OPPIJANSUOMEN JA -VIRON SÄHKÖISET TUTKIMUSAINEISTOT

Korpuksiin perustuvalla kielenoppimisen ja oppijankielen tutkimuksella on jo kahdenkymmenen vuoden perinteet indoeurooppalaisissa kielissä: esimerkiksi oppijanenglannin korpusta (International Corpus of Learner English, Granger ym. 2002) on alettu kerätä 1990-luvun alussa ja oppijanruotsista (esim. Svenska som Målspråk, Borin 2002) ja -norjasta (Norsk andrespråkskorpus, Tenfjord ym. 2006) on myös muokattu sähköisiä aineistoja 1990- ja 2000-luvuilla. Muistakin kielistä aineistoja on olemassa jo runsaahkosti (ks. Barlow 2005; Granger 2004; Eslon ja Metslang 2007), ja nyttemmin niitä on myös itämerensuomalaisista kielistä: muutaman viime vuoden ajan oppijankielen sähköisiä tekstiaineistoja on kerätty sekä suomesta että virosta.

Tähän katsaukseen olemme koonneet yhteen sekä oppijansuomea että oppijanviroa koskevat korpukset ja niitä koostavat hankkeet. Päätös katsauksen tekemisestä syntyi Vaasassa XXXV Kielitieteen päivillä pidetyssä työpajassa Korpukset ja oppijankieli: nykytilanne ja tulevaisuuden haasteet (Jantunen ja Sulkala 2008). Työpajan tavoitteina oli esitellä tämänhetkisiä oppijansuomen ja -viron korpusaineistoja ja pohtia tutkimuskysymyksiä ja -metodologiaa sekä oppijankielen sähköisten aineistojen ja niiden analyysin haasteita. Yhtenä toiveena työpajan lopuksi esitettiin, että nykyisistä aineistoista laadittaisiin kokoava katsaus ja aineistot tehtäisiin näkyviksi tutkijoille. Tämä kirjoitus on vastaus toiveeseen. Katsauksessa esitettävät tiedot kerättiin sähköpostikyselyllä syksyllä 2008. Yksittäisten aineistojen kuvaukset ja tiedot ovat peräisin seuraavilta henkilöiltä: Jarmo Harri Jantunen ja Saana Piltonen (ICLFI), Mirja Tarnanen (Yleisten kielitutkintojen korpus), Kirsti Siitonen ja Ilmari Ivaska (Edistyneiden suomenoppijoiden korpus), Maisa Martin (CEFLING-korpus), Pille Eslon ja Annekatrin Kaivapalu (EVKK) sekä Kristiina Praakli ja Kadri Sõrmus (Tarton yliopiston korpukset).

KANSAINVÄLINEN OPPIJANSUOMEN KORPUS (ICLFI)

Yksi suomalaisista oppijankielen korpuksista on Kansainvälinen oppijansuomen korpus eli International Corpus of Learner Finnish (ICLFI). Sitä on kerätty Oulun yliopiston johtamassa hankkeessa Korpustutkimus oppijankielen kielikohtaisista ja universaaleista ominaisuuksista vuodesta 2007 lähtien. Hankkeessa ovat mukana Oulun yliopiston suomi toisena ja vieraana kielenä -oppiaineen lisäksi Tallinnan yliopiston Eesti keele ja kultuuri instituut, Uumajan yliopiston Institutionen för moderna språk sekä suomen kielen ja kirjallisuuden laitokset Karjalan valtion pedagogisesta yliopistosta ja Petroskoin valtionyliopistosta. Hankkeessa tutkitaan yhtäältä niitä piirteitä, ▷

jotka yhdistävät kielenoppimista olipa oppijan äidinkieli mikä tahansa, ja toisaalta niitä piirteitä, jotka ovat tyypillisiä tietylle äidinkieliselle tai kohdekieliselle ryhmälle. Hanketta ovat rahoittaneet Pohjoismaiden ministerineuvosto sekä Oulun ja Uumajan yliopistot, ja sitä johtaa Jarmo Harri Jantunen.

ICLFI:n tekstejä on kerätty ulkomaisten yliopistojen suomen kielen ja kulttuurin opetuspisteissä, eli korpus on suomi vieraana kielenä -aineisto. Tekstejä keräämiä yliopistoja on hankkeessa mukana 19, ja vuoden 2008 loppuun mennessä aineistoa on saatu 13 eri äidinkielestä. Kirjoittamishetkellä aineistoa on eniten virolaisilta suomenoppijoilta: 21 % kokonaissanemäärästä. Puolalaisten opiskelijoiden tekstien sanemäärä on 14 %, venäläisten 13 % ja kiinalaisten, ruotsalaisten sekä saksalaisten kunkin 11 % koko aineistosta. Kirjoituksia on saatu lisäksi Espanjasta, Hollannista, Islannista, Italiasta, Itävallasta, Slovakiasta ja Tšekistä. Aineiston koko on vuoden 2008 lopussa noin 300 000 sanetta, ja tekstejä siinä on yhteensä noin 1700. Tarkoituksena on koota usean miljoonan sanan korpus, joka edustaa laajasti erikielisten ja eritasoisten suomenoppijoiden kieltä.

Kaikkiin teksteihin lisätään runsaasti metatietoa tekstintuottajasta, opetuskontekstista sekä itse tekstistä. Tekstintuottajasta aineistoon koodataan muun muassa tiedot äidinkielestä, taitotasosta, sukupuolesta, syntymävuodesta ja -paikasta. Oppimiskontekstista koodataan muun muassa tiedot suomenopettajan ja vanhempien äidinkielestä, käytössä olevista oppikirjoista ja oleskelusta Suomessa. Tekstistä puolestaan käyvät ilmi tekstityyppi, tehtävänanto, kirjoituspaikka ja kirjoitusajan rajaaminen sekä tekstin tuottamisessa apuna käytetyt välineet, kuten sanakirjat ja oppikirjat. Joissakin paperilla tulleissa teksteissä on epäselviä tai puuttuvia sanoja ja merkkejä, jotka on digitoidessa korvattu puuttumista tai epäselvyyttä kuvaavilla koodeilla. Myös teksteissä olevat hymiöt, kuvat, kokonaiset nimet, tarkat syntymäajat, osoitteet ja puhelinnumerot on muutettu koodeiksi ennen tekstien siirtämistä korpukseen.

Korpuksen nykyvaiheessa kirjoitukset ovat tekstityypiltään pääasiassa kuvauksia, esseitä, päiväkirjoja ja kertomuksia; näiden yhteenlaskettu osuus sanemäärästä on noin 70 %. Lisäksi aineistossa on muun muassa arvosteluja ja referaatteja. Kirjoittajat on jaettu opiskellun tuntimäärän perusteella kolmeen taitotasoryhmään: alkeis- ja keskitasoon sekä edistyneisiin. Eniten aineistoa korpuksessa on alkeistason suomenoppijoilta: 45 % koko sanemäärästä. Aineistosta 32 % on keskitasolta, ja edistyneiden osuus on 23 %. Lukuun ottamatta taustatietoja aineisto on niin sanottua raakatekstiä, eli siihen ei ole lisätty lingvististä tietoa esimerkiksi morfologisista tai syntaktisista ominaisuuksista. Annotoimaton raakatekstikorpus ei siten ole etukäteen analysoitua eikä ole myöskään sidoksissa ennen korpuksia luotuihin kategorioihin (ks. mm. Sinclair 1991, 2004; Hunston 2002). Esimerkiksi oppijankielen virhekoodaus on osoittautunut osittain ongelmalliseksi virhetapausten tulkintamahdollisuuksien moninaisuuden vuoksi (ks. esim. Barlow 2005: 340–342).

Oppijansuomen korpukset tarjoavat uusia tutkimusmahdollisuuksia oppijankielen tutkimuksen alalla. Vertailumahdollisuus indoeurooppalaisten ja ei-indoeurooppalaisten kielten välillä on sinänsä tärkeää, mutta myös vertailu oppijanviroon (ks. jäljempänä viron aineistot) on erittäin mielenkiintoista lähisukukielen oppimisen näkökulmasta. Hankkeessa aineiston avulla tutkimustaan tekevät ovat kiinnostuneet muun muassa (oppijan)kielen fraseologisuudesta, eräänlaisesta elementtirakenteisuudesta. Esimerkiksi Sisko Brunni tutkii väitöskirjatyössään (Brunni 2008), miten

oppijankielessä hallitaan verbit fraseologisina yksikköinä. Jarmo H. Jantunen on puolestaan tarkastellut oppijankielen avainsanaisuutta korpusvetoisin menetelmin ja havainnut oppijankielen fraseologisten yksiköiden poikkeavan melkoisesti natiivikielestä (Jantunen 2007, 2008a). Suomen kielen morfosyntaktiset piirteet kiinnostavat myös kielenoppimisen korpustutkimuksessa: Marianne Spoelman käyttää ICLFI-korpusta väitöskirjatutkimuksessaan, joka käsittelee partitiivin käyttöä erikielisten tuottamassa oppijankielessä (Spoelman 2008). Arja Roth Uumajasta tarkastelee niin ikään partitiivia ja myös Uumajassa opettava Tuija Määttä tutkii suomen kielen paikallissijojen oppimista. Tallinnan yliopiston dosentti Annekatrin Kaivapalu puolestaan pohtii, miten lähdekielen vaikutusta on mahdollista tutkia korpusten avulla universaalina ilmiönä (Kaivapalu 2007). Tietoa ICLFI-korpuksesta ja hankkeesta saa lisää kotisivulta http://www.oulu.fi/hutk/sutvi/oppijankieli/index.html.

YLEISTEN KIELITUTKINTOJEN KORPUS

Yleisten kielitutkintojen eli YKI-korpusta kootaan Jyväskylän yliopistossa. Yleiset kielitutkinnot on kansallinen vieraan ja toisen kielen kielitaidon tutkinto, jossa voi suorittaa taitotasotestin yhdeksässä kielessä (englanti, espanja, italia, ranska, ruotsi, saame, saksa, suomi ja venäjä) ja kolmella tutkintotasolla (perus-, keski- ja ylin taso). Muissa kielissä kuin suomessa ja ruotsissa suurin osa tutkinnon suorittajista on äidinkieleltään joko suomen- tai ruotsinkielisiä. Suomen ja ruotsin kielen tutkinnon suorittajilla on puolestaan kymmeniä eri äidinkieliä. Jokaisen kielen ja taitotason tutkinto sisältää viisi osakoetta: puhuminen, puheen ymmärtäminen, kirjoittaminen, tekstin ymmärtäminen sekä rakenteet ja sanasto. Jokaisesta osataidosta tulee taitotasoarvio tutkintotodistukseen. Tutkinto on tarkoitettu aikuisille, ja sen voi suorittaa riippumatta siitä, miten ja missä kielitaito on hankittu. Kielitaitotodistuksia käytetään tavallisimmin työ- tai opiskelutarkoituksiin sekä kansalaisuuden hakemiseen. Tutkinnon suorittajia on vuoteen 2008 mennessä ollut yli 40 000.

Yleisten kielitutkintojen korpus on luonteeltaan karttuva eli siihen tallennetaan lisää aineistoa kunkin testikerran jälkeen. Korpukseen on jo tallennettu 19 354 suorittajan yleistasoarviot ja taustatiedot sekä 3 226 kirjoittamisen ja 767 puhumisen suoritusta eri kielistä. Korpuksessa on eniten aineistoa suomen kielestä, koska suomen kielen tutkinto on suoritetuin. Aineistoa on tähän mennessä käytetty pro gradujen ja väitöskirjojen aineistona sekä koulutusmateriaalina opettajien täydennyskoulutuksessa. Korpusta hyödynnetään myös Jyväskylän yliopiston CEFLING-projektissa, jota rahoittaa Suomen Akatemia ja jossa tutkitaan toisen ja vieraan kielen oppijan kielitaidon kehittymistä taitotasolta toiselle (ks. lisää http://www.jyu.fi/hum/laitokset/kielet/cefling).

Aineisto on mielenkiintoinen niin tutkimusmielessä kuin kielipoliittisestikin. Korpusta voidaan hyödyntää tutkimus- ja opetustarkoituksissa, koska aineisto tarjoaa monipuolisesti tietoa kanta- ja uussuomalaisten aikuisten kielitaidosta. Mielenkiintoista suomen kielen aineistossa on myös se, että tutkinnon suorittajien koulutustaustat ovat hyvin heterogeenisiä. Korpukseen kerätty aineisto sisältää testinsuorittajien saamat taitotasoarviot, taustatiedot (muun muassa äidinkieli, sukupuoli, sosioekonominen asema, kohdekielen käyttöyhteydet ja -taajuus) sekä puhumisen ja kirjoittamisen suorituksia.

Kvantitatiivisten ja kvalitatiivisten aineistojen yhdistäminen on mahdollista ▷

kunkin testinsuorittajan saaman id-numeron avulla. Kirjoittamisen suorituksia voi siis etsiä esimerkiksi taitotasoarvion perusteella tai vaikkapa testinsuorittajan äidinkielen perusteella. Aineistot ovat kuitenkin erisuuruisia muun muassa siksi, että korpukseen viedään testiin osallistuneiden kirjoittamisen ja puhumisen suorituksia vain niiltä, jotka ovat antaneet siihen luvan, ja siksi, että aivan kaikki suorittajat eivät ole täyttäneet taustatietolomaketta. Kvalitatiivisen aineiston tarkennetulla haulla korpuksesta voi etsiä esimerkiksi kirjoittamisen suorituksia tekstilajin mukaan. Koska käyttöliittymä mahdollistaa hakujen tekemisen monilla eri muuttujilla, kannattaa ennen käytön aloittamista tutustua ohjeisiin ja esimerkkihakuihin. Korpuksessa on suomen- ja englanninkielinen käyttöliittymä.

Korpus on saatavilla verkkotietokantana, ja sen käyttöä koordinoi Yhteiskuntatieteellinen tietoarkisto. Tietokannan käyttö on maksutonta, mutta siihen vaaditaan käyttäjätunnus ja salasana, jotka voi tilata Yhteiskuntatieteellisestä tietoarkistosta toimittamalla tietoarkistoon käyttölupahakemus ja käyttöehtositoumus (ks. lisää http://www.fsd.uta.fi/aineistot/luettelo/FSD2324/meF2324.html).

EDISTYNEIDEN SUOMENOPPIJOIDEN KORPUS

Turun yliopistossa perustettiin vuonna 2007 korpushanke tutkimaan edistyneiden suomenoppijoiden kieltä. Hanketta ovat tähän asti ideoineet sen johtaja Kirsti Siitonen sekä tutkimusavustaja Ilmari Ivaska, joka on myös koodannut ja työstänyt materiaalia. Materiaalia on lisäksi työstänyt harjoittelijana Johanna Tiuraniemi. Hanke pyrkii osaltaan lisäämään tietoa suomesta toisena kielenä akateemisessa käytössä. Samalla se täydentää muiden oppijansuomen korpushankkeiden antamaa kuvaa oppijansuomen variaatioista.

Edistyneiden suomenoppijoiden korpuksen aineisto koostuu ensisijaisesti Turun yliopistossa suomen ja sen sukukielten maisteriohjelmassa opiskelevien ja opiskelleiden kirjallisista tuotoksista. Korpuksessa on useita tekstilajeja 2–3 vuoden ajalta informantista riippuen. Tämä mahdollistaa pitkittäistutkimukset ja välikielen dynamiikan tarkastelemisen. Tärkeimmät tekstilajit ovat opiskelijoiden tenttivastaukset, esseet ja tutkielmat. (LAS2.) Korpuksen yhteyteen koostetaan myös vertailuaineisto ensikieleltään suomalaisten opiskelijoiden teksteistä. Kuten Kalliokoski toteaa, tekstilaji on ennen kaikkea sosiokulttuurinen käsite (Kalliokoski 2006: 240). Tekstilajin hallinta on stereotyyppistä tietoa kielenkäyttäjän tavasta toimia kussakin tilanteessa eli se on yksi osa kielellistä sujuvuutta, jonka keskeisenä tekijänä on juuri variaatio (mts. 248). Näin ollen soveltuva vertailuaineisto on tutkimuksen kannalta hyvin tärkeää.

Tällä hetkellä osa digitoidusta aineistosta on jo koodattu ja koodaaminen etenee jatkuvasti. Koodaus toteutetaan Turun yliopiston Lauseopin arkiston mallin mukaisesti; teknisesti tämä tarkoittaa html-kielen kaltaisten tagien käyttöä, joiden avulla aineisto rakennetaan hierarkkisesti. Morfologinen koodaus tehdään jokaista sananmuotoa kohti vain kerran, ja kielen toisteisuus keventää koodausta aineiston karttuessa. Koodaus on Kotuksessa kehitetyn TEI-ohjeistuksen (The Text Encoding Initiative) muunnoksen mukainen (ks. esim. Inaba 2007: 151).

Koodatusta aineistosta ilmenee sanakohtaisesti sanaluokka, morfologinen status sekä sanan funktio lauseessa. Lisäksi koodattu aineisto on hakusanoitettu ja jaettu rakenneyksiköihin sana- ja lausetasosta alkaen. Syntaktista koodausta tehtäessä aineistossa esiintyvät virheet on myös kom-

mentoitu ja kommentoinnin perusteella on muodostettu virhetyypittely. Tämä metodi takaa aineistolähtöisen luokituksen. Virhekoodaus on kuitenkin aina subjektiivista tulkintaa, ja sen tärkeintä antia on äidinkielisen kielenkäyttäjän intuitiosta poikkeavien ratkaisujen havaitseminen. Korpus antaa kuvan oppijansuomen ominaisluonteesta ja sen suhteesta äidinkielisesti käytettyyn suomeen myös silloin, kun käytetty kieli on oikeakielisyysnormien mukaista. Koodaus mahdollistaa oppijankielen variaation ja sen muutoksen tarkastelun niin syntaktiselta, morfologiselta kuin leksikaaliseltakin kannalta, minkä lisäksi se tarjoaa välineitä kohdekielen normeista poikkeavien välikielten ja niiden muutoksen tarkasteluun.

Korpuksen käyttöliittymä tulee perustumaan Turun yliopiston Lauseopin X-arkiston käyttöliittymän malliin.[1] Tässä verkkoselaimella käytettävässä käyttöliittymässä haut voidaan kohdistaa leksikaalisiin, morfologisiin tai syntaktisiin ominaisuuksiin tai näiden ominaisuuksien yhdistelmiin. Tietyt seikat kieltävien hakuehtojen käyttö ja sanajärjestyksen huomioiminen on niin ikään mahdollista. (LA.)

Aineistoa kartutetaan jatkuvasti niin määrällisesti kuin laadullisestikin. Nykyisellä aineistonkeräämismetodilla korpuksen tenttivastausten ja esseiden vuotuinen kertymä on 15 000–20 000 saneen luokkaa kummassakin osiossa, tutkielmien osalta se on noin 60 000 sanetta.

CEFLING-KORPUS

CEFLING on Suomen Akatemian rahoittama tutkimushanke Jyväskylän yliopistossa. Tutkimushankkeen tavoitteena on tutkia Kielten oppimisen, opettamisen ja arvioinnin yhteisen eurooppalaisen viitekehyksen (Common European Framework of Reference for Languages, CEFR) kuvaamia kielitaitotasoja ja näillä tasoilla esiintyviä oppijakielen piirteitä suomen ja englannin kielissä. Hankkeessa tarkastellaan, miten viitekehyksessä esitetyt »can do» -tyyppiset taitotasokuvaukset suhtautuvat aitojen kirjoitustehtävien oppijakielen piirteisiin (ks. lisää http://www.jyu.fi/hum/laitokset/kielet/cefling).

Hankkeessa on kerätty kirjoitelmia yläkoululaisilta suomi toisena kielenä -oppijoilta ympäri Suomen. Tekstilajeja on neljä: epämuodollinen sähköpostiviesti, muodollinen sähköpostiviesti, mielipide ja kertomus. Suomen kielen korpus sisältää 893 kirjoitelmaa ja oppilaiden taustatiedot. Kirjoitelmat on kirjoitettu tekstitiedostoiksi ja koodattu .chat-tiedostoiksi käyttäen CHILDES Child Language Data Exchange System -ohjelmaa (http://childes.psy.cmu.edu/), mikä mahdollistaa osittain koneellisen analyysin. Teksteistä on koodattu morfosyntaktisin koodein paikallissijat, objektit, *olla*-verbi, infinitiivit, konditionaalit ja potentiaalit, kiellot, relatiivilauseet, passiivilauseet ja muut nominisubjektittomat lauseet.

Aineistosta tutkitaan parhaillaan muun muassa taivutusilmiöitä, relatiivilauseita ja negaatiota (Maisa Martin), lausetyyppejä, erityisesti transitiivisuutta (Nina Reiman), paikallissijojen käyttöä (Sanna Mustonen) ja yleistämisen keinoja, kuten passiivia ja geneerisiä lauseita (Marja Seilonen). Pro gradu -tutkielmia on tehty esimerkiksi *olla*-verbistä ja verbiketjuista (ks. lisää http://www.jyu.fi/hum/laitokset/kielet/cefling/en/Subprojects).

Hankkeessa myös verrataan yläkoululaisten kirjoitelmien kieltä aikuisten samaa taitotasoa edustavien kirjoitelmien kieleen.

[1] Mallin on suunnitellut Lauseopin arkiston tutkija Nobufumi Inaba.

Aikuisten korpus on koottu YKI-aineistosta. Tästä syystä tehtävät ovat pitkälti hyvin samankaltaisia.

Kansainvälisestä oppijansuomen korpuksesta (ICLFI) tämä oppijankirjoitelmien korpus eroaa siten, että kyseessä ovat nuoret suomi toisena kielenä -oppijat, jotka siis pääosin oppivat suomea elinympäristöstään, vaikka saavat myös opetusta. Koululaisille kirjoittaminen ei ole vain osa suomen kielen opetusta, vaan he kirjoittavat kaikissa oppiaineissa suomeksi, eivät äidinkielellään. Korpus on myös koodattu, vaikka tekstiversioita on tietenkin myös mahdollista käsitellä sellaisilla korpustyökaluilla, jotka eivät vaadi ennakkokoodausta.

Peruskoulussa tavoitteena ovat B-tason kielitaidot, joten CEFLING-aineistossa on luonnollisesti hyvin vähän C-tason suorituksia. Turun Edistyneiden suomenoppijoiden korpus taas edustaa akateemisen koulutuksen saaneiden suomenoppijoiden kielitaitoa, joten myös nämä korpukset täydentävät toisiaan.

Poikkeukselliseksi CEFLING-korpuksen tekee kansainvälisestikin se, että jokainen suoritus on arvioitu ja peruskorpukseen on hyväksytty vain ne suoritukset, joista kolme arvioijaa ovat yksimielisiä tai vain yksi heistä poikkeaa kahdesta muusta yhden taitotasoportaan verran. CEFR-asteikon lisäksi suoritukset on arvioitu myös opetussuunnitelmien perusteiden liitteenä olevan taitotasokriteeristön mukaisesti, joten aineisto mahdollistaa myös näiden asteikkojen vertailun.

VIRON VÄLIKIELEN KORPUS (EVKK)

Viron välikielen korpus (Eesti vahekeele korpus; http://evkk.tlu.ee) perustettiin Tallinnan yliopiston yleisen ja soveltavan kielitieteen laitoksessa vuonna 2004. Nykyään korpushanke toimii Viron kielen ja kulttuurin instituutin yhteydessä. Viron välikielen korpuksen kehittäminen on pohjautunut seuraaviin hankkeisiin: Eesti keelekeskkonna arengu analüüs, modelleerimine ja juhtimine (2003–2007), Koodivahetuse, vahe- ja lastekeele korpuste töötlemine ja haldamine (2005–2008), Koodivahetuse, eesti vahekeele ning lastekeele andmekorpuste koostamine ja üldkirjeldus (2005–2008) sekä Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine (2008–2010).

Viron välikielen korpus on verkkopohjainen kaikille avoin korpus, jonka tavoitteena on sekä tutkimustyö että viro toisena ja vieraana kielenä -opetuksen kehittäminen. Käyttöliittymä mahdollistaa korpuksen vapaan käytön Linux-ympäristössä; korpuksessa on konkordansseri, sana- ja muotofrekvenssin tilastot sekä lingvistinen virhetypologia. Tällä hetkellä korpuksessa on yli 700 000 sanetta, joista 500 000 on manuaalisesti koodattu lingvistisen virhetypologian mukaan (esimerkiksi leksikaaliset, leksikaalis-kieliopilliset, morfofonologiset, syntaktiset ja kommunikatiiviset virheet). Jokainen virhetyyppi jakaantuu hierarkkisesti alatyyppeihin.

EVKK oli alun perin viro toisena kielenä -korpus, venäjänkielisten vironoppijoiden kirjallisten tekstien kokoelma. Viime aikoina korpukseen on kuitenkin koottu myös eri lähdekielisten informanttien viro vieraana kielenä -aineistoa: suomen-, saksan-, englannin- ja unkarinkielisten vironoppijoiden tekstejä. Tekstilajeista korpuksessa on eniten esseitä, mutta myös referaatteja, henkilökohtaisia ja virkakirjeitä sekä artikkelianalyysejä; lisäksi aineistossa on harjoituksia ja käännöksiä. Korpuksessa on taustatiedot oppijoista (sukupuoli, ikä, äidinkieli, kotikieli, asuinalue, sosiaalinen tausta, koulutus, ammatti, viron kielen taitotaso) ja teksteistä (tekstilaji, pituus, sanojen ja lauseiden määrä), mutta myös tekstin käsittelijästä ja koodaajasta. Taus-

tatietoja on mahdollista hakea käyttöliittymän avulla.

Korpushankkeen puitteissa tehdään yhteistyötä Tallinnan yliopiston koulutusteknologiakeskuksen sekä Tarton yliopiston ja Viron kielen instituutin kieliteknologien kanssa. Kaksi vuotta sitten sai alkunsa tiivis yhteistyö Oulun yliopiston korpushankkeen kanssa. EVKK-hankkeessa on tällä hetkellä kaksi kehityssuuntausta: korpuksen laajentaminen viiteen miljoonaan sanaan asti ja kieliteknologinen kehitystyö. Tavoitteena on uusien osakorpusten lisääminen (valtakunnallisten viro toisena kielenä -tasokokeiden suoritukset, kielikylpyaineisto), standardisoitujen ohjelmien, tilastollisten menetelmien ja viron kieliteknologisten ohjelmien käyttäminen aineiston analyysissa, virhe-etsijän prototyypin kehittäminen ja metodin kehittäminen eri kielivarianttien vertailua varten. Korpusaineistoon pohjautuvan tutkimustyön päätarkoituksena on oppijoiden kielenkäytön leksikaalis-kieliopillisten mallien kuvaaminen sekä oppijankielen kielikohtaisten ja universaalien piirteiden selvittäminen. Tutkimuksen jatkuvan kehittämisen edellytyksenä on myös oppijanviron ja oppijansuomen korpusten synkronointi.

TARTON YLIOPISTON OPPIJANKIELEN KORPUKSET

Tarton yliopistossa on koottu kolmea korpusta, jotka on tarkoitus yhdistää yhdeksi viron oppijankielen korpukseksi vuonna 2010. Ensimmäinen niistä on Viron kielen rinnakkaiskorpus, jota työstivät vuosina 2006–2007 Raili Pool, Elle Vaimann ja Ingrid Rummo. Korpus on koottu yliopisto-opiskelijoiden kirjallisissa töissä esiintyvistä virheellisistä lauseista, joiden rinnalle on lisätty lauseiden korjatut versiot. Kokonaisia tekstejä samassa yhteydessä ei ole, mutta korjattuja lauseita voi olla yhtä virheellistä lausetta kohden useampi. Helmikuuhun 2008 mennessä korpuksessa on 9000 virheellistä lausetta ja 128 000 sanaa. Korjattuja lauseita on 9100, joissa sanoja on yhteensä 129 000. Jokaisen virheellisen lauseen yhteyteen on lisätty tiedot kirjoittajan sukupuolesta, kansalaisuudesta, äidinkielestä, asuinpaikasta ja kielenosaamisen tasosta. Lauseiden kielivirheitä ei ole nimetty, mutta uusi työryhmä työskentelee parhaillaan rinnakkaiskorpuksen materiaalin kontrolloimisen, analysoinnin ja täydentämisen parissa.

Toinen Tarton yliopistossa koottu korpus on Viron kielen tekstikorpus, joka koostuu yliopisto-opiskelijoiden kotiaineista, koeteksteistä, sähköpostiviesteistä, opinnäytteiden johdannoista ja tiivistelmistä sekä oppimispäiväkirjoista. Maaliskuusta 2008 lähtien Neeme Kahusk, Kristiina Praakli ja Kadri Sõrmus ovat analysoineet ja täydentäneet tekstikorpusta, mutta koska työ on kesken, ei korpuksen sane- tai lausemäärä ole vielä tiedossa.

Tartossa on koottu myös äidinkielisten oppijoiden korpusta. Näiden kolmen korpuksen yhdistämisen jälkeen yhdessä yhteisessä korpuksessa olisi siis materiaalia niin viroa toisena kielenä kuin äidinkielenäkin oppivilta kielenkäyttäjiltä. Tähän mennessä oppijankielen korpusten aineistosta on valmistunut yksi väitöskirja (Pool 2007), yksi pro gradu -tutkielma (Sõrmus 2008) sekä useita artikkeleita.

LOPUKSI

Oppijakorpusten kokoaminen ei-indoeurooppalaisista kielistä on erittäin tärkeää. Yhtäältä aineistot täydentävät kieliresursseja, jotka ovat voittopuolisesti indoeurooppalaisista kielistä. Tätä on pidettävä jo sinänsä tärkeänä tehtävänä kielen kuvauksen ja muuttumisen näkökulmasta. Toisaalta kielenoppimisen, sen haasteiden ▷

ja oppijankielen kuvauksen lähtökohdista on luonnollisesti olennaista, että aineistoja kerätään ja tutkimusta harjoitetaan viljalti myös ei-indoeurooppalaisista kielistä. Tässä tehtävässä on yllä kuvatuilla korpuksilla merkittävä rooli: niiden avulla tehtävä tutkimus voi problematisoida nyt olemassa olevan tutkimuksen tai täydentää sitä lisäämällä tietoa, joka ei liity valtakielten oppimiseen. Näin voidaan myös selvittää piirteitä, jotka ovat yleisiä kielenoppimisessa olipa oppijan äidinkieli tai opittava kohdekieli mikä tahansa. Yksi tällainen kaikkea kielenoppimista ja tuotosta yhdistävä eli ns. oppijankielen universaali piirre lienee epätyypilliset fraseologiset (kontekstuaaliset) rakenteet, jotka ilmenevät kohdekielestä poikkeavina leksikaalis-kieliopillisina ja leksikaalis-semanttisina myötäesiintyminä (aiheesta enemmän Jantunen 2007, 2008b). Uusien aineistojen myötä myös tutkimusmetodologia on monipuolistunut, kun jo melkeinpä perinteeksi muodostunut virheanalyysiin pohjautuva tutkimus (ks. esim. Dagneaux ym. 1998; Barlow 2005) on saamassa rinnalleen deskriptiivisen, korpusvetoisuutta painottavan tutkimusgenren (ks. Kallioranta 2009; Jantunen 2007, 2008a).

Suomessa ja Virossa on sähköisten oppijankielen aineistojen koonti ja analysointi vahvasti käynnissä ja uusi tutkimusala luo yhä tarkempaa kuvaa suomesta ja virosta; samalla eri kielikuntiin kuuluvia oppijankieliä ja eri äidinkielisten puhujien tuottamia kielivariantteja tarkastelemalla päästään yhä tarkempaan kuvaukseen (oppijan)kielen luonteesta. Näin saadaan tietoa myös sovellettavaksi oppimateriaalien ja sanakirjojen tekemisen sekä opetustyön avuksi.

Tutkimusyhteistyö on jo alkanut vilkkaasti lähisukukielten tutkijoiden kesken. Tästä ovat osoituksena lukuisat parin viime vuoden aikana pidetyt yhteiset seminaarit ja työpajat: Tallinnan yliopistossa on järjestetty kolme oppijankielen korpustutkimuksen teemaseminaaria vuosina 2007–2008, Joensuun Virsu-konferenssi 2007 sekä Kielitieteen päivät Oulussa 2007 ja Vaasassa 2008 ovat puolestaan nähneet aiheen ympärille kootut kansainväliset työpajansa. Tutkimusyhteistyö on ollut alusta alkaen erittäin innostunutta. Pitkälti yhteisten kysymysten ja tavoitteiden vuoksi on järkevää jatkaa ja syventää edelleen tutkimusta sekä suomalaisten että suomalaisten ja virolaisten hankkeiden kesken. Aineistoista ja harjoitettavasta tutkimuksesta tiedottaminen onkin kaiken tämän perusta. ■

JARMO HARRI JANTUNEN
etunimi.sukunimi@oulu.fi
SAANA PILTONEN
saanamoi[– –]@paju.oulu.fi

LÄHTEET

BARLOW, MICHAEL 2005: Computer-based analyses of learner language. – Rod Ellis & Gary Barkhuizen (toim.), *Analysing learner language* s. 335–357. Oxford: Oxford University Press.

BORIN, LARS 2002: CrossCheck project status report. –http://www.csc.kth.se/tcs/projects/xcheck/lagesrapport-021015.pdf. 14.12.2008.

BRUNNI, SISKO 2008: Kollokaatiot, kolligaatiot ja semanttiset myötäesiintymät verbirakenteiden opettamisen apuna. – Esitelmä työpajassa Descriptive corpus study on learner language and co-operation between EVKK and ICLFI, 2.–4. lokakuuta 2008, Helsinki.

Cefling-projektin kotisivut. http://www.jyu.fi/hum/laitokset/kielet/cefling. 10.12.2008.

DAGNEAUX, ESTELLE – DENNESS, SHARON – GRANGER, SYLVIANE 1998: Computer-aided error analysis. – *System: An*

international journal of educational technology and applied linguistics 26:2 s.163–174.

Eslon, Pille – Metslang, Helena 2007: Õppijakeel ja eesti vahekeele korpus. – *Eesti rakenduslingvistika ühigu aastaraamat 3. Estonian papers in applied linguistics* s. 99–116. Tallinn: Eesti keele sihtasutus.

Granger, Sylviane 2004: Computer learner corpus research: Current status and future prospects. – Ulla Connor & Thomas Upton (toim.), *Applied corpus linguistics: A multidimensional perspective* s. 123–145. Amsterdam: Rodopi.

Granger, Sylviane – Dagneaux, Estelle – Meunier, Fanny 2002: *International Corpus of Learner English. Handbook and CD-ROM*. Centre for English Corpus Linguistics. Presses Universitaires de Louvain.

Hunston, Susan 2002: *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Iclfi-hankkeen kotisivut. http://www.oulu.fi/hutk/sutvi/oppijankieli/index.html. 10.12.2008.

Inaba, Nobufumi 2007: Mikael Agricolan teokset tietokannan muodossa. – Kaisa Häkkinen & Tanja Vaittinen (toim.), *Agricolan aika* s. 147–162. Helsinki: BTJ-kustannus.

Jantunen, Jarmo H. 2007: Oppijansuomen piirteitä korpusvetoisesti. – Pirkko Muikku-Werner, Ossi Kokko & Hannu Remes (toim.), *Virsu 3. Suomalais-ugrilaisia kohdekieliä ja kontakteja* s. 69-83. Studies in Languages 42. Joensuu: Joensuun yliopisto.

—— 2008a: Corpus-driven analysis of co-textual units of meaning in learner language. Esitelmä konferenssissa New trends in corpus linguistics for language teaching and translation studies. In honour of John Sinclair. Granada 22.–24. syyskuuta 2008.

—— 2008b: Haasteita oppijankielen korpusanalyysille: Oppijankielen universaalit. – Pille Eslon (toim.), *Õppijakeele analüüs: võimalused, probleemid, vajadused* s. 67–92. Tallinna ülikooli eesti filoloogia osakonna toimetised 10.

Jantunen, Jarmo h. – Sulkala, Helena 2008: Korpukset ja oppijankieli: nykytilanne ja tulevaisuuden haasteet. Työpajan kuvaus. – *XXXV Kielitieteen päivät Vaasan yliopistossa 23.–24.5.2008. Tiivistelmävihko* s. 102. Vaasan yliopisto.

Kaivapalu, Annekatrin 2007: Lähdekielen vaikutuksen tutkimus korpusten pohjalta. Esitelmä XXXIV Kielitieteen päivillä Oulussa 24.–25. toukokuuta 2007.

Kalliokoski, Jyrki 2006: Tekstilajin taju ja toisella kielellä kirjoittaminen. – Anne Mäntynen, Susanna Shore & Anna Solin (toim.), *Genre – tekstilaji* s. 240–265. Tietolipas 213. Helsinki: Suomalaisen Kirjallisuuden Seura.

Kallioranta, Otto 20090: *Paljon*-adverbin kollokointi oppijansuomessa: korpusvetoinen tutkimus. Pro gradu -tutkielma. Oulun yliopiston suomen kielen oppiaine. http://oulu.fi/hutk/sutvi/oppijankieli/tutkimus/.

La = Lauseopin arkiston verkkosivut. – http://olaui.suo.utu.fi/~lauseopin_arkisto/ 20.11.2008.

Las2 = Edistyneiden suomenoppijoiden korpuksen verkkosivut – http://www.hum.utu.fi/oppiaineet/suomi/tutkimus/tutkimushankkeet/Tutkimushanke_Siitonen_Ivaska.html 24.11.2008.

Pool, Raili 2007: Eesti keele teise keelena omandamise seaduspärasusi täis- ja osasihitise näitel. Doktoritöö. Tartu: ▷