doi:10.5128/LV21.04

Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi

JARMO HARRI JANTUNEN
Oulun yliopisto

Tiivistelmä. Katsauksessa esitellään Kansainvälinen oppijansuomen korpus eli ICLFI (*International Corpus of Learner Finnish*). Korpus on yksi kolmesta suomalaisesta oppijankielen tutkimusaineistosta, ja sitä on koottu Oulun yliopistossa viiden yliopiston yhteishankkeessa vuodesta 2007 lähtien. Aineisto on suomi vieraana kielenä -materiaalia, sillä siihen kerätään suomen kielen opiskelijoiden tuottamia tekstejä yli 20 ulkomaisesta yliopistosta opetushenkilökunnan avustuksella. Katsauksessa esitellään ICLFI:n rakenne, koostamisperiaatteet ja taustamuuttujat. Lisäksi käsitellään myös korpuksen annotointia ja taitotasojaottelun periaatteita. Katsaus esittelee myös korpustypologian, jonka avulla oppijankielen sähköisiä tutkimusaineistoja voidaan kuvata ja luokitella.

Avainsanat: oppijankieli; korpus; korpustypologia; annotointi; taustamuuttujat

1. Johdanto

Se, mistä kielistä sähköisiä tutkimusaineistoja luodaan, heijastelee monia seikkoja. Kielentutkijoiden omaksumat teoriat, lähestymistavat ja menetelmät ja se, mitkä suuntaukset ovat saaneet prestiisiaseman tutkimusyhteisössä, ohjaavat aineistojen kokoamista. Olennaista on myös se, miten

nämä traditiot muuttuvat ja millaista variaatiota niissä on, millaisen aseman kieli on saanut tutkijoiden kiinnostuksen kohteena ja millainen on kielen asema käyttökielenä maailmanlaajuisesti. Lisäksi ratkaisevia seikkoja ovat mm. kääntämisen määrä kyseiselle kielelle ja myös se, miten paljon kieltä opiskellaan maailmassa. Edellä mainitut tekijät ovat luonnollisesti kytköksissä toisiinsa: laajasti käytettyä prestiisikieltä myös opiskellaan paljon, se on taajaan käännösten kohdekielenä ja sen tutkimusta harjoitetaan viljalti ja monipuolisesti.

Ei siten ole yllättävää, että maailmanlaajuisesti isoimmat korpukset onkin tehty englannista (ks. esim. Kennedy 1998; Lee 2010); asiantilan selittävät monet yllä mainituista seikoista. Myös muista indoeurooppalaisista kielistä on hyvin laajoja sähköisiä tekstiaineistoja, samoin lukuisista ei-indoeurooppalaisista kielistä. Suomalais-ugrilaisista kielistä laajoja korpusaineistoja on koottu etenkin suomesta, virosta ja unkarista (muista ks. mm. Suihkonen 2007; Antonsen ym. 2006).

Korpukset voivat olla kieltä laajasti kuvaavia yleiskorpuksia tai kapeamman kuvan antavia erikoiskorpuksia. Oman ryhmänsä kielikorpusten joukossa muodostavat oppijankielen erikoiskorpukset. Université Catholique de Louvainin korpustutkimuksen keskuksen (*Centre for English Corpus Linguistics*) "korpuskartaston" (Goossens & Granger 2011) avulla laskettuna maailmassa on tällä hetkellä runsaasti yli 100 oppijankielen korpusta, joskin on luultavaa, etteivät kaikki oppijankielen korpusaineistot ole tällä listalla. Näistä huomattavan osan muodostavat oppijanenglannin korpukset, joita on luettelossa yli 60. Lisäksi oppijanranskasta (11), -saksasta (8) ja -espanjasta (6) on useampia aineistoja. Korpuksista suurin ja tähän mennessä luultavasti myös eniten käytetty on Louvainin yliopistossa koottu *International Corpus of Learner English* (ICLE, Granger ym. 2002; 2009). Aineisto on nykyisellään 3,7 miljoonan saneen laajuinen, ja siinä on 16 osakorpusta äidinkielittäin. Seuraavassa katsauksessa esitellään Oulun yliopistossa koottava Kansainvälinen oppijansuomen korpus eli ICLFI (*International Corpus of Learner Finnish*) korpustypologian, taustamuuttujien ja annotaation näkökulmasta.

2. ICLFI tunnuslukuina

Oppijansuomea sisältäviä laajoja sähköisiä tekstiaineistoja on nykyisellään kolme: Oulussa koottava Kansainvälinen oppijansuomen korpus, Jyväskylän Yleisten kielitutkintojen YKI-korpus ja Turun Edistyneiden suomenoppijoiden korpus. Korpukset poikkeavat kokoamisperiaatteiltaan toisistaan ja myös täydentävät toisiaan. YKI-korpus on määriteltävissä suomi toisena kielenä -korpukseksi, samoin myös Turun korpus; ICLFI on puolestaan suomi vieraana kielenä -korpus. Turun aineisto sisältää edistyneiden suomen opiskelijoiden akateemisia tekstejä, YKI ja ICLFI puolestaan laajemman kirjon erityyppisiä tekstejä eri taitotasoilta. (Jantunen & Piltonen 2009.) Aineistot luovat mahdollisuuden verrata toisiinsa esimerkiksi toisena ja vieraana kielenä -opiskelijoiden tuottamien tekstien ominaispiirteitä, mikä on toistaiseksi yksi vähälle huomiolle jäänyt tutkimuskohde siinä missä pitkittäistutkimuskin (ks. Granger 2004; 2007; 2010).

Taulukkoon 1 on kuvattu ICLFI erilaisten tunnuslukujen ja -piirteiden avulla. Vertailun vuoksi rinnalla on esitetty ICLE:n vastaavat tiedot.

ICLFI on koostettu hierarkkisesti osakorpuksiin, joihin on sisällytetty tekstit esimerkiksi äidinkielen, taitotason ja tekstilajin perusteella. Osakorpushierarkia on kuvattu kuviossa 1.

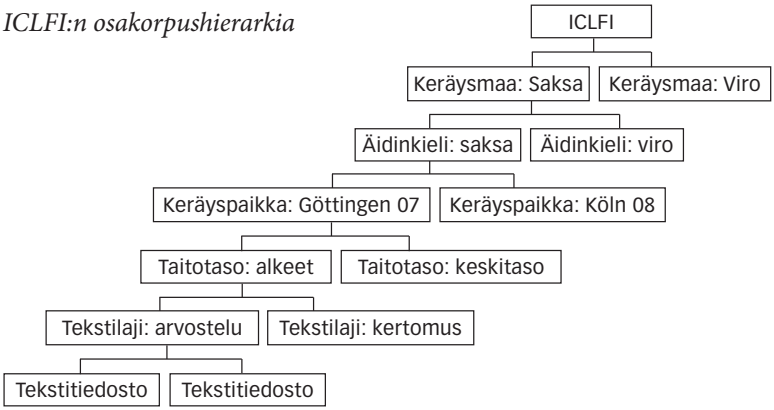
3. Korpustypologia

Seuraavaksi esitettävä kuvaus ICLFI:n typologiasta perustuu Atkinsin ym. (1992), Laviosa-Braithwaiten (1996), Grangerin (2007) ja soveltaen Lehtisen ym. (1995) kuvauksiin korpusten luokittelumahdollisuuksista. Typologialuokitteluun on otettu mukaan vain ne dimensiot, jotka ovat tarpeellisia ICLFI:n kuvauksessa – myös muita dimensioita on olemassa, esimerkiksi käännöskorpuksille. Luokittelussa on aluksi esitetty korpusten yleistä typologiaa koskevia dimensioita ja lopuksi sellaisia ominaisuuksia, jotka koskevat erityisesti oppijankieliaineistoja.

Taulukko 1. *ICLFI ja ICLE tunnuslukuina ja -piirteinä (tilanne 9/2011)*

	ICLFI	ICLE
Koko	1,0 miljoonaa sanetta, 5400 tekstiä	3,7 miljoonaa sanetta, 6 085 tekstiä
Koostaminen aloitettu	2007	1990
Äidinkieliä	22	16
Keräystapa	sähköinen ja manuaalinen	sähköinen
Tekstilajit	useita fiktiivisiä ja ei-fiktiivisiä tekstilajeja	argumentatiivinen essee
Tehtävänanto	opettajan määrittelemä	tutkimusryhmän määrittelemä
Tehtävän suoritus	opetukseen kuuluva harjoitustyö tai koe	korpusta varten
Taitotaso	alkeistaso, keskitaso ja edistyneet	ylempi keskitaso ja edistyneet
Annotointi	sanaluokka-annotointi osittain, morfosyntaktinen annotointi osittain	sanaluokka-annotointi kokonaan, virheannotointi kokonaan
Lemmatisointi	osittain	kokonaan

Kuvio 1.
ICLFI:n osakorpushierarkia



GENREDIMENSIO: yksitekstilajinen vs. monitekstilajinen korpus

ICLFI on monitekstilajinen korpus, sillä se sisältää tekstejä useista fiktiivisiä ja ei-fiktiivisiä tekstilajeja edustavista genrekategorioista. Monilajinen korpus antaa kattavamman kuvan kielivariantista kuin yksilajinen korpus, sillä jälkimmäisestä saadut tulokset voivat olla perua aineiston sisältämästä genrestä, eivätkä ole siten sinänsä oppijankielen ominaispiirre (ks. Barlow 2005). Toisaalta laaja kirjo eri tekstilajeja luo haasteita tutkimukselle, jos osakorpusten tekstilajit ovat vertailtavuuden näkökulmasta liikaa toisistaan poikkeavia.

TEEMADIMENSIO: yleiskorpus vs. terminologinen korpus

ICLFI on yleiskorpus, sillä sen sisältämien tekstien aihepiiriä ei ole rajattu toisin kuin terminologisen korpuksen, joka edustaa tietyn alan, esim. lääketieteen tekstejä.

REKISTERIDIMENSIO: kirjoitetun vs. puhutun kielen korpus

ICLFI:in on kerätty vain kirjoittamalla tuotettuja tekstejä. Vaihtoehtoisesti rekisteridimensiolla voidaan määritellä myös kielenkäyttötilanteen mukaan koostettua aineistoa, esimerkiksi asiakaspalvelutilanteista koostettua aineistoa.

KIELIDIMENSIO: yksikielinen, kaksikielinen (rinnakkais-) vs. monikielinen korpus

ICLFI on yksikielinen, vain suomeksi tuotettuja tekstejä sisältävä korpus.

VARIANTTIDIMENSIO: yksivarianttinen vs. verrannollinen korpus

Verrannollisuudella voidaan tarkoittaa joko ns. alkuperäiskielen (tai natiivikielen) ja sille verrannollisen kielivariantin sisältymistä samaan korpukseen.[1] Toisaalta se voi merkitä myös aineistoa, joka sisältää osakorpuksia, jotka varioivat muuttujien suhteen. ICLFI on verrannollinen jälkimmäi-

[1] Korpuspohjaisessa käännöskielen tutkimuksessa verrannollisuus on tarkoittanut alun perin jollakin kielellä kirjoitettujen tekstien ja tälle kielelle käännettyjen tekstien osakorpusten sisältymistä korpukseen (esim. alkuperäis- ja käännössuomi).

sen määritelmän perusteella, koska se sisältää toisilleen verrannollisia osakorpuksia, tärkeimpänä verrannollisuuden kriteerinä oppijoiden äidinkieli. Se ei kuitenkaan sisällä oppijankieliaineistolle verrannollisia natiivituotoksia.

KÄÄNNÖSDIMENSIO: ei-käännöskorpus (suoraan jollekin kielelle kirjoitettu) vs. käännöskorpus

ICLFI on ei-käännöskorpus, koska siihen ei ole otettu mukaan opiskelijoiden tuottamia käännösharjoituksia eli se ei sisällä varsinaisia käännöstekstejä. Tämä on tutkimuksen kannalta tärkeä valinta, koska itse käännösprosessi voi tuottaa kieleen omanlaisiaan piirteitä (ks. Mauranen & Kujamäki 2004), joiden erottaminen oppimisprosessin ilmiöistä olisi erittäin hankalaa. Luonnollisestikaan kaikkea käännösvaikutusta ei voitane kuitenkaan koskaan oppijankieliaineistoista poistaa, sillä onhan oman kielen ilmausten ja rakenteiden "kääntäminen" yksi tapa tuottaa kohdekielen ilmauksia etenkin opiskelun alkutaipaleella.

AIKADIMENSIO: synkroninen vs. diakroninen korpus

ICLFI on synkroninen korpus. Joistakin opetuspisteistä korpukseen kerätään kuitenkin tekstejä samoilta opiskelijoilta useamman vuosikurssin ajan, jolloin aineistoon syntyy näissä osakorpuksissa diakroninen ulottuvuus, mikä mahdollistanee ainakin jonkinasteisen pitkittäistutkimuksen.[2]

OTANTADIMENSIO: kokotekstikorpus vs. otekorpus

ICLFI on kokotekstikorpus, sillä aineiston tekstit ovat kokonaisia, eivät tekstifragmentteja.

MEDIUMDIMENSIO: sähköisenä kerätyt tekstit vs. käsin kirjoitettuna kerätyt tekstit

ICLFI sisältää sekä tekstejä, jotka on kirjoitettu suoraan tekstinkäsittelyohjelmalla sähköiseen muotoon, että tekstejä,

[2] Tätä kirjoittaessa diakronisen ulottuvuuden luovien tekstien määrää ei ole selvitetty.

jotka on kirjoitettu käsin ja digitoitu myöhemmin korpusta varten.

ANNOTAATIODIMENSIO: raakatekstikorpus vs. annotoitu korpus
ICLFI:n peruskorpus on raakatekstikorpus, joka ei sisällä lingvististä metatietoa. Korpusaineiston kieliopillinen annotointi (ja lemmatisointi) on kuitenkin meneillään.

Seuraavat dimensiot kuvaavat erityisesti oppijankielikorpuksia koskevia ominaisuuksia.

ÄIDINKIELIDIMENSIO: yksiäidinkielinen vs. moniäidinkielinen korpus
ICLFI on moniäidinkielinen aineisto: se sisältää kirjoitushetkellä tekstejä 22:lta eri äidinkieltä edustavalta oppijaryhmältä.

TAITOTASODIMENSIO: alkeistason, keskitason, edistyneen tason vs. monitaitotasoinen korpus
ICLFI on monitaitotasoinen korpus, sillä siinä on tekstejä eri taitotasoilta. Aineisto on jaoteltu taitotasoihin sekä annetun opetustuntimäärän että osittain eurooppalaisen viitekehyksen taitotasoasteikon mukaan.

OPPIMISKONTEKSTIDIMENSIO: toisen kielen vs. vieraan kielen korpus
ICLFI on vieraan kielen korpus, koska tekstit on tuotettu ulkomaisissa opetuspisteissä.

OPPIMISMENETELMÄDIMENSIO: oppimalla vs. omaksumalla hankitun kielitaidon korpus
ICLFI on suurimmaksi osaksi oppimalla hankittua kielitaitoa kuvaava aineisto, koska tekstien tuottajat ovat ulkomailla asuvia suomen kielen opiskelijoita. Oppimisen ja omaksumisen avulla saavutetun kielitaidon erottaminen ei kuitenkaan ole yksiselitteistä; tämä koskee erityisesti niitä tekstejä, joiden kirjoittajat ovat oleskelleet Suomessa tai joiden vanhemmista toinen tai joku sukulaisista hallitsee suomen kielen.

4. Taustamuuttujat

Tekstiaineiston tutkimuskäyttömahdollisuudet ovat sitä paremmat, mitä monipuolisemmin tekstien taustamuuttujat on kerätty ja mitä paremmin ne on dokumentoitu. Yksi korpukset muista tekstiarkistoista erottava seikka onkin juuri tarkka dokumentaatio keräämistavasta, tekstien tuottajista ja tekstien luomistilanteista. ICLFI sisältää runsaasti taustatietoa tekstien tuottajista, itse tekstistä ja oppimiskontekstista. Taustatietojen keräämisessä on käytetty hyväksi ICLE:n taksonomiaa (ks. Granger ym. 2002), jota on täydennetty sopivin osin. Taulukkoon (2) on kuvattu aineistoon kerätyt muuttujat jaoteltuna neljään kategoriaan. Aineisto sisältää yhteensä 22 taustamuuttujaa: 7 oppijaa koskevaa, 6 oppimiskontekstia koskevaa, 6 tekstiä koskevaa ja 3 muita ominaisuuksia koskevaa tietoa. Aineiston ollessa tarpeeksi laaja tekstejä on siis mahdollista valita tutkimukseen lukuisten taustamuuttujien perusteella.

Taulukko 2. *ICLFI:n tekstien taustamuuttujat*

Suomenoppija (tekstintuottaja)	• henkilötiedot: ikä, syntymäpaikka, sukupuoli, asuinpaikka • kielitaito: äidinkieli, muut hallitut kielet • taitotaso: opiskeluajan mukaan
Oppimiskonteksti (tekstin tuottamiskonteksti)	• kielelle altistuminen: vanhempien äidinkielet, suomen käyttö kotikielenä, sukulaisten antama suomen kielen opetus, oleskelu Suomessa, opettajan äidinkieli • oppikirjat
Teksti (tehtäväkohtaiset taustatiedot)	• tekstilaji • kirjoituksen tehtävänanto • ajankäyttö: rajattu, rajaamaton • testiluonteisuus: koe, harjoitustehtävä • apuvälineet: sanakirjat, oppikirjat, muu käytössä oleva materiaali • kirjoituspaikka: luokassa, kotona, muualla
Muut	• keräyspaikka ja -aika • medium: elektronisesti, manuaalisesti kirjoitettu

Käytetyimmät muuttujat ovat nykyisessä oppijankielen tutkimuksessa olleet oppijan äidinkieli ja taitotaso, joita on tarvittu mm. kieltenvälisen vaikutuksen ja kielitaidon kehittymisen tutkimuksessa. Mielenkiintoisia kielitaidon kehittymisen kannalta olisivat myös lukuisat oppimiskontekstiin liittyvät muuttajat: Miten esimerkiksi kohdekulttuurissa vietetty aika korreloi kielitaidon kehittymisen ja vaikkapa rekisterien hallinnan kanssa? Entä miten opettajan äidinkieli vaikuttaa kielen kehittymiseen?

Toistaiseksi ei ole tutkittu, millaisen jäljen natiivin ja ei-natiivin opettajan opetus jättää oppijan kielitaitoon. Voidaan ensinnä olettaa, että itsekin kieltä aikoinaan opiskellut opettaja tietää perusteellisesti opittavan kielen ongelmat ja karikot ja hänellä on näin arvokasta omakohtaista tietoa oppimisen näkökulmasta. Toiseksi voidaan olettaa, että ei-natiivilla opettajalla säilyy kielessä joitakin sellaisia piirteitä, jotka ovat tyypillisiä erittäin edistyneillekin kielenoppijoille. Tällainen piirre voi olla esimerkiksi epätavallinen fraseologisuus: opettajan käyttämien sanojen käyttöympäristöt saattavat poiketa leksikaalisesti, semanttisesti ja vaikkapa tekstilajivalintansa suhteen jossain määrin tyypillisestä kohdekielisestä käytöstä. Tämä ei ole kieliopillisuuden tai edes ymmärrettävyyden kannalta ongelma, vaan pikemminkin eräänlainen lainalaisuus ei-natiivin kielitaidossa. Niinpä esimerkiksi kielenoppijan fraseologiataitojen (kollokaatiot, kolligaatiot, semanttiset preferenssit jne., ks. Jantunen 2009) tutkimuksessa olisikin hyvä ainakin teoreettisesti pohtia myös opettajan äidinkieltä koskevan taustamuuttujan merkitystä, vaikka muiden taustamuuttujien ja kielenomaksumiseen vaikuttavien tekijöiden rajaaminen pois onkin äärimmäisen hankalaa.

Edellä esitellyt taustatiedot on lisätty jokaiseen tekstitiedostoon siten, että ne sijoittuvat tiedostossa alkuun ennen varsinaista oppijan kirjoittamaa tekstiä. Kuviossa 2 on annettu esimerkki korpuksen tekstitiedostosta taustatietoineen.

Kukin tietue on sijoitettu kulmasulkeisiin omalle rivilleen. Kulmasulkeiden käyttö mahdollistaa muun muassa sen, että korpusohjelmat, kuten *WordSmith Tools* (Scott 2008), jättävät tarvittaessa taustatiedot huomiotta ja analysoivat esimerkiksi aineiston saneiden frekvenssiä laskettaessa vain varsinaisen tekstin.

<keräyspaikka: Tartto>
<keräysvuosi: 2008>
<medium: sähköinen>
<koodi: xxxxxxx>
<syntymävuosi: 1988>
<sukupuoli: nainen>
<syntymäpaikka: Viro, Pärnu>
<asuinpaikka: Viro, Tartto>
<äidinkieli: viro>
<äidin äidinkieli: viro>
<isän äidinkieli: viro>
<puhutaanko kotona suomea: ei>
<läheiset opettaneet suomea: ei>
<opettajan äidinkieli: viro>
<asunut/ollut suomessa, vuosi, kesto, paikka: ei>
<oppikirja: Suomi selväksi>
<taso tuntimäärän mukaan: alkeistaso>
<taso CEFR:n mukaan (1. arvioija): B2>
<taso CEFR:n mukaan (2. arvioija): B1>
<opiskellut muita kieliä: englanti, venäjä, saksa>
<tekstityyppi: kuvaus>
<tehtävänanto: vapaa aihe>
<tentti/kirjoituskoe: ei>
<rajattu kirjoittamisaika: ei>
<missä kirjoitettu: kotona>
<käytetty apuvälineitä, mitä: kyllä: oppikirja, sanakirja>

Ruuat ja juomat
Minun lempiruokani on hapankaalikeitto ja liha. Myös pidän suklaasta. Pidän yleensa liian paljon makeisista. Aamulla syön tavallisesti voileipää ja juon kahvia. Päivällä ja illalla syön lämpiruokaa sekä syön paljon, koska minulla on iso ruokahalu. Tavallisesti syön paistia, pidän vähän keitosta paitsi hapankaalikeitosta. Pidän myös jälkiruuasta, erittäin olen kiinostunut suklaantäytekakuista. Vielä pidän marjoista, esimerkiksi kirsikoista, mustikoista ja mansikoista sekä hedelmistä kuten omenoista ja appelsiineista. Tavallisesti juon maitoa tai vettä, silloin tallöin myös mehua. En pidä erittäin väkevistä juomista, mutta kuin olen juhlissa, sitten juon vähän viiniä tai jotakin likööriä. Samoin pidän ruuanlaitosta. Minulla onnistuvat parhaiten kaikki leivokset ja pullat sekä vielä muutamat ruuat. Yleensä syön kotona, koska minulla ei ole niin paljon rahaa, että saisin syöda usein ulkona. Jos menen ulkona, sitten syön sitä mitä haluan, mutta sen täytyy olla edullinen. Syön melkein kaikkia ruokia, en ole dieetillä enkä ole myös kasvissyöjä.

Kuvio 2. *Esimerkki tekstitiedostosta taustatietoineen*

5. Oppijoiden ja tekstien taitotasot

Kun ICLFI-aineistoa alettiin kerätä, suunniteltiin aineiston taitotasoluokitus siten, että taitotaso määriteltiin oppijalle hänen saamansa opetustuntimäärän mukaan. Rajoiksi asetettiin arviot kontaktituntimääristä, jotka oppijat saavat tyypillisesti alkeis-, keski- ja edistyneellä tasolla. Taitotasot muodostuivat seuraaviksi:

- opetusta alle 200 tuntia alkeistaso
- opetusta 200–400 tuntia keskitaso
- opetusta yli 400 tuntia edistynyt taso.

Jako on kuitenkin karkea eikä perustu oppijan todelliseen kielitaidon tasoon. Lisäksi se on ongelmallinen monesta muustakin syystä: tuntimäärä ei ensinnäkään huomioi kielenoppimista ja -omaksumista luokkahuoneen ulkopuolella eikä myöskään kohdekielisessä maassa eli Suomessa saatua koulutusta. Sitä paitsi se jättää huomiotta myös aiemmat opinnot, sillä tuntimäärät lasketaan nimenomaan yliopistotasolla annetun opetuksen mukaan. Niinpä esimerkiksi nuoret, jotka ovat Karjalassa oppineet suomea jo koulussa ja siirtyneet sen jälkeen opiskelemaan sitä pääaineena yliopistoon Petroskoissa, on kuitenkin tuntijaon perusteella luokiteltu alkeistason oppijoiksi. On kuitenkin ilmeistä, että heidän kielitaidon tasonsa ei vastaa kontaktituntiperustaista luokittelua.

Oma ongelmansa liittyy myös äidinkieleltään virolaisten suomenoppijoiden taitotason luokitteluun: koska heillä on lähisukukielen tuoma etu etenkin opiskelun alkuvaiheessa, heidän kielitaitonsa kehittyy tyypillisesti alussa nopeasti ja saavuttaa helposti taitotason, jolle saman tuntimäärän opiskellut, ei-sukukieltä äidinkielenään puhuva ei vielä pääse.

Koska eri oppijaryhmien tuottamien tekstien vertailu on edellä mainituista syistä ongelmallista, on aineistoa alettu analysoida eurooppalaisen viitekehyksen (EVK) taitotasojen mukaan vuodesta 2010 lähtien. Kirjoittamishetkellä (9/2011) aineistosta on analysoitu EVK:n mukaan noin 32 %. Tämä vastaa noin 1 700:aa tekstiä ja 317 000:ta sanetta. Arviointi on aloitettu tutkimustarpeita silmällä pitäen Virosta, Venäjältä, Saksasta ja Hollannista koottujen tekstien osakorpuksista. Tällä hetkellä

edellä mainitun tekstimäärän on arvioinut kaksi koulutettua arvioijaa, eli teksteistä on kaksi toisistaan riippumatonta taitotasoarviota (ks. kuvio 1). Sitä, missä määrin nämä arvioinnit ovat yhdenmukaisia, ei ole vielä selvitetty. Aineisto antaa kuitenkin viitteitä siitä, että arviot voivat toisinaan poiketa toisistaan, yleensä kuitenkin vain yhden taitotason verran. Huomionarvoista on se, että kontaktituntiperustainen arvio koskee opiskelijaa ja jokainen teksti saa saman arvion keräysvuoden aikana (ellei tuntimäärä ylitä raja-arvoa). EVK-perustainen arviointi on puolestaan tekstikohtainen, ja saman tekstintuottajan tekstit voivat siten saada eri taitotasoarvion samankin keräysvuoden aikana.

Kontaktituntiperustainen ja EVK:een perustuva arviointi voivat antaa aineistosta erilaisen kuvan. Vertaillessaan vironkielisten tuottamien tekstien (n = n. 86 000) taitotasojen määräytymistä Spoelman (2010) havaitsi, että kontaktituntikriteerin perusteella tekstit jakautuivat seuraavasti: 56 % oli alkeistason (perustason), 11 % keskitason (itsenäisen kielenkäyttäjä tason) ja 33 % edistyneiden (taitavan kielenkäyttäjän tason) opiskelijoiden tuottamia tekstejä. Yhden EVK-perustaisen arviointikierroksen perusteella luvut jakautuvat vastaavasti seuraavasti: 6 %, 79 % ja 15 %. Vaikka Spoelmanin tekemät havainnot perustuvatkin vain yhden taitotasoarvioijan arviointiin, kertonevat luvut juuri edellä kuvatusta ongelmasta virolaisten opiskelijoiden kontaktituntiperustaisessa taitotasonmäärittelyssä. Tietoa siitä, millaisia erot ovat näiden kahden taitotasokriteerin välillä ei-sukukielisten tekstejä arvioitaessa, ei ole vielä käytettävissä tätä katsausta kirjoitettaessa.

Aineistossa säilytetään molemmilla tavoilla tehtävät kielitaidon arvioinnit (ks. kuvio 1). Tämä tekee mahdolliseksi yhtäältä opetustuntien määrään ja toisaalta viitekehykseen perustuvan arvion huomioon ottamisen ja vertailun analyysissa. Koska korpukseen on sisällytetty kaikkien arvioijien taitotasoarviot, voidaan kaikki arviot ottaa tarvittaessa huomioon.

6. Korpusaineiston annotointi

Oppijankielen korpusten annotoinnissa eli merkinnässä on ollut tavallista virheannotaatio (Barlow 2005: 341; ks. myös Granger 2004; 2007). Tämä johtuu ainakin osittain oppijankielen tutkimuksen virheanalyysipainotteisuudesta. Oppijan tuottamien virheiden koodauksella saadaankin tekstiin luotua runsaasti metatietoa, ja oppimisen ongelmakohtien analyysi voi tämän virhekoodauksen jälkeen perustua laajoihin tekstiaineistoihin. Virhekoodaus on kuitenkin pitkälti käsityötä ja siten työlästä tarkistusvaiheineen (ks. Dagneaux ym. 1998; Granger 2007), eikä koodaaminen ole läheskään aina yksiselitteistä virhetyyppien määrittelyn ongelmien vuoksi (ongelmista ks. Ellis & Barkhuizen 2005: 51–71).

Oppijankieliaineistoja on annotoitu myös kieliopillisesti. Virhekoodauksen sisältävä oppijanenglannin ICLE sisältää myös sanaluokkaannotaation, ja näin ollen aineistoa voidaan analysoida muullakin tavalla kuin virhelähtöisesti. Oppijankieliaineistojen kieliopillinen annotaatio on kaikkinensa erittäin haastavaa, sillä etenkään täysin automaattinen analyysi ei luonnollisestikaan anna yhtä hyvää tulosta oppijoiden tuottamasta materiaalista kuin natiiviaineistosta. Tähän ovat syynä muun muassa poikkeavat oikeinkirjoitus- ja taivutusmuodot. (De Haan 2000: 71; ks. myös van Rooy & Schäfer 2003: 836; Meunier & de Mönnink 2001). Oman haasteensa oppijankielen annotaatioon tuo myös opittavan kielen – tässä tapauksessa suomen – rikas morfologia: oppijat päätyvät leksikaalisia ja kieliopillisia morfeemeja yhdistellessään toistuvasti muotoihin, joita automaattinen analysaattori ei pysty tulkitsemaan oikein.

Toisaalta kuitenkin puhtaasti käsin tehtävä annotaatio olisi liian työläs toteutettavaksi. Niinpä ICLFI on koodattu ns. puoliautomaattisesti: tekstitiedosto on siirretty ensin *Word*-tekstinkäsittelyohjelmaan, jossa oikeinkirjoitusvirheet ja taivutusmuodoissa esiintyvät ongelmat (esim. väärät astevaihtelumuodot, kuten *kaktena* pro *kahtena*) on poistettu. Lauseke- ja virkerakenteita ei ole kuitenkaan korjattu oikeaan muotoon. Virhekorjauksen jälkeen tiedosto on siirretty Connexorin Fi-fdg-jäsen-

timeen (Fi-fdg)[3], joka lemmatisoi ja koodaa aineiston automaattisesti morfosyntaktisesti. Tämän jälkeen tiedostoon on palautettu tekstintuottajan kirjoittamat sananmuodot alkuperäisessä virheellisessä asussaan. Viimeisenä työvaiheena on morfosyntaktisen koodauksen tarkistaminen: automaattisen koodauksen jälkeen tekstiin jää jonkin verran virheitä, ja lisäksi jäsennin antaa usein monia vaihtoehtoisia koodauksia yhdelle muodolle, jolloin vaihtoehdoista on poimittava oikea manuaalisesti. Tämän prosessin tuloksena saadaan alkuperäinen teksti lemmatisoituna ja kieliopillisesti koodattuna. (Kuvio 3.)

1	Rakalle	rakas		@NH N SG ALL
2	Joulupukille	joulu#pukki		@NH N SG ALL
3	,	,		
4	Kiitos	kiitos		@ADVL N SG NOM
5	kirjelle	kirje		@NH N SG ALL
6	Rovaniemelta	rovaniemi	mod>5	@NH N SG ABL Prop
7	.	.		
8	<s>	<s>		
1	Tänä	tämä	attr>2	@PREMOD PRON SG ESS
2	Jouluna	joulu	tmp>9	@NH N SG ESS
3	meidän	me		@PREMOD PRON PL P1 GEN
4	käynnin	käynti		@PREMOD N SG GEN
5	Rovaniemella	rovaniemi		@NH N SG ADE Prop
6	jälkeen	jälkeen	goa>9	@ADVL ADV
7	Heidän	he	attr>8	@PREMOD PRON PL P3 GEN
8	kirje	kirje	subj>9	@NH N SG NOM
9	oli	olla	main>0	@MAIN V ACT IND PAST SG P3
10	hyvä	hyvä	attr>11	@PREMOD A SG NOM
11	yllatys	yllätys	comp>9	@NH N SG NOM
12	.	.		
13	<s>	<s>		

Kuvio 3. *Esimerkki annotoidusta ja lemmatisoidusta tekstistä ennen viimeistä korjausta. Esimerkissä on tekstin kaksi ensimmäistä virkettä (hollanti-alkeet)*

[3] Fi-fdg jäsentää tekstin dependenssikieliopin mukaisesti, mikä luonnollisesti vaikuttaa annotoinnin tulokseen ja tulkintamahdollisuuksiin.

Kuten edellä olevasta kuviosta huomaa, puoliautomaattinen annotaatio toimii hyvin, joskin vaatii vielä korjauksen: jäsennin on tulkinnut – ilmeisesti epätyypillisen sanajärjestyksen vuoksi – *jälkeen*-sanan adverbiksi. Esimerkistä näkee sen, että oppijan tuottama teksti on palautettu aineistoon kirjoitusvirheineen sellaisessa muodossa kuin se on alun perin kirjoitettu, vaikka automaattista annotaatiota varten esimerkissä olevat virheet onkin prosessin aikana kertaalleen korjattu. Automaattinen lauseenjäsennys on puolestaan oppijankieliaineistossa erittäin hankalaa ja luotettavuuden kannalta kyseenalaista morfologiseen analyysiin verrattuna. Tämän vuoksi aineiston annotoinnin tuloksesta on suunnitteilla yksinkertaistettu versio ilman lauseenjäsennystä; vaihtoehtona on myös käyttää jäsennintä, joka ei tee syntaktista annotointia.

7. Lopuksi

Yllä oleva kuvaus esittelee Kansainvälisen oppijansuomen korpuksen ja sen koostamisperiaatteet. Korpuksen koostaminen on jatkuva prosessi, ja esimerkiksi tässä esitetyt tunnusluvut muuttuvat jatkuvasti. Samoin annotaatioon ja taitotasojen määrittelyyn liittyvät seikat ovat kehittämisen kohteena ja siten muuttuvia. Korpuksen kokoaminen ja työstäminen onkin pitkä, monivuotinen prosessi. Prosessin aikana voi tulla vastaan muutos- ja kehittämistarpeita, jotka osaltaan lykkäävät aineistojen valmistumisaikaa myöhäisemmäksi. Vaikka korpuksen koostaminen voikin tuntua pitkään kestävältä ja haastavalta prosessilta, mahdollisimman tarkka koostamisperiaatteiden suunnittelu, kehitystyö aineiston keräämisen aikana ja aineiston pilotointi kannattavat. Valmis aineisto on silloin käyttäjänsä kannalta parempi. On myös tärkeää tehdä aineistolla varsinaista tutkimustyötä jo silloin, kun aineisto on vielä suhteellisen pieni: pilottiluontoisella tutkimuksella saadaan alustavaa tietoa tutkittavasta kohteesta, ja tuloksia voidaan tarkentaa myöhemmin aineiston kasvaessa ja ehkä myös tutkimuskysymysten tarkentuessa. Tutkimustyöllä on arvonsa myös itse aineiston kehittämistyössä.

Katsauksessaan oppijankielen korpustutkimuksesta Granger (2007) luettelee useita haasteita, jotka tulisi ottaa huomioon aineiston kokoamisessa ja tutkimuksessa. Näitä ovat korpusten kokoamiseen (aineistojen vähyys) ja analysoimiseen (kvantitatiivisen lisäksi myös kvalitatiivista tutkimusta) sekä monitieteisyyteen (toisen kielen oppimisen ja korpuslingvistiikan yhdistäminen) liittyvät seikat. Grangerin mainitseman aineistojen vähälukuisuuden lisäksi on kuitenkin kiinnitettävä yhä enemmän huomiota myös korpusten sisältöön ja laatuun. Etenkin oppijankielikorpusten kokoaminen on lähtenyt tyypillisesti liikkeelle aineistoa kokoavan tutkimusryhmän tarpeista.

Aineiston sisältöön voivat vaikuttaa muutkin seikat, kuten ICLFI:n osalta kontaktiopetuksessa tuotettujen tekstien määrä ja tekstilajit. Erityisesti tutkimustulosten vertailtavuuden kannalta olisi kuitenkin olennaista, että aineistot olisivat mahdollisimman samankaltaisia muun muassa taitotasojen (ja niiden määrittelykriteereiden) ja tekstilajien näkökulmasta. Pelkästään yksitekstilajisten ja yhtä taitotasoa sisältävien korpusten kokoaminen parantaa aineistojen vertailtavuutta, mutta toisaalta myös kaventaa oppijankielestä saatavaa kuvaa. Suomalaisia tutkimusaineistoja ajatellen pohdittavaksi kannattaisi nostaa nykyisten Suomessa kerättävien aineistojen keskinäinen vertailukelpoisuus ja kehittämistarpeet. Edelleen huomiota tulisi kiinnittää taustatietojen keräämiseen ja niiden dokumentoimiseen sekä aineistojen annotoinnin harmonisointiin. Näillä kaikilla keinoilla voidaan kehittää tutkimuskorpuksia ja luoda perusteet sille, että yhdessäkin tutkimuksessa voidaan hyödyntää useita aineistoja kuvauksen monipuolistamiseksi.

Lähteet

Atkins, Sue, Jeremy Clear, Nicholas Ostler 1992. Corpus design criteria. – Literary and Linguistic Computing 7 (1), 1–16. doi:10.1093/llc/7.1.1

Antonsen, Lene, Trond Trosterud, Ciprian-Virgil Gerstenberger 2006. Sámi giellatekno. http://giellatekno.uit.no/doc/lang/corp/corpus-sme.html (22.2.2011).

Barlow, Michael 2005. Computer-based analyses of learner language. – Rod Ellis, Gary Barkhuizen (Eds.). Analysing Learner Language. Oxford: Oxford University Press, 335–357.

Dagneaux Estelle, Sharon Denness, Sylviane Granger 1998. Computer-aided error analysis. – System: An International Journal of Educational Technology and Applied Linguistics 26 (2), 163–174.

de Haan, Pieter 2000. Tagging non-native English with the TOSCA-ICLE tagger. – Christian Mair, Marianne Hundt (Eds.). Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20). Amsterdam: Rodopi, 69–79.

Ellis, Rod, Gary Barkhuizen 2005. Analysing Learner Language. Oxford: Oxford University Press.

Fi-fdg: Connexor Machinese Syntax for Finnish. CSC, Tieteen tietotekniikan keskus. http://www.csc.fi/english/research/software/fi-fdg (19.1.2011).

Goossens, Diane, Sylviane Granger 2011. Learner corpora around the world. http://www.uclouvain.be/en-cecl-lcWorld.html (24.2.2011).

Granger, Sylviane 2004. Computer learner corpus research: Current status and future prospects. – Ulla Connor, Thomas A. Upton (Eds.). Applied Corpus Linguistics: A Multidimensional Perspective. Amsterdam: Rodopi, 123–145.

Granger, Sylviane 2007. A Bird's-eye view of learner corpus research. – Wolfgang Teubert, Ramesh Krishnamurthy (Eds.). Corpus Linguistics: Critical Concepts in Linguistics. Vol. 2. London, New York: Routledge, 44–72.

Granger, Sylviane 2010. Learner corpus research at Louvain: An overview. – Esitelmä. ASKeladden Network Seminar March 25th–26th 2010. University of Bergen.

Granger Sylviane, Estelle Dagneaux, Fanny Meunier 2002. The International Corpus of Learner English. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, Magali Paquot 2009. International Corpus of Learner English. Version 2. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Jantunen, Jarmo H. 2009. Minulla on aivan paljon rahaa. Fraseologiset yksiköt suomen kielen opetuksessa. – Virittäjä 113, 356–381.

Jantunen, Jarmo H., Saana Piltonen 2009. Oppijansuomen ja -viron sähköiset tutkimusaineistot. – Virittäjä 113, 449–458.

Kennedy, Graeme 1998. An Introduction to Corpus Linguistics. London: Longman.

Laviosa-Braithwaite, Sara 1996. The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation. Manchester: UMIST.

Lee, David 2010. Corpora, Collections, Data Archives. http://www.uow.edu.au/~dlee/corpora.htm (1.3.2011).

Lehtinen, Marja, Pirjo Karvonen, Tarmo Rahikainen 1995. Tekstikorpukset. Raportti tekstikorpusten koostamisperiaatteista ja nykysuomen tekstiaineistojen tarpeellisuudesta Kotimaisten kielten tutkimuskeskuksessa. Helsinki: Kotimaisten kielten tutkimuskeskus.

Mauranen, Anna, Pekka Kujamäki (Eds.) 2004. Translation Universals. Do They Exist? Amsterdam: Benjamins.

Meunier, Fanny, Inge de Mönnink 2001. Assessing the success rate of EFL learner corpus tagging. – Sylvie de Cock, Gaëtanelle Gilquin, Sylviane Granger, Stephanie Petch-Tyson (Eds.). Proceedings of the 22nd International Computer Archive of Modern and Medieval English Conference. ICAME 2001: Future Challenge for Corpus Linguistics. Louvain-la-Neuve, 16–20 May 2001. Louvain-la-Neuve: Centre for English Corpus Linguistics, 59–60.

Scott, Mike 2008. Developing WordSmith. – International Journal of English Studies 8 (1), 95–106.

Spoelman, Marianne 2010. The use of the partitive case in Finnish learner language: The operationalization of foreign language proficiency. – Esitelmä Metodit L2-korpustutkimuksessa -työpajassa AFinLAn syyssymposiumissa Vaasassa 12.–13.11.2010.

Suihkonen, Pirkko 2007. Computer corpora at the university of Helsinki corpus server. http://www.ling.helsinki.fi/uhlcs/data/corpora-combined.html#C1007 (22.2.2011).

van Rooy, Bertus, Lande Schäfer 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. – Dawn Archer, Paul Rayson, Andrew Wilson, Tony McEnery (Eds.). Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK), 28–31 March 2003. Vol. 16. Lancaster: UCREL, Lancaster University, 835–844.

Jarmo Harri Jantunen
Oulun yliopisto
suomi toisena ja vieraana kielenä
PL 1000
90014 Oulu, Finland
jarmo.jantunen@oulu.fi

International Corpus of Learner Finnish (ICLFI): typology, variables and annotation

JARMO HARRI JANTUNEN
University of Oulu

To date, computer-based analyses of learner language have mainly concentrated on studying the Indo-European languages, mainly English. However, when we aim to analyse the features that exist in the learner production no matter what their mother tongue backgrounds are (i.e. learner language universals), we should compile and analyse also other, non-Indo-European databases of learner language.

This overview introduces the International Corpus of Learner Finnish (ICLFI). The ICLFI is one of the major corpora representing Baltic-Finnic learner language, the others being the Estonian Interlanguage Corpus (EIC) and the Corpus of National Certificate Tests (CEFLING project). The compilation of the ICLFI started in 2007 in the project Corpus study on language-specific and universal features in learner language. The data is being compiled with the help of Finnish language teachers working at foreign universities. The corpus consists of the texts written by learners of Finnish, produced spontaneously in the language learning situations. So far (September 2011), the ICLFI contains circa one million words and texts from speakers of 22 different mother tongues. The text types vary from fiction to non-fiction.

In addition to the texts, the corpus includes a whole range of metatextual information on different variables; this information is encoded in the header of every text file. The variables are learner variables (e.g. age and mother tongue), learning context variables (e.g. teacher's mother tongue), and task variables (e.g. genre and reference tools). These data give the opportunity to study the effect of, for example, the mother tongue background and the proficiency level on the learner's learning process and production. The article discusses also corpus typology, according to which learner corpora can be described and categorized. Furthermore, also corpus annotation and proficiency level analysis are

discussed. The information retrieved from the ICLFI can be utilized in actual language teaching situations, as well as in textbook writing and dictionary compilation.

Keywords: learner language; corpus; corpus typology; annotation; variables