutkimuksen taustaa

Suomi on kansainvälistynyt muun maailman mukana, ja työvoimaa liikkuu maasta toiseen enemmän kuin aikaisemmin. Myös terveydenhuollon ammattihenkilöitä, kuten lääkäreitä, saapuu Suomeen muualta entistä enemmän. Vuonna 2010 Suomessa koulutettuja lääkäreitä laillistettiin kaikkiaan 665, kun samalla ulkomailla koulutuksen saaneita laillistettiin jo 374. (Paunio & Pelkonen 2012: 12.) Useampi kuin joka kolmas Suomessa vuonna 2010 laillistetuista lääkäreistä oli siis saanut koulutuksensa ulkomailla. Vuosina 1994–2009 Suomeen tuli lääkäreitä kaikkiaan 65 maasta, ja puolet tulijoista oli venäläisiä (Haukilahti & Virjo & Mattila 2010).

Kun kaksi valelääkäritapausta tuli julkisuuteen syksyllä 2011, keskusteluihin nousi huoli paitsi ulkomailta Suomeen saapuneiden lääkäreiden pätevyyden valvonnasta myös heidän kielitaidostaan. Esiin tuli muun muassa kokemuksia, joiden mukaan ulkomailta saapuneen lääkärin kielitaito ei olisi riittävä lääkärin ammatin harjoittamiseen Suomessa. Kielitaidosta johtuvia ongelmia on koettu paitsi vastaanotolla myös potilasasiakirjoja lukiessa (Paunio & Pelkonen 2012: 19). Esimerkiksi reseptejä on kirjoitettu jopa niin epäselvästi, että niitä tulkitsemaan on tarvittu toinen lääkäri (emt.). Varsinaisia kanteluita ei ole juuri tehty kielitaidottomuusepäilyjen takia, mutta joidenkin lääkäreiden potilaat ja kollegat ovat soitelleet huolestuneina valvontaa tekeville viranomaisille, ja jotkut ovat ottaneet yhteyttä myös Lääkäriliittoon (Sariola 2012: 926). Maija Tervolan (2017) toteuttamissa maahanmuuttajataustaisten lääkäreiden kanssa työskennelleiden terveydenhuollon ammattilaisten ryhmähaastatteluissa tuli ilmi, että maahanmuuttajalääkärien kielitaito koetaan usein puutteelliseksi ja sen katsotaan kuormittavan työyhteisöjä sekä vaarantavan jopa potilasturvallisuuden. Ongelmallisimmiksi tilanteiksi koetaan potilaiden ja maahanmuuttajalääkärien väliset keskustelut sekä puhelinkonsultointi (emt., 198–203). Tervola (emt., 203–205) katsoo, että ulkomailta tulevat lääkärit tarvitsisivat nykyistä enemmän tukea juuri kielitaitopuutteisiin liittyvissä asioissa.

Suomen lain mukaan terveydenhuollon ammattihenkilöllä täytyy olla hänen hoitamiensa tehtävien edellyttämä riittävä kielitaito (Finlex 1: 10 §, 13 §, 14 a §, 18 a §). Asetuksen mukaan riittävänä on pidettävä ”ammatinharjoittamisluvan tai -oikeuden edellyttämien tehtävien hoidon kannalta välttämätöntä kielitaitoa” (Finlex 2: 14 §). Potilaslain mukaan potilaalle on annettava selvitys hänen terveydentilastaan ja esimerkiksi hoidon merkityksestä sekä eri hoitovaihtoehdoista niin, että potilas ymmärtää sen sisällön riittävästi. Jos esimerkiksi lääkäri ei osaa potilaan käyttämää kieltä tai potilas ei voi tulla ymmärretyksi aisti- tai puhevian vuoksi, tulkitsemisesta on huolehdittava mahdollisuuk-

sien mukaan. (Finlex 3: 5–6 §.) Potilasasiakirjoja koskeva asetus (Finlex 4: 7 §) puolestaan koskee esimerkiksi epäselviä reseptejä. Sen mukaan potilasasiakirjoihin pitää merkitä potilaan hyvän hoidon järjestämisen, suunnittelun, toteuttamisen ja seurannan turvaamiseksi tarpeelliset ja laajuudeltaan riittävät tiedot. Lisäksi merkintöjen on oltava selkeitä ja ymmärrettäviä, ja niitä tehtäessä saa käyttää vain yleisesti tunnettuja ja hyväksyttyjä käsitteitä ja lyhenteitä (emt.).

Aiemmin Euroopan unioniin tai Euroopan talousalueeseen ETA:an kuuluvassa maassa koulutuksen saaneen lääkärin riittävän kielitaidon arviointi oli vain työnantajan vastuulla, kun taas ETA-maiden ulkopuolelta koulutettujen oli osoitettava riittävä kielitaito suorittamalla virallinen kielikoe (ks. Paunio & Pelkonen 2012: 16–17). Lääkärien laillistamismenettelyä ja ammattipätevyyksien valvontaa arvioinut terveydenhuollon valvontatyöryhmä esitti helmikuussa 2012, että käyttöön tulisi ottaa erityinen ammatillinen kielitaitokoe. Lisäksi työryhmä esitti, että EU ja ETA-alueen ulkopuolella tutkintonsa suorittaneen lääkärin kielitaito tulisi arvioida jo ennen lääkäriharjoittelijana toimimista, jota laillistaminen nykyisin edellyttää. (Paunio & Pelkonen 2012.)

Sittemmin tilanne on joiltain osin muuttunut. Kielitaitovaatimuksiin tehtäviä tiukennuksia alettiin valmistella Sosiaali- ja terveysalan lupa- ja valvontavirasto Valvirassa vuonna 2012 (Sariola 2012), ja lääkäreiden kielitaitoa koskeva lainsäädäntö kiristyi hieman vuoden 2016 alusta alkaen (Finlex 5). Nykyisin Valvira voi halutessaan vaatia kielitaidon todistamista myös EU- ja ETA-maissa tutkintonsa suorittaneilta lääkäreiltä ennen ammatinharjoittamisluvan myöntämistä (Finlex 5: 8 b §, Valvira 2018b). Lisäksi viime vuosina ulkomailta tulevien lääkäreiden ammatillista kielitaitoa on alettu kehittää mallilla, jossa yhdistyvät kielenoppiminen ja ammatillinen paneutuminen (Tampereen yliopisto 2015).

1.2 Lääkärien laillistamiskoe

Ulkomailla koulutuksensa saaneet ja tutkintonsa suorittaneet lääkärit eivät siis saa oikeutta ammattinsa harjoittamiseen Suomessa automaattisesti, vaan heidän täytyy erikseen hakea lääkärinoikeuksia Valviralta. EU- ja ETA-maiden ulkopuolelta Suomeen saapuvilta lääkäreitä vaaditaan suorittamaan lisäksi kolmiosainen laillistamiskoe. (Valvira 2018a, Valvira 2018b) Tämä koskee sekä ulkomaalaistaustaisia lääkäreitä että tutkintonsa ulkomailla suorittaneita suomalaisia. Maksullista koetta järjestää ainoana Suomessa Tampereen yliopiston lääketieteen yksikkö, tarkemmin sanottuna yleislääketieteen oppiaine. Näistä lääkärinammatin harjoittamisoikeutta hakevista käytän tässä tutkielmassani nimityksiä *kokeeseen osaa ottanut lääkäri*, *lääkäri* ja *koehenkilö* sekä muita vastaavia ilmauksia. Heidät on luokiteltu taustansa perusteella kahteen ryhmään: ulkomaalaistaustaisia kutsun

ei-natiiveiksi ja muita, esimerkiksi EU- ja ETA-maiden ulkopuolella tutkintonsa suorittaneita suomalaisia *natiiveiksi* lääkäreiksi tai koehenkilöiksi. Natiivien taustasta ei ole tässä yhteydessä tarkempaa tietoa. Heidän joukossaan saattaa olla yhtä lailla esimerkiksi kaksoiskansalaisia, jotka eivät ole aiemmin edes asuneet Suomessa, ja Suomesta maailmalle opiskelemaan lähteneitä.

Kuulusteluihin saapuessaan lääkäreillä on takanaan puolen vuoden harjoittelu suomalaisessa sairaalassa ja he ovat suorittaneet hyväksytysti suomen tai ruotsin kielen kielikokeen: joko Valtionhallinnon kielikokeen vähintään tyydyttävällä tasolla tai Yleisen kielitutkinnon vähintään taitotasolla kolme. He osallistuvat Tampereen yliopistossa kahteen kirjalliseen kokeeseen ja niiden hyväksytyn suorituksen jälkeen käytännön potilastenttiin. Ensimmäinen kirjallinen koe on kliininen kuulustelu, jonka kysymykset liittyvät potilastapauksiin. Toinen on suomalaisen terveydenhuollon kuulustelu, jonka osiot ovat sosiaalilääketiede, oikeuslääketiede ja reseptioppi. Käytännön potilastentissä lääkäri ottaa vastaan aitoja potilaita terveyskeskuksessa.

Tampereen yliopiston yleislääketieteen oppialan järjestämät tentit on vuosina 1994–2017 suorittanut hyväksytysti yhteensä 731 lääkäriä. Lisäksi Valvira (aiemmin Terveydenhuollon oikeusturvakeskus TEO) on tuona aikana laillistanut 122 lääkäriä, jotka ovat suorittaneet yhden tai kaksi tenttiä. Syynä osasuorituksiin ovat voineet olla maan liittyminen EU:hun tai potilastentin korvaaminen terveyskeskustyöskentelyllä, mikä aiemmin oli mahdollista. (Tampereen yliopisto 2018a.) Kuulusteluihin osallistuneista venäläisten suhteellinen osuus on kasvanut viime vuosina, sillä Virossa ja muissa EU:hun liittyneissä maissa lääkäriksi opiskelleiden ei ole enää tarvinnut ottaa osaa kokeeseen (Haukilahti ym. 2010).

Kuulusteluun osallistuvista huomattava osa, jopa 60 prosenttia, on viime vuosina hylätty (Suomen lääkärilehti 2012). Tämä tutkielma sai ideansa ja alkunsa Tampereen yliopiston tutkimushankkeesta, joka pyrki selvittämään, millainen yhteys ulkomaalaisten lääkäreiden kielitaidolla on heidän suoriutumiseensa laillistamiskokeessa. Tutkielman oli tarkoitus valmistua jo aiemmin niin, että sen havainnot ja tulokset olisivat olleet jo tutkimushankkeen toteuttajien käytettävissä, mutta kirjoittajan henkilökohtaisista syistä johtuen tutkielman valmistuminen on viivästynyt. Sillä välin hanke on tuottanut jo ainakin yhden julkaistun tutkimuksen (Tervola & Pajunen & Vainio & Honko & Mattila 2015) sekä pro gradu -tutkielman (Ruokolainen 2015), ja lisäksi Maija Tervolalta on tätä kirjoittaessa valmistumassa väitöskirja maahanmuuttajataustaisten lääkäreiden suomen kielen taidosta. Jo julkaistuista laillistamiskoetta koskevista tutkimuksista Tervolan ym. (2015) tutkimus on tämän tutkielman kannalta toistaiseksi relevantein ja ansaitsee erityistä huomiota.

Tervolan ym. tutkimus toteutettiin osittain samanaikaisesti tämän tutkielman kanssa ja osin samalla aineistolla. Sen tavoitteena oli selvittää, selittyvätkö laillistamiskokeen suuret hylkäysprosentin siihen osaa ottaneiden lääkäreiden kielitaidon puutteilla. Sitä varten Tervola ym. analysoivat

kokeen vastaustekstit kahdella kielitaitomittarilla. He erittelivät erityisesti kielen rakenteellista hahmottamista ja käytetyn sanaston yleisyyttä sekä vertasivat tuloksiaan koehenkilöiden kuulustelumenestykseen. Kielen rakenteellista hahmottamista varten he analysoivat kirjoitustaitoa sanahahmon hallinnan, lausehahmon hallinnan, ymmärrettävyyden ja koherenssin kriteereillä. Sanaston yleisyyttä he puolestaan tarkastelivat luokittelemalla käytetyn sanaston 11 yleisyystasoon, joista he muodostivat sanoille kolmiportaisen yleisyysasteikon. (Tervola ym. 2015: 340–341.)

Kielitaitoanalyysien tulokset korreloivat kuulustelumenestyksen kanssa niin, että kirjoitustaidoiltaan hyviksi arvioidut lääkärit menestyivät laillisuuskuulustelussa paremmin kuin ne, joiden taidot oli arvioitu heikoiksi. Kuulusteluissa hyvin menestyneet käyttivät vastauksissaan enemmän harvinaisia ja merkitykseltään täsmällisiä sanoja kuin heikosti menestyneet. Lopulta Tervola ym. päättelevät tutkimuksensa perusteella, että osa laillistamiskokeeseen osaa ottaneista lääkäreistä ei täyttänyt pakollisessa kielikokeessa määriteltyjä vähimmäisvaatimuksia, vaikka he olivat läpäisseet sen ennen laillistamiskoetta. He päättelevät, että lääkärien kielitaidon testaamista olisi syytä kehittää sen pätevyyden ja luotettavuuden näkökulmasta. (Emt., 342–344.)

1.3 Näkökulma ja tavoitteet

Oman mielenkiintoni erityinen kohde on ulkomailta Suomeen tulevien lääkäreiden kirjallisessa kokeessaan käyttämä sanasto. Sanaston eritteleminen tuo tärkeän näkökulman kielitaidon arviointiin, koska sanat ovat ne perusyksiköt, joiden varaan kieli ja kielitaito rakentuvat (esim. Read 2000: 1). Tavoitteeni on selvittää tarkemmin erityisesti sanaston yleisyyttä ja monimuotoisuutta, johon viitataan erilaisilla toisiaan lähellä olevilla käsitteillä, kuten leksikaalinen rikkaus, leksikaalinen diversiteetti ja leksikaalinen variaatio. Uskoakseni tämänkaltainen tarkastelu voi antaa lisävalaistusta esimerkiksi Tervolan ym. (2015) saamille tuloksille laillistamiskokeeseen osaa ottaneiden kielitaidosta ja suomen kielen taitojen yhteydestä kokeessa suoriutumiseen.

Tarkoitukseni on tehdä deskriptiivinen analyysi sanastosta, jota suomenkieliseen kokeeseen osallistuneet lääkärit ovat käyttäneet kokeen ensimmäisessä kirjallisessa osassa. Otteeni on kvantitatiivinen, enkä arvioi aineistossani olevia tekstejä holistisesti. Esittelen muun muassa tekstien sanaluokkajakauman ja yleisimmät käytetyt sanat sekä katson, kuinka suuren osan koko sanastosta kattaa yleisimpien sanojen kärki. Tarkastelen muun muassa vastauksissa käytetyn sanaston diversiteettiä ja suhteutan lekseemien esiintymistaajuutta Suomen sanomalehtikielen taajuussanastoon (CSC 2004). Tarkastelen myös lääkäreiden käyttämien sanastojen välistä variaatiota. Erittelen, millaisia sanastonkäytön eroja on havaittavissa natiiveiksi ja ei-natiiveiksi tulkittujen koehenkilöiden välillä.

Analyysi tuottaa sanastollisista taidoista kertovia tunnuslukuja, ja odotukseni on, että natiivien sanastotaidot ovat niiden perusteella paremmat kuin ei-natiivien. Lisäksi vertaan laskemiani tunnuslukuja siihen, kuinka koehenkilöt ovat menestyneet laillistamiskokeessa.

1.4 Tutkielman rakenne

Ensimmäisessä luvussa olen selostanut tutkielmani lähtökohtia ja hahmotellut sen näkökulmaa ja rakennetta. **Toisessa luvussa** esittelen tarkemmin tutkielman teoreettista pohjaa ja tärkeimmät käyttämäni käsitteet. Siinä otan esille keskeisiä näkökulmia ja esimerkkejä siitä, millä tavoin sanastotaitoja ja käytetyn sanaston monipuolisuutta on aikaisemmin tutkimuksissa tarkasteltu ja mitä sanastotaitoja tutkiessa tulisi ottaa huomioon.

Kolmannessa luvussa kerron pohjatietoja käyttämästäni laillistamiskoeaineistosta ja selostan, millä tavalla se on käsitelty analysoitavaan muotoon. Sen jälkeen esittelen omat tutkimuskysymykseni, näkökulmat, joista lähden laatimaan deskriptiivistä analyysiä, sekä tutkielmassa käyttämäni leksikaalisen diversiteetin tunnusluvut ja niiden laskutavat.

Neljäs luku on tutkielmani analyysiluku. Alaluvussa 4.1 esittelen aineistoa eri tavoin ja tarkastelen laillistamiskokeessa käytettyä kieltä erityisyysasteen, sanaluokkien osuuksien, sanaston kompleksisuuden, sanepituuksien ja käytetyn leksikon yleisyyden näkökulmasta. Alaluvussa 4.2 selostan, kuinka olen karsinut aineistoa sen tarkempaa analyysiä varten. Alaluvussa 4.3 pääsen kiinni tutkimukseni tarkemman analyysin osaan, jossa aineistosta on karsittu pois terveysalan ammattisanasto. Siinä tarkastelen vastaustekstien sane- ja lekseemimääriä, lasken suomen vastausteksteistä suomen frekventeimmän sanaston osuuksia ja esittelen käytetyn sanaston yleisyyttä kumulatiivisten prosenttien avulla. Alaluvussa 4.4 esittelen frekventeintä sanastoa sanaluokittain ja nostan esiin tärkeimpiä sanalistoihin liittyviä havaintojani. Tutkielmani ydin on nähdäkseni alaluku 4.5. Siinä esittelen käytetyn sanaston diversiteetin tunnuslukuja sekä taajuussanaston (CSC 2004) avulla vastauksista laskettuja suomen yleisimmän sanaston kumulatiivisia prosentteja. Havainnollistamisen vuoksi olen esittänyt kvantitatiivisen analyysini tuloksia luokittelemalla kokeeseen osaa ottaneet lääkärit viiteen tasoryhmään. Vertailen myös natiivien ja ei-natiivien ryhmien suoriutumista. Luvussa 4.6 analysoin tilastollisin menetelmin laillistamiskokeen aineiston leksikaalisen diversiteetin ja kokeessa menestymisen välistä yhteyttä.

Viides luku on tutkielman yhteenvetoluku. Siinä kertaan ensiksi tutkimukseni tärkeimmät tulokset. Toiseksi arvioin työni luotettavuutta ja merkitystä. Kolmanneksi pohdin, millä tavalla tutkimusta voisi jatkaa.

2 TEOREETTINEN VIITEKEHYS

2.1 Sana, lekseemi, sananmuoto, sane

Keskeisimmät käsitteet, joista lähden liikkeelle, ovat **sana**, **lekseemi**, **sananmuoto** ja **sane**. Näistä sana on primitiivikäsite, joka voi jo kirjoituksessa merkitä ainakin kolmen eri tason yksikköä: 1) sanaa abstrahoituna tyyppinä, 2) sanaa abstrahoituna taivutettuna muotona tai 3) sanaa tekstin yksikkönä, sanaesiintymänä (Niemikorpi 1991: 21; Penttilä 1963: 115–118). Penttilä (emt.) on käyttänyt näiden erottamiseksi kolmea termiä: sana, sananmuoto ja sane. Itse en halua aiheuttaa sekaannuksia viljelemällä joka yhteydessä yleiskielen primitiivikäsitettä sana, ja pyrin selvyyden vuoksi käyttämään termejä lekseemi, sananmuoto ja sane, joista lekseemi ja sane ovat tutkimukseni kannalta olennaisimpia. Toki sanoista puhun edelleen ajoittain tilanteissa, joissa katson, ettei väärinkäsityksen vaaraa ole.

Lekseemillä tarkoitan tässä tutkielmassani siis tietyn sanan abstraktiota, eli abstrahoitua tyyppiä, joka kattaa sanan kaikki taivutusmuodot. Toisin sanoen lekseemi on yhteen kuuluvien sananmuotoesiintymien luokka. Esimerkiksi englannin substantiivi voi toteutua neljänä eri graafisena muotona (*car, cars, car's, cars'*) ja ruotsin substantiivi kahdeksana muotona. Suomen substantiivi voi toteutua kielen morfotaktisten positioluokkien määrän vuoksi laskutavasta riippuen jopa 1500–2000 eri sananmuotona. Käytännön sanakirjoissa lekseemin kaikkia sananmuotoja edustaa sanakirjamuoto eli hakumuoto, joka suomessa on esimerkiksi nomineilla yksikön nominatiivi ja verbeillä ensimmäinen infinitiivi. (Karlsson 2008: 85–86, 187.) Tässä työssä sananmuodostuskeinoilla luodut sanat, kuten johdokset ja yhdyssanat, on määritelty omiksi lekseemeikseen. Muitakin lekseemiä vastaavia käsitteitä ja termejä on käytetty. Esimerkiksi englanninkielisessä tutkimuskirjallisuudesta puhutaan lekseemin sijaan usein lemmasta, kun viitataan lekseemin tavoin sanan abstrahoituun muotoon (esim. Read 2000: 18). Hongon (2013: 54) mukaan lemman määritelmä on tyypillisesti varsin lähellä lekseemin määritelmää. Tässä tutkielmassa suosin termiä lekseemi, ja käytännössä lemmoista puhuessani viittaan taajuussanastoissa listattuihin lekseemeihin ja lemmaamisella siihen sanastontutkimukselliseen työhön, jossa eri sanaesiintymät lasketaan saman lekseemin (tai lemman) esiintymiksi (ks. esim. Read 2000: 18).

Sane puolestaan on yksittäinen konkreettinen tekstitason yksikkö sekä tekstissä ilmenevä lekseemin ja yhtä aikaa sananmuodon esiintymä. (Niemikorpi 1991: 21–22; Karlsson 2008: 85–86.) Jos siis sama sana toistuu tekstissä eri muodoissa, tekstissä on vähemmän eri lekseemejä kuin eri saneita. Lekseemien ja saneiden välistä suhdetta tarkastelen laskemalla lääkärien laillistamiskokeen

vastausteksteistä muun muassa sana–sane-suhteet, eli käytännössä lekseemien ja saneiden osamäärät (type–token ratio, TTR). Kun lasken erilaisia diversiteettilukuja, lekseemi vastaa siis termiä *type* ja sane termiä *token* siinä mielessä kuin esimerkiksi McCarthy ja Jarvis (2010: 382) käyttävät noita termejä laskiessaan sanastosta kertovia tunnuslukuja. Analyysia varten olen laskenut myös muita sanastoa kuvaavia, lekseemien frekvenssiin perustuvia diversiteettilukuja, ja niistä olen poiminut tähän tutkielmaan Shannonin entropian (Shannon 1948) ja MTLD:n (McCarthy & Jarvis 2010) sekä kumulatiivisia prosentteja, jotka kertovat käytetyn sanaston yleisyysasteesta. Näitä erilaisia tunnuslukuja käsittelen erikseen tutkielman luvussa 3.

2.2 Toisen kielen oppiminen

Ihmisen ensimmäiseksi oppimaa kieltä kutsutaan tavallisesti äidinkieleksi tai ensikieleksi, ja sen omaksuminen on yhteydessä kielenoppijan yleiseen kehitykseen. Ensikieli opitaan vuorovaikutuksessa ympärillä olevan yhteisön kanssa, kuten muukin kulttuuri, ja kielen oppimisen keskeinen edellytys on päästä täysivaltaisena jäsenenä mukaan ympäristön sosiaalisiin käytänteisiin. (Sajavaara 1999: 73–74.) Muille kuin äidinkielisille suomen puhujille suomi on joko toinen kieli tai vieras kieli. Ilmaisua *toinen kieli* käytetään, kun puhutaan erilaisista Suomeen muualta muuttaneista tai muista vastaavanlaisista ihmisistä, jotka joutuvat opettelemaan suomea suomenkielisessä ympäristössä opetuksen avulla tai ilman opetusta. *Vieraasta kielestä* puhutaan silloin, kun suomi on esimerkiksi oppiaineena koulussa tai yliopistossa muussa kuin suomenkielisessä ympäristössä. Silloin opetus- tai itseopiskelutilanteet ovat ainoita tai lähes ainoita suomen kielen käyttöhetkiä. (Ks. esim. Martin 1999: 157, 172.) Nähdäkseni tutkimiani lääkärien laillistamiskokeita tehneet muut kuin suomenkieliset koehenkilöt opettelevat, puhuvat ja kirjoittavat suomea nimenomaan toisena kielenä.

Niihin, joille suomi on toinen kieli, viitataan usein esimerkiksi termillä S2-oppija. Universaalimmin toiseen kieleen viitataan vastaavalla tavalla tavallisesti lyhenteellä L2.

Siitä, kuinka äidinkielen jälkeen opitaan muita kieliä, vallitsee erilaisia teorioita ja toisistaan poikkeavia näkemyksiä (ks. esim. Sajavaara 1999: 76–97). Sajavaara (1999: 73–74) katsoo, että toinen kieli omaksutaan samalla tavalla kuin ensimmäinenkin, jos kielenoppija osallistuu samalla tavalla kielelliseen vuorovaikutukseen kuin ensikielellä. Martinin (1999: 161) mukaan voidaan kuitenkin olettaa, että ensimmäinen ja toinen kieli ovat olemukseltaan erilaisia ja että ne myös yleensä opitaan eri tavoilla. Hän otaksuu, että kielenoppijan oman kielen erot ja yhtäläisyydet opittavaan toiseen tai vieraaseen kieleen vaikuttavat siihen, kuinka helposti hän tuota kohdekieltä omaksuu. Tosin on huomattava, että osin oppijoiden välinen variaatio johtunee heidän ominaisuuksistaan sekä

aiemmista oppimiskokemuksistaan ja myös kulttuurinen etäisyys kohdekielen puhujiin vaikuttaa siihen, kuinka tuttuja kielenkäyttötavat ja ilmausten merkitykset ovat (emt.: 167–168).

Martin (1999) otaksuu, että kielenoppijan oman kielen erot ja yhtäläisyydet opittavaan toiseen tai vieraaseen kieleen vaikuttavat siihen, kuinka helposti hän tuota kohdekieltä omaksuu. Tosin on huomattava, että oppijoiden välinen variaatio johtunee osittain heidän ominaisuuksistaan ja aiemmista oppimiskokemuksistaan ja että myös kulttuurinen etäisyys kohdekielen puhujiin vaikuttaa siihen, kuinka tuttuja kielenkäyttötavat ja ilmausten merkitykset ovat (emt.: 167–168). Martinin (emt.) mukaan uudesta kielestä oppii helpoimmin ne asiat, jotka lähtökielessä ilmaistaan usealla mutta opittavassa kielessä vain yhdellä tavalla sekä luonnollisesti ne, joita opittavassa kielessä ei tarvitse huomioida lainkaan. Esimerkiksi suomen kielessä ei ole tarvetta opetella artikkelijärjestelmiä tai kieliopillisia sukuja eikä suomessa ole paljon erilaisia konsonantteja. Vaikka suomen kieli mielletään usein hankalaksi oppia, Martin (emt.) huomauttaa sen olevan melko läpinäkyvää sikäli, että oppimisen edistyessä pitkistä, aluksi vaikeaselkoisista sanoista alkaa erottaa yhdyssanojen osia ja johdoksia, mikä helpottaa ymmärtämistä. Toisaalta hankaluuksia tuo esimerkiksi morfologisten muotojen runsaus ja taivutusvartaloiden vaihtelu (emt.).

Jo arkikokemukset osoittavat, ettei ihminen välttämättä omaksu uutta kieltä helposti, vaikka hän eläisi maahanmuuttajana muiden kielenpuhujien joukossa. Tähän voi olla eri syitä. Sajavaara (1999: 74–75) huomauttaa, että toisen kielen käyttöalue saattaa jäädä ensimmäistä suppeammaksi siksi, että rajallinenkin kielitaito riittää yleensä elämän perustarkoituksiin. Lisäksi kielen oppimisen rajat riippuvat vuorovaikutuskäytänteiden laadusta, ja esimerkiksi toisen kielen käyttöalue saattaa jäädä paljon ensikieltä suppeammaksi ja yhteydet natiivipuhujiin voivat jäädä vähäisiksi (emt.).

Yleensä kieli kehittyy nuoruudessa ja aikuisuudessa sen perustan varaan, joka on opittu aiemmin lapsena. Kielen varhainen kehitys eroaa myöhemmästä sen nopeudeltaan, tärkeydeltään ja sisällöltään. Lapsuuden jälkeen kielenoppija oppii ymmärtämään ja käyttämään laajemmin esimerkiksi kuvallisia ilmaisuja sekä frekvenssiltään harvinaisempia tapoja sanoa. (Nippold 2007: 11; 1993.) On huomattava, että niillä kielenoppijoilla, jotka alkavat opetella toista kieltä vasta nuorena tai aikuisena, ei ole vastaavaa lapsena opittua kielen pohjaa, jonka varaan rakentaa, ja se tuo heille omat haasteensa. Esimerkiksi maahanmuuttaja voi joutua opettelemaan kielen perusteet verrattain nopeasti ja jopa yhtä aikaa, kun hän tavallaan joutuu alkamaan elää kielellisestikin suoraan aikuisten maailmassa.

2.3 Sanaston oppiminen ja sen merkitys

Kielenoppijan sanavarasto karttuu kolmella tavalla. Ensinnäkin esimerkiksi vanhempi tai opettaja voi opettaa kielenoppijalle uuden sanan ja sen merkityksen suoraan. Toiseksi oppija voi itse päätellä oudon sanan merkityksen kontekstista, kuten muista sanoista ja käyttöympäristöstä. Kolmanneksi oppija voi analysoida sanan osia ja päätellä merkityksen sanan sisältämistä morfeemeista ja niiden merkityksistä. (Nippold 2007: 29–35.) Siitä on tehty eritasoisia tulkintoja siitä, milloin oppijan katsotaan osaavan sanan, mutta esimerkiksi erään varsin tiukan määritelmän mukaan sana osataan vasta sitten, kun sen merkitys hallitaan sekä kontekstissa että ilman ja kun sitä osataan käyttää sujuvasti sekä tilanteeseen sopivalla tavalla (Channell 1988: 94–95).

Sanastonhallintaa on perusteltua tarkastella osana kielitaidon arviointia, sillä sanat ovat kielen peruspalikoita, sellaisia merkityksen yksiköitä, joista muodostuvat niitä laajemmat rakenteet, kuten lauseet, kappaleet ja kokonaiset tekstit (Read 2000: 1). Nationin (1993) mukaan sanastotaidot ja kielenkäyttö muodostavat kielitaitoa kasvattavan kehän: sanastotaidot mahdollistavat kielenkäytön, kielen käyttäminen puolestaan mahdollistaa sanastotaitojen kasvun. Niiranen (2004) on havainnut, että sanaston laajuudella on yhteys esimerkiksi taitoihin taivuttaa sanoja.

Tekstinymmärtämisen taidot liittyvät läheisesti myöhempään kielellisten taitojen kehitykseen ja aivan erityisesti sanaston hallintaan. Tätä koskevia tutkimuksia on käynyt läpi Marilyn A. Nippold, myös toisen kielen oppijoiden osalta (Nippold 2007: 1–14, 25–47). Esimerkiksi Iranissa toteutetussa tutkimuksessa (Beleghizadeh & Golbin 2010) todettiin, että opitun vieraan kielen sanaston laajuus oli yhteydessä luetun ymmärtämisen tasoon. Burundissa tehdyssä tutkimuksessa taas osoitettiin selvästi, että englantia opiskelevien sanastolliset taidot korreloivat sen kanssa, kuinka hyvin he ylipäätään taitavat englantia (Nizonkiza 2011). Kellyn (1991) mukaan sen jälkeen, kun vieraan kielen oppijat ylittävät taidoissaan keskitason, suurimmaksi kuullun ymmärtämisen esteeksi nousevat nimenomaan puutteet sanastonosaamisessa.

2.4 Kuinka paljon sanastoa tarvitaan?

Suomen kielen perussanakirja, joka kattaa suomen sanaston ytimen, sisältää noin 100 000 hakusanaa, ja suomen murteiden sanakirjaan on arvioitu tulevan kaikkiaan 350 000 hakusanaa (Kotimaisten kielten keskus 2017). Kaikkia näistä ei toki tarvitse osata, jotta voisi ymmärtää ja tuottaa suomea sujuvasti, vaan olennaista on tuntea riittävästi yleisimmin tarvittavia sanoja.

Siitä, kuinka paljon sanastoa tarvitsee osata kieltä ymmärtääkseen, on tehty tutkimusta erityisen paljon englannin kielestä ja sen oppijoista. Esimerkiksi Lauferin ja Ravenhorst-Kalovskin (2010) mukaan 8000 yleisintä englannin sanaa kattaisi 98 prosenttia tavanomaisista englanninkielisistä teksteistä ja riittäisi hyvään luetun ymmärtämiseen. Minimissään luetun ymmärtäminen edellyttää heidän mukaansa 4000–5000 sanaa, joka kattaisi 95 prosenttia teksteistä. Luvut sisältävät myös erisnimet. Englanninkielisen kaunokirjallisuuden avulla analyysiä tehneiden Hirschin ja Nationin (1992) mukaan yleisimmät 2000 sanaa eivät riitä siihen, että tekstin lukeminen sujuisi miellyttävästi niin, että tuntemattomien sanojen pystyisi riittävän hyvin arvaamaan tai päättelemään. He katsovat, että se edellyttäisi 5000 yleisimmän sanan tuntemista, joka riittäisi siihen, että outoja sanoja tulisi vastaan noin kolmen–viiden rivin välein. Siinä tapauksessa outojen sanojen merkityskin olisi melko hyvin arvattavissa tai pääteltävissä. Hirschin ja Nationin (1992) analyysin perusteella 5000 yleisimmällä englannin kielen sanalla yltäisi 97–98 prosentin peittoon tavallisessa kaunokirjallisessa tekstissä. Nationin (2006) mukaan 8000–9000 sanaperheen tunteminen riittäisi kattamaan 98 prosenttia kirjoitetusta ja 6000–7000 sanaperheen tunteminen saman osuuden puhutusta englannin kielestä. Hänellä sanaperheet (word-families) kattavat kuitenkin myös jonkin verran johdoksia (Nation 2006: 66–67), jotka usein luetaan myös erillisiksi opittaviksi yksiköikseen.

Antero Niemikorpi (1991) on tiivistänyt vuonna 1979 julkaistun Suomen kielen taajuussanaston (Saukkonen & Haipus & Niemikorpi & Sulkala 1979) tietoja ja koonnut taulukkoon siitä kumulatiivisia prosentteja. Kyseinen taajuussanasto on koottu 1960-luvun kauno- ja tietokirjallisuudesta, vuonna 1967 ilmestyneistä sanoma- ja aikakauslehdistä sekä radion puheohjelmista vuosilta 1968–1969. Niemikorven (1991: 53) mukaan sata suomen yleisintä lekseemiä kattaa noin 35 prosenttia, 1000 yleisintä noin 65 prosenttia ja 2000 yleisintä noin 74 prosenttia käytetystä kielestä. 5000 yleisintä lekseemiä yltää laskelman perusteella vasta runsaaseen 84 prosenttiin. Edes 9000 yleisintä lekseemiä ei Niemikorven (emt.) mukaan riitä peittämään 90 prosenttia käytetystä suomesta. Yleisimmät 10 000 lekseemiä riittävät siihen lähestulkoon hädin tuskin. Ero esimerkiksi Lauferin ja Ravenhorst-Kalovskin (2010) englannin kieltä koskeviin vastaaviin lukuihin on huomattava, ja se voi tarjota yhden mahdollisen selityksen sellaisille suomenoppijoille, joista tuntuu vaikealta saada suomi kunnolla haltuunsa. Toki suomenoppijoille helpotusta voi tuoda se, että suuri osa suomen leksikosta koostuu johdetuista sanoista. Sekin on kuitenkin otettava huomioon, että englannin ja suomen kieltä koskevat kumulatiiviset prosentit eivät yleensä ole suoraan vertailukelpoisia. Suomessa niiden laskemisessa käytetään yleensä lekseemejä, kun taas englannin kielessä puhutaan sanaperheistä, joita suomen kielestä kaikkine johtamisen mahdollisuuksineen ei ole mielekästä edes yrittää hahmottaa (ks. esim. Honko 2013: 54–56).

Yleisesti on otaksuttu, että toista tai vierasta kieltä puhuvilla passiivinen sanavarasto on aktiivista laajempi, minkä vuoksi kielenkäyttäjien voi olettaa ymmärtävän tekstiä monipuolisemmin kuin he pystyvät sitä aktiivisesti tuottamaan. Oletuksen passiivisen sanavaraston suuremmasta laajuudesta osoittaa todeksi esimerkiksi Lauferin ja Baribakhtin (1998) tutkimus, jossa tarkasteltavana oli joukko eritasoisia englantia vieraana tai toisena kielenä opiskelevia aikuisia. Opiskelijoista osa asui ja opiskeli Israelissa ja osa Kanadassa. Tutkimuksessa osoittautui myös, että toisen tai vieraan kielen oppijoilla aktiivinen sanavarasto kehittyi eri tahtia kuin passiivinen. Erilaisissa ympäristöissä ja eri aikoina toteutetun tutkimuksen mukaan erityisesti vapaassa itseilmaisussa käytettävä aktiivinen sanavarasto laajeni hitaammin ja vähemmän ennalta arvattavasti kuin passiivinen sanavarasto. Kaikilla testatuilla passiivinen sanavarasto oli selvästi laajempi kuin aktiivinen, mutta niiden laajuuden ero oli pienempi englantia vieraana kielenä opiskelleilla kuin sitä toisena kielenä opiskelevilla. Englantia toisena kielenä opiskelevilla englanninkielisen elinympäristön hyödyt alkoivat ilmetä aktiivisen sanavaraston laajentumisena vasta noin kaksi vuotta alueelle muuton jälkeen. (Laufer & Baribakht 1998.) Jotta lääkäri saisi tehtyä työnsä kunnolla, ei riitä, että pelkkä passiivinen sanavarasto on laaja, sillä hänen pitää pystyä keskustelemaan asiantuntevasti esimerkiksi potilaan oireista ja selittämään hankaliakin asioita potilaalleen – siis tuottamaan produktiivisesti tavallisen ihmisen ymmärrettävissä olevaa asiantuntijapuhetta ja -tekstiä.

2.5 Sanastotaitojen tutkimuksesta

Kielenoppijoiden sanastoa voi tutkia muun muassa kvantitatiivisesti kielen ekonomian, rikkauden, luettavuuden, tiiviyden ja sanataajuuden näkökulmasta. Viime vuosikymmeninä kvantitatiivisen tutkimuksen tekeminen on helpottunut ja monipuolistunut tietokoneiden hyödyntämisen ansiosta. (Niemikorpi 1991: 24–57.) Sanastollisia taitoja voi tutkia muun muassa erilaisin testein. John Read on kirjassaan *Assessing Vocabulary* (Read 2000) tehnyt laajan katsauksen erilaisten sanastotestien rakentamisesta ja hyödyllisestä käytöstä kielitaidon arvioinnissa. Tapoja arvioida kielenoppijoiden sanastotaitoja heidän tuottamiensa tekstien perusteella ovat tutkineet ja kehitelleet monet, esimerkiksi Laufer ja Nation (esim. 1995), Jarvis (esim. 2002) sekä McCarthy ja Jarvis (esim. 2007; 2010).

Sanastotaitojen tutkimuksessa on monenlaisia haasteita. Esimerkiksi Mari Honko (2013) huomauttaa, että täsmällistä ja varmasti kattavaa tietoa yksilön sanastonhallinnasta voidaan saada vain käymällä läpi kaikki hänen tuottamansa verbaalinen aineisto, mikä on käytännössä joko vaikeaa tai mahdotonta. Lisäksi Hongon mukaan yksilön tuottamaa kielenaineista analysoimalla voidaan

helposti saavuttaa vain produktiivinen sanasto ja siitäkin vain se osa, jonka hän on sattunut tuottamaan. Sanan produktiivinen hallitseminen ei puolestaan takaa reseptiivistä. (Honko 2013: 111.)

Read (2000) korostaa sanaston laajuuden tarkastelun merkitystä arvioitaessa kielenoppijan sanastotaitoja. Sanaston laajuuden tarkastelu puolestaan edellyttää Readin mukaan, että arvioijalla on käytössään kielestä laadukasta informaatiota sen lekseemeistä ja niiden yleisyydestä. Parhaimmaksi avuksi ovat selkeät lemmatut sanalistat, joissa kunkin lekseemin frekvenssi on selkeästi esitetty. (Emt.: 224–231.) Käytetyn sanaston yleisyys asettuu niiden avulla siis yhdeksi sanastotaitojen mittariksi. Itse käytän laillistamiskokeessa käytetyn sanaston yleisyyden tarkastelussani apuna vuonna 2004 koottua Suomen sanomalehtikielen taajuussanastoa (CSC 2004).

Suomessa sanastotaitojen tutkimusta on tehty erityisesti koululaisten kirjoitelmia analysoimalla (esim. Saarela 1997; Laine-Leinonen 2013) sekä suomi toisena kielenä -oppijoiden teksteistä (esim. Malin 2012). Mari Honko (2013) on tutkinut alakouluikäisten lasten sanastotaitoja ja verrannut keskenään toisen polven maahanmuuttajien sekä äidinkielisten suomenpuhujien leksikaalista osaamista. Kielitaidon kompleksistumista koskevassa väitöstutkimuksessaan (emt.) hän käytti kielitaidon arvioinnissa muun muassa sanojen frekvenssiin ja leksikaaliseen diversiteettiin perustuvia menetelmiä, jotka ovat relevantteja omankin tutkimukseni kannalta. Lääkärien pätevöitymiskokeiden tekstejä ovat vastikään tutkineet esimerkiksi Tervola, Pajunen, Vainio, Honko ja Mattila, jotka osoittivat yhteyden kokeeseen ottaneiden suomen kielen taidon ja koemenestyksen välillä (Tervola ym. 2015). Samaa laillistamiskuulustelun aineistoa on tarkasteltu melko vasta myös norminmukaisuuden näkökulmasta. Lääkäreiden kirjoittaessaan tekemiä normipoikkeamia tutkinut Ruokolainen (2015) havaitsi, että kielivirheiden yleisyyden ja heikon koemenestyksen välillä oli selvästi havaittava korrelaatio. Niistä, jotka tekivät eniten kielivirheitä, suurin osa ei saanut suoritettua koetta hyväksytysti.

2.6 Leksikaalinen rikkaus, variaatio ja diversiteetti

Sanaston rikkautta on mitattu monilla eri tavoilla. John Readin (2000: 198) mukaan **leksikaalinen rikkaus** (lexical richness) on yleiskäsite, jolla voidaan viitata tekstin sanaston vaihtelevuuteen, eli leksikaaliseen variaatioon, tekstin ilmaisutiiviyteen ja sanastossa tehtyjen virheiden määrään yhdessä. Näin ollen leksikaalisella rikkaudella on Readin mukaan monta eri ulottuvuutta. Tekstiä pidetään siis leksikaalisesti rikkaana, jos siinä on paljon leksikaalista variaatiota, se on leksikaalisesti sofistikoitunutta (sophisticated), jos se on ilmaisultaan tiivistä ja jos sanaston käytössä tapahtuneiden virheiden määrä on pieni (Read 2000: 200–201).

Tekstissä on paljon **leksikaalista variaatiota**, kun se on rakennettu käyttämällä pikemminkin paljon eri lekseemejä kuin toistamalla jatkuvasti samoja. Oletus tämän ajatuksen takana on, että leksikaalisesti rikkaiden tekstien kirjoittajilla on laajat sanastolliset tiedot, jotka auttavat heitä välttämään toistoa esimerkiksi synonyymejä käyttäen. (Read 2000: 200.) Esimerkiksi lekseemien ja saneiden osamäärää eli sana–sane-suhdetta (engl. type–token ratio, TTR) käytetään kuvaamaan, kuinka suurta on tekstin leksikaalinen variaatio, mutta mittari on hyvin karkea sekä puutteellinen, ja sille on kehitetty monipuolisempiakin korvaajia. Käsittelen asiaa tarkemmin luvussa 3.3.2.

Leksikaalinen sofistikoituneisuus eli hienostuneisuus tai vivahteikkuus (lexical sophistication) tarkoittaa sitä, kuinka paljon tekstin kirjoittaja on osannut käyttää kontekstiin sopivalla tavalla perussanastoa harvinaisempaakin sanastoa. **Leksikaalinen tiheys** (lexical density) puolestaan viittaa siihen, kuinka paljon kieliopillisia sanoja eli funktiosanoja, siis konjunktioita ja artikkeleita, on suhteessa sisältösanoihin. Esimerkiksi kirjoitettu kieli on tässä mielessä yleensä ilmaisultaan puhuttua tiiviimpää. (Read 2000: 200–201.) Readin (emt.) mukaan myös sanojen käytössä tapahtuneet virheet, kuten väärät taivutusmuodot tai tyylipoikkeamat, voidaan huomioida sanastotaitoja arvioitaessa. **Virheiden määrän** huomioimista kannattaisi harkita jo siitä syystä, että muuten esimerkiksi sanastotaitoja arvioitaessa täysin kontekstiin sopimattomat, virheellisesti ymmärretyt sanat saattavat parantaa kuvaa tekstin tuottaneen henkilön sanastotaidoista.

Leksikaalista diversiteettiä on käytetty kielen variaatiota kuvaavissa tutkimuksissa jopa leksikaalisen rikkauden synonyymina, mutta termien erot ovat sittemmin tarkentuneet. Nykyisin leksikaalisella diversiteetillä viitataan varsin vakiintuneesti erilaisiin tunnuslukuihin, joiden laskemiseen on käytetty tekstissä esiintyvien lekseemien ja niitä edustavien sananmuotojen määriä ja tilastollisia suhteita. (Jarvis 2013, Honko 2013: 106–107.) Esimerkiksi TTR on siis yksinkertainen leksikaalisen diversiteetin tunnusluku. Tiivistäen muotoiltuna käsite leksikaalinen diversiteetti viittaa tekstissä käytettyjen eri lekseemien valikoiman laajuuteen (McCarthy & Jarvis 2010: 381). Hongon (2013: 107) mukaan on hedelmällisintä käsitellä leksikaalista diversiteettiä leksikaalisen rikkauden osatekijänä ja rajata se tarkoittamaan leksikaalista toisteisuutta. Suomessa leksikaalista diversiteettiä ovat tässä mielessä käyttäneet mittarinaan Hongon (emt.) lisäksi muun muassa Saarela (1997), Malin (2012) ja Laine-Leinonen (2013).

Laufer ja Nation huomauttavat, että kielenoppijan tekstistä näkyvään sanaston diversiteettiin voivat vaikuttaa muutkin asiat kuin sen kirjoittajan oman sanaston laajuus. Niihin voivat lukeutua esimerkiksi tekstissä käsiteltävän aiheen tuttuus, kirjoittajan kirjoitustaito ja tekstin laatimisen syy. Näin ollen arvioihin ihmisen sanastotaidoista ja hänen tuottamansa tekstin variaatiosta tai rikkaudesta voi vaikuttaa jo se, mitä aihetta arvion pohjana olevassa tekstissä käsitellään. (Laufer & Nation 1995: 308.) Nähdäkseni oman kiinnostukseni kohteena oleva lääkärien pätevöittämiskoe on tar-

jonnut sellaiset kontrolloidut olot, joista saadut tekstit antavat luotettavan kuvan niiden kirjoittajien sanastotaidoista. Esimerkiksi tarkasteltavien tekstien aihe on niiden kirjoittajille tuttu, ja laillistamiskuulustelussa heillä on painetta ja oletettavasti tavoite pyrkiä hyvään kieleen ja ymmärrettävään ilmaisuun.

3 TUTKIMUSAINEISTO, -KYSYMYKSET JA -MENETELMÄT

3.1 Tutkimuksen aineisto

Tässä alaluvussa kerron perustietoja tutkimukseni aineistosta ja sen käsittelemisestä varsinaisen analyysin edellyttämään muotoon. Tarkemmin kuvailen koehenkilöiden käyttämää sanastoa ja analyysia varten tekemiä rajauksiani analyysiluvussa (luku 4).

3.1.1 Yleistä aineistosta

Lääkärien laillistamiskokeen ensimmäisen osa on kirjallinen kuulustelu, jonka tavoitteena on testata lääkärin perustietoja kliinisen lääketieteen ja terveydenhuollon aloilta. Kirjalliset kysymykset koostuvat potilastapausselostuksista tärkeimmiltä lääkärin erikoisaloilta, ja vaatimustasoltaan ne vastaavat suomalaista lääketieteen lisensiaatin koulutuksen loppukuulustelua. (Valvira 2018c.) Analysoimani korpus koostuu 46 koehenkilön kirjallisista koevastauksista. Tampereen yliopiston yleislääketieteen laitoksen työvoima on purkanut ne ensin käsin paperille kirjoitetusta muodosta tietokoneelle Word-tekstitiedostoksi. Suomen kielen professori Anneli Pajunen on koostanut aineiston (Pajunen 2013) Excel-taulukkomuotoon, ja Pajusen pitämälle suomen kielen syventävien opintojen Kielitieteen metodit -kurssille osallistuneet opiskelijat ovat koodanneet siitä harjoitustyönään muun muassa tavoitemuodot, sanaluokat ja normipoikkeamat tai virheet.[1] Olin itse yksi harjoitukseen osallistuneista opiskelijoista.

Koehenkilöistä kahdeksan on natiiveja suomen kielen puhujia ja loput 38 ovat ei-natiiveja, eli he puhuvat äidinkielenään jotain muuta kieltä kuin suomea ja siis oppivat suomea toisena kielenä. Kokeen luonteen ja kontekstin vuoksi vastaustekstit ovat yleensä varsin epikriisimäisiä, eli ne muistuttavat lääkärien työnsä yhteydessä kirjoittamia hoitotiivistelmiä. Ne sisältävät paljon melko universaalia lääketieteen erikoisammattisanastoa, kuten latinankielisiä tai muita vieraskielisiä ilmauksia, erikoisia lääkekemikaalien nimiä ja alan lyhenteitä.

Koin joitain hankaluuksia aineiston saattamisessa analysoitavaksi soveltuvaan muotoon. Niitä käsittelen seuraavassa alaluvussa, jossa selostan aineiston käsittelyä. Samalla tiivistän laatimiani sääntöjä ja linjauksia, joita tein käsitelläkseni aineiston mahdollisimman yhdenmukaisella tavalla.

[1] Syksyllä 2012 järjestetylle kurssille osallistuivat opiskelijat Jussi Aaltonen, Annika Haaramo, Jarkko Hakala, Jenna Halinen, Jaana Heinisuo, Essi Järvinen, Heini Kaalamo, Jenni Koponen, Anna Kuisma, Tapani Kusnetsoff, Ville Laine, Riikka Lehtinen, Inge Määttä, Mari Nykänen, Tiina Rainaho, Jenna Ruokolainen, Elina Salmenoja, Kiti Salonen, Heljä Silvennoinen, Emilia Tuuri ja Aino Vasileva.

3.1.2 Aineiston käsittely

Aineisto oli sen saadessani luokiteltu alustavasti sanaluokkiin. Sananesiintymien rinnalle oli alustavasti purettu auki sanakirjamuodot niin, että kukin lekseemi oli yhdellä rivillä. Tässä vaiheessa aineistossa olivat mukana muun muassa sanaliittojen yhteen kirjoittamisesta aiheutuneet virheet sekä tekniset merkinnät, kuten tehtäväosioiden numero- ja kirjaintunnukset. Havaitsin, että lisäksi aineistoa purkaneet koodaajat olivat syystä tai toisesta kirjoittaneet omia **huomautuksiaan** tutkittavan aineiston eli lääkäreiden kirjoittaman tekstin sekaan. Tästä syystä päätin tarkistaa aineiston lukemalla sen läpi alusta loppuun ja poistaa vastauksiin kuulumattomat kohdat. Vaikka lopulta tarkastelen lekseemejä sanatasolla, lausekontekstista on ollut korvaamaton apu aineistoon kuulumattomien saneiden etsimisessä ja muun muassa epäselvien sanojen tulkitsemisessa.

Aineiston tarkistamisen yhteydessä karsin analyysin ulkopuolelle myös kaikki **lyhenteet**, jotka eivät ole yhdyssanan määriteosia (2672 kpl), tehtäväpapereiden osiin liittyvät kirjaintunnukset ja muut **tekniset merkinnät** sekä **symbolit**, joihin lukeutuvat esimerkiksi lääkäreiden piirtämät nuolet ja matemaattiset merkit. Poistin aineistosta myös kaikki **numeroin kirjoitetut luvut**, elleivät ne ole yhdyssanan osana (esim. *14-vuotias*). Lukusanoista aineistoon on siis otettu mukaan vain ne, jotka on kirjoitettu kirjaimin. Tähän on kaksi syytä. Ensinnäkin numeroiden osaaminen ei mielestäni sinänsä kerro mitään kielitaidosta, koska samat numerot ovat käytössä eri kielissä. Toiseksi teksteissä on runsaasti numeroita ja niiden ottaminen mukaan analyysiin vääristäisi tuloksia.

Raakamuodossa aineistossa oli rivejä kaikkiaan 55 259, mutta kun tarkistin aineiston ja poistin muun muassa tekniset merkinnät, koodaajien ylimääräiset huomautukset ja tyhjät rivit sekä yhdistin kahdelle eri riville tavuttuneet saneet, sain saneiden määräksi tässä vaiheessa 42 934. Raakaaineistossa oli siis huomattavan paljon sellaista ylimääräistä, joka olisi voinut vääristää tutkimustuloksia. Myöhemmässä vaiheessa tein analysoitavaan aineistoon vielä lisää rajauksia, joista kerron tarkemmin analyysin yhteydessä.

Excel-muotoisessa aineistossani yksi rivi vastaa yhtä sanetta. Kuitenkin koevastauksia kirjoittaessaan lääkärit ovat paikoin jakaneet sanoja kahdelle eri riville, minkä vuoksi ne ovat siirtyneet korpustiedostoon kahdessa osassa ja kahdelle eri riville, eli kahtena eri saneena. Aineistoa tarkistaessani olen siis yhdistänyt tällaiset **eri riveille jakaantuneet saneet** yhdelle samalle riville, jos ne on selvästi jaettu riveille tavuviivojen avulla (esim. *laboratorio-kokeet*, *LDL-statiinilääkitys*). Lisäksi selviltä vaikuttavissa tapauksissa en ole antanut yhdyssanojen vähäisten **yhteen ja erikseen kirjoittamisen virheiden** tai rivinloppuisen tavuviivan puuttumisen häiritä, vaan olen tulkinnut yhdeksi saneeksi esimerkiksi *erikois kirurgin* ja *kaksi osainen*. Vastaavasti virheellisesti yhteen kirjoitetut täysin ymmärrettävät ja kontekstin perusteella selvästi erilleen kuuluvat saneet olen korjannut omiksi saneikseen, jotta esimerkiksi kirjoittajan käsialasta johtuvat tulkintavirheet eivät vääris-

täisi analyysia (esim. *hänon*, *onselvitettävä*, *expiraationlopussa*). Tämänkaltaisia virheitä oli korjattava määrällisen analyysin onnistumisen vuoksi jonkin verran, mutta ne edustivat vain pientä murto-osaa aineistosta. Nähdäkseni näistä korjauksista ei ole haittaa, koska en ole tekemässä aineiston virheanalyysia. **Sanaliittojen** erilliset osat lasken omiksi saneiksi ja lekseemeikseen, mutta yhtä useammasta saneesta koostuvia erisnimiä käsittelen yksittäisinä saneina ja lekseemeinä, siis yhtenä erisnimenä. Kuitenkin esimerkiksi *Basetovin tauti* on tulkintani mukaan kaksi eri sanetta ja lekseemiä, siis erisnimi ja yleisnimi. Samaa periaatetta noudatan verbien liittomuodoissa.

Aineistotaulukossa jokaisen saneen rinnalle on kirjattu omiin sarakkeisiinsa niiden tavoitemuoto ja lekseemin hakumuoto, ja samassa yhteydessä olen korjannut hakumuotojen merkitsemisen yhteydessä **kirjoitusvirheitä** (esim. *mialgia* muotoon *myalgia*). Yhdyssanojen satunnaismuodostelmia en kuitenkaan ole ryhtynyt muuttamaan toisiksi hakumuotoja merkitessä. Täysin **tunnistamattomat** sanat olen pyrkinyt poistamaan korpuksesta itselleni luomani säännön mukaan: jos sana on kirjoitettu niin epäselvästi, että sen voi vain yrittää arvata kontekstin avulla, olen poistanut sen sijaan, että olisin yrittänyt osua arvaamalla oikein. Paikoin ongelma on, että ilmeisesti kopioitaessa alkuperäistä aineistoa tutkimuskäyttöön jotkut sanat ovat jääneet kopiossa näkyviin vain osittain. Useimmiten näin katkenneen sanan kontekstin avulla voi päätellä melko luotettavasti, mistä sanasta on kysymys, sillä useimmiten katkenneista sanoista vaikuttaa puuttuvan vain yksittäisiä kirjaimia. Jos kirjaimia näyttää puuttuneen enemmän kuin yksi tai kaksi niin, ettei ole selvää, mistä sanasta on kysymys, en ole lähtenyt arvailemaan, vaan olen poistanut katkenneen sanan aineistosta. Paikoin lääkärit ovat pyyhkimisen sijasta yliviivanneet sanoja ja kirjoittaneet sanan uudelleen tai ajatuksen eri sanoin. Näistä **yliviivatuista sanoista** otan analyysiin mukaan ne, jotka olen tulkinnut kokonaisiksi ja ymmärrettäviksi.

Paikoin olen korjannut koodaajien tai luokitteluja tehneiden **ylikorjaamista**. Ilmeisesti osa aineistoa koodanneista oli jopa korvannut lääkäreiden kuulustelussa kirjoittamia sanoja kokonaan toisilla tulkitessaan ja kirjatessaan ylös lääkäreiden tavoittelemia muotoja. Esimerkiksi *eläimenvaihtelen* oli korjattu aineistossa muotoon *elämäntapamuutos*. Jälkimmäistä merkitystä vastauksen kirjoittanut lääkäri on luultavasti hakenut, mutta sanaa hän ei ollut sillä tavalla kirjoittanut, vaan sen sijaan oman satunnaisyhdistelmänsä. Tässä tapauksessa muutin sanan tavoitemuodon muotoon *elämänvaihtelu*. Vastaavasti *Lähellä ajan* oli korjattu *lähiaikoina*, minkä niin ikään palautin alkuperäiseksi. Koodaaja oli vaihtanut jopa sanat *vatsan pesu* sanoihin *mahalaukun tyhjennys*, jotka korjasin takaisin. *Tupakonta* oli koodaajan mielestä ollut *hän tupakoi* ja *vasta* puolestaan *reagoida*. Mielestäni tutkimukseni antaisi väärän kuvan kokeeseen osaa ottaneiden lääkäreiden käyttämästä sanastosta, ellen puuttuisi tällaiseen ylikorjaamiseen. Kokonaisuuteen nähden ylikorjaamisen määrä oli kuitenkin lopulta hyvin vähäinen.

3.2 Tutkimuskysymykset

Tutkimuksen päätarkoitus on saada aikaan deskriptiivinen analyysi koehenkilöiden eli laillistamiskuulusteluun osaa ottaneiden lääkäreiden käyttämästä sanastosta. Tämän tutkimustehtävän toteuttamiseksi laadin joukon tutkimuskysymyksiä, jotka jäsentävät analyysia ja työvaiheita. Pääkysymys koskee koehenkilöiden aktiivista sanavarastoa ja kuuluu seuraavasti:

Millaisen suomen kielen sanaston ulkomailta tulleet lääkärit ovat hallinneet laillistamiskuulustelussa?

Pääkysymykseeni lähden hakemaan vastauksia näkökulmaani tarkemmin määrittelevien apukysymysten avulla. Ne vievät analyysiä sanastotaitojen arvioinnin suuntaan:

Millainen on käytetyn sanaston koostumus?
Mitkä ovat lääkäreiden yleisimmin käyttämät lekseemit?
Kuinka laaja vastausten sanasto on?
Kuinka yleistä suomen sanastoa lääkäreiden käyttämä sanasto edustaa?
Kuinka monimuotoista käytetty sanasto on?
Millaisia eroja yksilöiden ja kieliryhmien välillä on havaittavissa?

Analyysini lopuksi otan huomioni keskipisteeksi lääkäreiden laillistamiskokeessa saamat koepisteet ja kiinnitän huomioni siihen, millainen yhteys leksikaalisilla taidoilla ja koemenestyksellä mahdollisesti on. Tätä varten olen muotoillut myös oman tutkimuskysymyksen:

Kuinka vahvasti sanastotaidot ovat yhteydessä laillistamiskokeessa menestymiseen?

Tutkimuskysymyksiäni lähden vastaamaan määrällisin menetelmin muun muassa tilastoohjelma Exceliä apunani käyttäen. Luokittelen koehenkilöiden käyttämän sanaston muun muassa sanaluokkien, sanan kompleksisuuden ja sen perusteella, kuinka yleisiä lekseemit ovat suomen kielen taajuussanastossa. Vertailua teen paitsi yksittäisten koehenkilöiden välillä myös sen mukaan, ovatko heidät luokiteltu taustaltaan natiiveiksi vai ei-natiiveiksi suomen puhujiksi.

3.3 Tutkimusmenetelmät

3.3.1 Sanaston koostumuksen analysoiminen ja yleiskuvan esittely

Analyysissani tarkastelen kuulusteluun osallistuneiden lääkäreiden käyttämän sanaston koostumusta kvantitatiivisesti ja tilastollisia menetelmiä hyödyntäen. Yleiskuvan hahmottamiseksi aineistoa on luokiteltu useiden muuttujien avulla. Niitä ovat esimerkiksi sanaluokka, erityisyysaste (kuuluuko sana yleissanastoon, terveyssanastoon vai lääkärien erikoisammattisanastoon), kompleksisuus (perussana vai kompleksinen sana) ja sanapituus. Näiden perustietojen avulla esittelen aineistoa ja vertailen natiivien ja ei-natiivien lääkäreiden ryhmien tekstejä. Esittelen myös koko aineiston taajimmin esiintyviä lekseemejä. Rajattuani tarkasteltavaa aineistoa siirryn tarkastelemaan käytetyn sanaston yleisyyttä, mitä varten olen luokitellut sanaston eri yleisyysasteisiin.

Aluksi käsittelen yhtäältä kaikkia 46 tekstiä yhtenä kokonaisuutena ja toisaalta kieliryhmien vastausten muodostamia sanastokokonaisuuksia. Myöhemmässä vaiheessa etenen laskemaan yksilökohtaisia sanastoja kuvaavia lukuja ja niiden keskiarvoja, ja analyysin lukuja esitellessäni hahmottelen lääkäreitä myös tasoryhmiin.

Sanojen yleisyyden luokittelussa käytän apunani Tieteen tietotekniikan keskuksen CSC:n Suomen sanomalehtikielen taajuussanastoa (CSC 2004). Se on käytettävissä olevista laajoihin korpuksiin perustuvista taajuussanastoista tuorein. Sen sanalista on koottu vuonna 2004, ja sen lähdeaineistossa on ollut lähes 44 miljoonaa sanaa. Se esittää yleisyysjärjestyksessä Suomen sanomalehtikielen 9996 yleisintä perusmuotoonsa muutettua lekseemiä ja on saatavilla digitaalisessa muodossa. Sen avulla tutkimusta tehdessä on kuitenkin huomioitava, että sen korpuksen rajoittuminen sanomalehtiin tuo mukanaan joitain epävarmuustekijöitä. Esimerkiksi korpuksen uutisten toistuvat aiheet ja ajallinen konteksti ovat voineet vaikuttaa frekventeimpien sanojen kärjen koostumukseen: vaikkapa yleisimpien substantiivien joukossa ovat lekseemit uutisissa runsaasti viljelty *prosentti* ja jo historiaan jäänyt Suomen rahayksikkö *markka*. Uudesta kaunokirjallisuudesta tai sosiaalisen median teksteistä koottu taajuussanasto voisi näyttää hieman toisenlaiselta. Korpuksen kontekstin mahdollisesti mukanaan tuomia ongelmia vähentänee, että luokittelen analysoimani aineiston lekseemit riittävän suuriin taajuusluokkiin, esimerkiksi 50 yleisimmän, 100 yleisimmän tai 1000 yleisimmän luokkiin.

3.3.2 Sanaston diversiteetin analysoiminen

Myöhemmässä vaiheessa tarkastelen koehenkilöiden laatimien koevastausten leksikaalista diversiteettiä laskemalla niiden sanastosta siitä kertovia tunnuslukuja, joiden avulla voin vertailla keskenään yksittäisiä lääkäreitä sekä natiivien ja ei-natiivien suomen puhujien ryhmiä. Erilaisia leksikaa-

lista diversiteettiä kuvaavia tunnuslukuja on arvioinut ja kehitellyt muun muassa Scott Jarvis (mm. 2010; 2013), jonka kirjoittamaa Perl-komentosarjaa olen saanut omassa analyysissäni hyödyntää.[2]

Tässä työssä aion käyttää varsinaisina sanaston diversiteettiä kuvaavina lukuina Shannonin entropiaa (Shannon 1948; ks myös Saarela 1997: 51–52; Malin 2012) ja MTLD:tä (measure of textual language diversity, esim. McCarthy & Jarvis 2010; Jarvis 2013: 94). Nämä tekstin diversiteettiä kuvaavat tunnusluvut eli indeksit on syytä esitellä tarkemmin. Apunani olen käyttänyt Jarvisin (mm. 2010, 2013) esityksiä niistä sekä Malinin (2012: 26–32) oivaa tiivistystä erilaisten tunnuslukujen laskemisesta. Lähden liikkeelle yksinkertaisesta sana–sane-suhteesta, joka on paljon käytetty.

TTR-luku eli sana–sane-suhde lasketaan yksinkertaisesti jakamalla teksteissä esiintyvien eri lekseemien määrä kaikkien saneiden määrällä. Jakolaskussa ylös osoittajan puolelle lasketaan siis jokainen tutkittavassa tekstissä esiintynyt lekseemi vain yhden kerran ja alas nimittäjän puolelle kaikkien saneiden kokonaismäärä mukaan lukien toistuvat lekseemit. Yksinkertaista ja helposti laskettavaa TTR:ää käytetään usein. Se on kuitenkin osoittautunut liian karkeaksi leksikaalisen diversiteetin mittariksi (esim. Jarvis 2013: 91–95). Se ei pysty ilmaisemaan sanaston diversiteettiä luotettavasti ainakaan pienillä aineistoilla, eikä se sovi hyvin sellaisten aineistojen vertailemiseen, jotka ovat sanemääriltään erikokoisia. Siksi en itse sitä käytä kuin sen verran, mitä sitä tarvitsee avuksi luotettavampien indeksien laskemisessa. Kun esimerkiksi Laine-Leinonen (2013) tarkasteli koululaisten tekstien leksikaalista diversiteettiä, kirjoitelmien TTR-luvut eivät kasvaneet iän mukaisesti, vaikka oikeasti tekstit olivat keskimäärin sitä pidempiä ja rikkaampia mitä vanhempia olivat niitä laatineet koululaiset (Laine-Leinonen 2013: 12–14, 82–86). Peruskoululaisten sanaston kehittymistä tutkineen Saarelan (1997: 85) mukaan TTR-arvo antaa lyhyille teksteille tavallaan liian hyvän merkitsevyyden, koska lyhyissä teksteissä sanat eivät toistu niin paljon kuin pitkissä. TTR-arvon ongelma laskettaessa sanaston diversiteettiä on siis se, että se on liikaa sidoksissa otoksen kokoon.

Shannonin entropia eli Shannonin indeksi on amerikkalaisen matemaatikon ja insinöörin Claude Shannonin (1948) luoma tunnusluku, jonka laskukaavan hän kehitti alun perin informaation entropian kuvaamiseksi. Sillä analysoitiin alkujaan tekstikatkelmia kirjain kirjaimelta. Tunnusluku viittasi viestin epävarmuuden eli ennustamattomuuden asteeseen (Shannon 1948: 10–12; ks. myös Jarvis 2013: 93). Sittemmin Shannonin entropiaa on käytetty kielitieteen lisäksi jo aiemmin ekologian piirissä. Kaava kirjoitetaan eri yhteyksissä hieman erilaisin variaatioin. Tässä se on esitettynä lyhyessä muodossaan:

[2] Kiitokset tästä kuuluvat sekä Scott Jarvisille ja Mari Hongolle, joka auttoi minua yhteydenpidossa ja toimitti komentosarjan käyttööni.

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

Kaavassa H merkitsee siis entropiaa, joka on sitä pienempi, mitä useammin analysoitava informaatio koostuu keskenään samanlaisista osasista, ja suurimmillaan, jos kaikki osaset ovat erilaisia (Shannon 1948: 10–12). Esimerkiksi ekologian piirissä samanlaiset osaset voisivat olla eläinyksilöitä, jotka kuuluvat samaan lajiin. Omassa analyysissäni taas samanlaiset osaset ovat vastaavasti samaan lekseemiin kuuluvia saneita. *R* merkitsee tässä yhteydessä analysoitavassa tekstissä esiintyvien eri lekseemien määrää, ja p_i puolestaan viittaa lekseemin i osuuteen kaikista saneista. Kaavalla lasketaan H-arvo tekstin jokaiselle saneelle, ja sen lopputulos eli entropian arvo on kaikkien saatujen H-arvojen painotettu summa (emt.).

Koululaisten sanaston kehittymistä tutkinut Leena Saarela (1997) päätyi väitöstutkimuksessaan käyttämään tekstin diversiteetin mittarina Shannonin entropiaa TTR-arvon ongelmallisuuden vuoksi, ja hän nimittää Shannonin entropiaa nimenomaan diversiteetiksi (ks. Saarela 1997: 51–52). Saarelan mukaan Shannonin entropian laskukaava eliminoi suuritaajuisten sanojen vinouttavaa vaikutusta. Hän laski Shannonin entropian koululaisten kirjoitelmista oppilaskohtaisesti. Shannonin entropian laskukaavaan kuuluvan logaritmin ansiosta suuritaajuiset sanat, kuten olla-verbi tai ja-sana, eivät vääristä tuloksia. Se käy hyväksi monipuolisuusmittariksi myös siksi, etteivät yksittäisten tekstien lyhyydet tai pituudet vääristä sanemääriä. (Saarela 1997: 105.) Myös Malin (2012) toteaa kielenoppijoita koskevassa tutkimuksessaan Shannonin entropian luotettavaksi sanaston leksikaalisen diversiteetin indeksiksi. Hänen mukaansa sen antamat tulokset vaikuttavat täysin riippumattomilta tekstien määrästä tai pituudesta (Saarela 1997: 51).

MTLD eli tekstin leksikaalisen diversiteetin mittari (measure of textual language diversity) pyrkii ratkaisemaan ongelmat, jotka monessa muussa sanaston monimuotoisuuden mittarissa johtuvat tutkittavien tekstien pituuksien eroista (McCarthy & Jarvis 2010: 381). MTLD-tunnusluku kertoo, kuinka monta peräkkäistä sanetta tekstissä keskimäärin on ennen kuin sanaketjusta laskettu TTR-indeksi eli sana–sane-suhde laskee erikseen määritellyn arvon alapuolelle. Sanojen järjestyksellä on merkitystä. MTLD:tä lasketaan lukemalla teksti sana sanalta ensin alusta loppuun, sitten lopusta alkuun. Peräkkäisistä sanoista lasketaan alati TTR-indeksiä eli sana–sane-suhdetta, ja kun sanaketjun TTR-luku on alittamassa määrättyä arvoa, laskeminen katkaistaan ja aloitetaan alusta seuraavasta sanasta. Luotettavia tuloksia antavaksi arvoksi on testaamalla saatu 0,720 (McCarthy & Jarvis 2010: 385). McCarthy ja Jarvis (2010: 384) käyttävät MTLD:n laskemisesta esimerkkinään sanaketjua *of the people, by the people, for the people*. Siinä TTR pysyy arvossa 1,00 neljän en-

simmäisen sanan verran *of* (1,00) *the* (1,00) *people* (1,00) *by* (1,00), mutta sitten TTR alkaa pienentyä sanojen toistuessa: *the* (0,80) *people* (0,667) *for* (0,714) *the* (0,625) *people* (0,556). TTR lasketaan kuitenkin vain siihen saakka, että se alittaa arvon 0,720. Silloin tekstin faktoriarvo nousee yhdellä kokonaisella faktorilla ja TTR-arvojen laskeminen aloitetaan alusta seuraavasta sanasta: *of* (1,00) *the* (1,00) *people* (1,00) *by* (1,00) *the* (0,80) *people* (0,667) *for* (1,00) *the* (1,00) *people* (1,00). Lisäksi MTLD:n laskemisessa otetaan huomioon myös häntäfaktori, joka useimmiten jää, kun analysoitava teksti ei pääty juuri siihen kohtaan, jossa määritelty TTR:n raja-arvo (tässä 0,720) alittuu. Häntäfaktorin arvo lasketaan siitä, kuinka suuren osan se muodostaa kokonaisesta faktorista eli kuinka kauas se jää määritellystä raja-arvosta, ja sen arvo lisätään aiemmin saatuun faktorilukuun. Jos teksti sisältäisi esimerkiksi neljä kokonaista faktoria ja jäljelle jäävien saneiden TTR-luku olisi jäänyt 0,887:ään, ensin laskettaisiin TTR-lukeman etäisyys lähtökohdasta (1 - 0,887 = -0,113 → 0,113) ja se, kuinka suuren osan se kattaa raja-arvon ja lähtökohdan erotuksesta (1 – 0,720 = -0,280 → 0,280): 0,113 / 0,280 = 0,404. Tämä lisättäisiin tekstistä laskettujen kokonaisten faktorien määrään: 4 + 0,404 = 4,404. (McCarthy & Jarvis 2010: 384–385; ks. myös Malin 2012: 28–30.) Seuraavaksi analysoitavan tekstin kokonaissanemäärä jaetaan lopullisella faktorilla. Samoin toimitaan sekä alusta loppuun että lopusta alkuun lasketun faktoriarvon kanssa, jolloin tulokseksi saadaan kaksi välivaiheen MTLD-arvoa. Varsinainen MTLD-indeksi saadaan laskemalla näiden tekstiä etuperin ja takaperin lukien saatujen lukujen keskiarvo. Kun Jarvis ja McCarthy (2010) testasivat rinnakkain erilaisia leksikaalisen diversiteetin mittareiden validiteettia, he totesivat MTLD:n varsin luotettavaksi indeksiksi. Samaan päätelmään tulee myös suomenoppijoiden tekstejä erilaisilla mittareilla analysoinut Essi Malin, jonka mukaan Shannonin entropia ja MTLD osoittautuivat siinä mielessä luotettaviksi mittareiksi, että ne toimivat, vaikka tutkittavat ja vertailtavat tekstit olisivat melko eripituisia (2012: 50–51).

Varsinaisten leksikaalisen diversiteetin tunnuslukujen rinnalle lasken Suomen sanomalehtikielen taajuussanaston (CSC 2004) avulla koevastauksissa käytetyn sanaston **kumulatiivisia prosentteja** ja niiden keskiarvoja. Ne kertovat, kuinka suuren osan kokeessa käytetystä sanastosta tiettyyn yleisyysluokkaan ja sitä yleisempään sanastoon kuuluvat lekseemit muodostavat. Lasken kumulatiivisia prosentteja tarpeen mukaan samojen lekseemien toiston huomioiden tai sen pois karsien, ja silloin puhun joko saneiden kumulatiivisista prosenteista tai lekseemien kumulatiivisista prosenteista. Esimerkiksi Tervola ym. laskivat kustakin vastauksesta jokaisen lekseemin vain kerran (Tervola ym. 2015: 340–341), mutta omassa tutkimuksessani tarkastelen erikseen myös niiden toistoa.

4 LAILLISTAMISKOKEESSA KÄYTETTY SANASTO

4.1 Perustietoja käytetystä sanastosta

Tässä alaluvussa kuvailen koehenkilöiden laatimia vastaustekstejä yhdessä kokonaiskuvan saamiseksi. Aineiston saneet on luokiteltu sanaluokan, erityisyysasteen ja kompleksisuuden mukaan, ja luokitteluperusteet esitän muuttuja muuttujalta omissa alaluvuissaan. Tässä vaiheessa aineistosta on karsittu pois lyhenteet, englanninkielinen sanasto sekä numeroin kirjoitetut lukusanat. Mukana ovat vielä kaikki erityisyysasteet, ellei toisin ole mainittu.

4.1.1 Aineiston tunnuslukuja

Taulukkoon 1 on tiivistetty perustietoja tutkittavasta aineistosta. Johdokset olen laskenut tavanomaisen jaottelun mukaisesti omiksi lekseemeikseen. Erityisesti adjektiiveista muodostettuja, yleensä tapaa ilmaisevia sti-loppuisia adverbejä voisi kenties perustellusti pitää myös osana adjektiivien taivutusparadigmaa, mutta tulkitsen ne työssäni kielentutkimuksessa vakiintuneella tavalla omiksi lekseemeikseen. Esimerkiksi *hieno* ja *hienosti* ovat näin tulkiten kaksi eri lekseemiä. Taulukossa lyhimpien ja pisimpien sanemäärien viereisiin sulkeisiin merkityt numeroinnit (esim. *nro 5*) viittaavat koehenkilöille tässä tutkimuksessa annettuihin yksilöintinumeroihin.

Taulukko 1. Aineiston tunnuslukuja.

	kaikki	**ei-natiivi**	**natiivi**
vastaustekstien määrä	46	38	8
sanemäärä keskimäärin	933,34	976,00	730,75
pisin vastaus (sanetta)	1618	1618 (nro 5)	1294 (nro 13)
lyhin vastaus (sanetta)	256	382 (nro 23)	256 (nro 18)
pituuden mediaani, (sanetta)	972,50	997,00	729,50
pituuden keskihajonta (sanetta)	298,62	263,90	385,16
aineiston sanemäärä	42934	37088	5846
ilman lääkärien erikoisammattisanastoa	40051	34627	5424
ilman mitään terveydenhuollon ammattisanastoa	27307	23587	3720

Jokseenkin yllättävää oli, että sanemäärältään pisimmän vastauskokonaisuuden tuotti suomea toisena kielenään oppiva lääkäri ja lyhimmän natiiviksi tulkittu suomenpuhuja. Tässä vaiheessa ei kuitenkaan vielä ole tarkasteltu sitä, mitkä tarkastelluista vastauksista ovat johtaneet suorituksen hyväksymiseen. Vastausten pituuksia näin eriteltäessä on siis huomattava, että tekstien pituus ei

suoraan kerro sen sisällöllisestä laadusta eli esimerkiksi siitä, onko lääkäri läpäissyt sillä laillistamiskokeen osan vai ei.

4.1.2 Käytetyn sanaston erityisyys

Aineiston saneet on luokiteltu *yleissanastoon*, *terveyssanastoon* ja *erikoisammattisanastoon*. Näistä erikoisammattisanasto tarkoittaa lääkäreiden ja muun terveysalan henkilöstön omaa sanastoa, joka on maallikolle vierasta. Terveyssanasto puolestaan on maallikonkin ymmärrettävissä olevaa terveysalan ammattikieltä. Yleissanasto kattaa loput koehenkilöiden käyttämästä sanastosta.

Erikoisammattisanastoon luen kuuluvaksi muun muassa alan erikoistermit (esim. *thorax*, *hemothorax*, *leukosytoosi*, *sternaalinen*), useimpien lääkeaineiden kemialliset nimet ja latinankieliset ilmaukset (esim. *per*, *bursa*, *anterior*). Luokan ulkopuolelle jäävät yleiskielisempään terveyssanastoon tulkitsemani, tavallisemmat tai usein toistuvat lekseemit (esim. *palpaatio*, *tuseeraus*, *dementia*, *neurologinen*, *anamneesi*, *konsultaatio*, *depressio*, *penisilliini*, *tutkimus*), joista suuren osan sijoitin terveyssanastoon. Lisäksi erikoisammattisanastoon kuuluu selkeästi valtaosa aineiston erisnimistä. Niiden joukko koostuu lähes kokonaan lääkkeiden tuotenimistä (mm. *Aciclovir*, *Obsidan*, *Prednisolon*) sekä sairauksien nimityksiin tai erikoisiin kokeisiin sisältyvistä vierasperäisistä sukunimistä (esim. *Henoch-Schönleinin* tauti, *Romberg*). Siksi olen luokitellut erisnimet erikoisammattisanastoon ja siten jättänyt ne tarkemman analyysin ulkopuolelle muutamaa yksittäistä poikkeusta (esim. *Suomi*, *Minna*, *Mikkola*) lukuun ottamatta.

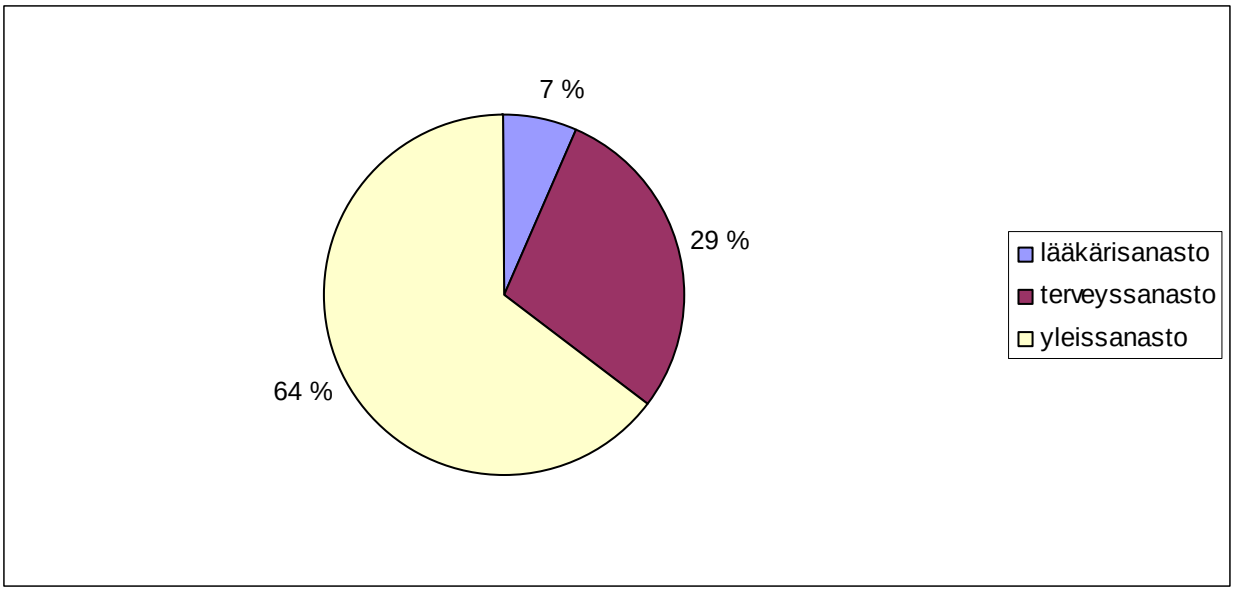
Yleissanaston ja **terveyssanaston** raja oli hankala määritellä aineistoa luokitellessa. Tämä johtui eri seikoista. Voi ajatella, että luokat ovat osittain päällekkäisiä: yhtäältä suomalaiseen arkikieleen on vakiintunut paljon terveyssanastoa ja myös alan erikoissanastoa, kun toisaalta terveysalalla tarvitsee käyttää huomattavan paljon esimerkiksi elinten ja ruumiinosien nimityksiä, jotka kuuluvat usein suomen frekventeimpään sanastoon (CSC 2004). Lisäksi jotkut sanat, jotka eivät alun perin ole mitenkään erityisesti liittyneet terveyteen, ovat saaneet terveysalalla oman erityismerkityksensä tai vakiintuneet erityisesti sen käyttämiksi ja esimerkiksi terveyskeskustyössä välttämättömiksi termeiksi.

Päädyin luokittelemaan esimerkiksi terveyskeskustyössä välttämättömän sanaston arkikielisimpiäkin lekseemejä terveyssanastoon, jos ne olivat selvästi tulkittavissa (ainakin) terveyssanaston alueelle kuuluviksi. Terveyssanastoon luokittelin yleiskielisistä lekseemeistä muun muassa ihmisen anatomiaan liittyvät sanat, kuten sisäelimet (esim. *maksa*, *haima, siitin*), aineiston kontekstissa terveyteen liittyviä erityismerkityksiä kantavat lekseemit ja terveysalan termit (esim. *ruusu*, *leikkaus*, *sieni, viljely*), sairaudet ja niiden oireet (esim. *yskä*, *kipu*, *sairaus*, *aamujäykkyys*, *masennus*, *mustelma*) sekä tyypillisesti diagnooseissa toistuvan ja esimerkiksi oireita tyypittelevän sanaston (esim.

paikallinen, *ulkoinen*, *verenvuotoinen*, *vointi*, *yleiskunto*, *hengitysvaikeus*, *tykyttely*). Rajatapauksista monet asettuivat terveyssanaston puolelle lähemmässä tarkastelussa lopulta luontevasti, koska tässä kontekstissa niitä käytettiin yleensä nimenomaan terveyteen tai terveysalan työhön liittyvässä erityismerkityksessä.

Kaiken kaikkiaan koehenkilöt käyttivät vastauksissaan paljon maallikolle hankalia tai käsittämättömiä lyhenteitä ja erikoisammattisanastoa sekä muuta lääkärintyössä tarvittavaa terveyssanastoa. Paikoin he paikkailivat suomen kielen taitojensa aukkoja käyttämällä suomen sijaan englannin sanoja. Englanninkielisten saneiden määrä on kuitenkin kokonaisuuteen nähden niin pieni, että prosenttiluku pyöristyy nollaan. Kuvaaja 1 pyrkii antamaan kokonaiskuvan käytetyn leksikon erityisyyden jakaumasta. Kuvaaja on tehty lyhenteiden, englanninkielisen sanaston ja numeroin kirjoitettujen lukusanojen poiston jälkeen.

Kuvaaja 1. Sanaston erityisyysasteiden osuudet.



Saneiden erityisyysasteiden jakauma oli melko tarkasti samanlainen natiiveiksi ja ei-natiiveiksi tulkituilla koehenkilöillä. Taulukosta 2 voi lukea erojen olevan niin vähäisiä, että kun prosenttiluvut pyöristää yhden prosentin tarkkuuteen, natiivien ja ei-natiivien saamat erityisyysluokkien osuudet ovat samansuuruiset. Erikoinen yhteensattuma on, että terveyssanaston saamat prosenttiosuudet ovat kieliryhmillä samat jopa kahden desimaalin tarkkuudella.

Taulukko 2. Koko käytetyn sanaston erityisyysasteet.

Ei-natiivit		
erityisyys	sanemäärä	%
lääkärisanasto	2461	6,64 %
terveyssanasto	10639	28,69 %
yleissanasto	23987	64,68 %
yhteensä	37087	100,00 %
Natiivit		
erityisyys	sanemäärä	%
lääkärisanasto	422	7,22 %
terveyssanasto	1677	28,69 %
yleissanasto	3747	64,10 %
yhteensä	5846	100,00 %
Kaikki		
erityisyys	sanemäärä	%
lääkärisanasto	2883	6,72 %
terveyssanasto	12316	28,69 %
yleissanasto	27734	64,60 %
yhteensä	42933	100,00 %

4.1.3 Sanaluokkajakauma

Taulukko 3 esittää aineiston saneiden sanaluokkien jakauman. Sen luvuissa ovat mukana kaikki erityisyysasteet. Saneet olen luokitellut seuraaviin sanaluokkiin: substantiivi, adjektiivi, pronomini, numeraali, verbi, adverbi, konjunktio ja adpositio. Verbien partisiippimuodot olen jakanut adjektiiveihin ja verbeihin sen mukaan, käytetäänkö niitä verbimäisesti vai adjektiivin tapaan. Verbimäisesti käytettynä partisiippi saa itse esimerkiksi verbin täydennyksiä tai tilannetta kuvaavia määritteitä (esim. *stressiin liittyvä*), kun adjektiivimaisesti käytettynä se kertoo luonnehdittavan asian ominaisuudesta (esim. *osaava mies*) (ISK § 632).

Taulukko 3. Sanemäärät sanaluokittain.

Sanaluokka	ei-natiivi	%	natiivi	%	kaikki	%
substantiivi	18976	51,16 %	3010	51,49 %	21986	51,21 %
verbi	7833	21,12 %	1217	20,82 %	9050	21,08 %
konjunktio	2913	7,85 %	392	6,71 %	3305	7,70 %
adjektiivi	2772	7,47 %	446	7,63 %	3218	7,50 %
adverbi	2136	5,76 %	475	8,13 %	2611	6,08 %
pronomini	1725	4,65 %	194	3,32 %	1919	4,47 %
adpositio	409	1,10 %	58	0,99 %	467	1,09 %
erisnimi	265	0,71 %	39	0,67 %	304	0,71 %
numeraali	59	0,16 %	15	0,26 %	74	0,17 %
Yhteensä	37088	100,00 %	5846	100,00 %	42934	100,00 %

Partikkelit olen luokitellut adverbien luokkaan hyödyntämäni Sanomalehtikielen taajuussanaston (CSC 2004) mallin mukaisesti, ja sitä mukaillen olen luokitellut myös erisnimet omaksi joukokseen. Analyysissa myöhemmin hyödyntämäni tietokoneskriptin toiminta edellyttää tätä yhdenmu-

kaisuutta, ja lisäksi olen halunnut pitää mahdollisuuden avoimena esimerkiksi taajuussanaston ja laillistamiskoeaineiston sanaluokkaosuuksiin perustuvalle vertailulle. Numeraalien määrän pienuus johtuu siitä, että valtaosa vastauksissa olleista luvuista oli kirjoitettu numeroin ja jo aineistoa tarkistettaessa on poistettu kaikki numeroin kirjoitetut luvut. Aineistoon ovat jääneet siksi vain ne lukusanat, jotka koehenkilöt olivat kirjoittaneet auki kirjaimin ja suomeksi (esim. *yksi*, *kolmas*, *pari*).

Kuten taulukosta käy ilmi, ei-natiivien ja natiiveiksi tulkittujen lääkäreiden välillä ei ole suuria eroja siinä, mitä sanaluokkia he ovat käyttäneet. Silmiinpistävin ero on se, että natiivit koehenkilöt käyttivät selvästi enemmän abverbejä kuin ei-natiivit. Lisäksi ei-natiivit käyttivät tekstissään suhteessa selvästi enemmän konjunktioita ja pronomineja kuin natiivit.

Kun sanaluokkien jakaumaan tarkastelee erityisyysasteittain, on nähtävissä, että lääkärien erityisammattisanasto ja terveyssanasto koostuvat valtaosin substantiiveista. Sanaluokkien jakauman ero yleissanastoon on huomattava. Tämä käy ilmi alla olevasta taulukosta 4.

Taulukko 4. Sanaluokkien jakauma erityisyysasteittain.

Sanaluokka	lääkäri	%	terveys	%	yleis	%	Yhteensä	%
substantiivi	2262	78,46 %	11625	91,22 %	8099	29,66 %	21986	51,21 %
verbi	27	0,94 %	439	3,44 %	8584	31,44 %	9050	21,08 %
konjunktio	7	0,24 %	0	0,00 %	3298	12,08 %	3305	7,70 %
adjektiivi	336	11,65 %	549	4,31 %	2333	8,54 %	3218	7,50 %
adverbi	19	0,66 %	78	0,61 %	2514	9,21 %	2611	6,08 %
pronomini	0	0,00 %	0	0,00 %	1919	7,03 %	1919	4,47 %
adpositio	2	0,07 %	0	0,00 %	465	1,70 %	467	1,09 %
erisnimi	230	7,98 %	53	0,42 %	21	0,08 %	304	0,71 %
numeraali	0	0,00 %	0	0,00 %	74	0,27 %	74	0,17 %
Yhteensä	2883	100,00 %	12744	100,00 %	27307	100,00 %	42934	100,00 %

Terveyssanastossa substantiivien osuus on jopa yli yhdeksän kymmenystä (91,22 %). Erikoisammattisanastossa korostuvat substantiivien (78,46 %) lisäksi adjektiivit (11,65 %) sekä erisnimet (7,98 %), joita esiintyi muun muassa tautien nimityksissä. Terveyssanastoon luokitellun sanaston joukossa on myös hieman verbejä (3,44 %) sekä adjektiiveja (4,31 %) Muiden luokkien sanojen osuus oli terveys- ja erikoisammattisanastossa vähäinen. Yleissanastossa puolestaan suurimman luokan muodostavat verbit lähes kolmanneksen osuudellaan (31,44 %). Lisäksi yleissanaston sanaluokkajakauma on huomattavasti tasaisempi. Kaiken kaikkiaan taulukko osoittaa, kuinka merkittävä vaikutus erityisyysasteiden karsimisella pois tarkemmasta analyysista on sanaluokkajakaumaan.

4.1.4 Sanaston kompleksisuus

Taulukot 5 ja 6 osoittavat, että koevastauksissa käytetty sanasto koostuu valtaosin perussanoista (63,33 %), jos lekseemien toisto lasketaan mukaan. Käytetyistä saneista kompleksisia sanoja eli

yhdyssanoja ja johdoksia oli kaikkiaan runsas kolmannes (35,63 %). Natiivien ja ei-natiivien ryhmien sanaston kompleksisuuden välillä oli havaittavissa selviä eroja. Natiiveiksi tulkitut koehenkilöt käyttivät suhteessa hieman enemmän johdoksia ja yhdyssanoja kuin ei-natiivit. Näitä taulukoiden lukuja lukiessa on kuitenkin huomattava, että ne kertovat kummankin vastaajaryhmän käyttämistä saneista yhtenä kokonaisuutena sekä kaikista tarkasteltujen vastausten saneista yhdessä, eivät yksittäisten vastaajien käyttämistä saneista. Taulukoissa näkyvän pienen luokittelemattomien saneiden joukon muodostavat lähes kokonaan lääkeaineiden yleis- ja erisnimet, jotka lukeutuvat muutamaa poikkeusta lukuun ottamatta lääkärien erikoisammattisanastoon.

Taulukko 5. Saneiden kompleksisuus, kaikki erityisyysasteet mukana.

Kompleksisuus	ei-natiivi	%	natiivi	%	kaikki	%
perussana	23828	64,25 %	3363	57,53 %	27191	63,33 %
johdos	7442	20,07 %	1394	23,85 %	8836	20,58 %
yhdyssana	5424	14,62 %	1037	17,74 %	6461	15,05 %
luokittelematon	394	1,06 %	52	0,89 %	446	1,04 %
Yhteensä	37088	100,00 %	5846	100,00 %	42934	100,00 %

Taulukko 6. Saneiden kompleksisuus yhdistetyin luokin, kaikki erityisyysasteet mukana.

Kompleksisuus	ei-natiivi	%	natiivi	%	kaikki	%
perussana	23828	64,25 %	3363	57,53 %	27191	63,33 %
kompleksinen	12866	34,69 %	2431	41,58 %	15297	35,63 %
luokittelematon	394	1,06 %	52	0,89 %	446	1,04 %
Yhteensä	37088	100,00 %	5846	100,00 %	42934	100,00 %

Hieman toisenlaisen näkökulman koehenkilöiden käyttämään sanastoon antaa tarkastelutapa, jossa jokainen käytetty lekseemi lasketaan vain yhden kerran. Taulukko 7 osoittaa, kuinka asetelma on kääntynyt päälaelleen verrattuna taulukkojen 5 ja 6 saneiden tarkasteluun: käytettyjen lekseemien joukosta selvästi yleisimpiä ovatkin yhdyssanat ja toiseksi yleisimpiä johdokset, kun taas perussanat muodostavat luokista pienimmän.

Taulukko 7. Lekseemien kompleksisuuden aste, kaikki erityisyysasteet mukana

	ei-natiivi	%	natiivi	%	kaikki	%
yhdyssana	2098	39,11 %	640	31,71 %	2466	40,20 %
johdos	1704	31,77 %	676	33,50 %	1983	32,33 %
perussana	1351	25,19 %	667	33,05 %	1457	23,75 %
luokittelematta	211	3,93 %	35	1,73 %	228	3,72 %
yhteensä	5364	100,00 %	2018	100,00 %	6134	100,00 %

Taulukoiden luvuissa voi yllättää, että esimerkiksi koko lääkärijoukon yhdyssanojen prosenttiosuus lekseemeistä on suurempi (40,20 %) ja toisaalta perussanojen osuus (23,75 %) pienempi

kuin kummallakaan osaryhmällä erikseen. Tämä kuitenkin on seurausta siitä, että samalla, kun kieliryhmän edustajat ovat käyttäneet paljon erilaisia yhdyssanoja, perussanojen valikoima on ollut niillä samankaltaisempi. Yhdyssanojen kirjo ja perussanavalikoiman samankaltaisuus näkyvät taulukossa siis tällä tavalla. Yhdyssanojen ja johdoksien yhteenlaskettu osuus nousee koko koehenkilöjoukon käyttämistä lekseemeistä jopa 72,53 prosenttiin (Taulukko 8).

Taulukko 8. Lekseemien kompleksisuuden aste yhdistetyin luokin.

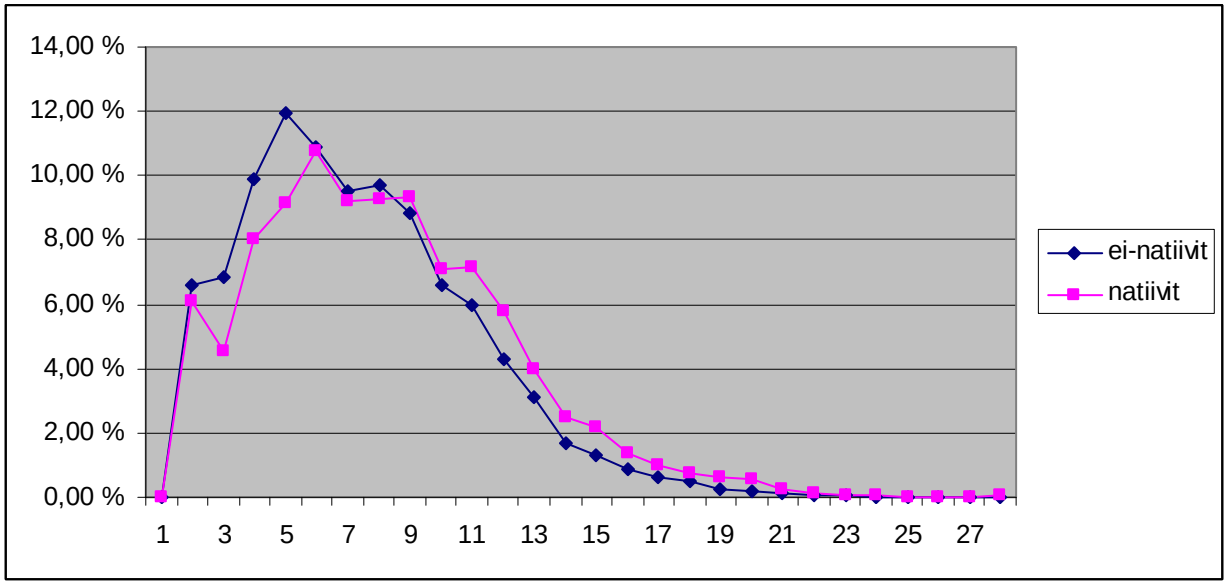
	ei-natiivi	%	natiivi	%	kaikki	%
kompleksinen	3802	70,88 %	1316	65,21 %	4449	72,53 %
perussana	1351	25,19 %	667	33,05 %	1457	23,75 %
luokittelematta	211	3,93 %	35	1,73 %	228	3,72 %
yhteensä	5364	100,00 %	2018	100,00 %	6134	100,00 %

Verrattaessa ryhmätasolla natiivien ja ei-natiivien koehenkilöiden käyttämien lekseemien kompleksisuutta näyttäisi ensi alkuun siltä, että ei-natiivit käyttäisivät enemmän kompleksisia lekseemejä kuin natiiveiksi tulkitut. Tällainen ei kuitenkaan ole taaskaan luotettava päätelmä, koska yhdyssanojen, johdosten ja perussanojen osuutta ei ole laskettu yksittäisistä vastaajista vaan kokonaisista ei-natiivien ja natiivien ryhmistä. Lukujen väliset erot olisivat luultavasti toisenlaiset, jos natiivien ja ei-natiivien vastausten määrät olisivat yhtä suuria.

4.1.5 Saneiden pituus

Aineiston saneista runsas kolmannes (34,27 %) oli alle kuusi merkkiä pitkiä, yli puolet (54,62 %) alle kahdeksan merkkiä pitkiä ja neljä viidennestä (79,83 %) korkeintaan kymmenen merkkiä pitkiä. Natiiveiksi katsottujen koehenkilöiden käyttämät saneet ovat keskimäärin hieman pidempiä kuin ei-natiiveilla. Ei-natiivin tuottama sane on keskimäärin 7,39 merkkiä pitkä, kun taas natiiveille keskipituus on selvästi pidempi, 8,17 merkkiä. Kaikkien tässä tarkasteltujen saneiden pituuden keskiarvo on 7,50 merkkiä. Ryhmien tuottamien saneiden pituuseroja havainnollistaa kuvaaja 2, joka muodostaa käytettyjen saneiden pituuksien profiilit.

Kuvaaja 2. Eripituisten saneiden osuudet kieliryhmittäin.



Kuvaajassa erimittaisten saneiden osuuksia kuvaava käyrä on natiiveilla varsin samanmuotoinen kuin ei-natiiveilla, mutta hieman oikeammalla. Ei-natiiveilla lyhyet sanat siis korostuvat enemmän, kun taas natiiveilla joukossa on enemmän pitkiä sanoja. Tämä on linjassa sen aiemman havainnon kanssa, että natiivien käyttämät saneet ovat hieman kompleksisempia kuin ei-natiivien, eli joukossa on suhteessa enemmän yhdyssanoja ja erilaisia johdoksia.

4.1.6 Yleisimmät lekseemit

Yleisimmät laillistamiskokeessa käytetyt lekseemit (Taulukko 9) ovat paljolti samoja kuin taajuussanaston (CSC 2004) frekventeimmät sanat, mutta järjestys on toinen. Sekä natiiveilla että ei-natiiveilla kaikkein yleisimpiä ovat sanat *olla* ja *ja* niin kuin taajuussanastossa. Esimerkiksi taajuussanaston kolmantena oleva *että* on kuitenkin laillistamiskokeen aineistossa vasta sijalla 26, eikä se ole natiiveiksi luettujen ryhmän listassa edes 40 yleisimmän sanan joukossa.

Koekontekstin vaikutuksesta yleisimmissä sanoissa on paljon lääkärin vastaanotolla ja koevastauksissa tarvittavaa sanastoa, joka ei taajuussanastossa ole yltänyt yleisimmille sijoille. Vastaustekstien 40 yleisimmän lekseemin kärjessä ovat muun muassa *potilas*, *kipu* ja *sydän*, jotka ovat taajuussanastossa vasta sijoilla 847, 3043 ja 1232.

Lekseemin *jos* sijoitus aineiston yleisimpien sanojen listauksessa on uskoakseni osoitus siitä, kuinka kuulustelutilanne on vaikuttanut käytettyyn kieleen. Se on aineistossa jopa kolmanneksi yleisin lekseemi, vaikka taajuussanastossa se yltää vasta sijalle 42. Ero selittyy suoraan kokeen

luonteella: siihen osallistuvien on selostettava, kuinka he toimisivat tehtävänannossa kuvaillun potilaan tapauksessa. Kokeen luonteen vuoksi aineistossa on mukana esimerkiksi tällaisia virkkeitä:

Jos on mikrosyttinen anemia aikuisella sen tarvitsee tutkia ja löytää syy.
Jos potilaalla on astmapäiväkirja, pyydän se.

Taulukko 9. Yleisimmät lekseemit kieliryhmittäin, kaikki erityisyysasteet.

	kaikki	%	ei-natiivi	%	natiivi	%
1	olla	7,06 %	olla	7,24 %	olla	5,92 %
2	ja	3,01 %	ja	2,99 %	ja	3,11 %
3	jos	1,84 %	jos	1,89 %	jos	1,51 %
4	voida	1,35 %	potilas	1,43 %	ei	1,37 %
5	potilas	1,33 %	voida	1,39 %	voida	1,06 %
6	tai	1,26 %	tai	1,34 %	tai	0,75 %
7	ei	1,11 %	ei	1,07 %	potilas	0,72 %
8	hän	0,92 %	hän	1,06 %	kipu	0,65 %
9	kipu	0,69 %	se	0,74 %	muu	0,63 %
10	se	0,68 %	kipu	0,70 %	mahdollinen	0,60 %
11	mikä	0,60 %	mikä	0,64 %	tulla	0,56 %
12	sydän	0,59 %	sydän	0,62 %	myös	0,51 %
13	palpaatio	0,55 %	palpaatio	0,56 %	joka	0,51 %
14	muu	0,53 %	oire	0,52 %	tarvita	0,48 %
15	oire	0,51 %	muu	0,51 %	ottaa	0,46 %
16	hoito	0,46 %	hoito	0,49 %	sekä	0,46 %
17	iho	0,46 %	keuhko	0,47 %	sydän	0,44 %
18	tarvita	0,44 %	iho	0,46 %	koholla	0,44 %
19	keuhko	0,44 %	tarvita	0,44 %	palpaatio	0,43 %
20	anemia	0,41 %	anemia	0,42 %	iho	0,43 %
21	auskultaatio	0,39 %	auskultaatio	0,41 %	oire	0,41 %
22	vatsa	0,38 %	vatsa	0,39 %	syy	0,41 %
23	myös	0,37 %	että	0,37 %	kyse	0,39 %
24	lääke	0,34 %	lääke	0,36 %	mukaan	0,39 %
25	status	0,34 %	koska	0,35 %	tulehdus	0,38 %
26	että	0,33 %	myös	0,35 %	lapsi	0,36 %
27	tutkimus	0,32 %	tutkimus	0,34 %	infektio	0,36 %
28	koska	0,32 %	status	0,34 %	sairaus	0,34 %
29	syy	0,32 %	tehdä	0,32 %	merkki	0,34 %
30	tehdä	0,31 %	tarkistaa	0,31 %	aloittaa	0,34 %
31	jälkeen	0,30 %	tauti	0,31 %	se	0,33 %
32	tauti	0,29 %	jälkeen	0,31 %	anemia	0,33 %
33	lapsi	0,28 %	syy	0,30 %	verenpaine	0,33 %
34	ottaa	0,28 %	kysyä	0,29 %	lääkitys	0,33 %
35	tarkistaa	0,28 %	tilanne	0,28 %	mikä	0,31 %
36	sairaus	0,28 %	lapsi	0,27 %	hoito	0,31 %
37	kysyä	0,27 %	sairaus	0,27 %	vatsa	0,31 %
38	tilanne	0,27 %	väri	0,27 %	status	0,31 %
39	väri	0,27 %	pitää	0,27 %	muutos	0,29 %
40	kyse	0,25 %	murtuma	0,26 %	normaali	0,29 %

4.2 Tarkemman analyysin rajaaminen yleissanastoon

Jo lähtökohtaisesti olen rajannut aineiston tarkemmin analysoitavasta osasta pois koehenkilöiden käyttämän **englanninkielisen** sanaston, koska päämääräni on tarkastella suomen kielen sanastotaitoja. Poikkeuksina mukana ovat sellaiset alkujaan englanninkieliset sanat, jotka ovat löytäneet tiensä myös suomen kielen sanastoon vakiintuneina lainoina. Esimerkiksi käy suomen yleiskielessäkin varsin tavallinen sana *burnout*. Vastaavasti myöskään koevastauksissa käytetty **erikoisammattisanasto** eli ei nähdäkseni ole olennaista tutkittaessa suomen kielen sanastonhallintaa, sillä se on hyvin universaalia, lähinnä ammattilaisten ymmärtämää erikoiskieltä ja latinaa. Tässä tapauksessa erikoisammattisanaston laajuus ja monipuolisuus eivät siis nähdäkseni kerro ulkomailta tulevien lääkäreiden suomen sanastotaidoista, vaan pikemminkin sen huomioiminen analyysissa vääristäisi tuloksia. Näistä syistä erikoisammattisanasto on mukana tutkimuksessani lähinnä vain siinä yhteydessä, kun suhteutan sen määrää muuhun käytettyyn sanastoon, ja tarkemmassa analyysiosuudessa jätän sen surutta tarkasteluni ulkopuolelle. Lisäksi karsin tarkemmasta analyysistäni pois **terveyssanaston**. Päädyin ratkaisuun, koska nähdäkseni sekin kuuluu jo lähtökohtaisesti koehenkilöiden erityisosaamisalueelle. Sen aineistossa mukana pitäminen vääristäisi kuvaa koehenkilöiden yleisistä suomen kielen ja sanastotaidoista, joihin pyrin kohdentamaan fokukseni.

Myös **lyhenteet** on karsittu pois analyysista jo aiemmassa vaiheessa. Tarkemmin tarkasteltuna niistäkin ylivoimainen enemmistö kuuluisi luonteeltaan kuitenkin tässä vaiheessa karsiutuvaan lääkäreiden erikoisammattisanastoon, sillä ne ovat suurelta osin erilaisten laboratoriokokeiden ja niihin liittyvien arvojen lyhenteitä (esim. *PEF*, *S-krea*, *proBNP*). Kaltaiseni maallikon on monin paikoin hyvin vaikea tulkita, mihin lyhenteet viittaavat. Esimerkiksi pelkkä *EKG* tai *rtg* on karsiutunut aineistosta. Lyhenteet ovat silti edelleen mukana aineistossa niissä tapauksissa, joissa ne ovat selkeästi yhdyssanan määriteosina (esim. *EKG-tutkimus*, *rtg-kuva*).

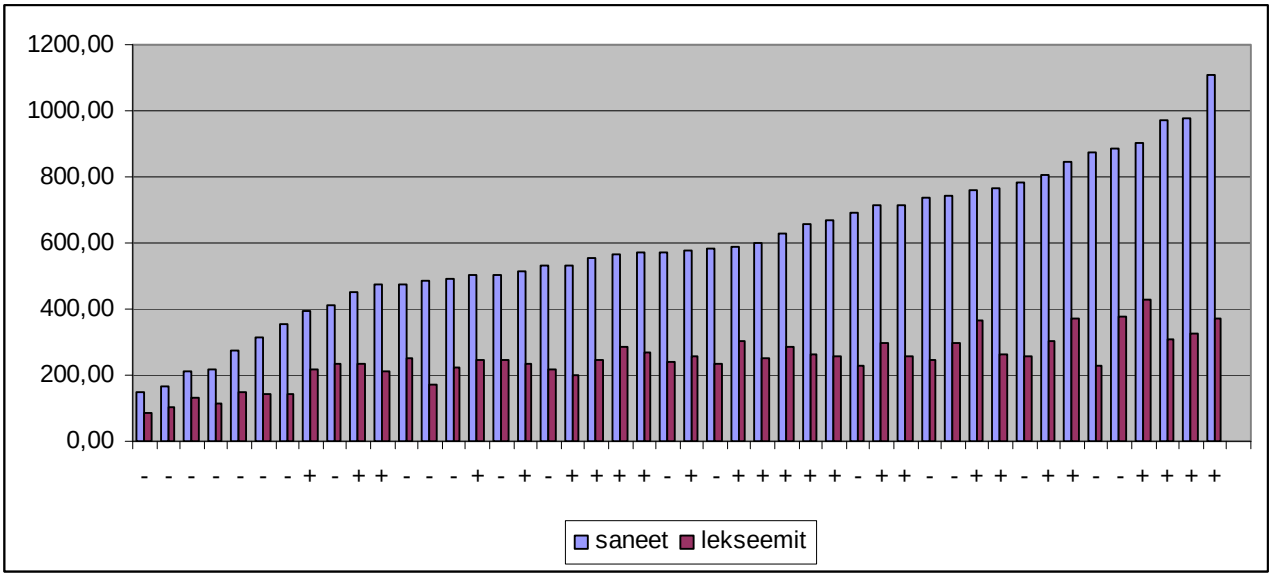
4.3 Saneet ja lekseemien frekvenssit

4.3.1 Tekstien pituus: saneiden ja lekseemien määrän vaihtelu

Kuvaajan 3 avulla voi hahmottaa yleiskuvan analysoitujen vastaustekstien pituuden vaihtelusta. Siinä pidemmät pylväät kuvaavat vastaajien sanemääriä ja lyhyemmät lekseemien määriä. Sanemäärä kertoo yleissanastoon luokiteltujen saneiden kokonaismäärän lekseemien toiston mukaan lukien, kun taas lekseemien määrässä jokainen lekseemi on laskettu vain yhden kerran. Koehenkilöt on järjestetty kuvaajaan tekstin pituuden mukaan. Saneilla mitaten analysoitavien tekstien pituus

vaihteli 147 saneesta 1111 saneeseen, lekseemien määrä puolestaan 85 lekseemistä 431 lekseemiin. Kuvaajan pylväiden alla olevat plusmerkit viittaavat siihen, että koehenkilö on yltänyt laillistamiskokeen osassa hyväksyttyyn suoritukseen, miinusmerkki puolestaan viestii päinvastaisesta. Siitä näkee, että hylätyt suoritukset painottuvat niille, jotka ovat kirjoittaneet kaikkein lyhyimmät vastaukset, mutta silti kaikkien paljon tekstiä tuottaneidenkaan suorituksia ei suinkaan ole hyväksytty.

Kuvaaja 3. Saneiden ja lekseemien määrät koehenkilöittäin sanemäärien mukaan.



Kuvaaja osoittaa, että vastauksissa lekseemien määrä ei kasva suoraan verrannollisesti sanemäärän kanssa. Sen sijaan samanpituisilla teksteillä saattaa olla huomattava ero lekseemien määrässä ja pidemmässä tekstissä saattaa olla vähemmän lekseemejä kuin lyhyemmässä. Toisin sanoen toiset koehenkilöt ovat toistaneet samoja sanoja verraten enemmän kuin toiset.

4.3.2 Yleissanaston yleisimmät lekseemit

Taulukko 10 esittää lääkäreiden eniten käyttämien lekseemien lukumäärät, prosenttiosuudet analyysissa mukana olleesta yleissanastosta ja kyseiseen lekseemiin asti kertyneet kumulatiiviset prosentit eli sen, kuinka suuren osuuden lekseemi ja sitä yleisemmin käytetyt lekseemit kattavat analyysin aineistosta. Listat ja prosenttiosuudet kuvaavat yleisimpien sanojen osuuden vastaajaryhmittäin niin, että ne kertovat koko joukon käyttämästä leksikosta, eivät yksittäisistä vastaajista suoraan.

Taulukko 10. Käytetyn yleissanaston 50 yleisintä lekseemiä koehenkilön kielen mukaan.

	ei-natiivi				natiivi			
	lekseemi	**lkm**	**%**	**kum. %**	**lekseemi**	**lkm**	**%**	**kum. %**
1	olla	2684	11,38 %	11,38 %	olla	346	9,30 %	9,30 %
2	ja	1110	4,71 %	16,09 %	ja	182	4,89 %	14,19 %

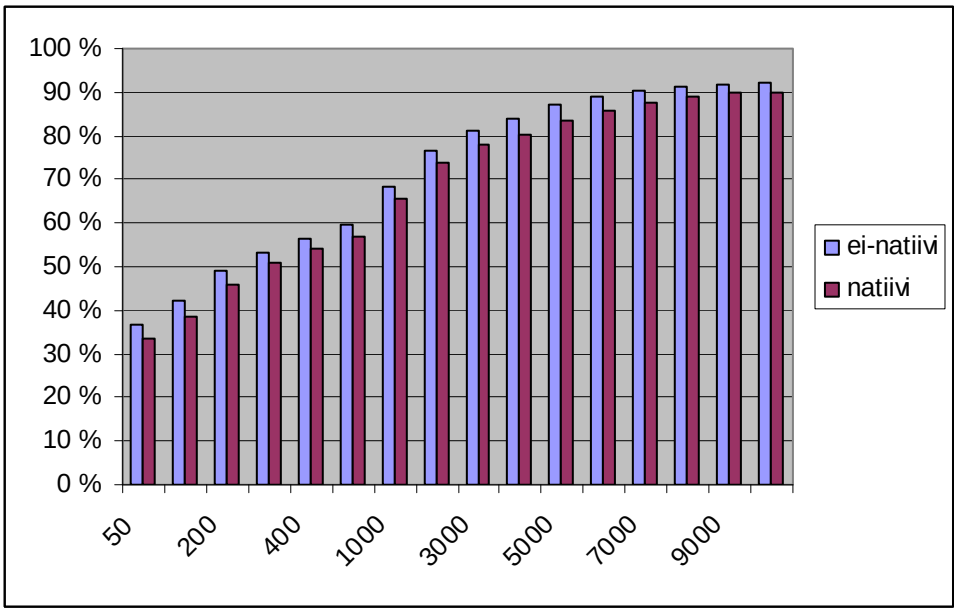
3	jos	702	2,98 %	19,06 %	jos	88	2,37 %	16,56 %
4	voida	517	2,19 %	21,25 %	ei	80	2,15 %	18,71 %
5	tai	498	2,11 %	23,36 %	voida	62	1,67 %	20,38 %
6	ei	398	1,69 %	25,05 %	tai	44	1,18 %	21,56 %
7	hän	394	1,67 %	26,72 %	muu	37	0,99 %	22,55 %
8	se	274	1,16 %	27,88 %	mahdollinen	35	0,94 %	23,49 %
9	mikä	239	1,01 %	28,90 %	tulla	33	0,89 %	24,38 %
10	muu	190	0,81 %	29,70 %	joka	30	0,81 %	25,19 %
11	iho	172	0,73 %	30,43 %	myös	30	0,81 %	25,99 %
12	tarvita	163	0,69 %	31,12 %	tarvita	28	0,75 %	26,75 %
13	vatsa	145	0,61 %	31,74 %	ottaa	27	0,73 %	27,47 %
14	että	136	0,58 %	32,31 %	sekä	27	0,73 %	28,20 %
15	koska	131	0,56 %	32,87 %	iho	25	0,67 %	28,87 %
16	myös	129	0,55 %	33,42 %	syy	24	0,65 %	29,52 %
17	tehdä	119	0,50 %	33,92 %	kyse	23	0,62 %	30,13 %
18	tarkistaa	116	0,49 %	34,41 %	mukaan	23	0,62 %	30,75 %
19	jälkeen	114	0,48 %	34,90 %	lapsi	21	0,56 %	31,32 %
20	syy	113	0,48 %	35,38 %	aloittaa	20	0,54 %	31,85 %
21	kysyä	107	0,45 %	35,83 %	merkki	20	0,54 %	32,39 %
22	tilanne	102	0,43 %	36,26 %	se	19	0,51 %	32,90 %
23	lapsi	101	0,43 %	36,69 %	mikä	18	0,48 %	33,39 %
24	pitää	99	0,42 %	37,11 %	vatsa	18	0,48 %	33,87 %
25	väri	99	0,42 %	37,53 %	muutos	17	0,46 %	34,33 %
26	miten	94	0,40 %	37,93 %	niin	17	0,46 %	34,78 %
27	ottaa	94	0,40 %	38,33 %	normaali	17	0,46 %	35,24 %
28	tämä	93	0,39 %	38,72 %	tämä	16	0,43 %	35,67 %
29	paino	91	0,39 %	39,11 %	viitata	16	0,43 %	36,10 %
30	aika	89	0,38 %	39,48 %	ainakin	15	0,40 %	36,51 %
31	kyse	86	0,36 %	39,85 %	tehdä	15	0,40 %	36,91 %
32	kanssa	83	0,35 %	40,20 %	väri	15	0,40 %	37,31 %
33	kuinka	83	0,35 %	40,55 %	kertoa	14	0,38 %	37,69 %
34	silmä	79	0,33 %	40,89 %	vuoksi	14	0,38 %	38,06 %
35	mutta	75	0,32 %	41,20 %	jälkeen	13	0,35 %	38,41 %
36	katsoa	73	0,31 %	41,51 %	käyttö	13	0,35 %	38,76 %
37	hyvä	72	0,31 %	41,82 %	tilanne	13	0,35 %	39,11 %
38	joka	72	0,31 %	42,12 %	hyvä	12	0,32 %	39,44 %
39	päivä	72	0,31 %	42,43 %	johtua	12	0,32 %	39,76 %
40	tulla	70	0,30 %	42,73 %	nämä	12	0,32 %	40,08 %
41	paljon	69	0,29 %	43,02 %	saada	12	0,32 %	40,40 %
42	kun	68	0,29 %	43,31 %	yhteys	12	0,32 %	40,73 %
43	olkapää	68	0,29 %	43,60 %	mutta	11	0,30 %	41,02 %
44	äiti	68	0,29 %	43,88 %	ne	11	0,30 %	41,32 %
45	käyttää	64	0,27 %	44,16 %	viikko	11	0,30 %	41,61 %
46	koko	63	0,27 %	44,42 %	aiheuttaa	10	0,27 %	41,88 %
47	mahdollinen	63	0,27 %	44,69 %	katsoa	10	0,27 %	42,15 %
48	minkälainen	63	0,27 %	44,96 %	lämpö	10	0,27 %	42,42 %
49	alkoholi	62	0,26 %	45,22 %	poissulkeminen	10	0,27 %	42,69 %
50	sitten	62	0,26 %	45,48 %	sitten	10	0,27 %	42,96 %

Kuten kaikki erityisyysasteet kattavassa listassakin (Taulukko 9), frekventeimpiä lekseemejä ovat *olla*, *ja* ja *jos*. Tälläkin kertaa (Taulukko 10) erityisesti olla-verbin osuus on huomattavan suu-

ri, yli kymmenesosa vastaustekstien kaikista erityisyysasteeltaan yleissanastoon luokitelluista saneista. Myös konjunktion *ja* viittä prosenttia lähestyvä osuus erottuu joukosta. Konjunktio *jos* sijoittuu listalla jo kolmanneksi, vaikka taajuussanastossa (CSC 2004) se on vasta 42. sijalla. Uskoakseni koekonteksti on kasvattanut sen taajuutta. Listan lekseemien yleisyysprosenttien rinnalle laskemistani kumulatiivisista prosenteista puolestaan voi nähdä, että ei-natiiveilla frekventeimmät lekseemit kattavat jatkuvasti pari prosenttia suuremman osan kaikista saneista kuin natiiveilla. Ero syntyy kymmenen tiuhimpaan käytetyn lekseemin joukossa ja erityisesti sillä, että ei-natiivit käyttävät *olla*-verbiä selvästi enemmän kuin natiivit.

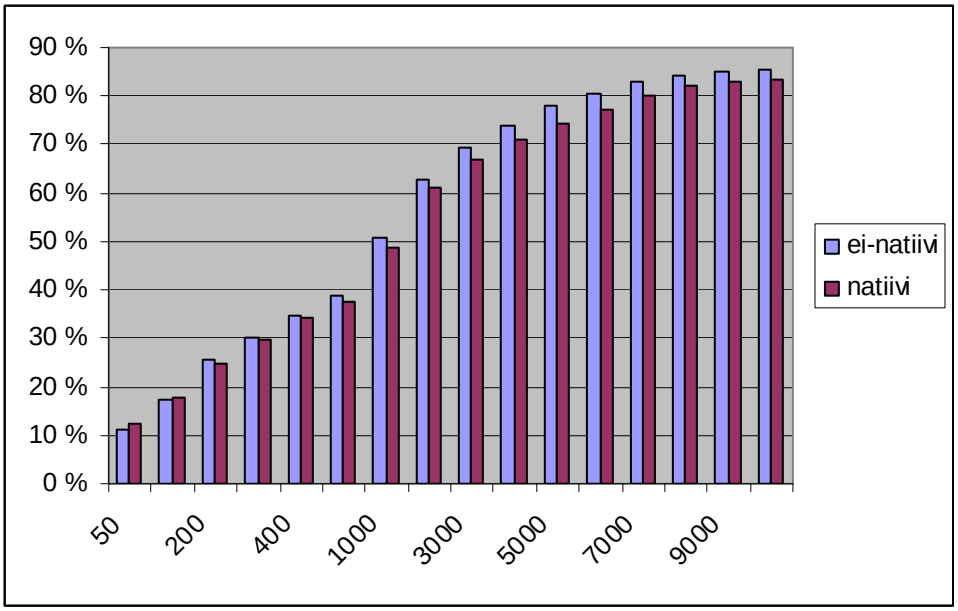
Taajuussanastoa hyödyntämällä saa laajempaa kokonaiskuvaa siitä, kuinka yleistä kokeen aineiston leksikko on suhteessa kokeen ulkopuolella käytettyyn suomen kieleen. Esimerkiksi 50 suomen kielen yleisintä lemmaa (CSC 2004) kattaa saneista keskimäärin jo yli 36 prosenttia eli selvästi yli kolmanneksen yksittäisten koevastausten koko **sane**määrästä (Kuvaaja 4). Tuhat yleisintä lemmaa puolestaan kattaa koko joukolla keskimäärin lähes 68 prosenttia käytetyistä saneista. 7000 yleisintä lemmaa yltää keskimäärin jo 90 prosenttiin koevastausten sanemäärästä. Kuvaaja 4 havainnollistaa, että ei-natiiveilla koehenkilöillä kumulatiiviset prosentit ovat kaikissa sanaston yleisyysluokissa hieman suurempia kuin natiiveiksi tulkituilla. Toisin sanoen ei-natiivit näyttävät käyttävän keskimäärin yleisempää sanastoa kuin natiivit, mikä on odotuksenmukaista. Heillä yleiset sanat toistuvat enemmän kuin natiiveilla koehenkilöillä.

Kuvaaja 4. Käytettyjen yleissanaston saneiden kumulatiivisten prosenttien keskiarvo kielen ja taajuussanaston yleisyyksien mukaan.



Vastaava käytettyjen yleissanaston lekseemien ja kokonais**lekseemi**määrien analyysi (Kuvaaja 5) tarjoaa selvästikin pienempiä prosenttilukuja, koska siinä tarkastellaan käytettyä sanastoa huomioimatta saman lekseemin toistoa. Tässäkin jokainen lekseemi on laskettu kustakin vastauksesta siis vain kerran. Esimerkiksi 50 taajuussanaston (CSC 2004) yleisintä lemmaa kattaa keskimäärin runsaat 11 prosenttia yksittäisten koehenkilöiden käyttämistä lekseemeistä. Sen 1000 yleisintä lemmaa riittää kattamaan yli 50 prosenttia ja 9996 yleisintä noin 85 prosenttia koevastauksista.

Kuvaaja 5. Käytettyjen yleissanaston lekseemien kumulatiivisten prosenttien keskiarvot kielen ja taajuussanaston yleisyyksien mukaan.



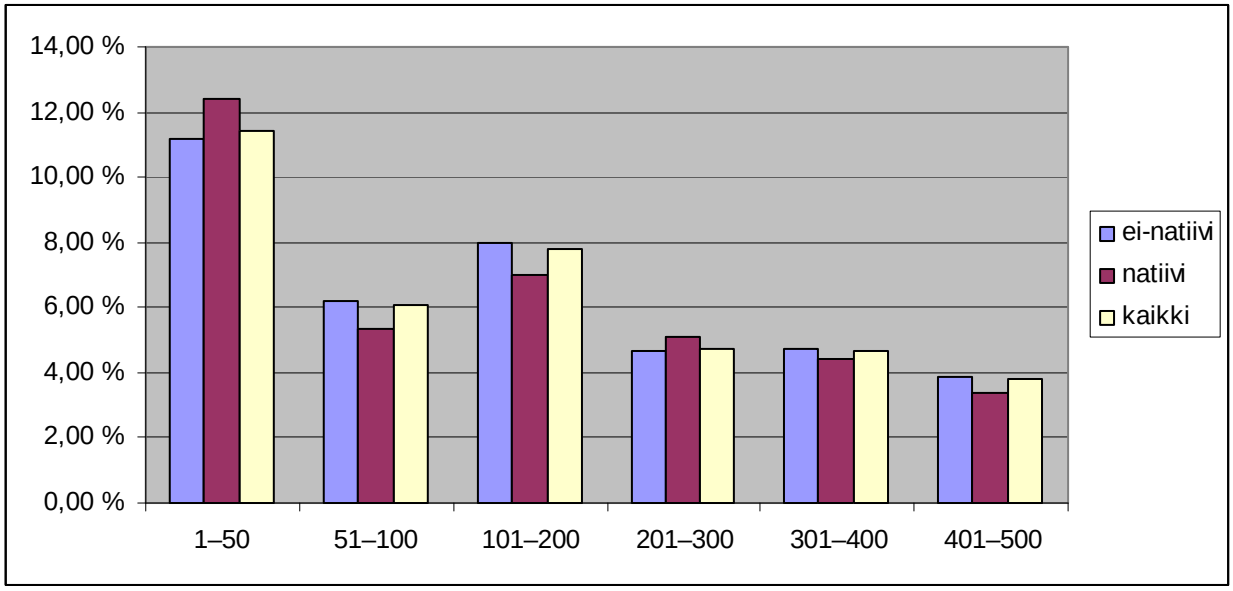
Kuvaaja 5 havainnollistaa kahden lääkäriryhmän välisiä eroja, jotka vaikuttavat osin yllättäviltä: analyysin mukaan natiiveilla koehenkilöillä kaikkein yleisimpien lekseemien osuudet vaikuttavat olevan suurempia kuin ei-natiiveilla. Jo 300 frekventeimmällä lemmalla laskettuna ei-natiivien kumulatiiviset prosentit ovat kuitenkin korkeampia kuin natiiveilla, ja ero kasvaa parin prosentin mittaluokkaan laajennettaessa tarkastelua aina 9996 yleisimpään lemmaan saakka.

4.3.3 Suomen yleisimpien lekseemien osuus yleissanastosta

Laillistamiskokeeseen osallistuneiden lääkäreiden käyttämistä lekseemeistä huomattavan suuri osa lukeutuu Sanomalehtikielen taajuussanaston yleisimpien lemmojen joukkoon. Kuten kuvaajasta 6 on nähtävissä, keskimäärin yli 11 prosenttia koehenkilöiden käyttämistä lekseemeistä kuuluu 50 yleisimmän kärkeen. Kuvaajasta on huomattava, että sen kaksi ensimmäistä frekvenssiryhmää käsittävät kumpikin vain 50 taajuussanaston kärkipään lekseemiä, kun taas seuraavat kukin sata lekseemiä. Sadan yleisimmän lekseemin kärki muodostaa yhteensä 17,40 prosentin osuuden koehenkilöi-

den käyttämästä yleissanastosta. Pylväät esittävät yksittäisistä koevastauksista laskettujen prosenttiosuuksien keskiarvoja.

Kuvaaja 6. Taajuussanaston yleisimpien lekseemien osuuksia koehenkilöiden käyttämän yleissanaston lekseemeistä, koehenkilöiden keskiarvo.



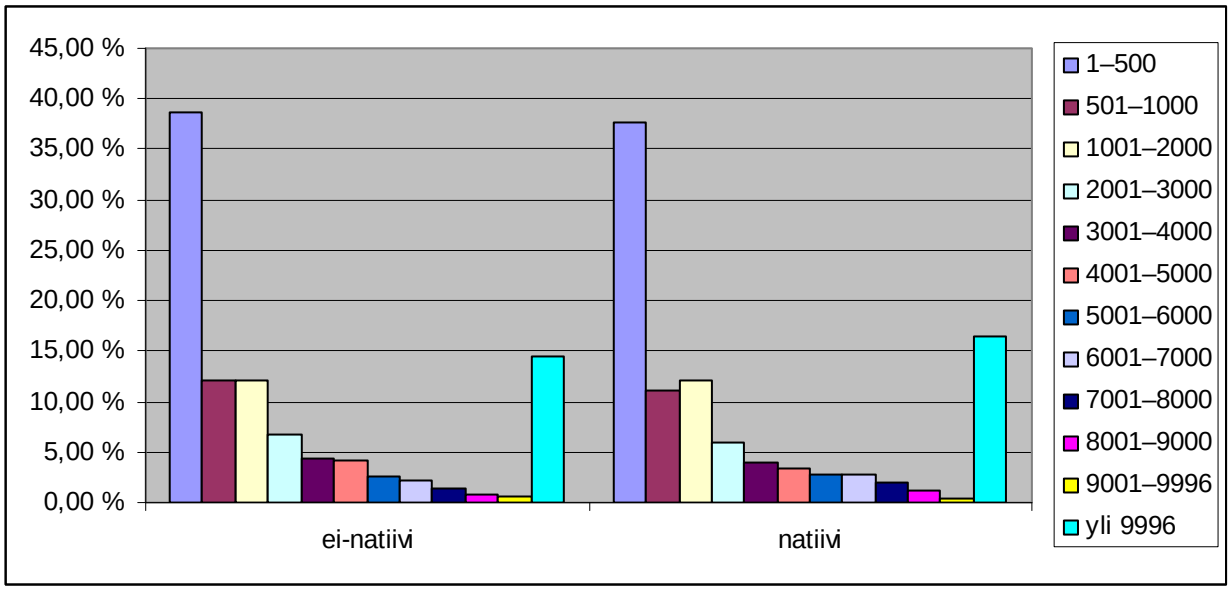
Kuvaaja 6 osoittaa, että natiiveiksi suomen puhujiksi tulkittujen lääkäreiden vastauksissa suomen kielen 50 yleisimmän lekseemin osuus on suurempi kuin ei-natiiveilla. Tätä voi siis luonnehtia yllätykseksi. Odotuksenmukaisesta olisi ollut, että ei-natiivit koehenkilöt olisivat tukeutuneet enemmän kaikkein yleisimpään sanastoon, joka oletettavasti on heille tutuinta, ja että suomea oletettavasti taitavammin käyttävillä aivan yleisimpien lekseemien osuus olisi vastaavasti pienempi. Ryhmien välinen ero on kuitenkin sen verran vähäinen, että se on voinut syntyä esimerkiksi tässä tutkimuksessa tehtyjen aineiston rajausten vuoksi. Analysoitavasta aineistosta on aiemmin karsittu pois muun muassa terveyssanasto ja lyhenteet, joita natiivit puhujat mahdollisesti osaavat käyttää muunkielisiä laajemmin. Toki ero voi johtua myös vastausten laadusta, sillä vastaustekstien pituuserot ja niiden saamat pistemääräerot ovat suuria paitsi koko aineistossa myös natiiveiksi tulkittujen koehenkilöiden välillä.

Yksi selitys natiivien suuremmalle 50 yleisimmän lekseemin osuudelle voi olla myös kielitaidon kehittyneisyys. Honko (2013: 352) havaitsi, että koululaisten kielitaidon kehittyessä alakoulun aikana tekstin tuottaminen aikuismaistui, mikä merkitsi käytännössä muun muassa virkerakenteen kompleksistumista. Virkerakenteiden kompleksistuminen puolestaan voi merkitä suurempaa konjunktioiden käyttöä, koska konjunktioilla yhdistellään lauseita ja merkitään niiden välisiä suhteita. Taajuussanaston 50 yleisimmän joukkoon mahtuu kahdeksan frekventeintä konjunktiota (*ja, että, mutta, kun, kuin, tai, sekä, jos*). Kuitenkin Saarela (1997: 80–82, 181) havaitsi, että esimerkiksi tois-

luokkalaisilla koululaisilla funktiosanojen, kuten konjunktioiden, osuus teksteissä on suurempi kuin yläkoulun kahdeksasluokkalaisilla. Kummallakin esimerkiksi konjunktioita on paljon, mutta pienillä koululaisilla määrä selittyi ketjuvirkkeillä ja asioiden yhdistelemisellä *ja*-sanan avulla, kun taas kielenkäyttötaidoissaan edistyneemmillä kahdeksasluokkalaisilla konjunktioiden määrää selittää virkkeiden monimutkaisuus ja sivulauseiden käyttö (emt.). Laillistamiskoeaineiston natiivien ja ei-natiivien yleisimpien lekseemien käytön erojen syiden varmistaminen vaatisi näin ollen tarkempaa perehtymistä yksittäisten koehenkilöiden sanastoon ja myös virkerakennetason tarkastelua.

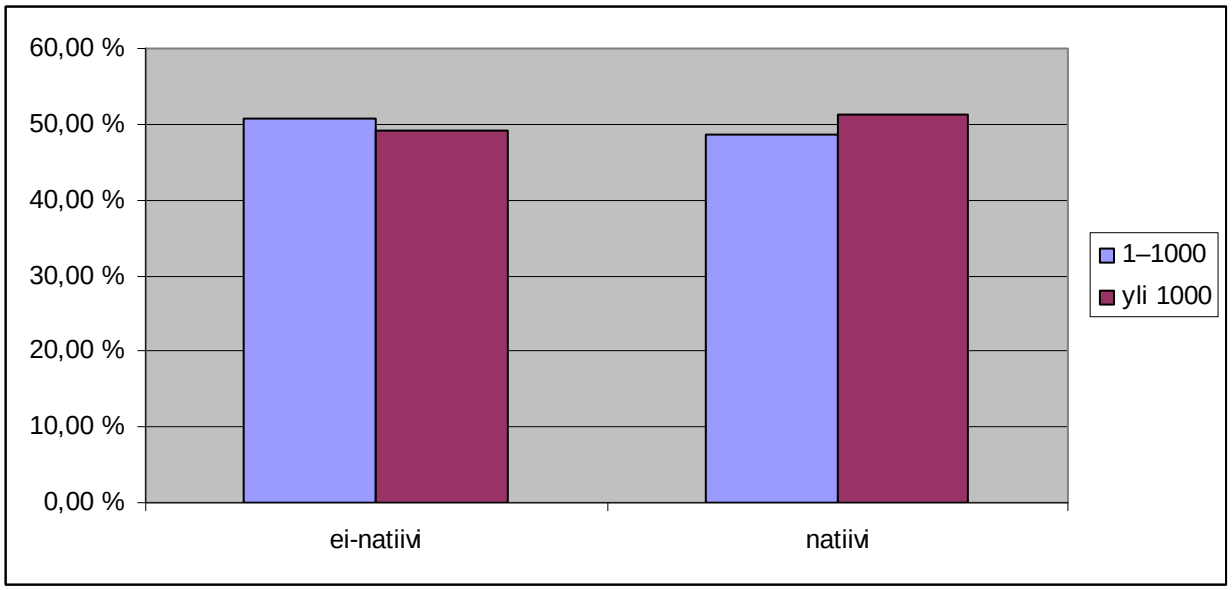
Huomattava on silti myös se, että kun yhtäältä natiivit koehenkilöt ovat käyttäneet ei-natiiveja enemmän suomen kielen yleisimpien lekseemien kärkeä, he ovat toisaalta käyttäneet enemmän myös aineiston harvinaisimpiin kuuluvia lekseemejä. Tämä on nähtävissä selvästi kuvaajasta 7, joka kertoo, kuinka paljon koehenkilöryhmät ovat käyttäneet eri yleisyysluokkiin kuuluvia yleissanaston lekseemejä. Kuvaajan pylvässarjojen kaikkein vasemmanpuoleiset eli korkeimmat pylväät kattavat samat taajuussanaston 500 yleisintä lemmaa kuin koko kuvaaja 6.

Kuvaaja 7. Koehenkilöiden käyttämien lekseemien määrät yleisyysasteittain Sanomalehtikielen taajuussanaston mukaan, yksittäisten koehenkilöiden keskiarvo.



Keskimäärin noin puolet käytetyistä lekseemeistä kuuluu suomen tuhannen frekventeimmän joukkoon (Kuvaaja 8). Tässä suhteessa natiivien ja ei-natiivien käyttämässä sanastossa ei ole suuria eroja. Natiiveilla koehenkilöillä tuhannen yleisimmän lekseemin osuus jää kuitenkin keskimäärin hieman 50 prosentin alapuolelle, kun taas ei-natiiveilla niiden osuus kipuaa hienoisesti 50 prosenttia suuremmaksi.

Kuvaaja 8. Tuhannen yleisimmän ja sitä harvinaisempien lekseemien käyttö koehenkilön kielen mukaan, yksittäisten koehenkilöiden keskiarvo.



Taajuussanaston 9996 lekseemin ulkopuolelle jääneiden eli suomen kielen harvinaisimpien lekseemien määrä osoittautuu analysoidussa yleissanastoaineistossa yllättävän suureksi (keskimäärin 14,83 % lekseemeistä). Tähän on erilaisia syitä. Joukossa on esimerkiksi paljon sellaista kuvailevaa sanastoa, joka on luonteeltaan yleistä ja helposti ymmärrettävää ja arkisessa käytössä muttei kuitenkaan frekventtiä lukumääriä tarkastellen (esim. *kellertävä*, *rahina*, *vinkuna*, *kumista*). Joukossa on myös ei-natiivien lääkäreiden omia muodosteita (esim. *yöaamu*, *ohimeno*). Lähempi tarkastelu osoittaa myös, että taajuussanastoon lukeutumattomien lekseemien määrä johtuu osin tässä tutkimuksessa käytetyistä sanojen erityisyysasteiden (yleissanasto, terveyssanasto, lääkärien erikoisammattisanasto) luokitteluperusteista: yleissanaston joukkoon on päätynyt paljon sellaista sanastoa, joka on tyypillistä erityisesti terveysalalla, vaikkakin luonteeltaan muuten melko yleistä (esim. *hengitys*, *lantio*, *kalpea*, *alkoholinkäyttö*). Mahdollisesti luokitteluperusteita olisi kannattanut johdonmukaistaa vielä enemmän ennen analyysivaihetta.

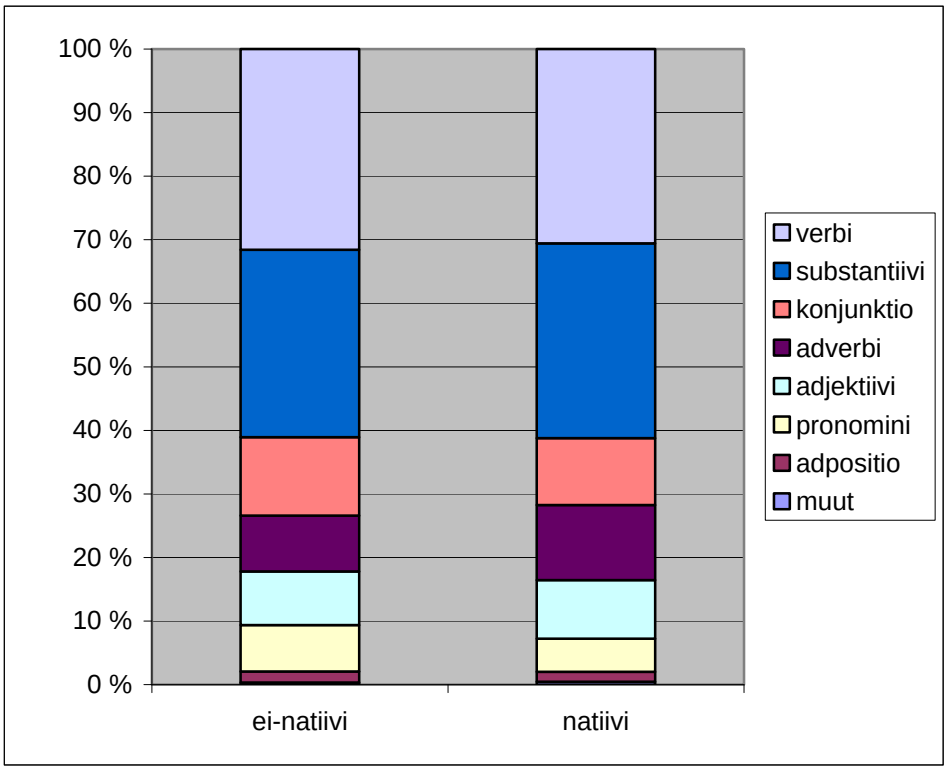
4.4 Yleisimmät lekseemit sanaluokittain

Tässä alaluvussa esittelen koehenkilöiden tuottamien vastaustekstien yleissanaston sanaluokkajakaumaa ja kunkin sanaluokan yleisimpien lekseemien kärkeä kieliryhmittäin. Vertaan ei-natiivien ryhmän sanastoa kokonaisuutena natiivien ryhmänä käyttämään sanastoon. Tähän ratkaisuun päädyin erityisesti siitä syystä, että sekä koehenkilöittäin että sanaluokittain tarkastelemista ajatellen vastaustekstit voivat nähdäkseni olla osin aivan liian lyhyitä.

Yleissanaston sanaluokkajakaumien perusteella ei-natiivien ja natiivien koehenkilöiden väliset erot eivät ryhmittäin tarkasteltuina ole kovin suuria. Kummallakin ryhmällä selvästi yleisimpiä ovat verbit ja substantiivit, minkä näkee hyvin kuvaajasta 9, ja substantiivien ja verbien yhteinen osuus yleissanastosta on molemmilla ryhmillä noin 61 %. Ei-natiiveilla verbien ja substantiivien suhde kallistuu kuitenkin aavistuksen verbivoittoisammaksi (verbejä 31,57 % – substantiiveja 29,50 %) kuin natiiveilla (substantiiveja 30,65 % – verbejä 30,56 %).

Muiden sanaluokkien käytössä on havaittavissa selvempiä eroavaisuuksia. Ei-natiivit ovat käyttäneet enemmän konjunktioita (12,32 %) ja pronomineja (7,31 %) kuin natiivit (10,54 % ja 5,22 %). Natiivien kielenkäytössä taas adverbit (11,80 %) ovat selvästi tavallisempia kuin ei-natiiveilla (8,80 %). Havainnot vihjaavat tarkasteltujen lääkäriryhmien kielen leksikaalisen tiheyden (ks. esim. Read 2000: 200) eroista: natiivien teksteissä sisältösanojen osuus on suurempi ja siksi siinä lienee enemmän sisältöä suhteessa sanamäärään.

Kuvaaja 9. Yleissanastoon kuuluvien saneiden sanaluokkajakaumat koehenkilöiden kielen mukaan.



Sanaluokkakohtaisiin frekventeimpien lekseemien listoihin (taulukot 11–17) olen kirjannut rinnakkain natiivien, ei-natiivien ja koko aineiston yleisimmin käytetyt lekseemit yleisyysjärjestyksessä. Taulukon sarakkeista voi lukea lekseemin, sen esiintymien määrän sekä osuuden samaa sanaluokkaa edustavista lekseemeistä. Listan alta on luettavissa, kuinka suuren osuuden muodostavat 3, 10 ja

30 sanaluokan yleisintä lekseemiä. Jos sarakkeita koskevassa aineistossa, esimerkiksi koko natiivien yleissanastossa, on ollut vähemmän kuin 30 samaa sanaluokkaa edustavaa lekseemiä, lukumäärä on merkitty 30:n paikalle taulukon alle. Tiiviyssyistä olen lyhentänyt niiden sanaluokkien listoja, joissa eri lekseemejä on ollut vähiten. Niiden häntäpäässä oli useita vain korkeintaan muutaman esiintymän lekseemejä.

Ilman erillistä laillistamiskokeen tehtävänantojen analyysiäkin on ilmeistä, että tehtävänannot ovat vaikuttaneet käytettyyn sanastoon. Nähdäkseni vaikutus on silminnähtävä esimerkiksi frekventeimpien substantiivien listassa: eri vastauksissa toistuvat eniten esimerkiksi sellaiset ruumiinosat tai elimet, joista tai joihin liittyvistä vaivoista kokeen kysymyksissä on selvästikin puhuttu. Joukkoon kuulunevat muun muassa lekseemit *iho*, *vatsa* ja *lapsi*.

Yleisimpien sanojen sanaluokkakohtaisia listoja kootessani havaitsin joitain käsittelyvaiheessa aineistoon jääneitä lyönti- tai luokitteluvirheitä. Esimerkiksi *vaiko* eli *ko*-liitepartikkelilla varustettu konjuntio *vai* oli erillään muista *vai*-esiintymistä. Konjunktioiden joukossa oli myös koko *ikään kuin* -rakenne yhteen kirjoitetussa muodossa. Muutama yksittäinen sanaesiintymä oli myös luokiteltu väärään sanaluokkaan, esimerkiksi substantiivi *taite* konjunktioihin. Tämänkaltaiset aineiston puutteet osoittautuivat kuitenkin yksittäisiksi ja jäävät kokonaisuuteen nähden vaikutuksiltaan käytännössä merkityksettömiksi.

Substantiiveilla yleisimpien lekseemien kärki kattaa selvästi pienemmän osuuden kaikista saman sanaluokan sanoista kuin muilla sanaluokilla, eli toisin sanoen yleisimpiä substantiiveja käytetään tasaisemmin. Kaikkein yleisin substantiivi (*iho*) ei kata kuin pari prosenttia käytetystä substantiivien joukosta ja kolmen kärkikin (*iho*, *vatsa*, *syy*) vain kuutisen prosenttia. 30 yleisimmänkin kärki kattaa vain runsaat 30 prosenttia yleissanaston kaikista substantiiveista.

Taulukko 11. Käytetyn yleissanaston frekventeimmät substantiivit.

	ei-natiivi			natiivi			kaikki		
	lekseemi	**kpl**	**%**	**lekseemi**	**kpl**	**%**	**lekseemi**	**kpl**	**%**
1	iho	172	2,47 %	iho	25	2,19 %	iho	197	2,43 %
2	vatsa	145	2,08 %	syy	24	2,11 %	vatsa	163	2,01 %
3	syy	113	1,62 %	kyse	23	2,02 %	syy	137	1,69 %
4	tilanne	102	1,47 %	lapsi	21	1,84 %	lapsi	122	1,51 %
5	lapsi	101	1,45 %	merkki	20	1,75 %	tilanne	115	1,42 %
6	väri	99	1,42 %	vatsa	18	1,58 %	väri	114	1,41 %
7	paino	91	1,31 %	muutos	17	1,49 %	kyse	109	1,35 %
8	aika	87	1,25 %	väri	15	1,32 %	paino	99	1,22 %
9	kyse	86	1,24 %	tilanne	13	1,14 %	aika	90	1,11 %
10	silmä	79	1,14 %	käyttö	13	1,14 %	silmä	88	1,09 %
11	päivä	72	1,03 %	yhteys	12	1,05 %	päivä	81	1,00 %
12	olkapää	68	0,98 %	viikko	11	0,96 %	merkki	73	0,90 %
13	äiti	68	0,98 %	lämpö	10	0,88 %	olkapää	73	0,90 %
14	alkoholi	62	0,89 %	poissulkeminen	10	0,88 %	äiti	71	0,88 %
15	tyttö	59	0,85 %	silmä	9	0,79 %	käyttö	70	0,86 %
16	viikko	58	0,83 %	päivä	9	0,79 %	muutos	69	0,85 %
17	käyttö	57	0,82 %	tupakointi	9	0,79 %	viikko	69	0,85 %
18	koko	56	0,80 %	poissulku	9	0,79 %	alkoholi	68	0,84 %
19	käsi	53	0,76 %	toteaminen	9	0,79 %	koko	61	0,75 %
20	merkki	53	0,76 %	paino	8	0,70 %	käsi	60	0,74 %
21	kaula	52	0,75 %	tulos	8	0,70 %	kaula	59	0,73 %
22	muutos	52	0,75 %	muisti	8	0,70 %	tyttö	59	0,73 %
23	epäily	51	0,73 %	vanhempi	8	0,70 %	paikka	57	0,70 %
24	liike	51	0,73 %	tila	8	0,70 %	liike	56	0,69 %
25	paikka	51	0,73 %	tarve	8	0,70 %	ongelma	54	0,67 %
26	ruoka	51	0,73 %	käsi	7	0,61 %	pää	54	0,67 %
27	ongelma	50	0,72 %	kaula	7	0,61 %	ruoka	53	0,65 %
28	alue	48	0,69 %	asia	7	0,61 %	alue	52	0,64 %
29	pää	48	0,69 %	annos	7	0,61 %	epäily	52	0,64 %
30	rouva	47	0,68 %	määrä	7	0,61 %	väsymys	52	0,64 %
	3 yleisintä		6,18 %	**3 yleisintä**		6,32 %	**3 yleisintä**		6,14 %
	10 yleisintä		15,45 %	**10 yleisintä**		16,58 %	**10 yleisintä**		15,24 %
	30 yleisintä		31,36 %	**30 yleisintä**		31,58 %	**30 yleisintä**		30,58 %

Natiivien ja ei-natiivien lääkäreiden yleissanaston yleisimpien substantiivien joukossa on enimmäkseen samoja lekseemejä (Taulukko 11). Eri kieliryhmien 30 yleisimmän substantiivin listoista kuitenkin näkee, että natiiveilla useimmiten esiintyvien substantiivien kärkeen on päätynyt keskimäärin hieman pidempiä ja hieman kompleksisempia sanoja, esimerkiksi sellaisia kuin *tupakointi*, *poissulku*, *toteaminen* ja *vanhempi*.

Taulukko 12. Käytetyn yleissanaston frekventeimmät verbit.

	ei-natiivi			natiivi			kaikki		
	lekseemi	**kpl**	**%**	**lekseemi**	**kpl**	**%**	**lekseemi**	**kpl**	**%**
1	olla	2683	36,03 %	olla	346	30,43 %	olla	3029	35,29 %
2	voida	517	6,94 %	ei	80	7,04 %	voida	579	6,75 %
3	ei	398	5,34 %	voida	62	5,45 %	ei	478	5,57 %
4	tarvita	163	2,19 %	tulla	33	2,90 %	tarvita	191	2,23 %
5	tehdä	119	1,60 %	tarvita	28	2,46 %	tehdä	134	1,56 %
6	tarkistaa	116	1,56 %	ottaa	27	2,37 %	ottaa	121	1,41 %
7	kysyä	107	1,44 %	aloittaa	20	1,76 %	tarkistaa	121	1,41 %
8	pitää	99	1,33 %	viitata	16	1,41 %	kysyä	116	1,35 %
9	ottaa	94	1,26 %	tehdä	15	1,32 %	tulla	103	1,20 %
10	katsoa	73	0,98 %	kertoa	14	1,23 %	pitää	103	1,20 %
11	tulla	70	0,94 %	saada	12	1,06 %	katsoa	83	0,97 %
12	käyttää	64	0,86 %	johtua	12	1,06 %	saada	71	0,83 %
13	alkaa	61	0,82 %	katsoa	10	0,88 %	käyttää	69	0,80 %
14	saada	59	0,79 %	aiheuttaa	10	0,88 %	alkaa	67	0,78 %
15	syödä	56	0,75 %	kysyä	9	0,79 %	viitata	62	0,72 %
16	lähettää	55	0,74 %	kuulua	9	0,79 %	lähettää	61	0,71 %
17	epäillä	51	0,68 %	sopia	8	0,70 %	syödä	61	0,71 %
18	antaa	50	0,67 %	antaa	7	0,62 %	aloittaa	59	0,69 %
19	käydä	49	0,66 %	arvioida	7	0,62 %	antaa	57	0,66 %
20	auttaa	47	0,63 %	liittyä	7	0,62 %	epäillä	52	0,61 %
21	todeta	47	0,63 %	olemassa	7	0,62 %	auttaa	51	0,59 %
22	viitata	46	0,62 %	alkaa	6	0,53 %	aiheuttaa	50	0,58 %
23	arvioida	41	0,55 %	lähettää	6	0,53 %	käydä	50	0,58 %
24	aiheuttaa	40	0,54 %	esiintyä	6	0,53 %	todeta	49	0,57 %
25	mennä	40	0,54 %	ellei	6	0,53 %	arvioida	48	0,56 %
26	selvittää	40	0,54 %	tarkistaa	5	0,44 %	mennä	44	0,51 %
27	aloittaa	39	0,52 %	käyttää	5	0,44 %	laskea	43	0,50 %
28	laskea	38	0,51 %	syödä	5	0,44 %	selvittää	43	0,50 %
29	täytyä	35	0,47 %	laskea	5	0,44 %	kertoa	35	0,41 %
30	poissulkea	32	0,43 %	seurata	5	0,44 %	poissulkea	35	0,41 %
	3 yleisintä		48,31 %	**3 yleisintä**		42,92 %	**3 yleisintä**		47,60 %
	10 yleisintä		58,67 %	**10 yleisintä**		56,38 %	**10 yleisintä**		57,96 %
	30 yleisintä		71,56 %	**30 yleisintä**		69,31 %	**30 yleisintä**		70,65 %

Verbeistä yleisimpiä ovat *olla*, *voida* ja *ei*, joista jälkimmäiset ovat eri kieliryhmiin kuuluvilla lääkäreillä eri järjestyksessä (Taulukko 12). Ne kattavat verbiesiintymistä lähes puolet (47,60 %) Yksin verbin *olla* osuus on kaikilla kokeeseen osaa ottaneilla jopa yli 35 prosenttia. Sen käytön ero natiivien ja ei-natiivien välillä on huomattava: ei-natiiveilla *olla*-verbi kattaa yli 36 prosentin osuuden verbeistä, kun taas natiiveilla osuus jää noin 30 prosenttiin. Juuri tästä syystä ei-natiiveilla myös kolmen yleisimmän verbin osuus (48,31 %) on suurempi kuin natiiveilla (42,92 %). Koko lääkärijoukolla 10 yleisintä verbiä kattaa noin 58 prosenttia kaikista verbiesiintymistä ja 30 yleisintä 71 prosenttia.

Kokeen frekventeimmät **konjunktiot** *ja*, *jos* ja *tai* muodostavat muista selvästi erottuvan kolmen kärjen. Kummallakin kieliryhmällä ne kattoivat neljä viidesosaa konjunktioesiintymistä, mikä

on laskettavissa taulukosta 13. Natiivit ovat käyttäneet kaikkiaan vain 16:ta ja ei-natiivit yhteensä 21 eri konjunktiota. Tässä yhteydessä on kuitenkin muistettava, että konjunktiot muodostavat suomen kielessä ylipäätään pienen sanaluokan. Huomiota listoissa herättää lisäksi se, että natiivit ovat käyttäneet selvästi vähemmän konjunktioita *että* ja *koska*. Natiiveilla kymmenen yleisimmän konjunktion joukossa ovat mukana myös *sekä* ja *joten*, mikä saattaa olla vihje esimerkiksi kielen ei-natiiveja sofistikoituneemmasta kielenkäytöstä eli vivahteikkuudesta ja kompleksisemmasta lauserakenteesta, mutta päätelmien tekeminen edellyttäisi tarkempaa vastaustekstien analyysiä ja käytännössä omaa tutkimusta.

Taulukko 13. Käytetyn yleissanaston frekventeimmät konjunktiot.

	ei-natiivi			natiivi			kaikki		
	lekseemi	**kpl**	**%**	**lekseemi**	**kpl**	**%**	**lekseemi**	**kpl**	**%**
1	ja	1109	38,16 %	ja	182	46,43 %	ja	1291	39,14 %
2	jos	702	24,16 %	jos	88	22,45 %	jos	790	23,95 %
3	tai	498	17,14 %	tai	44	11,22 %	tai	542	16,43 %
4	että	136	4,68 %	sekä	27	6,89 %	että	142	4,31 %
5	koska	130	4,47 %	mutta	11	2,81 %	koska	137	4,15 %
6	mutta	75	2,58 %	koska	7	1,79 %	mutta	86	2,61 %
7	kun	68	2,34 %	vai	7	1,79 %	kun	74	2,24 %
8	vai	54	1,86 %	että	6	1,53 %	vai	61	1,85 %
9	eli	30	1,03 %	kun	6	1,53 %	sekä	56	1,70 %
10	kuin	29	1,00 %	joten	4	1,02 %	kuin	32	0,97 %
	3 yleisintä		79,46 %	**3 yleisintä**		80,10 %	**3 yleisintä**		79,53 %
	10 yleisintä		97,42 %	**10 yleisintä**		97,45 %	**10 yleisintä**		97,36 %
	21 yleisintä		100,00 %	**16 yleisintä**		100,00 %	**25 yleisintä**		100,00 %

Adjektiiveista frekventeimmät ovat *mahdollinen* (4,20 %), *hyvä* (3,60 %) ja *normaali* (2,83 %). Neljäntenä seuraa interrogatiivinen proadjektiivi *minkälainen* (2,70 %), joka tosin yhdessä lyhyemmän synonyyminsä *millainen* kanssa yltäisi toiselle sijalle (yhteensä 4,16 %). Natiivien ja ei-natiivien listojen kärki koostuu pääosin samoista lekseemeistä, mutta natiiveilla adjektiivi *mahdollinen* toistuu huomiota herättävän paljon: se kattaa yli kymmenesosan (10,20 %) heidän käyttämistään adjektiiveista, kun taas ei-natiiveilla lekseemin osuus jää runsaaseen kolmeen prosenttiin. Tämä voi johtua vastaajaryhmissä eri tavoilla painottuvista tavoista ilmaista epävarmuutta. Ylipäätään aineistossa käytetään melko paljon esimerkiksi *voi olla* -rakennetta, ja yleisimpien verbien listaus (Taulukko 14) osoittaa, että ei-natiivit ovat käyttäneet natiiveja useammin verbiä *voida*. Tämän yhteyden ja selityksen todentaminen vaatisi kuitenkin erillistä analyysiä.

Taulukko 14. Käytetyn yleissanaston frekventeimmät adjektiivit.

	ei-natiivi			natiivi			kaikki		
	lekseemi	**kpl**	**%**	**lekseemi**	**kpl**	**%**	**lekseemi**	**kpl**	**%**
1	hyvä	72	3,62 %	mahdollinen	35	10,20 %	mahdollinen	98	4,20 %
2	mahdollinen	63	3,17 %	normaali	17	4,96 %	hyvä	84	3,60 %
3	minkälainen	63	3,17 %	hyvä	12	3,50 %	normaali	66	2,83 %
4	tärkeä	50	2,51 %	millainen	9	2,62 %	minkälainen	63	2,70 %
5	normaali	49	2,46 %	pieni	8	2,33 %	tärkeä	51	2,19 %
6	todennäköinen	45	2,26 %	säännöllinen	7	2,04 %	todennäköinen	50	2,14 %
7	oikea	41	2,06 %	korkea	7	2,04 %	oikea	45	1,93 %
8	säännöllinen	37	1,86 %	pitkä	6	1,75 %	säännöllinen	44	1,89 %
9	vaikea	37	1,86 %	suuri	6	1,75 %	korkea	42	1,80 %
10	korkea	35	1,76 %	todennäköinen	5	1,46 %	vaikea	40	1,71 %
11	poikkeava	34	1,71 %	heikko	5	1,46 %	poikkeava	37	1,59 %
12	vasen	33	1,66 %	kohonnut	5	1,46 %	vasen	37	1,59 %
13	pitkä	26	1,31 %	aiempi	5	1,46 %	millainen	34	1,46 %
14	kuiva	26	1,31 %	oikea	4	1,17 %	pitkä	32	1,37 %
15	millainen	25	1,26 %	vasen	4	1,17 %	pieni	29	1,24 %
16	viime	25	1,26 %	oma	4	1,17 %	viime	27	1,16 %
17	paikallinen	23	1,16 %	sosiaalinen	4	1,17 %	kuiva	26	1,11 %
18	seuraava	22	1,11 %	aikainen	4	1,17 %	seuraava	25	1,07 %
19	pieni	21	1,06 %	vaikea	3	0,87 %	paikallinen	24	1,03 %
20	samanlainen	20	1,01 %	poikkeava	3	0,87 %	samanlainen	23	0,99 %
21	kalpea	20	1,01 %	seuraava	3	0,87 %	suuri	23	0,99 %
22	aiheellinen	19	0,95 %	samanlainen	3	0,87 %	aiheellinen	20	0,86 %
23	lievä	18	0,90 %	matala	3	0,87 %	kalpea	20	0,86 %
24	symmetrinen	18	0,90 %	iso	3	0,87 %	lievä	20	0,86 %
25	suuri	17	0,85 %	kylmä	3	0,87 %	matala	20	0,86 %
26	matala	17	0,85 %	syvä	3	0,87 %	yleinen	19	0,81 %
27	yleinen	17	0,85 %	vakava	3	0,87 %	symmetrinen	18	0,77 %
28	asiallinen	16	0,80 %	positiivinen	3	0,87 %	asiallinen	17	0,73 %
29	sopiva	16	0,80 %	ensisijainen	3	0,87 %	oma	17	0,73 %
30	väsynyt	16	0,80 %	tihentynyt	3	0,87 %	sopiva	17	0,73 %
	3 yleisintä		9,95 %	**3 yleisintä**		18,66 %	**3 yleisintä**		10,63 %
	10 yleisintä		24,72 %	**10 yleisintä**		32,65 %	**10 yleisintä**		24,99 %
	30 yleisintä		46,28 %	**30 yleisintä**		53,35 %	**30 yleisintä**		45,78 %

Koko joukolla kolme yleisintä adjektiivia kattavat runsaan kymmenesosan (10,63 %), kymmenen yleisintä neljänneksen (24,99 %) ja 30 yleisintä jonkin verran alle puolet (45,78 %) kaikista adjektiiviesiintymistä. Erityisesti lekseemin *mahdollinen* runsas toisto kasvattaa natiivien listan yleisimpien adjektiivien osuutta verrattuna ei-natiivien listaan. Vaikka varsinainen terveyssanasto on pyritty karsimaan tässä vaiheessa pois analysoitavasta aineistosta, lääkärien potilaiden diagnosoimisessa käyttämä kieli tuntuu huokuvan yleisimpien adjektiivien listoista: koehenkilöiden yleisimmät adjektiivit ovat olleet *mahdollinen*, *hyvä*, *normaali*, *minkälainen*, *tärkeä*, *todennäköinen*, *oikea*, *säännöllinen*, *korkea* ja *vaikea*, kun taas taajuussanastossa (CSC 2004) kymmenen kärki koostuu adjektiiveista *uusi*, *hyvä*, *suuri*, *oma*, *viime*, *koko*, *pieni*, *ensi*, *vanha* ja *tärkeä*. Kuulustelu-

aineiston yleisin adjektiivi *mahdollinen* on Suomen sanomalehtikielen taajuussanastossa frekvenssiltään vasta 23. adjektiivi.

Adverbien käytössä on nähtävissä lääkäriryhmien välistä vaihtelua. Yhteistä natiiveille ja ei-natiiveille on, että kummallakin *myös* on yleisin adverbi. Natiiveilla sen osuus on 15,95 % ja ei-natiiveilla 14,75 %. Sen jälkeen frekventeimpien adverbien listoissa (Taulukko 15) on suuria eroja. Tämä näkyy myös laskemissani adverbien kumulatiivisissa prosenteissa siten, että kaikkien kokeeseen osallistuneiden käyttämien adverbien kolmen kärki kattaa pienemmän osuuden (13,84 %) adverbeista kuin kummallakaan ryhmällä erikseen (14,75 % ja 15,95 %).

Taulukko 15. Käytetyn yleissanaston frekventeimmät adverbit.

	ei-natiivi			natiivi			kaikki		
	lekseemi	**kpl**	**%**	**lekseemi**	**kpl**	**%**	**lekseemi**	**kpl**	**%**
1	myös	129	6,22 %	myös	30	6,83 %	myös	159	6,32 %
2	miten	94	4,53 %	mukaan	23	5,24 %	miten	100	3,98 %
3	kuinka	83	4,00 %	niin	17	3,87 %	kuinka	89	3,54 %
4	paljon	68	3,28 %	ainakin	15	3,42 %	mukaan	81	3,22 %
5	sitten	62	2,99 %	sitten	10	2,28 %	paljon	77	3,06 %
6	mukaan	58	2,80 %	paljon	9	2,05 %	sitten	72	2,86 %
7	kotona	51	2,46 %	jo	9	2,05 %	kotona	58	2,31 %
8	milloin	49	2,36 %	aiemmin	9	2,05 %	milloin	55	2,19 %
9	hyvin	48	2,31 %	kotona	7	1,59 %	hyvin	54	2,15 %
10	aikaisemmin	43	2,07 %	kuluttua	7	1,59 %	missä	46	1,83 %
11	vielä	42	2,02 %	yleensä	7	1,59 %	aikaisemmin	46	1,83 %
12	missä	41	1,98 %	miten	6	1,37 %	vielä	45	1,79 %
13	pois	36	1,73 %	kuinka	6	1,37 %	pois	39	1,55 %
14	usein	35	1,69 %	milloin	6	1,37 %	usein	37	1,47 %
15	vain	30	1,45 %	hyvin	6	1,37 %	jo	36	1,43 %
16	nyt	29	1,40 %	varten	6	1,37 %	nyt	32	1,27 %
17	heti	28	1,35 %	juuri	6	1,37 %	vain	32	1,27 %
18	jo	27	1,30 %	tiedossa	6	1,37 %	heti	31	1,23 %
19	todennäköisesti	25	1,20 %	missä	5	1,14 %	niin	29	1,15 %
20	ensin	23	1,11 %	kuitenkin	5	1,14 %	kuluttua	29	1,15 %
21	kuluttua	22	1,06 %	muuten	5	1,14 %	todennäköisesti	28	1,11 %
22	joskus	21	1,01 %	tarpeen	5	1,14 %	ensin	27	1,07 %
23	kauan	21	1,01 %	ensin	4	0,91 %	kuitenkin	26	1,03 %
24	kuitenkin	21	1,01 %	myöhemmin	4	0,91 %	varten	24	0,95 %
25	mielestä	21	1,01 %	kotiin	4	0,91 %	yleensä	23	0,91 %
26	siis	21	1,01 %	tuskin	4	0,91 %	aiemmin	21	0,84 %
27	varten	18	0,87 %	helposti	4	0,91 %	joskus	21	0,84 %
28	erityisesti	18	0,87 %	taas	4	0,91 %	kauan	21	0,84 %
29	lisäksi	17	0,82 %	aikaisemmin	3	0,68 %	mielestä	21	0,84 %
30	ehkä	16	0,77 %	vielä	3	0,68 %	siis	21	0,84 %
	3 yleisintä		14,75 %	**3 yleisintä**		15,95 %	**3 yleisintä**		13,84 %
	10 yleisintä		33,01 %	**10 yleisintä**		30,98 %	**10 yleisintä**		31,46 %
	30 yleisintä		57,69 %	**30 yleisintä**		53,53 %	**30 yleisintä**		54,89 %

Frekventeimpien adverbien kärki on paljolti erilainen kuin taajuussanastossa (CSC 2004). Aineiston 10 yleisintä adverbia ovat *myös*, *miten*, *kuinka*, *mukaan*, *paljon*, *sitten*, *kotona*, *milloin*, *hyvin* ja *missä*, kun taas taajuussanaston (CSC 2004) yleisimmät adverbit ovat *myös*, *jo*, *nyt*, *vain*, *paljon*, *vielä*, *niin*, *kuitenkin*, *noin* ja *hyvin*.

Taulukon 16 listat osoittavat, että natiivien ja ei-natiivien ryhmät ovat käyttäneet toisistaan poikkeavilla tavoilla myös **pronomineja**. Koko joukolla kolme frekventeintä pronominia ovat *hän* (20,64 %), *se* (14,49 %) ja *mikä* (13,29 %). Kärki on tällainen erityisesti ei-natiivien ryhmän vaikutuksesta. Sillä *hän* kattaa pronominiesiintymistä jopa 22,78 %, *se* 15,01 % ja *mikä* 13,74 %. Ei-natiiveilla pronominin *hän* käyttö siis korostuu, mutta natiiviryhmällä sen käyttö on siihen verrattuna olematonta. Esiintymiä on vain kolme, mikä kattaa puolitoista prosenttia kaikista pronomineista.

Taulukko 16. Käytetyn yleissanaston frekventeimmät pronominit.

	ei-natiivi			natiivi			kaikki		
	lekseemi	**kpl**	**%**	**lekseemi**	**kpl**	**%**	**lekseemi**	**kpl**	**%**
1	hän	393	22,78 %	muu	37	19,07 %	hän	396	20,64 %
2	se	259	15,01 %	joka	30	15,46 %	se	278	14,49 %
3	mikä	237	13,74 %	se	19	9,79 %	mikä	255	13,29 %
4	muu	190	11,01 %	mikä	18	9,28 %	muu	227	11,83 %
5	tämä	93	5,39 %	tämä	16	8,25 %	tämä	109	5,68 %
6	joka	72	4,17 %	nämä	12	6,19 %	joka	102	5,32 %
7	minä	50	2,90 %	ne	11	5,67 %	ne	59	3,07 %
8	ne	48	2,78 %	jokin	8	4,12 %	minä	53	2,76 %
9	moni	45	2,61 %	kaikki	6	3,09 %	moni	49	2,55 %
10	joku	38	2,20 %	itse	5	2,58 %	joku	42	2,19 %
11	itse	35	2,03 %	toinen	5	2,58 %	itse	40	2,08 %
12	kaikki	33	1,91 %	moni	4	2,06 %	kaikki	39	2,03 %
13	molemmat	33	1,91 %	joku	4	2,06 %	toinen	36	1,88 %
14	toinen	31	1,80 %	hän	3	1,55 %	jokin	34	1,77 %
15	kuka	30	1,74 %	minä	3	1,55 %	molemmat	33	1,72 %
16	jokin	26	1,51 %	eri	3	1,55 %	kuka	30	1,56 %
17	sama	25	1,45 %	muutama	2	1,03 %	sama	26	1,35 %
18	te	23	1,33 %	mikään	2	1,03 %	nämä	23	1,20 %
19	he	18	1,04 %	nuo	2	1,03 %	te	23	1,20 %
20	nämä	11	0,64 %	sama	1	0,52 %	he	19	0,99 %
21	muutama	9	0,52 %	he	1	0,52 %	muutama	11	0,57 %
22	mikään	7	0,41 %	jokainen	1	0,52 %	mikään	9	0,47 %
23	sinä	7	0,41 %	eräs	1	0,52 %	sinä	7	0,36 %
24	me	3	0,17 %				eri	3	0,16 %
25	usea	2	0,12 %				jokainen	3	0,16 %
26	jokainen	2	0,12 %				me	3	0,16 %
27	mihin	1	0,06 %				nuo	2	0,10 %
28	tässä	1	0,06 %				usea	2	0,10 %
29	kukin	1	0,06 %				eräs	1	0,05 %
30	kumpi	1	0,06 %				mihin	1	0,05 %
	3 yleisintä		51,54 %	**3 yleisintä**		44,33 %	**3 yleisintä**		48,41 %
	10 yleisintä		82,61 %	**10 yleisintä**		83,51 %	**10 yleisintä**		81,81 %
	30 yleisintä		99,94 %	**23 yleisintä**		100,00 %	**30 yleisintä**		99,79 %

Natiivien käytössä puolestaan korostuivat pronominit *muu*, *joka* ja *se*. Jo frekventeimmät *muu* ja *joka* vastaavat yli kolmanneksesta natiivien pronominiesiintymistä. Ei-natiiveilla niiden prosenttiosuudet taas ovat huomattavasti pienemmät. Vastaajaryhmien pronominiluetteloissa on muuten paljon samaa, joskin natiivien pronominilista on lyhyempi oletettavasti ainakin osittain tekstien pienemmän määrän vuoksi.

Adpositioiden osalta eniten käytettyjen listassa (Taulukko 17) korostuu tietty pieni sanajoukko vielä vahvemmin kuin pronomineilla. Kolme yleisintä kattaa selvästi yli puolet (56,99 %) kaikista adpositioiden esiintymistä ja 10 yleisintä jo yli 90 prosenttia. Eniten aineistossa on käytetty adpositioita *jälkeen*, *kanssa*, *aikana*, *vuoksi*, *takia* ja *ilman*. Natiivien ja ei-natiivien ryhmien selvin ero vaikuttaa olevan se, että natiivit ovat käyttäneet huomattavasti enemmän syyn ilmaisussa käytettävää lekseemiä *vuoksi*. Ainakin päämerkitykseltään identtisten adverbien takia ja vuoksi osuus on natiiveilla lähes 28 prosenttia, kun taas ei-natiiveilla niiden osuus jää siitä lähes puoleen.

Taulukko 17. Käytetyn yleissanaston frekventeimmät adpositiot.

	ei-natiivi			natiivi			kaikki		
	lekseemi	**kpl**	**%**	**lekseemi**	**kpl**	**%**	**lekseemi**	**kpl**	**%**
1	jälkeen	114	28,01 %	vuoksi	14	24,14 %	jälkeen	127	27,31 %
2	kanssa	83	20,39 %	jälkeen	13	22,41 %	kanssa	91	19,57 %
3	aikana	42	10,32 %	kanssa	8	13,79 %	aikana	47	10,11 %
4	takia	34	8,35 %	ennen	8	13,79 %	vuoksi	43	9,25 %
5	vuoksi	29	7,13 %	aikana	5	8,62 %	takia	36	7,74 %
6	ennen	26	6,39 %	takia	2	3,45 %	ennen	34	7,31 %
7	ilman	25	6,14 %	lisäksi	2	3,45 %	ilman	26	5,59 %
8	kautta	7	1,72 %	ilman	1	1,72 %	kautta	7	1,51 %
9	päästä	4	0,98 %	muassa	1	1,72 %	lisäksi	4	0,86 %
10	kuluttua	4	0,98 %	pohjalta	1	1,72 %	päästä	4	0,86 %
	3 yleisintä		58,72 %	**3 yleisintä**		60,34 %	**3 yleisintä**		56,99 %
	10 yleisintä		90,42 %	**10 yleisintä**		94,83 %	**10 yleisintä**		90,11 %
	30 yleisintä		99,02 %	**13 yleisintä**		100,00 %	**30 yleisintä**		98,71 %

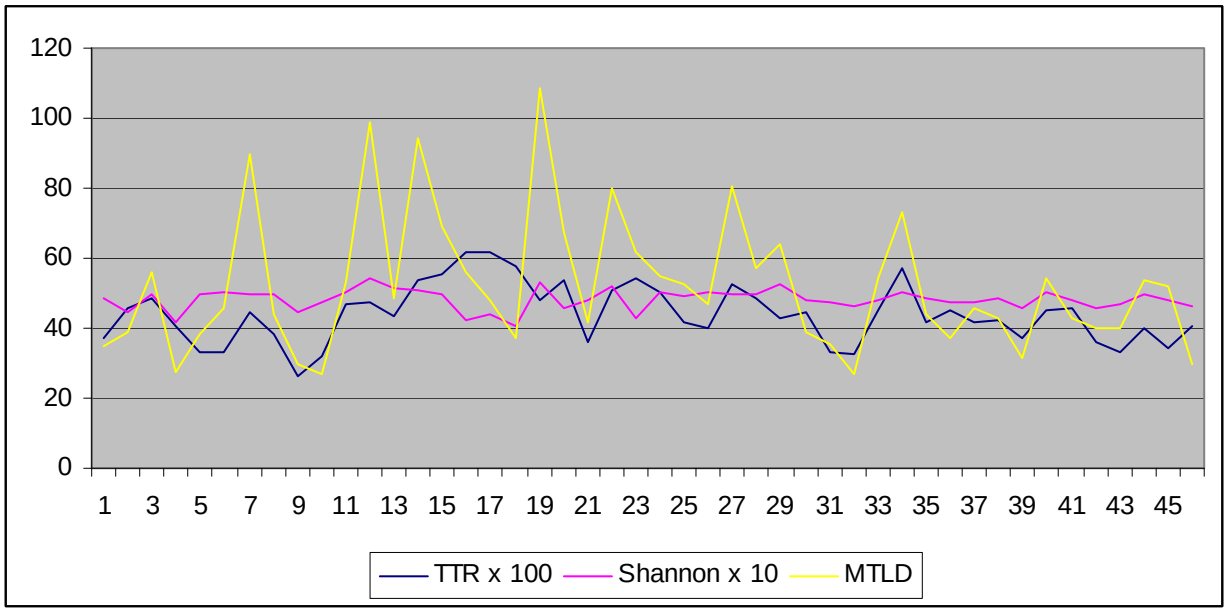
Muiden sanaluokkien esiintymät ovat jääneet pieniksi. Aineiston karsimisen jälkeen analysoitavassa yleissanastossa on yhteensä 74 kirjaimin kirjoitettua numeraalia, ja joukossa on vain 12:ta eri lekseemiä. Määrä on niin vähäinen, ettei niitä ole tarpeen listata erilliseen taulukkoon. Frekventeimpiä niistä ovat *ensimmäinen* (22 esiintymää), *kaksi* (18) ja *yksi* (15). Niitä seuraavat *kolme* (4), *pari* (4), *kuusi* (3), *kolmas* (2) ja *viisi* (2). Numeraaleilla *toinen*, *eka*, *kymmenen* ja *neljä* on kullakin yksi esiintymä. Erisnimeksi luokiteltuja saneita on vieläkin vähemmän kuin numeraaleja, vain 21. Niistä valtaosa (14) koostuu selvästikin tehtävänannoissa mainituista potilaiden etu- ja sukunimistä (esim. *Maija Meikäläinen*). *Suomi* mainitaan kaksi kertaa, *Ivalo* ja *Rovaniemi* kumpikin kerran. Loput kolme erisnimeä osoittautuvat lähemmässä tarkastelussa virheellisesti kirjoitetuksi ja luokitel-

luksi lääkeaineen nimeksi (*Museo*, p.o. *Muse*) tai lääketieteellisten termien nimityksiksi tai niiden osiksi (*Rinne*-testi, *Tannerin* asteikko).

4.5 Leksikaalisen diversiteetin tunnusluvut

Analyysin perusteella laillistamiskokeen vastausten leksikaalinen diversiteetti vaihtelee huomattavasti. Tämä käy ilmi keskeisistä tunnusluvuista. Alla oleva kuvaaja 10 antaa yleiskuvaa vastaajien yleissanaston monipuolisuudesta kolmella eri mittarilla. Mukana ovat Shannonin entropia eli Shannonin indeksi, MTLD ja TTR eli sana–sane-suhde. Kuvaajan x-akselilla ovat 46 koehenkilöä tässä tutkielmassa käytetyn numeroinnin mukaisessa järjestyksessä, ja yläpuolella kulkee kolme eri käyrää kuvaamassa tunnuslukujen vastaajakohtaista vaihtelua. Jotta kolmen eri tunnusluvun vaihtelua ja niiden eroja voisi visualisoida mielekkäästi yhdessä ja samassa kuvaajassa, käytössä on viivadiagrammi ja TTR-indeksi on kerrottu tässä kuvaajassa sadalla ja Shannonin indeksi kymmenellä.

Kuvaaja 10. Shannonin indeksi, MTLD ja TTR-luku vastaajittain.



Kuvaajasta näkyvät eri tavalla laskettujen tunnuslukujen laskemistavoista johtuvat eroavaisuudet. Niiden vuoksi indeksien vaihteluvälien skaalat ovat erilaiset, eli mikä näkyy tässä kuvaajassa silminnähden siinä, että esimerkiksi eri koehenkilöiden MTLD-lukujen erot ovat suhteessa suurempia kuin suhteelliset erot Shannonin indeksissä. Kuitenkin kummankin kuvaajissa huiput ja notkot, eli suurimmat ja pienimmät indeksilukemat, sattuvat pääosin samojen vastaajien kohdalle, vaikka selviä erojakin on havaittavissa. Tarkastellessa Shannonin indeksin ja MTLD:n välistä hajontakuviota ja korrelaatiota on havaittavissa selvä trendi, ja tarkempi analyysi vahvistaa, että Shan-

nonin indeksin ja MTLD:n välillä on selvä positiivinen korrelaatio, joka on tilastollisesti erittäin merkitsevä ($r = 0{,}58$, $p < 0{,}001$).

Vertailun vuoksi mukana oleva, käytetyn sanaston diversiteetistä karkeammin kertova TTR vaikuttaa myös isoilta osin mukailevan kahta muuta käytettyä mittaria. Aineistosta lasketut MTLD ja TTR muodostavat hajontakuvion, jossa on selvä trendi, ja niiden välillä on selvä positiivinen korrelaatio, joka on tilastollisesti erittäin merkitsevä ($r = 0{,}55$, $p = 0{,}001$). Tämä korrelaatio ei yllätä, koska monimutkaisempi tunnusluku MTLD lasketaan hyödyntäen sana–sane-suhteen laskutapaa (ks. luku 3.3.2). TTR ja Shannonin entropia eivät kuitenkaan korreloi vastaavalla tavalla.

Kokeen suorittajien suoritusten vertailemiseksi ja erojen havainnollistamiseksi olen jakanut heidät Shannonin indeksin perusteella viiteen tasoryhmään, jotka olen nimennyt numeroin 1–5 heikoimmasta vahvimpaan (Taulukko 18). Katson Shannonin entropian olevan sopiva mittari jaon perusteeksi ja vertailun rungoksi, koska se on toisaalla (esim. Saarela 1997; Malin 2012) todettu toimivaksi leksikaalisesta diversiteetistä kertovaksi tunnusluvuksi. Tasoryhmiin jako on tehty mekaanisesti, mutta ryhmät eivät ole tasasuuruisia. Ryhmissä 1–4 on yhdeksän ja ryhmässä 5 kymmenen koehenkilöä. Tämä oli mielekäs ratkaisu, koska suurimpiin Shannon-lukemiin yltäneiden ryhmässä peräti neljän lääkärin indeksi pyöristyi kahden desimaalin tarkkuudella samaan lukuun, 5,03:een, ja seuraavaksi parhaaseen ero oli hieman suurempi. Käyttäen tuota kohtaa jakolinjana parhaan ryhmän kooksi tuli kymmenen henkilöä. Muiden ryhmien rajakohdissa erot olivat hieman selvempiä.

Taulukko 18. Tasoryhmien laillistamiskokeessa suoriutumista kuvaavia lukuja.

tasoryhmä	saneet, lkm	leks., lkm	Shannon	MTLD	TTR	kum. leks. 50	kum. leks. 500	pisteet	läpäisy	natiivien osuus
5	716	324	5,15	71,52	0,47	10,22 %	35,63 %	66,88	60 %	30 %
4	601	265	4,98	60,19	0,46	10,19 %	34,62 %	67,28	89 %	22 %
3	618	251	4,85	46,19	0,41	10,74 %	38,95 %	61,14	67 %	0 %
2	652	240	4,69	39,04	0,39	11,55 %	40,90 %	54,75	44 %	0 %
1	367	147	4,36	41,16	0,47	14,40 %	42,73 %	45,72	11 %	33 %

Taulukko 18 kertoo tasoryhmien keskiarvot yleissanaston käytöstä laskemistani tunnusluvuista, ja niiden avulla sanastotaitojen vaihtelu tulee ilmi selvästi. Lisäksi olen sijoittanut taulukon oikean laidan sarakkeisiin prosenttilukuina tasoryhmäkohtaisen kokeen läpäisyprosentin ja natiivien lääkäreiden osuuden. Vertailun vuoksi myös sana–sane-suhde eli TTR on taulukossa. On muistettava TTR:n heikkous sanaston diversiteetin mittarina: eripituisten tekstien saamat TTR-luvut eivät ole keskenään vertailukelpoisia. Vain sitä käyttämällä esimerkiksi lyhyet tekstit voisivat saada pitkiä helpommin suhteettoman hyviä arvosanoja sanastonsa monipuolisuudesta. Yleissanastoaineistosta lasketut TTR-luvut ovat voimakkaassa ja tilastollisesti erittäin merkitsevässä käänteisessä korrelaa-

tiossa ($r = -0{,}7466$, $p < 0{,}001$) sanemäärään, eli mitä lyhyemmän tekstin koehenkilö on kirjoittanut, sen helpommin tekstin TTR on ollut korkea. Tämä on linjassa esimerkiksi Jarvisin (2013) ja Saarelan (1997) TTR:n ongelmia koskevien havaintojen ja huomioiden kanssa.

Myös Shannonin entropia korreloi pituuslukujen kanssa, mutta päinvastoin kuin TTR:llä, korrelaatio ei ole negatiivinen vaan positiivinen. Shannon on selvästi yhteydessä sanemäärään ($r = 0{,}5828$, $p < 0{,}001$) ja vielä selvemmin lekseemien määrään ($r = 0{,}8742$, $p < 0{,}001$). Kummassakin tapauksessa korrelaatio on tilastollisesti erittäin merkitsevä. Toisin sanoen pitkien vastaustekstien sanasto on Shannonin indeksin perusteella tyypillisesti monipuolisempaa kuin lyhyiden.

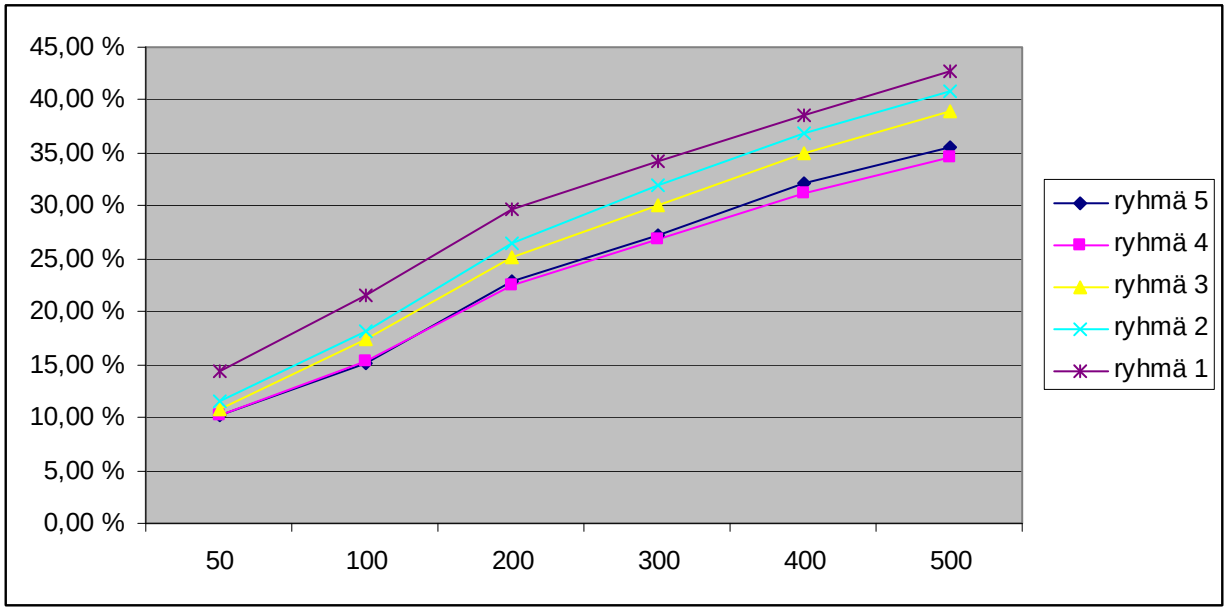
Shannonin entropian perusteella heikoimman tasoryhmän vastaukset ovat muihin verrattuna huomattavan lyhyitä, mutta silti pelkän TTR:n mukaan sanasto olisi yhtä rikasta kuin parhaalla ryhmällä (Taulukko 18). Luvut selittyvät siis sillä, että hyvin lyhyissä vastauksissa yleisimpien lekseemien toistoa ei ehdi tapahtua niin paljon kuin pidemmissä. Tasoryhmien keskimääräiset Shannonin indeksit ja MTLD-luvut kuitenkin mukailevat paremmin esimerkiksi tasoryhmien keskimääräisiä koepistemääriä. Koemenestyksen ja tunnuslukujen välistä korrelaatiota analysoin tarkemmin seuraavassa, sille omistetussa luvussa (luku 4.6).

Vertaileminen muihin leksikaalista diversiteettiä tarkastelleisiin tutkimuksiin voi antaa mielikuvaa siitä, mitä koehenkilöiden teksteistä lasketut tunnusluvut merkitsevät. Esimerkiksi suomalaisten peruskoululaisten käyttämää sanastoa tutkinut Saarela (1997: 105–106) laski toisluokkalaisten Shannonin indeksin keskiarvoksi 3,50, neljäsluokkalaisten keskiarvoksi 3,81, kuudesluokkalaisten keskiarvoksi 4,36 ja kahdeksasluokkalaisten keskiarvoksi 4,24. Muodostamistani tasoryhmistä heikoimmankin tekstit ovat leksikaaliselta diversiteetiltä keskimäärin samaa tasoa kuin Saarelan aineiston kuudes- ja kahdeksasluokkalaisen kirjoittajien tekstit, vaikka laillistamiskoeaineistosta on tässä vaiheessa karsittu pois varsinainen lääkärien ammattisanasto lyhenteineen. Malinin toteuttamassa (2012: 50–52) eritasoisten 7.–9-luokkalaisten S2-koululaisten ja aikuisten S2-oppijoiden tekstien analyysissä Shannonin indeksi vaihteli arvojen 5,55 ja 6,65 välillä ja oli sitä suurempi, mitä edistyneemmistä suomen taitajista oli kyse. Malinin aineistostaan laskemat MTLD-arvot puolestaan vaihtelivat välillä 38,41–125,97. Malin (emt.) kuitenkin huomauttaa itse, että hänen analyysinsa on tehty eri taitotasoja edustavista tekstikimpuista ja etteivät hänen laskemansa tunnusluvut ole siksi vertailukelpoisia sellaisten tutkimusten lukujen kanssa, joissa tarkastellaan yksittäisten vastaajien yksittäisiä tekstejä. Sellaisiksi Saarelan (1997) analysoimat koululaisten tekstit ja oman aineistoni koevastauskokonaisuudet voi mieltää.

Olen liittänyt taulukkoon tunnuslukujen joukon rinnalle käytetystä yleissanastosta laskemiani kumulatiivisten prosenttien keskiarvoja (sarakkeet *kum. leks. 50* ja *kum. leks. 500*). Ne kertovat, kuinka suuri osa ryhmien käyttämistä lekseemeistä edustaa suomen yleisintä sanastoa. Lekseemien

kumulatiiviset prosentit voivat nähdäkseni auttaa hahmottamaan kuvaa siitä, millaista sanastoa kokeeseen osaa ottaneet lääkärit ovat käyttäneet sekä kuinka rikasta ja laajaa heidän hallitsemansa sanastonsa on. Ne viestivät siitä, kuinka paljon he ovat pysytelleet perussanastossa ja kuinka pian aktiivisen sanaston rajat alkavat tulla heillä vastaan. Niillä näyttää olevan jonkinlainen korrelaatio myös kokeen pistemääriin. Tarkastelen myös lekseemejä koskevien kumulatiivisten prosenttien yhteyttä laillistamiskokeessa menestymiseen muiden tunnuslukujen rinnalla tarkemmin seuraavassa luvussa (luku 4.6). Sitä ennen on syytä eritellä kumulatiivisista prosenteista tehtyjä havaintoja.

Kuvaaja 11. Yleissanaston frekventeimpien lekseemien käyttö tasoryhmittäin, kumulatiiviset prosentit taajuussanaston 500 yleisimmän sanan perusteella.

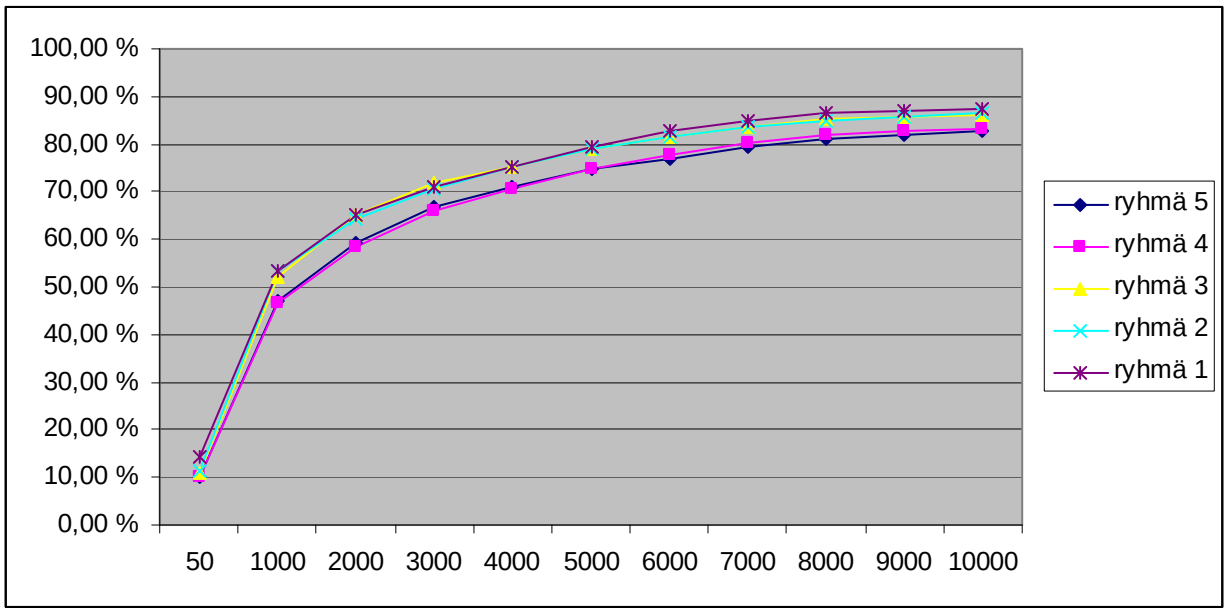


Tasoryhmien erot yleisimmän sanaston käytössä tulevat selvästi esiin kuvaajasta 11. Se näyttää tasoryhmien keskimääräiset prosentit siitä, kuinka suuren osuuden niiden käyttämästä leksikosta muodostavat 50, 100, 200, 300, 400 ja 500 suomen yleisintä sanaa. Kumulatiiviset prosentit on laskettu käyttäen apuna suomen sanomalehtikielen taajuussanastoa (CSC 2004). Heikoin ryhmä erottuu muista jo 50 yleisimmän sanan kohdalla: niiden osuus on jopa 14,40 prosenttia, kun muilla osuus vaihtelee 10,19–11,55 prosentissa. 500 yleisintä suomen sanaa kattaa kahdella heikoimmalla ryhmällä yli 40 prosenttia käytetyistä lekseemeistä, kun taas kahdella parhaalla ryhmällä niiden osuus jää noin 35 prosenttiin. Tasoryhmien sijoitus siinä, kuinka paljon ne käyttävät tässä vertailtujen yleisyysluokkien lekseemejä, ei juuri vaihtele. Poikkeuksen tekevät ryhmät 4 ja 5, joilla osuudet ovat alati lähes identtiset.

Kuvaajassa 12 tarkastellaan vastaavalla tavalla laillistamiskokeessa käytettyjen lekseemien kumulatiivisia prosentteja yleisyysluokista, joihin on laskettu suomen 50, 1000, 2000, 3000, 4000,

5000, 6000, 7000, 8000, 9000 ja 9996 (kuvaajassa merkitty luvulla 10 000) yleisintä sanaa. Sen mukaan tasoryhmät jakaantuvat käytännössä kahteen osaan. Kolmella heikoimmalla ryhmällä 50 prosentin osuus käytetyistä lekseemeistä tulee vastaan jo ennen 1000 suomen yleisintä lemmaa ja 2000 yleisimmän lemman kohdalla kumulatiivinen prosentti on jo 65:n luokkaa. Kahdella parhaalla ryhmällä kumulatiiviset prosentit ovat kauttaaltaan useamman prosentin pienempiä. Suurimmilla yleisyystasoilla tasoryhmien 1, 2 ja 3 kumulatiiviset osuudet tasoittuvat 87 prosentin tuntumaan, kun taas ryhmät 4 ja 5 jäävät alle 83 prosenttiin. Merkille pantavaa on, että nämä kumulatiiviset prosentit tasoittuvat etenkin kolmella heikoimmalla ryhmällä varsin lähelle Niemikorven (1991: 53; ks. luku 2.4) suomen kielen käytöstä laskemia kumulatiivisia prosentteja. Niemikorven (emt.) mukaan 9000–10 000 yleisintä lekseemiä kattaa noin 90 prosenttia käytetystä suomen kielestä.

Kuvaaja 12. Yleissanaston frekventeimpien lekseemien käyttö tasoryhmittäin, kumulatiiviset prosentit koko taajuussanaston mukaan.



Kahdeksan natiiviksi suomenpuhujaksi katsotun lääkärin sijoittuminen tasoryhmiin herättää hämmästystä. Heistä viisi on Shannonin entropian perusteella joko parhaassa tai toiseksi parhaassa ryhmässä, mutta kolme sijoittuu yllättäen kaikkein heikoimpaan joukkoon. Syitä tähän voi olla erilaisia. Kolmikon alkuperäiset vastaustekstit ovat osin hyvin tiiviitä ja luettelomaisia sekä sisältävät paljon lääkärien erikoisammattisanastoa ja terveyssanastoa, jotka eivät ole mukana tässä analyysin vaiheessa. Silti on huomattava, että noiden kolmen koehenkilön lyhyet vastaukset ovat saaneet heikot pistemäärät eivätkä ne ole yltäneet hyväksyttyyn suoritukseen. Vika ei välttämättä ole niinkään kielitaidossa, vaan kokeeseen osaa ottaneilla voi olla puutteita myös lääkärintyöhön tarvittavassa

osaamisessa. Aivan vastaavasti pitkät ja rikasta kieltä käyttävät koevastaukset eivät välttämättä ole yhteydessä niiden kirjoittajien ammattitaitoon lääkäreinä.

Taulukkoon 19 on poimittu kieliryhmittäin lähes samat keskiarvotiedot kuin aiemmin tasoryhmien vertailussa ja lisäksi kokeen läpäisyprosentit kieliryhmittäin. Siitä näkee, että yleissanaston saneilla mitaten natiivit kirjoittivat keskimäärin lyhyemmät tekstit kuin ei-natiivit, mutta lekseemien toistoa oli natiiveilla vähemmän. Tästä huolimatta Shannonin entropialla mitattuna kieliryhmien käyttämä sanasto on jotakuinkin yhtä monimuotoista.

Taulukko 19. Laillistamiskokeessa suoriutumista kuvaavia keskiarvoja kieliryhmittäin.

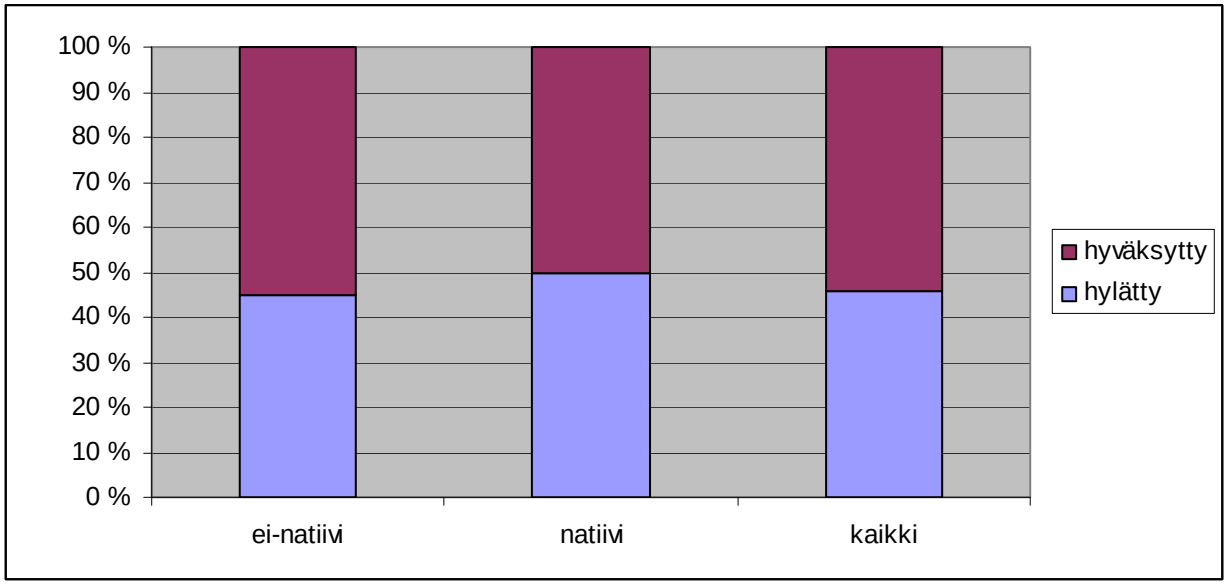
	saneet, lkm	leks., lkm	Shannon	MTLD	TTR	pist.	kum. leks. 50	kum. leks. 500	kum. leks. 9996	läpäisy
EI-NATIIVI										
keskiarvo	621	250	4,82	49,72	0,42	60,30	11,18 %	38,67 %	85,52 %	55,26 %
mediaani	624	254	4,79	46,28	0,40	59,03	11,44 %	42,48 %	88,40 %	
keskihajonta	196	59	0,24	18,02	0,07	16,83	2,00 %	4,80 %	3,43 %	
NATIIVI										
keskiarvo	465	233	4,80	63,14	0,54	54,66	12,42 %	37,70 %	83,52 %	50,00 %
mediaani	436	237	5,01	54,54	0,55	54,00	11,63 %	35,40 %	84,58 %	
keskihajonta	276	117	0,46	20,97	0,06	19,82	3,24 %	5,91 %	3,66 %	
KAIKKI										
keskiarvo	594	247	4,81	52,05	0,44	59,32	11,39 %	38,50 %	85,17 %	54,35 %
mediaani	573	246	4,85	47,67	0,44	61,00	10,63 %	38,80 %	85,14 %	
keskihajonta	222	74	0,30	19,44	0,08	17,54	2,35 %	5,06 %	3,59 %	

Myös tästä taulukosta voi tehdä odotusten vastaisen havainnon, että natiiveiksi luokitellut lääkärit eivät ole suoriutuneet kokeesta keskimäärin ei-natiiveja paremmin. Lisäksi lekseemien kumulatiivisissa prosenteissa näkyy luvun 4.3 kuvaajista 5 ja 6 tehty havainto: natiiveilla suomen kaikkein yleisimmät lekseemit muodostavat suuremmat osuuden käytetystä yleissanastosta kuin ei-natiiveilla, mutta kun tarkastelua laajennetaan harvinaisempien sanojen suuntaan, esimerkiksi 500 sanaan, natiivien kumulatiiviset prosentit jäävät jo pienemmiksi. Myös koko taajuussanastolla (CSC 2004) laskettu lekseemien kumulatiivinen prosentti jää natiiveilla hieman pienemmäksi.

4.6 Leksikaalisen diversiteetin yhteys laillistamiskokeessa menestymiseen

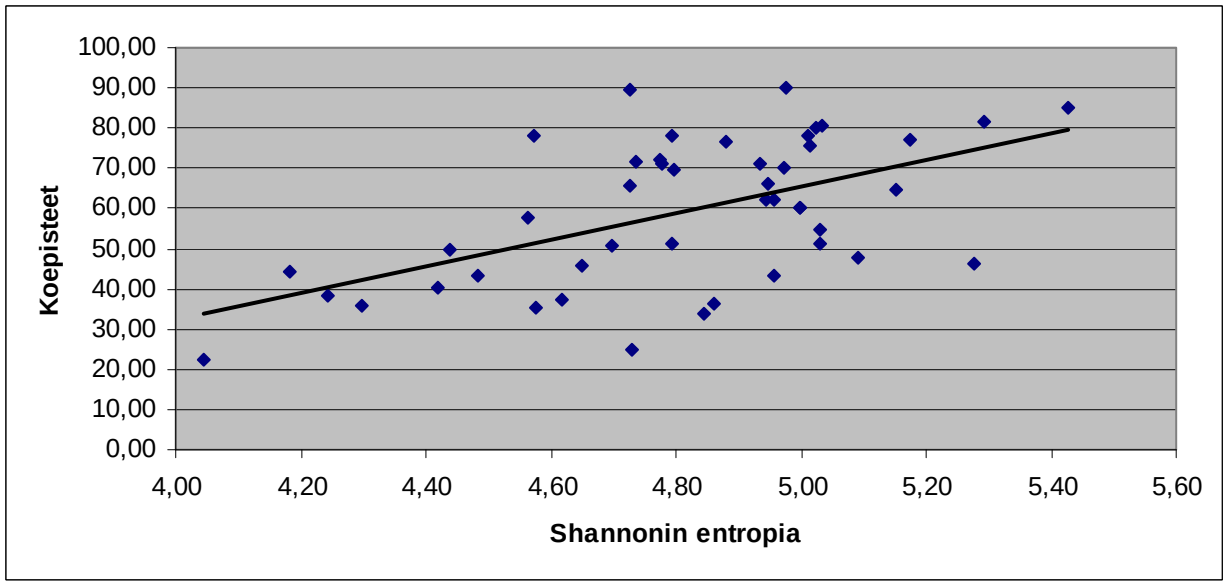
Tarkastelemassani koeaineistonipussa hyväksyttyjen suoritusten osuus on kaikkiaan noin 54 ja hylättyjen noin 46 prosenttia. Kahdeksasta natiivin koehenkilön vastauksesta joka toinen on siivittänyt kirjoittajansa hyväksyttyyn suoritukseen, kun taas 38 ei-natiivista lääkäristä hyväksytyt vastaukset on kirjoittanut 21 lääkäriä eli yli 55 prosenttia joukosta. (Kuvaaja 13.)

Kuvaaja 13. Hyväksyttyjen ja hylättyjen koehenkilöiden osuudet kielen mukaan.

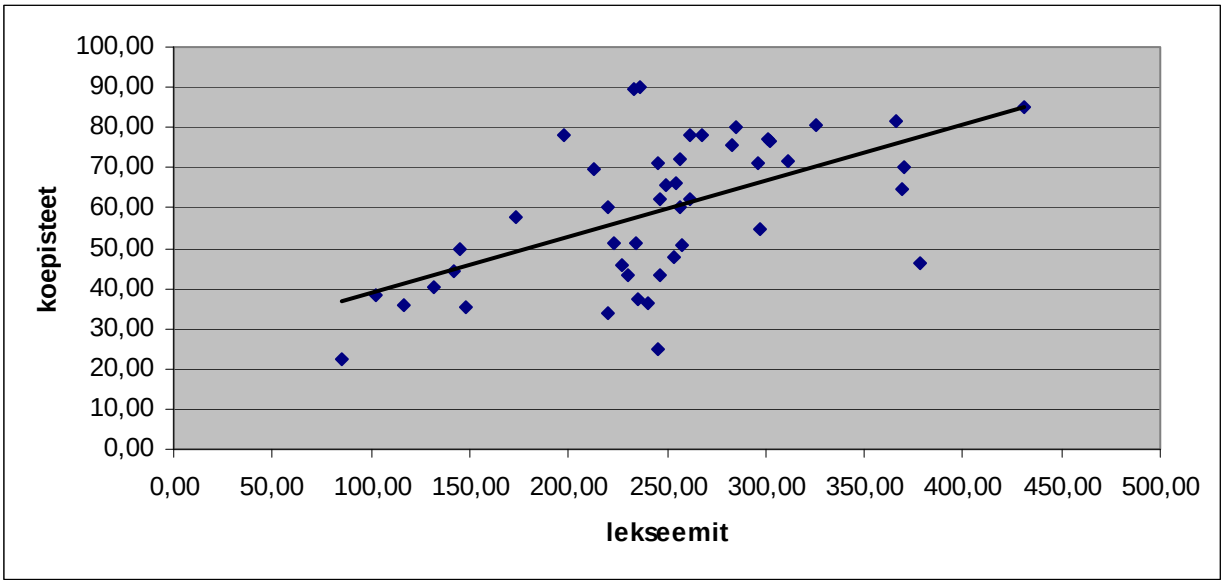


Kun tarkastellaan koehenkilöiden saamia koepisteitä ja yleissanaston diversiteettiä Shannonin indeksin perusteella, niillä on havaittavissa selvä yhteys (Kuvaaja 14). Niiden korrelaatiokerroin (Pearson, r) on 0,56, ja korrelaatio on kaksisuuntaisen p-arvon perusteella tilastollisesti erittäin merkitsevä ($p < 0{,}001$), kun luottamustasoksi oli määritettynä 95 %. Näin ollen voi sanoa, että mitä monimuotoisempaa lääkäreiden sanasto on ollut, sitä todennäköisemmin he ovat yltäneet laillisuuskuulustelussa hyviin pistemääriin.

Kuvaaja 14. Koepisteet ja Shannonin indeksit.

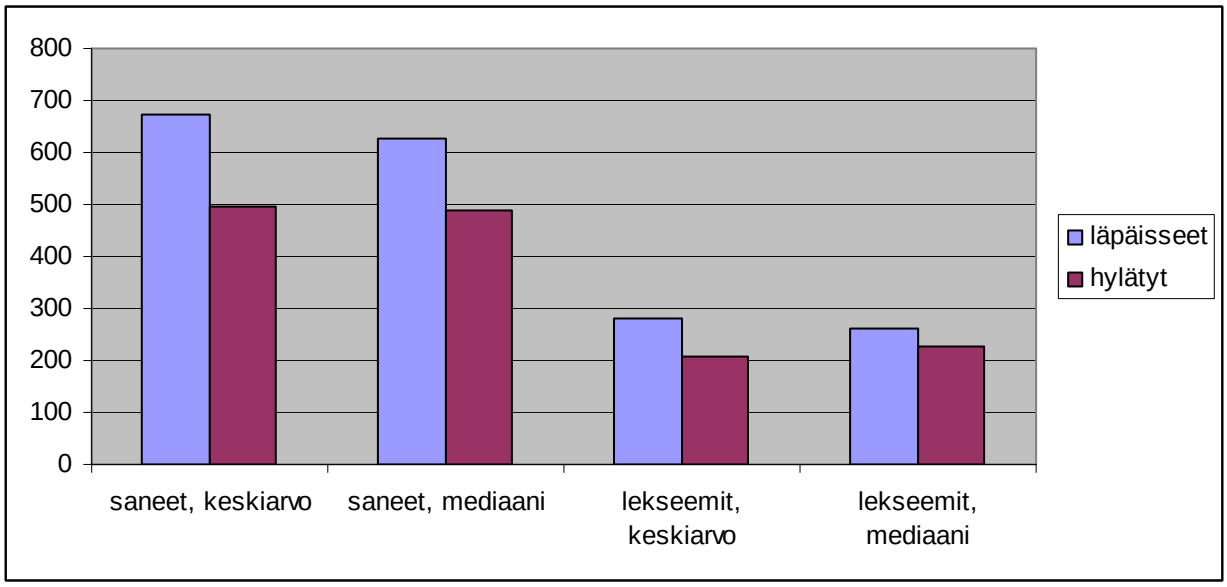


Kuvaaja 15. Koepisteet ja lekseemien määrät.



Yleissanaston lekseemien ja saneiden suhde ei korreloi aineistossa koemenestyksen kanssa, mutta sen sijaan pelkkä lekseemien runsaus on yhteydessä koepistemääriin. Kuvaaja 15 kertoo lekseemimäärän ja koepisteet vastaajittain. Vaikka siitä näkyy jonkin verran hajontaa, muuttujien välinen korrelaatio on jopa 0,68, ja se on tilastollisesti erittäin merkitsevä ($p < 0{,}001$). Korrelaation merkitsevyydestä kertova kaksisuuntainen p-arvo pyöristyy tässä yhteydessä 0,000:aan.

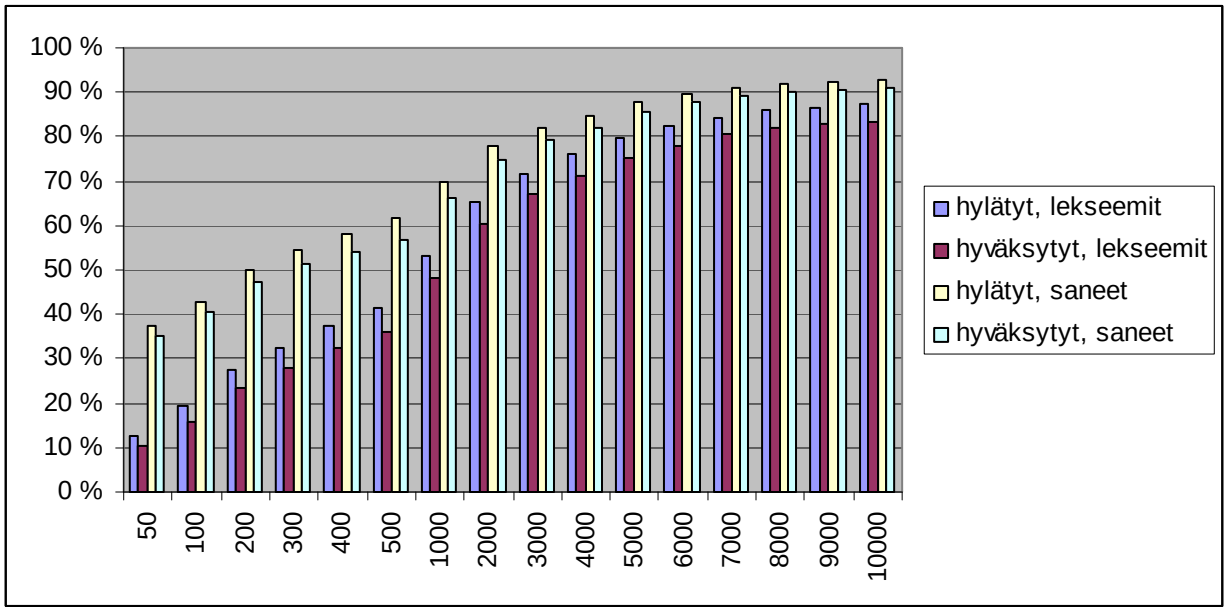
Kuvaaja 16. Saneiden ja lekseemien määrät koehenkilöiden koemenestyksen mukaan.



Kokeessa menestyneet eli sen läpäisseet lääkärit ovat paitsi kirjoittaneet keskimäärin pidemmät vastaukset myös käyttäneet keskimäärin laajempaa sanastoa kuin kokeessa hylätyt. Läpäisseiden ja hylättyjen saneiden ja lekseemien määrien erot käyvät selvästi ilmi kuvaajasta 16, josta on luettavissa keskiarvot ja mediaanit.

Tarkasteltaessa rinnakkain hyväksyttyjä ja hylättyjä suorituksia sekä niiden yleissanaston saneiden ja lekseemien käytöstä laskettuja kumulatiivisia prosentteja, on selvästi havaittavissa, että hyväksyttyyn suoritukseen yltäneillä kumulatiiviset prosentit ovat kauttaaltaan pienempiä kuin kokeessa hylätyillä. Tämä käy havainnollisesti ilmi kuvaajasta 17.

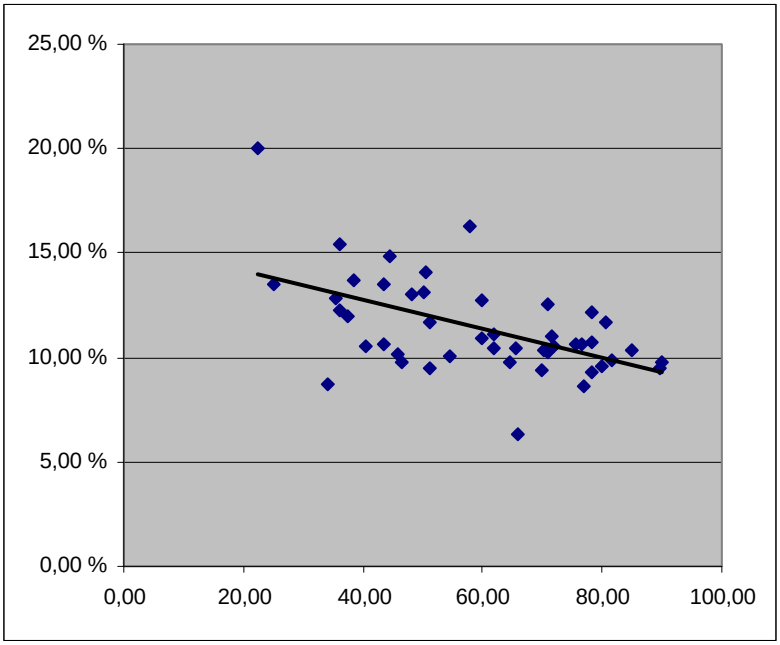
Kuvaaja 17. Hyväksyttyjen ja hylättyjen käyttämän sanaston kumulatiivisia prosentteja.



Suomen kielen kaikkien frekventeimmän sanaston käytön määrä on käänteisessä korrelaatiossa kokeen pistemääriin. Voi sanoa, että mitä suurempi osuus lääkärien käyttämästä sanastosta on ollut suomen kaikkein yleisintä, sitä todennäköisemmin he ovat jääneet heikoille pistemäärille. Tämä käy ilmi esimerkiksi kuvaajasta 18, jossa on tarkasteltu suomen 50 frekventeimmän lekseemin osuuden (y-akseli) ja kokeen pistemäärän (x-akseli) yhteyttä. Kyseisten muuttujien välinen korrelaatio on -0,54 ja erittäin merkitsevä ($p < 0,001$). Riippuvuussuhde on melkein yhtä vahva suomen 500 yleisimmän lekseemin osuuden ja koepisteiden välillä ($r = -0,53$, $p < 0,001$). Lekseemien kumulatiivisten prosenttien ja koepistemäärien välinen korrelaatio on kauttaaltaan tilastollisesti merkitsevää ($p < 0,01$) tai erittäin merkitsevää ($p < 0,001$). Kuitenkaan tarkastellessa vastaavalla tavalla saneiden käyttöä riippuvuutta ei ole havaittavissa yhtä laajasti tai vahvasti: saneiden kumulatiivisista prosenteista 500 yleisimmän suomen sanan osuuden ja koepisteiden käänteinen korrelaatio on

tilastollisesti melkein merkitsevä ($r = -0{,}33$, $p < 0{,}05$), ja samoin käy suomen 400 yleisimmän sanan ja saatujen koepisteiden välisessä käänteisessä riippuvuudessa ($r = -0{,}30$, $p < 0{,}05$).

Kuvaaja 18. Suomen 50 yleisimmän lekseemin osuus ja koehenkilöiden koepistemäärät.



Kaiken kaikkiaan havainnot tukevat hypoteesia, että koehenkilöiden sanastotaidot ovat yhteydessä heidän menestymiseensä tarkastelun kohteena olleessa laillistamiskokeen kirjallisessa osassa. Ne vahvistavat Tervolan ym. (2015) analyysin tulokset, joiden mukaan ulkomailta saapuneiden lääkärien kielitaitotaso korreloi laillisuuskuulustelun menestyksen kanssa. Nähdäkseni on kuitenkin huomattava, että koemenestystä ei pysty varmasti päättelemään pelkästä käytetyn sanaston analyysistä, koska myös täysin väärä koevastaus voisi olla esimerkiksi pitkä ja sanastollisesti monimuotoinen. Yhtä hyvin voi olla, että asiantunteva lääkäri kykenee vastaamaan laillistamiskokeen kysymyksiin tarkasti ja napakasti niin, että hänen käyttämänsä sanasto jää lopulta niukemmaksi kuin väärin tai huonosti vastanneella. Tämänkaltaisista varauksista huolimatta sanastoanalyysin tulokset näyttävät olevan samansuuntaisia laillistamiskokeessa annettujen pistemäärien kanssa.

5 YHTEENVETO

5.1 Johtopäätökset

Laillistamiskuulusteluaineistossa käytetystä leksikosta 64 prosenttia oli yleissanastoa, 29 prosenttia terveysalan sanastoa ja 7 prosenttia lääkärien erikoisammattisanastoa. Sanaluokista runsaimmin käytettiin substantiiveja, jotka muodostivat yli puolet tarkastellusta aineistosta. Valtaosa substantiiveista (63 %) lukeutui ammattisanastoon, joka karsiutui pois tämän tutkielman analyysin tarkemmasta vaiheesta. Runsas viidennes kaikesta käytetystä sanastosta oli verbejä. Natiiveiksi lasketut lääkärit käyttivät kaiken kaikkiaan huomattavasti useammin adverbejä, ei-natiiveilla puolestaan teksteissä esiintyi hieman natiiveja useammin konjunktioita ja pronomineja.

Käytetyistä saneista kaikkiaan 63 prosenttia oli perussanoja, 21 prosenttia johdoksia ja 15 prosenttia yhdyssanoja. Kun toisto karsitaan, yhdyssanojen ja johdosten eli kompleksisten sanojen joukkoon voidaan lukea lähes kolme neljästä (73 %). Natiivit lääkärit käyttivät niitä suhteessa hieman enemmän kuin ei-natiivit. Aineiston saneista runsas kolmannes oli alle kuusi merkkiä pitkiä, yli puolet alle kahdeksan merkkiä pitkiä ja neljä viidennestä korkeintaan kymmenen merkkiä pitkiä. Natiivien käyttämät saneet olivat keskimäärin hieman pidempiä kuin ei-natiivien, koska ei-natiiveilla lyhyet sanat korostuvat enemmän, kun taas natiiveilla joukossa on enemmän pitkiä sanoja. Koko aineiston yleisimmät lekseemit olivat *olla*, *ja*, *jos*, *ei* ja *voida*. Yleisimmän leksikon kärkipäässä korostui lääkärin vastaanotolla tarvittava sanasto, mikä oli odotuksenmukaista.

Yleissanaston käyttöä koskevan tarkemman analyysin mukaan 50 taajuussanaston (CSC 2004) yleisintä lemmaa kattoi keskimäärin runsaat 11 prosenttia yksittäisten lääkäreiden käyttämistä lekseemeistä, 1000 yleisintä riitti kattamaan yli 50 prosenttia ja 9996 yleisintä noin 85 prosenttia koevastauksista. Natiiveilla koehenkilöillä suomen kaikkein frekventeimpien lekseemien osuudet olivat suurempia kuin ei-natiiveilla, mutta taajuussanaston 300 frekventeimpään lekseemiin mennessä ei-natiivien kumulatiiviset prosentit ylittivät natiivien prosentit. Kaiken kaikkiaan natiiveiksi luetut lääkärit käyttivät ei-natiiveja harvinaisempaa sanastoa. Kieliryhmien yleisimpien lekseemien vertaileminen tuo esiin joitain selviä eroja, joiden syiden selvittäminen vaatisi tarkempaa ja myös laadullista analyysia.

Tätä tutkielmaa varten tehtyjen kvantitatiivisten analyysien perusteella on selvää, että laillistamiskokeeseen osaa ottaneiden lääkäreiden sanastotaidot vaihtelevat huomattavasti ja että ne ovat yhteydessä koemenestykseen. Käytetyn yleissanaston perusteella kokeessa heikoimmin suoriutuneilla sanasto oli muita köyhempää ja frekventimpää, hyvin suoriutuneilla taas yleensä monimuo-

toisempaa. Esimerkiksi suomen kaikkein yleisimmän sanaston käytön määrä on käänteisessä korrelaatiossa kokeen pistemääriin, eli mitä suurempi osuus kokeeseen osallistuneen lääkärin käyttämästä sanastosta oli suomen kaikkein yleisintä, sitä todennäköisemmin hän jäi heikoille pisteille. Kokeen hyväksytysti läpäisseet taas käyttivät keskimäärin hylättyjä harvinaisempaa sanastoa. Kiinnostava on myös havainto, että jo yksittäisten koehenkilöiden käyttämien lekseemien määrä korreloi vahvasti koemenestyksen kanssa.

Kaiken kaikkiaan tämän analyysin tulokset asettuvat linjaan sen kanssa, mitä Ruokolainen (2015) havaitsi analysoidessaan kielen norminmukaisuutta samassa koeaineistossa: kielivirheiden yleisyyden ja heikon koemenestyksen välillä vallitsi selvä riippuvuus, eikä esimerkiksi suurin osa eniten virheitä tehneistä yltänyt hyväksyttyyn suoritukseen. Lisäksi tämä tutkimus antaa vahvistusta Tervolan ym. (2015) saamille tuloksille. He osoittivat, että laillistamiskokeessa ne, joilla suomen kielen taito oli kielitaitoanalyysien perusteella parempi, menestyivät yleensä myös laillisuuskuulustelussa paremmin kuin ne, joiden suomen kielen taidot arvioitiin heikommiksi. Tämänsuuntaiset tutkimustulokset herättävät huomiota jo siksi, että laillistamiskokeen tarkoitus on testata lääkärin perustietoja kliinisen lääketieteen ja terveydenhuollon aloilta eikä enää kielitaitoa (Valvira 2018c).

Tutkielman tulosten joukkoon voi lukea myös analyysimenetelmien käytöstä ja toimivuudesta tehdyt havainnot. Varsinaisina leksikaalisen diversiteetin tunnuslukuina käytetyt Shannonin indeksi ja MTLD antoivat keskenään samansuuntaisia tuloksia, ja lisätukea niille antoi käytetyn sanaston tarkastelu kumulatiivisten prosenttien avulla. Shannon ja MTLD vaikuttavat toimivilta apuvälineiltä kielitaidon arvioinnissa, kun taas sana–sane-suhteelle eli TTR:lle ei kannattaisi antaa juuri painoarvoa sanastotaitoja määriteltäessä, kuten aiemmissakin tutkimuksissa on todettu (ks. esim. Jarvis 2013; Laine-Leinonen 2013; Malin 2012; Saarela 1997). Shannonin indeksin ja MTLD:n rinnalla varsin kelvollisia ja helppokäyttöisiä sanaston käyttöä kuvaavia lukuja vaikuttavat olevan myös sanaston frekvenssiluokituksia hyödyntävät kumulatiiviset prosentit. Näiden analyysitapojen yhdistelmä antaa käytetyn sanaston monimuotoisuudesta ja yleisyydestä jo melko laajan käsityksen.

5.2 Tutkimuksen luotettavuuden arviointia

Laillistamiskokeeseen osallistuneiden lääkäreiden sanastollisten taitojen arvioinnissa hankaluutta tuo aineiston kontekstin vaikutus käytettyyn leksikkoon. Vaikka analyysista rajaa pois erityisimmän erikoisammattisanaston, se koostuu isolta osin terveyskeskuksessa tarvittavista sanastosta, joka voi hyvin olla vastaajille kaikkein tutuinta suomen sanastoa. Siksi koehenkilöt voivat vaikuttaa taidoil-

taan etevämmiltä kuin he todellisuudessa ovat. Kun taas terveyssanaston rajaa pois analyysista, kokeeseen osaa ottaneiden sanastotaidot voivat vaikuttaa huomattavastikin todellisuutta heikommilta.

On huomattava, että tämän tutkimuksen varsinaisessa analyysiosassa on tarkasteltu vain yleissanaston käyttöä, ja siksi siinä saadut tunnusluvut eivät ole suoraan vertailukelpoisia mistä tahansa tekstistä laskettujen diversiteettilukujen kanssa. Kuitenkin esimerkiksi vastaajien välinen vertailu ja sanastotaitojen ja koemenestyksen välisen yhteyden tarkasteleminen on nähdäkseni mahdollista, ja karsitulla aineistolla voi tehdä joitain päätelmiä kokeeseen osaa ottaneiden lääkärien sanastotaidoista yleisemminkin.

Analysoiduista koevastauksista ei kontekstinsa vuoksi pysty päättelemään aukottomasti koehenkilöiden suomen kielen taitotasoa tai välttämättä edes sanastotaitojen tasoa. Ne kuitenkin piirtävät kielitaidosta ja sanastotaidoista suuntaa-antavan kuvan, ja niiden perusteella voidaan tehdä joitain päätelmiä. Kielitaitotasojen määrittelemiseksi tarkemmin tarvittaisiin avuksi muun muassa vastaajakohtaista virheanalyysiä, ja käytettyä kieltä olisi eriteltävä vastauskohtaisesti esimerkiksi Eurooppalaisen viitekehyksen (2003) tarjoamien kriteerien avulla. Kattava leksikaalisten taitojen arvioiminen edellyttäisi myös kielitaidon arvioimista eri kielenkäyttötilanteissa, kun taas tässä tutkimuksessa tarkastellut tekstit olivat tietystä erityiskontekstista, jossa tuotettua sanastoa vielä karsittiin tarkemman analyysin ja kielenkäyttäjien vertailemista varten.

Olen pyrkinyt tekemään kaikki työskentelyvaiheet niin tarkasti kuin mahdollista ja välttämään ja korjaamaan aineiston käsittelyssä ja analyysissä niin systemaattiset virheet kuin yksittäiset lyöntivirheet. Silti analyysin yhteydessä huomasin, että täysin virheettömästi en kymmenientuhansien saneiden käsittelystä ole suoriutunut. Ensinnäkin aineiston käsittelyssä on tapahtunut jonkin verran esimerkiksi lyöntivirheitä, mutta moninkertaisen tarkistamisen ansiosta uskon saaneeni karsittua ne lähes täysin pois. Jäljelle mahdollisesti jääneiden lyöntivirheiden merkityksen uskon olevan mitätön. Toiseksi analyysin loppuvaiheessa muutaman yksittäisen tekstin sanamäärissä on havaittavissa yhden tai kahden saneen tai lekseemin suuruisia virheitä, jotka havaitsin verratessani käyttämäni Perl-komentosarjan antamaa lekseemimäärää aineistosta itse poimimaani lukuun. Myös aineistonkuvausvaiheessa kompleksisuudesta rakentamissani taulukoissa lekseemien kokonaismäärissä näytti olevan muutaman lekseemin poikkeama. Näissä tapauksissa kyse on luultavasti taulukkolaskentaohjelmassa sanarivien määrän laskemisen yhteydessä tapahtuneista näppäilyvirheistä. Havaitsemani lukumäärien eroavaisuudet ovat kuitenkin niin vähäisiä ja harvinaisia, ettei horjunnalla ole tutkimustulosten kannalta merkitystä. Suurempi merkitys on ollut esimerkiksi niillä säännöillä, joilla rajasin lopulliseen analyysiin päätyvän aineiston eli joilla vedin rajan yleissanaston ja terveys- sekä lääkärien erikoisammattisanaston välille. Esimerkiksi lekseemit *symmetrinen* ja *paikallinen* eivät rajautuneet erikoissanaston mukana pois tarkemmasta analyysin vaiheesta, vaikka siihenkin ratkai-

suun olisi ollut perusteita. Vaikka analysoitavan rajauksen voisi tehdä myös erilaisin ja ehkä suoraviivaisemminkin perustein, uskon saamieni tulosten antavan analysoidusta aineistosta kokonaisuutena tarkastellen luotettavan kuvan.

Tutkimuksen luotettavuuteen voi vaikuttaa myös sen toteuttamisen ja kirjoittamisen pitkäksi venynyt aikataulu. Olen edistänyt tutkimusta napakan aloituksen jälkeen paljolti sivutoimisesti, ja työ on ollut välillä pitkiä aikoja tauolla. On mahdollista, että esimerkiksi tekemissäni luokitteluissa esiintyy hienoista horjuntaa. Matkan varrella olen myös tarkentanut tutkimukseni kysymyksenasettelua. Olen kuitenkin pyrkinyt pitämään varsin tarkasti kirjaa kaikista ratkaisuistani ja linjauksistani, jotta tutkimustyön pariin olisi aina helppo palata niin, ettei esimerkiksi aineiston käsittelyn tai luokittelun linja muutu. Hyvien muistiinpanojen ansiosta uskon välttäneeni tutkimusta vääristävät linjanmuutokset.

Jos tiedot analysoitujen koevastausten saamista pistemääristä olisivat olleet käytettävissäni jo työskentelyn varhaisemmassa vaiheessa, tutkimuksesta olisi muotoutunut hieman erilainen. Esimerkiksi jatkuva natiiveiksi ja ei-natiiveiksi tulkittujen lääkäreiden vertaaminen keskenään vaikutti sitä kyseenalaisemmalta ratkaisulta, mitä pidemmälle työ eteni. Onneksi sain käyttööni tarkemmat tiedot kuulusteluun osallistuneiden koemenestyksestä, kun työni alkoi lähestyä loppuvaihetta.

5.3 Tutkimuksen merkitys ja mahdollisia jatkotutkimuksen aiheita

Sanaston deskriptiivinen analyysi ja analyysivälineiden testaus ja kehittely ovat perustason tutkimusta, joka toimii materiaalina muille tutkijoille ja antaa heille eväitä jatkaa aiheessa syvemmälle. Tätä tutkimusta voitaneen hyödyntää esimerkiksi maahanmuuttajataustaisten lääkärien täydennyskoulutuksen suunnittelussa. Tutkielmani vähittäisen valmistumisen aikana ulkomailta Suomeen tulleiden lääkäreiden kielenoppimista on pyritty tehostamaan Tampereen yliopiston koordinoimassa hankkeessa, ja heille on kehitetty omaa harjoittelujaksoihin kytkeytyvää koulutusmallia, joka taipuu eritasoisille osaajille (Tampereen yliopisto 2018b). Lisäksi laillistamiskokeeseen menossa oleville ja myös jo ammattiaan Suomessa harjoittaville lääkäreille tarjotaan muun muassa suomen kielen preppauskursseja (Tampereen yliopisto 2018a). Tämän tutkielman havainnot ja tulokset voivat osaltaan auttaa selvittämään, kuinka suuri merkitys kielitaitojen ja erityisesti sanastotaitojen puutteilla on ollut lääkärien laillistamiskokeen suuriin hylkäysprosentteihin. Ne myös saattavat auttaa tekemään johtopäätöksiä siitä, mihin suuntaan ulkomailta tulevien lääkäreiden kielitaitovaatimuksia ja lisäkoulutustarjontaa tulisi kehittää erityisesti sanastotaitojen osalta.

Pro gradun laajuisessa työssä paljon jää auttamatta käsittelemättä. Laajasta ja monipuolisesta aineistosta heräsi useita ideoita pienistä ja suurista tutkimuksen laajentamisen, jatkamisen ja täydentämisen mahdollisuuksista. Jo vertailuaineistojen hyödyntäminen täydentäisi tässä tutkimuksessa syntyvää kuvaa laillistamiskuulusteluun osallistuneiden sanastosta. Esimerkiksi yksittäisten maahanmuuttajalääkäreiden sanastoprofiilien rakentaminen ja vertaileminen taidoiltaan eritasoisten kollegoidensa sanastoprofiileihin voisi tuottaa arvokasta tietoa asiantuntijoiden toisen kielen sanaston oppimisesta. Tarvetta voisi olla myös sanastotaitojen ja yleensä suomen kielen taidon kehityksestä kertovalle pitkittäistutkimukselle. Sanastoon liittyviä taitoja voisi hyvin verrata muiden maahanmuuttajien kuin lääkäreiden osaamiseen, ja ylipäänsä laillistamiskokeeseen ja lääkärien kielitaitoon liittyvien hankkeiden kokemuksista voi ammentaa paljon muidenkin maahanmuuttajien kielikoulutukseen.

Tämän tutkimuksen lääkäreiden sanasto-osaamisesta antamaa kuvaa voisi parantaa analysoimalla tarkemmin yleisimpien sanojen luetteloita ja ottamalla tutkimuksessa paremmin huomioon vastaajien vain kerran käyttämät eli niin sanotut hapaks legomenon -sanat. Lekseemien kompleksisuutta käytin lopulta vain yhtenä muuttujana kuvatessani tutkimuksen kokonaisaineistoa, ja analyysin syvemmälle menevässä osassa keskityin käytetyn sanaston diversiteettilukuihin ja kumulatiivisiin prosentteihin. Kuitenkin myös kompleksisuutta voisi hyödyntää esimerkiksi eri tavoin kokeessa menestyneiden lääkäreiden käyttämän sanasto-osaamisen vertailussa. Kokonaan itsenäisen tutkimuksensa voisi tehdä monestakin yksityiskohdasta – vaikkapa siitä, millä tavoin kielenoppijat ilmaisevat ajatustensa epävarmuutta ja miten keinojen kirjo kasvaa kielitaidon kehittyessä.

Analyysini toista vaihetta tekemäni rajaus pelkkään yleissanastoon jätti hieman epävarmuuden tunnetta erityisesti aineiston rajaamiseen liittyvien pitkällisten pohdintojeni takia. Nyt analyysi jäi aiemmin suunnittelemaani suppeampaan muotoon ensisijaisesti pro gradun rajoitusten vuoksi. Siksi tätä tutkimusta kannattaisi ehkä laajentaa toteuttamalla analyysin syventävä osuus sisällyttämällä sen aineistoon myös kokeessa käytetty terveyssanasto. Se voisi tarjota paitsi kiinnostavaa lisätietoa myös varmennusta tässä tutkielmassa saaduille tuloksille ja havainnoille. Samassa yhteydessä olisi harkittava myös paljon käytettyjen lyhenteiden ottamista mukaan, sillä nekin voisivat täydentää kuvaa kokeeseen osallistuneiden sanastotaidoista ja auttaa erottamaan osaajat paremmin.

LÄHTEET

Kirjat ja artikkelit

AKBARIAN, IS'HAAQ 2010: The relationship between vocabulary size and depth for ESP/EAP learners. *System* 3/2010 s. 391–401.

BALEGHIZADEH, SASAN – GOLBIN, MOHAMMAD 2010: The Effect of Vocabulary Size on Reading Comprehension of Iranian FTL Learners. *Linguistic and Literary Broad Research and Innovation* 2/2010 s. 33–46.

CARTER, RONALD – MCCARTHY, MICHAEL (toim.) 1988: *Vocabulary and Language Teaching.* London: Longman.

CHANNELL, JOANNA 1988: Psycholinguistic considerations in the study of L2 vocabulary acquisition. – Carter, Ronald & McCarthy, Michael (toim.) *Vocabulary and Language Teaching* s. 83–96. London: Longman.

Eurooppalainen viitekehys = Eurooppalainen kielten oppimisen, opettamisen ja arvioinnin yhteinen viitekehys. Helsinki: WSOY 2003.

HAUKILAHTI, RIITTA-LIISA – VIRJO, IRMA – MATTILA, KARI 2010: ETA-alueen ulkopuolella lääkäriksi valmistuneet. Lääkärien kuulustelujärjestelmä ja siihen osallistuneet vuosina 1994–2009. *Suomen lääkärilehti* 41/2010 s. 3315–3321.

HIRSH, DAVID – NATION, PAUL 1992: What Vocabulary Size is Needed to Read Unsimplified Texts for Pleasure. *Reading in a Foreign Language* 2/1992 s. 689–696.

HONKO, MARI 2013: *Alakouluikäisten leksikaalinen tieto ja taito. Toisen sukupolven suomi ja S1-verrokit.* Acta Universitatis Tamperensis 1865. Tampere: Tampereen yliopisto.

ISK = HAKULINEN, AULI – VILKUNA, MARIA – KORHONEN, RIITTA – KOIVISTO, VESA – HEINONEN, TARJA-RIITTA – ALHO, IRJA 2004: *Iso suomen kielioppi.* Helsinki: Suomalaisen Kirjallisuuden Seura.

JARVIS, SCOTT 2002: Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing* 1/2002 s. 57–84.

JARVIS, SCOTT 2013: Capturing the Diversity in Lexical Diversity. *Language Learning* 63 (Supplement 1) s. 87–106.

KELLY, PETER 1991: Lexical ignorance. The main obstacle to listening comprehension with advanced foreign language learners. *International Review of Applied Linguistics in Language Teaching.* 2/1991 s. 135–147.

LAINE-LEINONEN, JAANA 2013: *Koulukorpuksen leksikko. 1.–6.-luokkalaisten aktiivinen sanavarasto.* Suomen kielen pro gradu -tutkielma. Tampere: Tampereen yliopisto.

LATOMAA, SIRKKU (toim.) 2004: Äidinkieli ja toiset kielet. Pohjoismainen kaksikielisyystyöpaja Tampereella 18.–20.10.2002 s. 79–96. Tampere Studies in Language, Translation and Culture. Series B 1. Tampere: Tampere University Press.

LAUFER, BATIA – NATION, PAUL 1995: Vocabulary Size and Use. Lexical Richness in L2 Written Production. *Applied Linguistics* 3/1995 s. 307–322.

LAUFER, BATIA – PARIBAKHT, T. SIMA 1998: The Relationship Between Passive and Active Vocabularies: Effects of Language Learning Context. *Language Learning* 3/1998 s. 365–391.

LAUFER, BATIA – RAVENHORST-KALOVSKI, GEKE C. 2010: Lexical threshold revisited. Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* 1/2010 s. 15–30.

LIU, NA – NATION, I. S. P. 1985: Factors Affecting Guessing Vocabulary in Context. *RELC Journal* 1/1985 s. 33–42.

LU, XIAOFEI 2012: The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal* 2/2012 s. 190–208.

LÄMSÄ, RIIKKA – MANDERBACKA, KRISTIINA – KUUSIO, HANNAMARIA – AALTO, ANNA-MARI – ELOVAINIO, MARKO 2012: Ylilääkärien ja ulkomaalaistaustaisten lääkärien kokemuksia ammatinharjoittamisluvista. *Suomen lääkärilehti* 37/2012 s. 2555–2560.

MALIN, ESSI 2012: *Suomi toisena kielenä -oppijoiden sanaston kehittyminen taitotasolta toiselle siirryttäessä.* Pro gradu -tutkielma. Jyväskylän yliopiston kielten laitos.

MARTIN, MAISA 1999: Suomi toisena ja vieraana kielenä. Teoksessa Sajavaara, Kari & Piirainen-Marsh, Arja (toim.) *Kielenoppimisen kysymyksiä* s. 157–178. Jyväskylä: Jyväskylän yliopisto.

MCCARTHY, PHILIP – JARVIS, SCOTT 2010: MTLD, vocd-D, and HD-D. A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 2/2010 s. 381–392.

MOINZADEH, AHMAD – MOSLEHPOUR, ROGHAIEH 2012: Depth and Breadth of Vocabulary. Knowledge. Which Really Matters in Reading Comprehension of Iranian EFL Learners? *Journal of Language Teaching and Research* 5/2012 s. 1015–1026.

NATION, I. S. P. 1993: Vocabulary size, growth, and use. – Schreuder, Robert & Weltens, Bert (toim.) *The Bilingual Lexicon* s. 115–134. Studies in Bilingualism 6. Amsterdam/Philadelphia: John Benjamins Publihing Company.

NATION, I. S. P. 2006: How Large a Vocabulary is Needed for Reading and Listening? *The Canadian Modern Language Review* 1/2006 s. 59–82.

NIEMIKORPI, ANTERO 1991: *Suomen kielen sanaston dynamiikkaa.* Acta Wasaensia 26. Vaasa: Vaasan yliopisto.

NIIRANEN, LEENA 2004: Oppimisen kohteena suomen verbit. Kaksikielisten oppilaiden ja luokkahuoneoppijoiden verbisanaston ja verbintaivutuksen vertailua. – Latomaa, Sirkku (toim.) *Äidinkieli ja toiset kielet: Pohjoismainen kaksikielisyystyöpaja Tampereella 18.–20.10.2002* s. 79–96. Tampere Studies in Language, Translation and Culture. Series B 1. Tampere: Tampere University Press.

NIPPOLD, MARILYN A. 2007: *Later language development: school-age children, adolescents, and young adults.* 3rd edition. Austin (TX): Pro-Ed.

NIZONKIZA, DÉOGRATIAS 2011: The relationship between lexical competence, collocational competence, and second language proficiency. *English Text Construction* 1/2011 s. 113–146.

PAUNIO, RIITTA-LEENA – PELKONEN, RISTO 2012: *Terveydenhuollon valvontatyöryhmän mietintö.* Sosiaali- ja terveysministeriön raportteja ja muistioita 2012:8. Helsinki: Sosiaali- ja terveysministeriö.

PENTTILÄ, AARNI 1963: *Suomen kielioppi.* Porvoo: WSOY.

READ, JOHN 2000: *Assessing Vocabulary*. Cambridge: Cambridge University Press.

RUOKOLAINEN, JENNA 2015: *Soitin korvassa ja keuhkossa nesteetta? Kielen norminmukaisuus ulkomaalaistaustaisten lääkärien laillistamiskuulustelussa.* Pro gradu -tutkielma. Tampereen yliopiston kieli- käännös- ja kirjallisuustieteiden yksikkö.

SAARELA, LEENA 1997: *Peruskoululaisten kirjoitelmien kehittyminen sanastontutkimuksen valossa.* Acta Universitatis Ouluensis B Humaniora 25. Oulu: Oulun yliopisto.

SAJAVAARA KARI 1999: Toisen kielen oppiminen. Teoksessa Sajavaara, Kari & Piirainen-Marsh, Arja (toim.) *Kielenoppimisen kysymyksiä* s. 73–102. Jyväskylä: Jyväskylän yliopisto.

SAJAVAARA, KARI – PIIRAINEN-MARSH, ARJA (toim.) 1999: *Kielenoppimisen kysymyksiä.* Jyväskylä: Jyväskylän yliopisto.

SARIOLA, SUVI 2012: Kielitaidosta pitää puhua. *Suomen lääkärilehti* 12/2012 s. 926–929.

SAUKKONEN, PAULI – HAIPUS, MARJATTA – NIEMIKORPI, ANTERO – SULKALA, HELENA 1979: *Suomen kielen taajuussanasto*. Porvoo: WSOY.

SHANNON, CLAUDE E. 1948: A Mathematical Theory of Communication. Reprinted with corrections. *The Bell System Technical Journal* 27 s. 379–423, 623–656.

TERVOLA, MAIJA – PAJUNEN, ANNELI – VAINIO, SEPPO – HONKO, MARI – MATTILA, KARI 2015: Maahanmuuttajataustaisten lääkärien suomen kielen taito laillistamiskuulustelussa. *Duodecim* 131 s. 339–346.

TERVOLA, MAIJA 2017: Työelämän näkökulma maahanmuuttajataustaisten lääkärien kielitaitoon. *Sosiaalilääketieteellinen Aikakauslehti* 3/2017 s. 196–208.

Verkkolähteet

CSC 2004 = TIETEEN TIETOTEKNIIKAN KESKUS CSC 2004: *Suomen sanomalehtikielen taajuussanasto*. Tekstikorpus. Kielipankki. http://urn.fi/urn:nbn:fi:lb-201405272

FINLEX 1 = *Laki terveydenhuollon ammattihenkilöistä 28.6.1994/559.* http://www.finlex.fi/fi/laki/ajantasa/1994/19940559 (4.4.2018)

FINLEX 2 = *Asetus terveydenhuollon ammattihenkilöistä 28.6.1994/564.* http://www.finlex.fi/fi/laki/ajantasa/1994/19940564 (4.4.2018)

FINLEX 3 = *Laki potilaan asemasta ja oikeuksista 17.8.1992/785.* http://www.finlex.fi/fi/laki/ajantasa/1992/19920785 (4.4.2018)

FINLEX 4 = *Sosiaali- ja terveysministeriön asetus potilasasiakirjoista 30.3.2009/298*. http://www.finlex.fi/fi/laki/ajantasa/2009/20090298 (4.4.2018)

FINLEX 5 = *Laki terveydenhuollon ammattihenkilöistä annetun lain muuttamisesta 1659/2015.* https://www.finlex.fi/fi/laki/alkup/2015/20151659 (5.4.2018)

KOTIMAISTEN KIELTEN KESKUS 2017: *Tervetuloa Suomen murteiden sanakirjaan.* Suomen murteiden sanakirjan etusivu. http://kaino.kotus.fi/sms/ (4.4.2018)

SUOMEN LÄÄKÄRILEHTI 2012: *Tampereen yliopisto tutkii ulkomaalaisten lääkärien kielitaitoa*. Verkkouutinen. http://www.laakarilehti.fi/uutinen.html?opcode=show/news_id=12138/type=1 (26.3.2018)

TAMPEREEN YLIOPISTO 2015: Ulkomaisten lääkäreiden ammatillista kielitaitoa kehitetään uudella mallilla. Verkkouutinen. http://www.uta.fi/ltl/ilmoitus.html?id=111328 (26.3.2018)

TAMPEREEN YLIOPISTO 2018a: *EU-/ETA-alueen ulkopuolella koulutettujen lääkärien kuulustelut.* Tampereen yliopiston ylläpitämä, lääkärien laillistamiskokeeseen liittyvistä asioista tiedottava verkkosivu. http://www.uta.fi/med/tutkimus/tutkimusryhmat/yleislaaketiede/laakarien_kuulustelut.html (26.3.2018)

TAMPEREEN YLIOPISTO 2018b: *Maahanmuuttajalääkärin koulutuspolku* -hankkeen sivusto. https://research.uta.fi/maahanmuuttajalaakarit/tausta-ja-tavoitteet/ (26.3.2018)

VALVIRA 2013: *Ohje lääkäreille.* http://www.valvira.fi/luvat/ammattioikeudet/hakemusohjeet/eu_eta_maiden_ulkopuolella_koulutetut/laakarit (25.1.2013)

VALVIRA 2018a: *EU/ETA-valtioiden ulkopuolella koulutetut lääkärit*. http://www.valvira.fi/terveydenhuolto/ammattioikeudet/hakemusohjeet/eu_eta_valtioiden_ulkopuolella_koulutetut/laakarit (26.3.2018)

VALVIRA 2018b: *Ohje lääkäreille.* http://www.valvira.fi/terveydenhuolto/ammattioikeudet/hakemusohjeet/eu_eta-valtioissa_koulutetut/laakarit (5.4.2018)

VALVIRA 2018c: *EU/ETA-alueen ulkopuolella koulutetuilta lääkäreiltä vaadittavat kuulustelut.* http://www.valvira.fi/terveydenhuolto/ammattioikeudet/hakemusohjeet/eu_eta_valtioiden_ulkopuolella_koulutetut/laakarit/kuulustelut (8.5.2018)

Aineistolähde

PAJUNEN, ANNELI 2013: *Lääkärien pätevöitymisaineisto.* Opiskelijoiden harjoitustyö lääkärien laillistamiskokeen vastausaineistosta Tampereen yliopistossa professori Anneli Pajusen syksyllä 2012 pitämällä Kielitieteen metodit -kurssilla. Excel-aineisto.