Oppijansuomen sähköiset tutkimusaineistot

Nykytilanne

Jarmo Harri Jantunen ja Silja Pirkola

1 Johdanto

Korpukset tarjoavat kielentutkijoille mahdollisuuden tutkia kieltä laajojen sähköisten aineistojen avulla. Nykyisin suomen kielen tutkijoiden hyödynnettävissä on paitsi korpuksia, jotka koostuvat äidinkielisten suomenpuhujien kielestä, myös useita sellaisia korpuksia, jotka sisältävät oppijansuomea. Maailmanlaajuisesti oppijankieliaineistot ovat yhä kasvava korpusaineistojen muoto, ja aineistoja on syntynyt runsaasti myös muista kielistä kuin englannista, joka on pitkään ollut ja on edelleenkin korpusten valtakieli. Niin sanottuun korpusatlakseen (Dumont & Granger 2014) on tähän mennessä lueteltu lähes 140 erilaista oppijankielikorpusta, joissa kohdekielinä ovat englannin lisäksi muun muassa espanja, ranska, italia, saksa, arabia, unkari, viro ja suomi. Näistä aineistoista suurin osa on kirjoitetun kielen aineistoja, jotka sisältävät tyypillisesti muun muassa esseitä ja sähköpostiviestejä ilman diakronista ulottuvuutta, mutta sekä puhuttujen että diakronisten korpusten määrä kasvaa jatkuvasti. Oppijansuomen sähköisiä aineistoja on esitelty *Virittäjässä* jo aiemmin (Jantunen & Piltonen 2009), mutta tuosta koonnista on kulunut aikaa kuusi vuotta, joten aineistot ovat sen jälkeen kehittyneet. Useimmat korpukset ovat laajentuneet, ja niitä on myös annotoitu, eli niihin on lisätty esimerkiksi morfosyntaktista metatietoa. Lisäksi on koottu kokonaan uusia aineistoja, jotka osaltaan tarjoavat lisää mahdollisuuksia oppijansuomen korpustutkimukselle.

Tässä katsauksessa kokoamme yhteen ajantasaiset tiedot seitsemästä oppijansuomen korpuksesta. Aiemmin (Jantunen & Piltonen 2009) käsiteltyjen ICLFI-, LAS2-, YKI- ja Cefling-korpusten lisäksi esittelemme uudemmat Topling-, Dialuki- ja Long Second -korpukset, jotka on koostettu edellisen katsauksen jälkeen. Katsauksessa kuvailemme kunkin korpuksen perustiedot ensin lyhyesti ja sen jälkeen vertailemme aineistoja toisiinsa erilaisten luokittelupiirteiden, taustamuuttujien sekä annotoinnin näkökulmasta. Olemme koonneet korpusten tiedot sähköpostikyselyiden vastauksista, hankkeiden verkkosivuilta, katsauksessa mainituista artikkeleista sekä tieto-

kannoista, joihin on aiemmin kerätty korpuksia koskevia tietoja hankkeissa työskentelevien käyttöön.[1]

2 Korpusten esittely

Kansainvälinen oppijansuomen korpus eli ICLFI (*International Corpus of Learner Finnish*) on suomi vieraana kielenä -korpus[2], jota on koostettu vuodesta 2007 alkaen Oulun yliopiston johtamassa hankkeessa Korpustutkimus oppijankielen kielikohtaisista ja universaaleista ominaisuuksista. Hanke oli viiden yliopiston yhteishanke, ja korpuksen koostamista ovat rahoittaneet Riksbankens Jubileumsfond, Oulun yliopisto ja Fin-Clarin-konsortio. Korpus koostuu suomen kielen opiskelijoiden kirjoittamista teksteistä, jotka on kerätty yli 20 ulkomaisesta yliopistosta opetushenkilökunnan avustuksella. Tekstintuottajat ovat opiskelleet suomen kieltä yliopistossa pää- tai sivuaineena tai yksittäisinä kursseina. Korpuksen koko kasvaa jatkuvasti, sillä tekstejä kerätään korpukseen lisää; uusien tekstien keräystä varten on suunniteltu erityistä verkkosivustoa. Aineistoon on tehty kieliopillinen annotaatio ja pieneltä osin myös virheannotaatio (annotoinnista tarkemmin luvussa 4).

Edistyneiden suomenoppijoiden korpusta (LAS2) koostetaan Turun yliopistossa. Hankkeen tavoitteena on muun muassa lisätä tietoa suomesta toisena kielenä akateemisessa käytössä: korpus sisältää Turun yliopiston suomen ja sen sukukielten maisteriohjelman opiskelijoiden kirjoittamia akateemisia tekstejä. (LAS2.) Tekstit ovat toistaiseksi kielitieteisiin keskittyneiden humanististen alojen opiskelijoiden kirjoittamia, mutta hankkeen seuraavassa vaiheessa aineistoa kerätään myös muilta tieteenaloilta (Ivaska 2014a: 25). Korpus jakaantuu kolmeen eri osakorpukseen, jotka sisältävät keskenään eri tekstilajeja. Oppijansuomen lisäksi korpus sisältää natiivisuomea, sillä kustakin osakorpuksesta on koottu samat tekstilajit sisältävä verrannollinen natiivikorpus. Lisäksi LAS2 sisältää diakronista aineistoa: siihen on kerätty samoilta informanteilta tekstejä koko opiskeluajalta (1–4 vuotta). Korpus on osin annotoitu kieliopillisesti, ja siihen on lisätty myös mahdollisuus myöhempään virheannotointiin. Korpuksen koostaminen on aloitettu vuonna 2007, ja se jatkuu edelleen. (Mas. 23–24.)

Jyväskylän yliopiston Yleisten kielitutkintojen korpus (YKI-korpus) koostuu Yleisten kielitutkintojen testiaineistosta. Yleiset kielitutkinnot on kielitaitotesti, jonka voi suorittaa kaikkiaan yhdeksässä kielessä: englannissa, espanjassa, italiassa, ranskassa, ruotsissa, saamessa, saksassa, venäjässä ja suomessa (Solki a), ja korpuksessa on aineistoja jokaisesta testikielestä. Suomen (ja ruotsin) kielen tutkinnon suorittajat ovat pääasiassa henkilöitä, jotka tarvitsevat kielitaitotodistuksen esimerkiksi kansalaisuuden hakemista varten.

---

1. Osa tässä katsauksessa esitetyistä tiedoista on kerätty sähköpostikyselyillä keväällä 2014. Kiitämme tietojen antamisesta seuraavia henkilöitä: Sisko Brunni ja Valtteri Airaksinen (ICLFI), Ilmari Ivaska ja Kirsti Siitonen (LAS2), Tuija Hirvelä ja Sari Ahola (YKI), Maisa Martin (Cefling ja Topling), Riikka Ullakonoja, Ari Huhta ja Jaana Alila (Dialuki) sekä Maria Kela (Long Second).

2. Noudatamme tässä jakoa, jossa suomi vieraana kielenä -termillä viitataan ulkomailla tapahtuvaan suomen kielen opiskeluun ja suomi toisena kielenä -termillä puolestaan Suomessa tapahtuvaan suomen oppimiseen.

Aineistossa on vain tutkinnon hyväksytysti suorittaneiden ja tutkimusluvan antaneiden puhumisen ja kirjoittamisen suorituksia, ja se sisältää testin suorittajien taitotasoarviot ja taustatiedot. Puhumisen suoritukset ovat korpuksessa äänitiedostoina, ja niitä on jokaiselta suorittajalta yksi (Solki a). Suomen kielen kirjoittamisen suorituksia on noin 3 700 suorittajalta jokaiselta kolme kappaletta, ja korpukseen lisätään kunkin testikierroksen jälkeen lisää aineistoa. Kaikki tasoarviotiedot ja tausta-aineisto lisätään, mutta itse suorituksia lisätään otostaen. Aineiston kokoaminen on aloitettu vuonna 2002.

Cefling-korpus on koostettu vuosina 2007–2009 Jyväskylän yliopiston Cefling-hankkeessa, jota rahoitti Suomen Akatemia. Korpus koostuu yläkouluikäisten suomen ja englannin oppijoiden kirjoitelmista. Hankkeessa tutkittiin toisen ja vieraan kielen taidon kehittymistä taitotasolta toiselle, ja siinä myös verrattiin yläkoululaisten ja aikuisten oppijoiden suoriutumista kirjoittamistehtävissä; aikuisten aineistona käytettiin YKI-korpuksen tekstejä. (Cefling.) Hankkeessa kerättiin myös äidinkielisten suomenpuhujien suorituksia samoista kirjoittamistehtävistä, joten aineisto mahdollistaa oppijansuomen ja natiivisuomen vertailun. Lisäksi hankkeen yhteydessä kerättiin sananjohtamistehtäviä (ks. Penttinen 2010), mutta nämä tehtävät ja tulokset eivät ole yleisesti saatavilla. Cefling-aineistoon on tehty kieliopillinen annotaatio.

Topling-hankkeen korpus on niin ikään koottu Jyväskylän yliopistossa. Topling-hanke oli jatkoa Cefling-hankkeelle, se oli käynnissä vuosina 2010–2013 ja sitä rahoitti Suomen Akatemia. Hankkeessa tutkittiin, miten suomi toisena kielenä -oppijoiden sekä englannin ja ruotsin oppijoiden kirjoittamistaidot kehittyvät suomalaisessa koulutusjärjestelmässä. (Topling.) Korpus on diakroninen: sen aineisto on kerätty vuosina 2010–2012 kolmella eri keräyskierroksella. Aineiston suomenkieliset tekstit ovat alakoululaisten, yläkoululaisten ja lukiolaisten kirjoittamia, eli kyseessä on Cefling-korpuksen tavoin koululaisaineisto. Topling- ja Cefling-korpukset sisältävät eniten alimmille kielitaidon tasoille arvioituja tekstejä, eli ne mahdollistavat erityisesti kielenoppimisen alkuvaiheiden tutkimisen.

Myös Dialuki-korpus painottuu alimpien kielitaitotasojen teksteihin. Korpus koostuu niin ikään kouluikäisten oppijoiden kirjoitelmista, ja sekin on koottu Jyväskylän yliopistossa. Vuosina 2010–2013 käynnissä ollutta Dialuki-hanketta rahoittivat Suomen Akatemia ja Iso-Britannian Economic and Social Research Council (ESRC). Hankkeessa tutkittiin luku- ja kirjoitustaidon kehittymistä toisessa ja vieraassa kielessä ja pyrittiin selvittämään, mitkä kognitiiviset tekijät ennustavat kielenoppijan vahvuuksia ja heikkouksia luku- ja kirjoitustaidossa. Korpus sisältää venäjänkielisten oppijoiden suomenkielisiä kirjoitelmia ja suomenkielisten oppijoiden englannin kirjoitelmia. (Solki b, ks. myös Ullakonoja ym. 2012.) Korpuksen S2-aineistoa on kolmea tyyppiä: alakouluikäisten tekstit, yläkouluikäisten tekstit sekä pitkittäisaineisto, jossa osa alakouluikäisistä oppijoista teki saman tehtävän uudelleen noin kahden vuoden kuluttua ensimmäisestä keräyskerrasta. Korpus on siis osin diakroninen. Osana Dialuki-hanketta samat oppijat suorittivat laajan joukon myös muita kielitaitoa mittaavia tehtäviä, kuten äidinkielen ja toisen kielen lukemis- ja sanastotehtäviä, äidinkielen kirjoitustehtäviä ja psykolingvistisiä tehtäviä. Tämä aineisto sisältää myös puhuttua kieltä, sillä psykolingvistiset tehtävät tallennettiin äänitiedostoiksi. Lisäksi oppijansuomen suoritusten ohella hankkeessa kerättiin suorituksia myös äidinkielisiltä suomen- ja venäjänpuhujilta.

Korpuksista uusin on oppijansuomen Long Second -korpus (Long Second), jota on koottu vuodesta 2011 alkaen Helsingin ja Tallinnan yliopistojen yhteistyöhankkeena. Aineisto on uraauurtava suomalaisten oppijankorpusten joukossa, sillä edellä käsitellyistä, pääosin kirjoitetuista tekstiaineistoista poiketen se sisältää videoituja, monikielisiä luokkahuonetilanteita. Ne on tallennettu kahdesti viikossa yhden lukuvuoden ajan syyskuusta toukokuuhun. Joka toinen nauhoite tehtiin ryhmätyötunnilla ja joka toinen frontaaliopetustunnilla, mutta sosiaalimuodon vaihtelun lisäksi aineiston naturalistisuutta ei keruuvaiheessa suitsittu millään varsinaisilla testausasetelmilla. Aineisto on kerätty helsinkiläisen alakoulun valmistavassa luokassa; pääosallistujat edustavat viroa ja venäjää, mutta aineistossa on myös kurdin-, makedonian-, latvian- ja portugalinkielisiä lapsia. Lisäksi pääosallistujille tehtiin videointijakson lopuksi toukokuussa 2012 yksilöhaastattelut, jotka on myös videoitu. Aineistoon kuuluvat myös lapsille tehdyt sosiometriset mittaukset (syksyllä 2011 ja keväällä 2012) sekä kahden opettajan haastattelut keväältä 2012. Muista oppijansuomen aineistoista Long Second eroaa myös siinä, että se on ensisijaisesti pitkittäistutkimukseen tarkoitettu aineisto, joka perustuu multimodaaliseen ja naturalistiseen pienryhmävuorovaikutukseen, ja sen osallistujat ovat muuttaneet Suomeen vain hieman ennen keräyksen alkua.

Käsillä olevista korpuksista kuusi on suomi toisena kielenä -korpuksia, eli niiden materiaalin tuottajat oppivat suomen kieltä Suomessa. Ainoastaan siis ICLFI on suomi vieraana kielenä -korpus: sen tekstit on kerätty ulkomailla asuvilta suomen kielen oppijoilta. Cefling-, Topling- ja Dialuki-korpusten tekstit on kirjoitettu alun perin juuri korpusta varten, kun taas muiden neljän aineistot on tuotettu alkuaan muita tarkoituksia varten. ICLFI- ja LAS2-korpusten oppijat ovat kirjoittaneet tekstit oman opiskelunsa yhteydessä, ja YKI-korpuksen tekstit taas on tuotettu oppijoiden kielitaitotason arvioimista varten. Long Second -aineisto on syntynyt puolestaan luonnollisissa oppimistilanteissa.

Useimmissa oppijansuomen korpushankkeissa oppijat ovat kirjoittaneet tekstit käsin ja tekstit on siirretty sitten sähköiseen muotoon, mutta ICLFI- ja LAS2-korpuksissa on myös paljon alun perin tekstinkäsittelyohjelmilla kirjoitettuja tekstejä. Suurin osa käsiteltävistä korpuksista sisältää ainoastaan kokonaisia tekstejä, ei tekstikatkelmia. LAS2-korpuksessa on kuitenkin kokonaisten tekstien ohella myös tutkielmien lukuja (Ivaska 2014a: 25). Tärkeää on myös se, onko korpusten tekstit kirjoitettu suoraan suomeksi vai käännetty jostain muusta kielestä. Tässä käsiteltävistä kirjoitetun kielen korpuksista mikään ei sisällä käännöstekstejä, vaan aineistojen materiaali on tuotettu suoraan suomeksi.

Taulukossa 1 (ks. s. 92–93) esitetään kootusti perustietoja näistä seitsemästä oppijansuomen korpuksesta. Siinä on luokiteltu korpuksia erilaisten dimensioiden mukaan Jantusen (2011: 90–92) luokittelutapaan pohjautuen, ja lisäksi siinä on esitetty korpusten laajuus sekä mainittu esimerkkijulkaisuja kustakin korpuksesta. Korpuksista YKI, Cefling, Topling ja Dialuki sisältävät myös muita kuin suomenkielisiä tekstejä, samoin Long Second -aineisto myös muita kieliä kuin puhuttua suomea (joskaan aina ei ole mahdollista määritellä oppijan tuottaman vuoron kieltä), mutta taulukon laajuus-, genre- ja taitotaso-osioissa on ilmoitettu vain korpusten suomenkielisen aineiston määrät. Lisäksi Cefling-aineiston osalta on esitetty ainoastaan perusaineiston laajuus.

Taulukko 1.
Oppijansuomen korpusten perustiedot.

	ICLFI	LAS2	YKI	CEFLING	TOPLING	DIALUKI	LONG SECOND
Laajuus	4 850 tekstiä 920 000 sanetta	775 tekstiä 657 000 sanetta	11 200 kirjoittamisen suoritusta (780 000 sanetta), 1482 puhumisen suoritusta	527 tekstiä 26 000 sanetta	2 548 tekstiä 101 000 sanetta	307 tekstiä (josta pitkittäisaineistoa 61 tekstiä) 12 000 sanetta	36 oppituntia, joista 11 kokonaan ja 22 puolittain litteroitu (tilanne 12/2014)
Genre	kertomuksia, kuvauksia, esseitä, päiväkirjoja, arvosteluja, referaatteja, mielipidekirjoituksia, vastineita, uutisia, sähköpostiviestejä, kirjeitä, satiireja, työhakemuksia	akateemisia tekstejä: 43 % ajallisesti rajoitettuja tekstejä (tenttivastauksia), 30 % julkaistavaksi tarkoitettuja tekstejä (tutkielmien lukuja ja artikkelien käsikirjoituksia), 27 % ei-julkaistavaksi tarkoitettuja tekstejä (esim. esseitä)	kultakin kirjoittajalta kolme eri tekstiä: esim. epämuodollinen viesti, puolivirallinen kirjoitelma ja mielipidekirjoitus	epämuodollisia viestejä 34 %, muodollisia viestejä 23 %, kertomuksia 22 %, mielipidekirjoituksia 21 %	epämuodollisia viestejä 51 %, kertomuksia 18 %, muodollisia viestejä 16 %, mielipidekirjoituksia 14 %	narratiivisia, mielipiteen ilmaisua vaativia tekstejä 90 %, epämuodollisia viestejä 10 %	luokkahuonevuorovaikutus, josta 50 % opetuskeskusteluja ja 50 % vapaata keskustelua (esim. kertominen, tarinointi, vitsit, lauleskelu, argumentointi), lisäksi haastatteluita ja sosiometrisiä mittauksia
Teema	yleiskorpus	toistaiseksi humanistisen alan, erityisesti kielitieteen, tekstejä	yleiskorpus	yleiskorpus	yleiskorpus	yleiskorpus	yleiskorpus (erityisesti luokkahuonekorpus)
Rekisteri	kirjoitettu kieli	kirjoitettu kieli	kirjoitettua ja puhuttua kieltä	kirjoitettu kieli	kirjoitettu kieli	kirjoitettu kieli (sekä tarkkaan rajattuja puhumisen suorituksia)	puhuttu, monenkeskinen kieli

Tekstien kieli	suomi	suomi	suomi ja 8 muuta kieltä	suomi ja englanti	suomi, englanti ja ruotsi	suomi, englanti ja venäjä	suomi, venäjä, viro, englanti ja jonkin verran muita kieliä
Suomenkielisen aineiston variantit	oppijansuomi	oppijansuomi ja natiivisuomi	oppijansuomi	oppijansuomi ja natiivisuomi	oppijansuomi	oppijansuomi ja natiivisuomi	oppijansuomi ja natiivisuomi
Aika	synkroninen, osin diakroninen	synkroninen, osin diakroninen	synkroninen	synkroninen	diakroninen	synkroninen, osin diakroninen	diakroninen
Annotaatio	kieliopillinen annotaatio (tehty 100 %), virheannotaatio (tehty 5 %)	kieliopillinen annotaatio (tehty 61 %), mahdollisuus myöhempään virheannotointiin	ei annotaatiota	kieliopillinen annotaatio (tehty 100 %), virheannotaatiota ei tehty	ei annotaatiota	ei annotaatiota suomenkielisessä aineistossa (engl.kielisestä osa annotoitu)	ei-kielellinen annotaatio litteroinnin ohessa
Oppijoiden äidinkieli	22 äidinkieltä	15 äidinkieltä	16 äidinkieltä	yli 20 äidinkieltä	yli 20 äidinkieltä	venäjänkielisiä tai muuten venäjätaustaisia	venäjä, viro, portugali (myös kurdi, latvia, makedonia)
Taitotaso	A2: 7 % B1: 43 % B2: 36 % C1: 12 % C2: 2 %	B1: 4 % B2: 32 % C1: 62 % C2: 3 %	Perus-, keski- ja ylimmän tason suorituksia, joista keskitason suorituksia suhteessa eniten.	A1: 22 % A2: 38 % B1: 35 % B2: 6 %	alle A1: 1 % A1: 19 % A2: 43 % B1: 26 % B2: 10 % C1: 1 %	alle A1: 0 % A1: 30 % A2: 44 % B1: 21 % B2: 4 % C1: 0%	muuttuu ajassa: lähtötaso alle A1, päättötaso A1.3/A2.1
Julkaisuja	Jantunen 2011; Spoelman 2013.	Ivaska 2014a, 2014b.	Toivola & Tossavainen 2011; Tarnanen 2007.	Martin ym. 2010; Huhta ym. 2014.	Toropainen, Härmälä & Lahtinen 2012; Palviainen, Kalaja & Mäntylä 2012.	Alderson ym. 2015; Nieminen ym. 2011.	

Ceflingin perusaineistolla tarkoitetaan sitä hankkeen aineistoa, joka on arvioitu yhdenmukaisesti (arvioinnista tarkemmin luvussa 3). Hankkeen materiaaleihin kuuluu lisäksi muuta aineistoa: kaiken kaikkiaan Cefling-hankkeessa on kerätty 893 suomenkielistä S2-oppijoiden kirjoittamaa tekstiä ja lisäksi natiivisuomen aineisto. Dialukikorpuksen osalta taulukossa on esitetty S2-aineistoa koskevat lukumäärätiedot, mutta hankkeessa on kerätty myös natiiviaineistoa noin 1 000 tekstin verran ja lisäksi siis muita kielenoppimiseen liittyviä tehtäviä, joista osa sisältää puhuttua kieltä.

3 Taustatiedot

3.1 Yleisimmät taustamuuttujat

Korpuksia koostettaessa oppijoista, heidän suomen oppimisensa kontekstista sekä korpukseen koottavista tuotoksista on kerätty erilaisia metatietoja. Oppijankieltä tutkittaessa voidaan siis tarkastella monien eri taustamuuttujien vaikutusta kieleen sen mukaan, mitkä muuttujat käytettävässä aineistossa on otettu huomioon. Tutkimuksissa kiinnitetään yleisimmin huomiota taitotasoon ja oppijan äidinkieleen, ja niistä onkin kerätty tieto kaikissa seitsemässä käsiteltävässä korpuksessa. Taitotasojakaumat on esitetty taulukossa 1, ja taulukosta nähdään myös, että korpusten kielenoppijat edustavat monia eri äidinkieliä. Ainoastaan Dialukin S2-aineistossa äidinkielen suhteen ei ole vaihtelua, sillä kaikki oppijat ovat venäjänkielisiä tai muuten venäjätaustaisia. LAS2-, YKI- ja Ceflingkorpuksissa suurin äidinkieliryhmä on venäjä ja ICLFI-korpuksessa viro.

Tuotoksen taitotason ja oppijan äidinkielen lisäksi aineistoista on mahdollista tutkia myös monien muiden taustamuuttujien vaikutusta oppijankieleen ja tarkastella eri tekstilajeja, joita oppijat tuottavat. Long Secondin ”tekstilajit” poikkeavat luonnollisesti kirjoitetun kielen genreistä: niitä ovat frontaaliopetus ja ryhmätyöskentely (ks. taulukkoa 1, s. 92–93). Aineiston tekstilajit tuotetaan sosiaalimuodon (pulpettirivi vs. ryhmätyöpöytä), vuorovaikutusrakenteen (opetuskeskustelu vs. epämuodollinen jutustelu) ja tunnilla tapahtuvan toiminnan (yksilö- vs. ryhmätyö) kautta.

Korpuksissa yleisimmin huomioon otetut taustamuuttujat on esitetty alla olevassa luettelossa. Kunkin muuttujan perässä on mainittu ne korpukset, joista kyseinen taustamuuttuja puuttuu. Taustamuuttujien luokittelussa on käytetty osittain hyödyksi Jantusen (2011: 93) esitystapaa, jossa taustatiedot on luokiteltu oppijaa, oppimiskontekstia ja tuotosta koskeviin muuttujiin.

Suomenoppija
Henkilötiedot
– Ikä
– Sukupuoli (ei Cefling)
Kielitaito
– Äidinkieli
– Muut oppijan hallitsemat kielet (ei YKI)
– Suomen opiskelu vuosina (ei Dialuki)

– Taitotaso itsearvioituna kielen eri alueilla (ei ICLFI, YKI, Long Second eikä Dialukin alakouluaineisto)

Oppimiskonteksti

– Suomen käyttö kotikielenä
– Suomen käyttö oppimistilanteiden ulkopuolella (ei ICLFI)

Teksti/tuotos

– Taitotaso EVK:n mukaan
– Tekstilaji
– Kirjoituksen tehtävänanto (ei LAS2)
– Keräysaika

Taustatiedot ovat eri korpuksissa hyödynnettävissä eri muodossa. ICLFI-korpuksessa kaikki taustatiedot on lisätty korpuksen nykyversiossa kunkin tekstin yhteyteen. LAS2-korpuksessa puolestaan jokaisen tekstin yhteyteen on merkitty joitakin tietoja: oppijan ID-numero, tekstin genre, tekstin ID-numero, oppijan äidinkieli, oppijan saama ylin ja alin taitotasoarvio sekä mahdollisesti aika, joka on kulunut oppijan edellisen tekstin keräämisestä. Muut taustatiedot on linkitetty kuhunkin tekstiin. (Ivaska 2014a: 26–28.) Suoraan tekstin yhteyteen merkityt tiedot ovat siis hieman vaivattomammin hyödynnettävissä suoraan tekstitiedostosta, kun aineistoa tarkastelee teksteittäin, mutta toisaalta taustatietojen etsiminen koko aineistosta on hankalaa ilman selkeää kansio- tms. -rakennetta tai hakutoimintoa. YKI- ja Dialuki-korpusten tiedot on lisätty samantapaisella periaatteella: testinsuorittajien ID-numerot yhdistävät taustatiedot ja tekstit toisiinsa ja tekevät mahdolliseksi hakujen tekemisen kahden eri aineiston välillä (Solki a). Topling- ja Cefling-hankkeiden taustatietolomakkeita puolestaan ei ole digitoitu, vaan niitä säilytetään paperilomakkeina, joten taustatiedot eivät ole saatavissa suoraan korpuksesta. Aineistojen liittäminen osaksi Kielipankkia FinClarin-konsortion puitteissa lähitulevaisuudessa muuttanee metatietojen linkityksiä teksteihin jossain määrin.

3.2 Oppijaa ja oppimiskontekstia koskevat taustamuuttujat

Oppijan taustasta ja oppimiskontekstista voidaan korpuksia koostettaessa kerätä lisäksi monia muitakin tietoja kuin luettelossa kuvattuja yleisimpiä tietoja; jotkin metatiedot ovat siis muuttujina esimerkiksi vain yhdessä tai kahdessa korpuksessa. Tällainen metatieto on esimerkiksi se, millainen sosioekonominen asema oppijalla on. Se on tiedossa YKI- ja Dialuki-korpuksista. Siihen, mitä oppikirjaa oppija on käyttänyt suomea opiskellessaan, kiinnitetään huomiota puolestaan ICLFI- ja Topling-aineistoissa. Korpusten avulla voidaan tutkia myös esimerkiksi sitä, millainen vaikutus oppijan vanhempien tai opettajien äidinkielellä on oppijan tuottamaan kieleen. Vanhempien äidinkielet ovat tiedossa ICLFI- ja Dialuki-aineistoissa ja opettajan äidinkieli vain ICLFI-korpuksessa. Dialuki-aineiston tekstintuottajilta on puolestaan selvitetty myös heidän motivaatiotaan suomen kielen opiskeluun (Solki c).

Dialuki-hankkeessa keskityttiin kouluikäisten luku- ja kirjoitustaidon kehittymisen tutkimiseen, joten hankkeen aineistossa on joitakin aiheeseen liittyviä muuttujia, joita

ei ole muissa korpuksissa. Hankkeen taustatietoina oppilailta on kysytty esimerkiksi lukemaan oppimisen ikää sekä vanhempien ja muiden sukulaisten lukemisvaikeuksia. Myös oppilaiden kouluarvosanoja suomen kielessä on kysytty Dialuki-hankkeessa, kuten myös Topling- ja Cefling-hankkeissa, joiden aineistot niin ikään sisältävät siis nimenomaan kouluikäisten oppijoiden kieltä.

ICLFI-, LAS2- ja YKI-korpukset sisältävät aikuisten oppijoiden kieltä. Näistä LAS2-korpuksen oppijoiden kieli on lähtökohtaisesti edistyneellä tasolla, ja niinpä tekstintuottajilta on tiedusteltu muun muassa heidän itse antamaansa suomen kielen opetusta, joka ei ole muuttujana missään muussa korpuksessa. LAS2-korpuksessa on ylipäätään useita muuttujia, joita ei oteta huomioon muissa korpuksissa. Tällaisia ovat esimerkiksi suoritetut suomen kielen opinnot (esim. perus- tai aineopinnot) ja tieto siitä, onko oppijalla sellaisia tuttavia, joiden kanssa hän puhuu vain suomea.

Suomen käyttö oppimistilanteiden ulkopuolella on metatietona viidessä korpuksessa, mutta muuttuja otetaan niissä huomioon keskenään hieman eri tavoin. Metatietona voi olla esimerkiksi se, kuinka usein oppija käyttää suomea vapaa-ajallaan. Tiedot voivat kuitenkin olla yksityiskohtaisempiakin: LAS2-, Cefling-, Topling- ja Dialuki-korpuksista selviää esimerkiksi, millaisia tekstilajeja oppija kirjoittaa vapaa-ajallaan. Dialuki-aineiston oppijoilta on kysytty myös heidän omaa suhtautumistaan lukemiseen ja kirjoittamiseen vapaa-ajalla ja lisäksi sitä aikaa, jonka he päivässä käyttävät lukemiseen ja kirjoittamiseen. LAS2-korpuksessa taustamuuttujana on luetun suomenkielisen kirjallisuuden määrä sivuina sekä se, kuinka usein oppija lukee suomenkielistä sanomalehteä.

3.3 Taitotaso taustamuuttujana

Korpuksissa on useanlaista tietoa taitotasoista. Kaikkien korpusten tuotokset on ensinnäkin luokiteltu eurooppalaisen viitekehyksen (EVK) kielitaitotasojen mukaan. Koululaisaineistojen tuotokset on arvioitu toisekseen myös opetussuunnitelman mukaan, mikä tarjoaa mahdollisuuden verrata eri arviointitapoja toisiinsa. Lisäksi kustakin oppijasta on tiedossa se, kuinka kauan tämä on opiskellut suomea. Dialuki-aineistossa on tosin kysytty vain, milloin oppija on muuttanut Suomeen ja miltä luokalta alkaen käynyt Suomessa koulua. Sen lisäksi että ICLFI-korpuksessa jokainen yksittäinen teksti on arvioitu EVK-taitotasojen mukaan, tekstintuottajat on jaettu alkeis-, keski- ja edistyneen tason oppijoihin sen mukaan, montako tuntia suomen opetusta he ovat arviolta saaneet. Tämä tieto näkyy taustatiedoista, mutta aineistoa ei ole luokiteltu opiskelijan vaan tekstin taitotason mukaan.

ICLFI-korpuksessakin kullekin oppijalle voidaan myös määritellä taitotaso niiden suoritusten perusteella, joita oppija on tuottanut korpukseen. Näin on tehty Dialuki- ja YKI-korpuksissa. Dialuki-aineiston jokaiselle tekstintuottajalle on arvioitu luetun ymmärtämisen sekä kirjoittamisen taitotaso. YKI-korpuksessa jokaiselle tutkinnon suorittajalle on puolestaan määritelty tasoarvio neljästä eri osataidosta: tutkinto mittaa kirjoittamista, puheen ymmärtämistä, puhumista sekä tekstin ymmärtämistä. Oppija saa arvion siis jokaisesta osataidosta. YKI-korpuksessa on kolmen eri tutkintotason suorituksia: perustasolla tutkinnon suorittaja voi saada taitotasoarvion alle 1, 1 tai 2,

keskitasolla alle 3, 3 tai 4 ja ylimmällä tasolla puolestaan alle 5, 5 tai 6. (Solki a; OPH.) Tasot on linkitetty eurooppalaisen viitekehyksen taitotasoille.

Taitotason arviointi on toteutettu hieman eri periaattein eri korpuksissa. LAS2-korpuksen tekstit on arvioinut yksi arvioija, ja arvio on tehty vähintään kahdesta informantin tuottamasta tekstistä. Samoin YKI-korpuksen kunkin suoritukset on arvioinut yksi henkilö, mutta tarvittaessa on käytetty useampia arvioijia (Solki d). ICLFI-korpuksen kunkin tekstin on taas arvioinut kaksi arvioijaa, ja jos nämä arviot poikkeavat toisistaan, arvion on antanut lisäksi kolmas henkilö. Cefling-, Topling- ja Dialuki-hankkeissa arvioijia on puolestaan ollut kolme. Cefling-hankkeen perusaineistoon on otettu mukaan vain ne tekstit, joiden taitotason arvioijat ovat arvioineet keskenään riittävän samalla tavoin: vähintään kahden kolmesta arvioijasta on täytynyt arvioida teksti samalle EVK-tasolle, ja kolmas arviointi on saanut poiketa tästä korkeintaan yhden EVK-tason. Jos tekstin taitotason arvioinnissa on ollut enemmän hajontaa eri arvioijien kesken, tekstiä ei ole otettu mukaan aineistoon. Topling- ja Dialuki-hankkeissa tekstien arviointi on analysoitu Facets-ohjelman tilastollisilla menetelmillä. Ohjelman avulla voidaan esimerkiksi ottaa huomioon erot siinä, kuinka ankarasti kukin arvioija on arvioinut suoritukset. Topling-hankkeessa Facets-analyysin perusteella on jätetty aineistosta pois ne tekstit, joiden arvioinnit poikkeavat toisistaan merkittävästi.

Taitotasoa tarkasteltaessa voidaan ottaa huomioon myös oppijan tekemä itsearviointi omasta kielitaidostaan. LAS2-, Topling- ja Cefling-korpusten tekstintuottajat ovat arvioineet oman suomen kielen taitonsa kielen eri alueilla, kuten sanaston hallinnassa, kirjoittamisessa, puheen ymmärtämisessä ja puhumisessa. LAS2-korpuksen tekstintuottajat ovat arvioineet kielitaitonsa asettamalla järjestykseen, minkä kielen osa-alueista he hallitsevat parhaiten, minkä toiseksi parhaiten ja niin edelleen. Topling- ja Cefling-korpusten oppijat puolestaan ovat tehneet itsearvioinnin kouluarvosanoin tai hymiöiden avulla kielen eri osa-alueilla. Dialukin yläkouluaineiston oppijoiden tekemä arviointi on puolestaan toteutettu kyllä–ei-väittämin, ja se keskittyy ainoastaan kirjoittamisen ja lukemisen arvioimiseen, mutta arviointi on siinä yksityiskohtaisempaa ja perustuu eurooppalaiseen viitekehykseen.

3.4 Tekstejä koskevat taustamuuttujat

Korpuksissa on oppijoita ja oppimiskontekstia koskevien taustamuuttujien lisäksi myös tekstiä koskevia muuttujia, joista tekstin taitotasoa käsiteltiin edellä. Alaluvun 3.1 luettelossa on esitetty yleisimmät tekstiä koskevat taustamuuttujat, mutta niiden lisäksi erityisesti ICLFI-korpuksessa on muitakin tekstiin liittyviä muuttujia. ICLFI-korpuksessa tekstien kesken on vaihtelua paitsi luettelossa mainittujen neljän muuttujan myös seuraavien tekijöiden suhteen: onko kirjoittamiseen käytetty aika rajattu vai rajaamaton, onko kirjoittamistilanne luonteeltaan testimäinen, onko kirjoittamisessa käytetty sanakirjoja, oppikirjoja tai muita apuvälineitä, onko kirjoituspaikkana koti, koulu vai jokin muu paikka, missä kaupungissa teksti on kerätty ja onko teksti kirjoitettu alun perin käsin vai tekstinkäsittelyohjelmalla? ICLFI-korpuksessa nämä tekijät on otettu huomioon taustamuuttujina ja merkitty erikseen jokaisen tekstin yhteyteen.

Myös LAS2-korpuksessa näissä tekijöissä on vaihtelua eri tekstien kesken (lukuun ottamatta keräyspaikkaa, joka on kaikkien tekstien osalta Turku), mutta tietoja ei kuitenkaan ole merkitty teksteihin, eli ne eivät ole aineistossa varsinaisina taustamuuttujina. Siitä huolimatta nämä seikat ovat jossain määrin tiedossa LAS2-korpuksen teksteistä, sillä esimerkiksi keräystapa on pääosin tiedossa osakorpuksittain. Ajallisesti rajoitetut tekstit on nimittäin kirjoitettu pääosin käsin, julkaistavaksi tarkoitetut tekstit tekstinkäsittelyohjelmalla ja myös ei-julkaistavaksi tarkoitetut tekstit pääosin tekstinkäsittelyohjelmalla. Sen sijaan YKI-, Cefling-, Topling- ja Dialuki-korpuksissa monet mainituista tekijöistä ovat vaihtelemattomia korpuksen eri tekstien kesken. Näissä korpuksissa nimittäin kaikki tekstit on esimerkiksi kirjoitettu tilanteessa, jossa sanakirjoja tai muita apuvälineitä ei ole käytetty, ja samoin niiden tekstit on kirjoitettu ajallisesti rajatuissa olosuhteissa. Näiden korpusten osalta ei ole siis mahdollista tutkia, miten näiden tekijöiden vaihtelu vaikuttaisi oppijan tuottamaan kieleen.

Tekstiä koskevista taustatiedoista kirjoitusten tehtävänantoa ei ole liitetty LAS2-korpukseen, sillä aineistoon on kerätty ainoastaan itse opintosuoritukset. YKI-korpuksessa varsinaiset tehtävänannot taas ovat salassa pidettäviä, mutta kuhunkin tehtävään on kuitenkin liitetty tieto tehtävätyypistä: tehtävät ovat joko otsikkopohjaisia tai ohjattuja. Otsikkopohjaisissa tehtävissä annetaan muutama otsikko tai väittämä, joiden pohjalta laaditaan kirjoitelma. Ohjatuissa tehtävissä puolestaan mainitaan, kenelle kirjoitetaan ja millaisessa tilanteessa. Tehtävässä voidaan antaa kirjoitelman sisältöä koskevia ohjeita tai lyhyehkö teksti, jonka pohjalta teksti laaditaan. (Solki e.)

Eri korpusten tehtävänannot poikkeavat toisistaan siltä osin, kuka ne on määritellyt. Tehtävänannot ovat LAS2- ja ICLFI-korpuksissa opettajien määrittelemiä, kun taas Cefling-, Topling- ja Dialuki-aineistoissa ne ovat tutkijoiden määrittelemiä. YKI-korpuksen tehtävänannot puolestaan määritellään yleisissä kielitutkinnoissa. Eri korpusten tehtävänannot poikkeavat toisistaan myös sisällöltään, ja lisäksi yhden korpuksen sisälläkin voi olla lukuisia eri tehtävänantoja. Erityisesti ICLFI-korpuksessa on paljon erilaisia tehtävänantoja: oppijoita on pyydetty muun muassa kertomaan opiskelijaarjesta, kirjoittamaan kirje ystävälle, laatimaan mielipidekirjoitus artikkelin pohjalta, analysoimaan katsomaansa elokuvaa ja kirjoittamaan yhteenveto radiokuunnelman pohjalta. Cefling-aineistossa tehtävänantoja on puolestaan viisi erilaista: tehtävänä on ollut kirjoittaa sähköpostiviesti ystävälle, opettajalle ja verkkokauppaan sekä laatia mielipidekirjoitus ja kertomus. Topling-hankkeen tehtävät perustuvat Cefling-hankkeen tehtäviin. Edelleen Dialuki-aineiston tehtävänannot (n. 90 %) pohjautuvat Cefling- ja Topling-hankkeissa kehitettyihin tehtävänantoihin.

4 Annotointi

YKI-, Topling- ja Dialuki-korpusten tekstejä ei ole koodattu eli annotoitu, vaan korpukset sisältävät ainoastaan raakatekstiä eli sen materiaalin, minkä oppijat ovat tuottaneet. ICLFI-, LAS2-, Cefling- ja Long Second -korpuksia puolestaan on annotoitu. Annotoinnilla tarkoitetaan eri yhteyksissä hieman eri asioita, mutta perinteisesti sillä

tarkoitetaan lingvistisen tiedon lisäämistä korpukseen (Gries 2009: 9–10; Heikkinen, Lounela & Voutilainen 2012: 374; Leech 2004). Voidaan ajatella, että annotointi tekee korpuksesta käytettävämmän, sillä annotointi mahdollistaa aineistojen automaattisen analysoinnin myös kieliopillisten piirteiden osalta. Jokaiseen sanaesiintymään voidaan esimerkiksi merkitä kyseisen lekseemin sanaluokka. Sanaluokan koodaus (*part-of-speech tagging*, *POS tagging*) tekee esimerkiksi homonyymien erottamisen korpuksen käyttäjälle helpommaksi: eri sanaluokkiin kuuluvien homonyymien frekvenssejä tai muita piirteitä voidaan tällöin tarkastella korpuksessa erikseen. Kaikkiaan korpusten annotaatio tulisi aina toteuttaa siten, että alkuperäisestä tekstistä ei kadoteta mitään ja että raakateksti olisi sekin tutkijoiden saatavilla, sillä kaikille korpuksen käyttäjille annotoinnista ei ole hyötyä. (Leech 2004.) Kaikista käsillä olevista annotoiduista korpuksista on säilytetty myös raakateksti.

Korpuksiin voidaan lisätä sekä kielioppiin että kielivirheisiin liittyvää tietoa, ja näitä kahta prosessia (kuten myös niiden lopputuloksia) kutsutaan kieliopilliseksi annotoinniksi sekä virheannotoinniksi. Kieliopilliseen annotointiin voi sisältyä morfologista ja syntaktista analyysia sekä lemmatisointia (Heikkinen ym. 2012: 375). Annotointi tehdään ICLFI-korpuksessa automaattisesti Connexorin Fi-fdg-jäsentimellä käyttäen hyödyksi Tieteen tietotekniikan keskuksen, CSC:n, etäpalvelinta, mutta tulos tarkistetaan manuaalisesti jälkeenpäin. Korpuksen annotoinnissa käytettävä jäsennin on kehitetty alun perin natiivikielen analyysia varten (annotoinnista tarkemmin Jantunen, Brunni, Lehto & Airaksinen 2014). LAS2-korpuksessa annotointi tehdään niin ikään jäsentimellä ja tarkistetaan manuaalisesti jälkikäteen, mutta sen jäsennin on kehitetty varta vasten oppijankielen analyysiin, juuri LAS2-korpuksen tarpeisiin (Ivaska 2014a: 28). Cefling-aineiston tekstit on puolestaan koodattu manuaalisesti: koodaajat ovat merkinneet teksteihin tietyt morfosyntaktiset piirteet. Tekstit on koodattu CHAT-tiedostoiksi (*Codes for Human Analysis of Transcripts*), analysoitu CLAN-ohjelmalla (*Computerized Language Analysis*) ja tallennettu CHILDES-tietokantaan.

Oppijansuomen korpuksiin on tehty enemmän kieliopillista kuin virheannotointia. Sekä ICLFI-, LAS2- että Cefling-korpuksiin on koodattu esimerkiksi paikallissijat eli merkitty paikallissijoissa olevien sanojen yhteyteen, mistä sijamuodosta on kyse (LAS2:n tieto: Ivaska 2014a: 27). Toisaalta korpukset kuitenkin hiukan poikkeavat toisistaan sen suhteen, mitä eri morfologisia ja syntaktisia piirteitä niihin on koodattu. Annotoiduista kolmesta kirjoitetun kielen korpuksesta kaikkia on morfosyntaktisen annotoinnin lisäksi lemmatisoitu. Cefling- ja ICLFI-aineistot on lemmatisoitu kokonaan ja LAS2-korpus osittain; tarkemmat tiedot esitetään taulukossa 1 (s. 92–93). Lemmatisoinnissa jokaisen korpuksen sanan yhteyteen lisätään kyseisen sanan perusmuoto eli lemma[3] (Gries 2009: 10). Tämän koodauksen ansiosta sanan kaikki taivutusmuodot voidaan hakea korpuksesta yhdellä haulla. Oppijankielen korpuksissa lemmatisointi on erityisen oleellista, sillä oppijoiden tekemien erilaisten virheiden vuoksi erilaisten sananmuotojen määrä on suuri. (Jantunen ym. 2014.)

3. Suomen kielen tutkimuksessa on usein eroteltu toisistaan eri sananmuotojen muodostama abstraktio ja sanan yksittäinen esiintymä käyttämällä termiparia *sana – sane* tai termiä *lekseemi*. Muun muassa korpustutkimuksen myötä on kuitenkin alettu käyttää myös termiä *lemma*.

Virheannotoinnissa tekstiin koodataan esimerkiksi oikeinkirjoitus- ja kielioppivirheet. Virheannotointi mahdollistaa sellaisten piirteiden analysoimisen, jotka ovat tyypillisiä oppijankielelle, mutteivät natiivipuhujien kielelle. Virhekoodatusta korpuksesta voidaan löytää paitsi odotuksenmukaisia myös ennakoimattomia kielenpiirteitä. Lisäksi virheiden mukaan koodatussa aineistossa päästään käsiksi esimerkiksi tapauksiin, joissa tekstintuottaja on jättänyt käyttämättä esimerkiksi pronominia, konjunktiota tai muuta tarvittavaa sanaa. (Dagneaux, Dennes & Granger 1998: 172; Granger 2002: 14; Jantunen ym. 2014.) Kieliopillinen annotointi ei mahdollista tällaista analyysia.

Virheannotoinnin etuja ei kuitenkaan voida toistaiseksi kovin laajalti hyödyntää oppijansuomen tutkimuksessa. Ensinnä nimittäin Cefling-korpuksessa ei ole lainkaan tehty varsinaista virheannotointia. Aineistoon on kuitenkin merkitty omalla koodillaan kohdat, joista selvästi puuttuu jokin kielellinen aines, esimerkiksi *olla*-verbi, objekti tai paikallissijan pääte, joten sikäli Cefling-aineistosta on mahdollista pieniltä osin analysoida oppijankielen virheitä. Myöskään LAS2-korpukseen ei ole tehty varsinaista virheannotaatiota, mutta siihen on kuitenkin varattu mahdollisuus virheiden merkitsemiseen: kieliopillisen annotoinnin yhteyteen on lisätty kommenttiosio, johon virheet voidaan myöhemmin koodata (Ivaska 2014a: 27). Oppijansuomen korpuksista toistaiseksi ainoastaan ICLFI-korpukseen on tehty systemaattista virheannotointia. Sen tekemistä varten ICLFI-hankkeessa on luotu virheluokitus, joka sisältää yhdeksän eri virhekategoriaa; näitä ovat esimerkiksi ortografiset, morfologiset ja leksikaaliset virheet. Virheannotointisysteemin luominen on aloitettu vuoden 2013 alussa, ja toistaiseksi virheannotointi on tehty noin viiteen prosenttiin korpuksen aineistosta. (Jantunen ym. 2014.)

Long Second -aineistoon on litteroinnin yhteydessä tehty puheaineiston käytettävyyden kannalta välttämättömäksi katsottu ei-kielellinen annotointi. Tähän mennessä systemaattisesti annotoidut seikat liittyvät kielenvalintaan, prosodiikkaan ja melodisiin jaksoihin. Annotointiin on käytetty ELAN-litterointiohjelmassa olevaa erillistä kommenttiraitaa ja merkinnät on tehty englanniksi, ajatellen kansainvälistä tutkijayhteisöä. Litteraatteihin on merkitty puheenvuoron kieli (*English*, *Estonian*, *German*, *Russian*, *English/Estonian*, *English/Finnish*, *English/Russian*, *Finnish/Estonian*, *Finnish/Russian*, *Gibberish*, *unclear language*), erilaiset toistuvat äännähdykset (*burping*, *coughing*, *explosion sound*, *farting sound*, *laughing*, *sighing*, *sniffling*, *whining voice*, *yawning*), melodiset äänet (*lilting*, *singing*, *whistling*) ja prosodiset erikoisuudet (*high pitch*, *palatalization*, *quiet voice*, *syllabifying*, *ultra falsetto*, *whispering*). Erityisen haastavaa on ollut kielen määrittely. Esimerkiksi luokan monikielisin oppilas puhuu viroa, suomea, englantia ja venäjää välillä samassa puheenvuorossa, ja usein on mahdotonta määritellä, onko jokin lausuma viroa vai suomea. Siksi on päädytty merkintään *Estonian/Finnish* tai *Finnish/Estonian*, sen mukaan kumpi kieli on litteroijan mielestä puheenvuorossa ollut voitolla. Summittaisestakin kielimerkinnästä on kuitenkin välitöntä hyötyä tutkijoille, sillä hakutoiminnon avulla voi nyt nostaa esimerkiksi kaikki litteroidut viro/suomi-puheenvuorot tarkempaan analyysiin, jolloin tutkija voi itse tarkentaa litteraattia oman näkemyksensä mukaiseksi. Viron- ja englanninkielistä puhetta ei ole suomennettu eikä venäjänkielistä puhetta käännetty englanniksi. Tällaiset jäävät kielentutkijoiden itsensä tehtäviksi tai käännätettäviksi.

5 Lopuksi

Kuten edellä on esitetty, oppijansuomen korpukset ovat keskenään erilaisia esimerkiksi dimensiopiirteiden, taustamuuttujien ja annotoinnin suhteen. Muun muassa tutkimusaiheesta riippuu, mikä aineisto on mihinkin tutkimukseen tarkoituksenmukaisin. Koska aineistoja on kerätty ilman yhteistä koordinointia ja jokaisen tutkimusryhmän omista intresseistä lähtien, eivät ne ole yhteismitallisia ja verrannollisia, mikä tuo omat ongelmansa tutkimukseen. Esimerkiksi tehtävänantojen erilaisuus korpusten välillä vaikuttaa selvästi sanastotutkimuksen onnistumiseen. Oppimiskontekstivertailu (suomi toisena ja vieraana kielenä) kohtaa omat hankaluutensa sekä aineistojen tehtävänantojen erilaisuuden että taitotasojen epätasaisuuden vuoksi. Omanlaisensa ja omalla tavallaan vakavakin ongelma on myös se, ettei oppijanaineistoille verrannollisia natiiviaineistoja ole kerätty kuin kolmelle edellä mainituista korpuksista; jos tavoitteena on selvittää, miten kielenoppijan tuotokset poikkeavat niin sanotuista natiiviteksteistä, olisi vertailu voitava tehdä mahdollisimman verrannollisiin teksteihin, mitä tulee muun muassa tekstintuottajien ikään, koulutukseen ja tekstien tekstilajeihin, tuottamisprosesseihin ja tehtävänantoihin.

ICLFI-, YKI- ja LAS2-korpuksia kartutetaan jatkuvasti, ja ICLFI- ja LAS2-korpusten annotointeja jatketaan edelleen. Kaikkia aineistoja ei ole annotoitu eikä annotointia ole näköpiirissä, mikä osaltaan rajaa tutkimuskysymyksiä tai ainakin menetelmiä. Osa korpuksista siis kehittyy paraikaa. Pitkittäisaineistojen ja puhekieltä sisältävien aineistojen vähäistä määrää korjaa uusi Long Second -korpus, mutta vastaavia aineistoja olisi saatava tutkijoiden käyttöön enemmän.

Kaikki käsitellyt seitsemän korpusta ovat tutkijoiden käytettävissä ja tarjoavat mahdollisuuden tutkia oppijan tuottamaa suomea laajojen aineistojen pohjalta, joissa materiaali on sähköisessä muodossa ja siten helposti tutkijan ulottuvilla. Edellä käsiteltyjen aineistojen käytettävyys ja saatavuus myös helpottuu, kun ne siirretään pois yliopistojen omilta palvelimilta. Käynnissä onkin aineistojen siirtäminen osaksi Kielipankkia Fin-Clarin Content -rahoituksen avulla.

Lähteet

Alderson, J. Charles – Haapakangas, Eeva-Leena – Huhta, Ari – Nieminen, Lea – Ullakonoja, Riikka (tulossa 2015): *The diagnosis of reading in a second or foreign language*. New Perspectives in Language Assessment Series. New York: Routledge.

Cefling. Linguistic Basis of the Common European Framework for L2 English and L2 Finnish. https://www.jyu.fi/hum/laitokset/kielet/tutkimus/hankkeet/paattyneet-hankkeet/cefling/suom (30.5.2014).

Dagneaux, Estelle – Dennes, Sharon – Granger, Sylviane 1998: Computer-aided error analysis. – *System* 26 s. 163–174.

Dumont, Amandine – Granger, Sylviane 2014: *Learner corpora around the world*. Louvain-la-Neuve: Université catholique de Louvain, Centre for English Corpus Linguistics. http://www.uclouvain.be/en-cecl-lcworld.html (2.12.2014).

Granger, Sylviane 2002: A bird's-eye view of learner corpus research. – Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (toim.), *Computer learner corpora, second language*

acquisition and foreign language teaching s. 3–33. Amsterdam: John Benjamins.
Gries, Stefan Th. 2009: What is corpus linguistics? – *Language and Linguistics Compass* 3 s. 1–17.
Heikkinen, Vesa – Lounela, Mikko – Voutilainen, Eero 2012: Automaattinen analysaattori tekstilajitutkimuksessa. – Vesa Heikkinen, Eero Voutilainen, Petri Lauerma, Ulla Tiililä & Mikko Lounela (toim.), *Genreanalyysi. Tekstilajitutkimuksen käsikirja* s. 372–391. Kotimaisten kielten keskuksen julkaisuja 169. Helsinki: Gaudeamus.
Huhta, Ari – Alanen, Riikka – Tarnanen, Mirja – Martin, Maisa – Hirvelä, Tuija 2014: Assessing learners' writing skills in a SLA study. Validating the rating process across tasks, scales and languages. – *Language Testing* 31 s. 307–328. http://ltj.sagepub.com/content/early/recent.
Ivaska, Ilmari 2014a: The corpus of advanced learner Finnish (LAS2). Database and toolkit to study academic learner Finnish. – Jarmo H. Jantunen, Sisko Brunni & Marianne Spoelman (toim.), Learner language, learner corpora. From corpus compilation to data analysis. – *Apples – Journal of Applied Language Studies* 8 (special issue 3) s. 21–38. http://apples.jyu.fi/.
—— 2014b: Edistyneen oppijansuomen avainrakenteita. Korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin. – *Virittäjä* 118 s. 161–192.
Jantunen, Jarmo Harri 2011: Kansainvälinen oppijansuomen korpus (ICLFI). Typologia, taustamuuttujat ja annotointi. – Annekatrin Kaivapalu, Johanna Laakso, Pirkko Muikku-Werner, Pirkko & Maria-Maren Sepper (toim.), *Lähivõrdlusi. Lähivertailuja* 21 s. 86–105. http://dx.doi.org/10.5128/LV21.04.
Jantunen, Jarmo Harri – Brunni, Sisko – Lehto, Liisa-Maria – Airaksinen, Valtteri 2014: Oppijankieliaineistojen annotointi. Esimerkkinä ICLFI:n annotoinnin prosessit, ongelmat ja ratkaisut. – Maarit Mutta, Pekka Lintunen, Ilmari Ivaska & Pauliina Peltonen (toim.), *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 7 s. 60–80. http://ojs.tsv.fi/index.php/afinla/article/view/48160 (2.12.2014).
Jantunen, Jarmo Harri – Piltonen, Saana 2009: Oppijansuomen ja -viron sähköiset tutkimusaineistot. – *Virittäjä* 113 s. 449–457.
LAS2. Edistyneiden suomenoppijoiden korpus. http://www.utu.fi/fi/yksikot/hum/yksikot/suomi-sgr/tutkimus/tutkimushankkeet/las2/Sivut/home.aspx (30.5.2014).
Leech, Geoffrey 2004: Adding linguistic annotation. – Martin Wynne (toim.), *Developing linguistic corpora. A guide to good practice* s.17–29. Oxford: Oxbow Books. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm (30.5.2014)
Long Second. Long Second: Suomen kielen kehittyminen alakoulun valmistavalla luokalla. http://blogs.helsinki.fi/kielen-ja-kirjallisuuden-didaktiikan-tutkimus/tutkimushankkeet/long-second (2.12.2014).
Martin, Maisa – Mustonen, Sanna – Reiman, Nina – Seilonen, Marja 2010: On becoming an independent user. – Inge Bartning, Maisa Martin & Ineke Vedder (toim.), *Communicative proficiency and linguistic development, intersections between SLA and language testing research* s. 57–80. EUROSLA Monographs Series 1. European Second Language Association. http://www.eurosla.org/monographs/EM01/EM01home.php.
Nieminen, Lea – Huhta, Ari – Ullakonoja, Riikka – Alderson, J. Charles 2011: Toisella ja vieraalla kielellä lukemisen diagnosointi. Dialuki-hankkeen teoreettisia ja käytännöllisiä lähtökohtia. – Esa Lehtinen, Sirkku Aaltonen, Merja Koskela, Elina Nevasaari & Mariann Skog-Södersved (toim.), *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 3 s. 102–115. http://ojs.tsv.fi/index.php/afinla/article/view/4470/4216 (30.5.2014).
OPH = Opetushallitus: Tietoa kielitutkinnoista. http://www.oph.fi/koulutus_ja_tutkinnot/kieli-

tutkinnot/yleiset_kielitutkinnot/tutkintoesite (17.6.2014).
Palviainen, Åsa – Kalaja, Paula – Mäntylä, Katja 2012: Development of L2 writing. Fluency and proficiency. – Lea Meriläinen, Leena Kolehmainen & Tommi Nieminen (toim.), *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 4 s. 47–59. http://ojs.tsv.fi/index.php/afinla/article/view/7037.
Penttinen, Kati 2010: *Voisitko apua? Suomi toisena kielenä -oppijoiden sananmuodostustaitojen jäljillä*. Pro gradu -tutkielma. Jyväskylän yliopisto, kielten laitos, suomen kieli.
Solki a = Yleiset kielitutkinnot – korpus. Jyväskylä: Soveltavan kielentutkimuksen keskus. http://yki-korpus.jyu.fi/etusivu.html (23.6.2014).
Solki b = DIALUKI: Toisen tai vieraan kielen lukemisen ja kirjoittamisen diagnosointi. Jyväskylä: Soveltavan kielentutkimuksen keskus. https://www.jyu.fi/hum/laitokset/solki/tutkimus/projektit/dialuki/su (28.5.2014).
Solki c = Motivaatiokysely. Jyväskylä: Soveltavan kielentutkimuksen keskus. https://www.jyu.fi/hum/laitokset/solki/tutkimus/projektit/dialuki/su/motivaatiokysely (30.5.2014).
Solki d = Tietoa yleisistä kielitutkinnoista. Jyväskylä: Soveltavan kielentutkimuksen keskus. https://www.jyu.fi/hum/laitokset/solki/yki/yleista/tietoakielitutkinnoista/index_html (15.5.2014).
Solki e = Kirjoittaminen. Jyväskylä: Soveltavan kielentutkimuksen keskus. https://www.jyu.fi/hum/laitokset/solki/yki/yleista/osat_aihealueet/kirjoittaminen (28.5.2014).
Spoelman, Marianne 2013: *Prior linguistic knowledge matters. The use of the partitive case in Finnish learner language.* Acta Universitatis Ouluensis B Humaniora 111. Oulu: Oulun yliopisto.
Tarnanen, Mirja 2007: Testiaineistosta kielenoppijakorpukseksi. – Olli-Pekka Salo, Tarja Nikula & Paula Kalaja (toim.), *Kieli oppimisessa. Language in learning* s. 197–213. *AFinLAn vuosikirja* 65. Jyväskylä: Suomen soveltavan kielitieteen yhdistys AFinLA ry.
Toivola, Sari – Tossavainen, Henna 2011: Opiskelijoiden käsityksiä yleisten kielitutkintojen korpuksen käyttömahdollisuuksista. – Esa Lehtinen, Sirkku Aaltonen, Merja Koskela, Elina Nevasaari & Mariann Skog-Södersved (toim.), *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 3 s. 158–169. http://ojs.tsv.fi/index.php/afinla/article/view/4466 (30.5.2014).
Topling. Toisen kielen oppimisen polut. https://www.jyu.fi/hum/laitokset/kielet/tutkimus/hankkeet/topling (30.5.2014).
Toropainen, Outi – Härmälä, Marita – Lahtinen, Sinikka 2012: Kaksi asteikkoa, kaksi eri tilannetta. Äidinkielellä ja vieraalla kielellä kirjoitettujen tekstien kriteeripohjaisen arvioinnin haasteita. – Lea Meriläinen, Leena Kolehmainen & Tommi Nieminen (toim.), *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 4 s. 60–79. http://ojs.tsv.fi/index.php/afinla/article/view/7038 (30.5.2014)
Ullakonoja, Riikka – Nieminen, Lea – Haapakangas, Eeva-Leena – Huhta, Ari – Alderson, Charles 2012: Kaksikieliset oppilaat suomea ja venäjää kirjoittamassa. Minun rakkaus väri – valeasininen ja violetti. – Lea Meriläinen, Leena Kolehmainen & Tommi Nieminen (toim.), *Monikielinen arki* s. 113–134. *AFinLAn vuosikirja* 70. Jyväskylä: Suomen soveltavan kielitieteen yhdistys AFinLA ry.

Kirjoittajien yhteystiedot:
Jarmo Harri Jantunen: etunimi.h.sukunimi@jyu.fi
Silja Pirkola: etunimi.t.sukunimi@student.jyu.fi