Mutta, M., P. Lintunen, I. Ivaska & P. Peltonen (toim.) 2014. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2014 / n:o 7. 60–80.

Jarmo Harri Jantunen[1], Sisko Brunni[2], Liisa-Maria Lehto[2] & Valtteri Airaksinen[2]

[1]Jyväskylän yliopisto, [2]Oulun yliopisto

Oppijankieliaineistojen annotointi – esimerkkinä ICLFI:n annotoinnin prosessit, ongelmat ja ratkaisut

This article illustrates the grammatical and error annotation of learner language with the help of the *International Corpus of Learner Finnish* (ICLFI). In particular, we will focus on issues arising from handling with at least semi-automatic methods a morphologically rich language. What makes this corpus special compared to, for example, English-language material, is the frequent variation in different forms and related errors, both due to the rich morphology of the target language. This article begins with a description of the design and implementation process of both the grammatical and error annotation, followed by a brief introduction to the material for which the annotations were designed. Finally, we outline some of the problems that have arisen during the annotation process and their solutions.

Keywords: corpus study, learner language corpora, annotation, error annotation

1 Johdanto

Artikkelissamme esittelemme sähköisten oppijankieliaineistojen kieliopilliseen ja virheannotointiin liittyviä haasteita ja niiden ratkaisuja. Oppijankielikorpuksia on annotoitu virheanalyysin viitekehyksessä jo aivan niiden keräämisen alkumetreiltä alkaen, ja sittemmin myös kieliopillinen annotointi on tullut yhä tavallisemmaksi. Koska kielenoppijoiden tuotos poikkeaa monella tapaa äidinkielisten puhujien tuottamasta kielestä, kieliopillisessa annotoinnissa on tavallista enemmän työtä. Virheannotointikategoriat, jotka on luotu esimerkiksi indoeurooppalaisten oppijankieliaineistojen virheannotoinnin pohjaksi, eivät puolestaan sellaisenaan sovellu oppijansuomen aineistojen kategorisoinnin perustaksi. Näitä seikkoja käsittelemme seuraavassa Kansainvälisen oppijansuomen korpuksen (ICLFI, Jantunen 2011) avulla. Luomme aluksi yleiskatsauksen oppijankieliaineistojen annotointiprosesseihin, minkä jälkeen esittelemme ICLFI-aineiston ja siihen luotuja kieliopilliseen ja virheannotointiin liittyviä ratkaisuja.

2 Korpusaineistojen annotointi

Korpusaineistojen käytettävyyden parantamiseksi teksteihin lisätään usein metadataa, kuten taustatietoja ja erilaisia lingvistisiä koodauksia (prosessista ks. esim. Heikkinen, Lounela & Voutilainen 2012). Tekstien tuottajiin, itse teksteihin ja keräystilanteeseen liittyvät taustatiedot ovat oleellista informaatiota erityisesti oppijankielen kaltaisissa erikoiskorpuksissa, sillä niistä tehtävä tutkimus on usein vertailevaa ja vertailuun otettavat tekstit valitaan tyypillisesti juuri taustamuuttujien perusteella. Aineiston käytettävyyttä voidaan parantaa myös lisäämällä tekstiä ja sen elementtejä selittävää lingvististä tietoa. Sanoihin voidaan lisätä esimerkiksi merkitsin (*tag*), joka kertoo sanan sanaluokan tekstikontekstissaan (*part-of-speech-tagging* eli *POS tagging*). Tätä prosessia kutsutaan annotoinniksi, kuten myös vaikkapa morfologisen informaation lisäämistä. Samaa termiä (sekä myös *annotaatiota*) käytetään myös prosessin lopputuloksesta eli sähköisistä kielimateriaaliin liitetyistä tai sitä selittävistä lingvistisistä merkitsimistä. (Leech 1997a: 2; Heikkinen ym. 2012.)

Laajojen sähköisten korpusten hyödyllisyys arvioidaan tyypillisesti sen perusteella, miten niistä saadaan ammennettua tietoa. Hyvin usein tiedon saaminen edellyttää kuitenkin ensin uuden tiedon lisäämistä niin sanottuihin raakateksteihin eli teksteihin, jotka eivät sisällä minkäänlaista metatietoa. Esimerkiksi homonyymiset ilmaukset voivat kuulua eri sanaluokkiin, ja korpuksen käyttäjän on lisättävä tämä tieto saamiinsa osumiin voidakseen hyödyntää tuloksia. Annotoituun aineistoon tämä tieto on lisätty etukäteen, joten oikean informaation löytyminen korpuksesta nopeutuu ja helpottuu.

Koska annotointi on kallista ja aikaa vievää toimintaa, ei sitä ole taloudellista tehdä kerta toisensa jälkeen uudelleen. Lisäksi kun aineisto on jo kertaalleen annotoitu, voi korpusta hyödyntää jatkossa entistä monipuolisemmin: aikaisempaa koodausta voidaan hyödyntää uusien annotointien tekemisessä tai korpusta voidaan käyttää useisiin tutkimuksiin. Kun aineistoon on lisätty esimerkiksi sekä kieliopillinen annotointi että virheannotointi, lisäävät ne korpuksen käytettävyyttä tukiessaan toisiaan halutun ilmiön etsimisessä. Sanaluokkien koodausta taas voidaan hyödyntää esimerkiksi leksikografiassa, lauseiden jäsennyksessä tai sanalistojen teossa. (Leech 1997a: 4.) Toisaalta Leech (2004) tähdentää, että korpuksen monikäyttöisyys ei välttämättä ole suorassa suhteessa korpukseen tehtyjen yleisten annotointien kanssa, vaan toisinaan tutkimuksen kannalta hyödyllisemmiksi voivat osoittautua erityisesti kutakin tutkimustilannetta varten suunnitellut annotoinnit. On lisäksi muistettava, että esimerkiksi tekstikorpuksissa varsinaisen aineiston muodostavat aina itse tekstit (ns. raakatekstit); annotoinnit antavat vain lisäinformaatiota (Leech 1997a: 4).

Jotta annotointi aidosti auttaisi tutkijoita työssään, on prosessissa noudatettava tiettyjä periaatteita: 1) Annotoitu aineisto tulee tallentaa niin, että raakaversio on aina otettavissa uudelleen käyttöön. Vastaavasti annotoinnit tulee pystyä irrottamaan itse korpuksesta ja tarvittaessa tallentamaan erikseen. 2) Annotointiprosessi tulee dokumentoida tarkasti. Dokumentaation tulee sisältää mm. käytetyn annotointijärjestelmän kuvaus, tieto annotoinnin tekopaikasta ja tekijästä. Lisäksi on dokumentoitava annotoinnin laatuun liittyviä seikkoja (virheiden mahdollisuus, miten annotoinnit on tarkastettu jne.). 3) Lisäksi annotointijärjestelmien tulisi olla muiden korpusten tekijöiden hyödynnettävissä, jottei työtä tarvitse aloittaa alusta. Tämän vuoksi järjestelmän tulisi perustua yleisesti hyväksyttyyn ja neutraaliin analyysiin, jotta se olisi mahdollisimman laajalti ja helposti ymmärrettävissä ja hyödynnettävissä. Mikään annotointijärjestelmä ei voi kuitenkaan asettautua absoluuttiseksi standardiksi muille korpusten koostajille, koska annotointitarpeet voivat vaihdella muun muassa korpukseen tarkoituksen, koon ja kielen mukaan. Tämä ei kuitenkaan tarkoita sitä, etteikö mahdollisimman suureen yhdenmukaistamiseen korpusten kesken tulisi pyrkiä. (Leech 1997a: 6–7).

Korpuksia voidaan annotoida monin eri tavoin. Esimerkiksi puhekorpuksiin voidaan lisätä pragmaattinen, diskursiivinen tai foneettinen annotaatio. Pragmaattisessa annotoinnissa kiinnitetään huomiota siihen, missä asemassa ilmaus toimii kontekstissaan; sama ilmaus voi toimia esimerkiksi käskynä tai pyyntönä. Diskurssiannotoinnissa kiinnitetään vuorostaan huomiota esimerkiksi pronominien viittaussuhteisiin. Foneettisessa annotoinnissa taas koodataan ilmausten lausumiseen, painoon ja intonaatioon liittyviä seikkoja. Myös ilmausten tyyliä voidaan annotoida. Syntaktisessa annotoinnissa puolestaan koodataan sanojen kieliopillisia suhteita lauseiden jäsennyksessä. Kaikissa korpuksissa oleellinen annotointitaso on sanan lemmatisointi. Erityisen oleellista se on

oppijankielen korpuksissa, joissa lemman sananmuotojen variaatio on suuri oppijoiden tekemien erilaisten taivutus- ja oikeinkirjoitusvirheiden vuoksi. Lemmatisoinnissa sanan tekstissä esiintyvään taivutettuun muotoon lisätään tieto sanan perusmuodosta eli lemmasta (esim. **kodussa* 'kodissa' > KOTI). Tämä tehostaa korpuksen käyttöä mahdollistamalla kaikkien taivutusmuotojen etsimisen yhdellä haulla. Teksti voidaan lisäksi annotoida semanttisesti, jolloin esimerkiksi homonyymisiin ilmauksiin lisätään tieto siitä, mihin semanttiseen kategoriaan sanat kuuluvat (disambiguoinnista esim. Heikkinen ym. 2012: 384–386). Tällöin hakua voidaan rajata koskemaan sananmuotoa tai lemmaa vain tietyssä merkityksessään. (Leech 2004; ks. myös Garside, Leech & McEnery 1997.) Oppijankielen korpuksiin usein lisätty virheannotaatio puolestaan mahdollistaa tuotettujen virheiden analyysin sekä oppijankielen ja natiivikielen vertailun: missä kohdin ja miten oppijoiden kieli eroaa natiivipuhujien kielestä (Granger 2002: 14).

2.1 (Oppijankieli)korpusten kieliopillinen ja syntaktinen annotointi

Oppijankielen annotointi ei voi rajoittua vain virheiden analysointiin. Toisen ja vieraan kielen tutkimuksessa kielen omaksuminen ja sen vaiheittaisuus ovat keskeisiä tutkimuskohteita (ks. Ellis 1994: 73–76; Pienemann 1998). Vaiheittaisuuden kuvaaminen perustuu pitkälti erilaisten kielellisten ilmiöiden välisiin suhteisiin, joiden tutkimiseen lukuisat oppijankieleen korpukset soveltuvat erinomaisesti. Aineistojen tehokas käyttö edellyttää kuitenkin, että niihin on lisätty erilaisia kieliopillisia annotaatioita. Yleisin lingvistinen annotaatio, joka oppijakielen aineistoihin lisätään, on sanojen sanaluokkien kuvaus (ks. Rooy & Schäfer 2003; Schmidt 1994; Granger 2002). Grangerin (2002: 17) mukaan oppijankielen aineistoissa sanaluokkakoodaus lisää selvästi korpuksen arvoa ja eri korpusten vertailtavuutta; sanaluokkakoodauksen etuja ovat myös niissä käytettävien ohjelmien automaattisuus, laaja saatavuus ja halpuus.

Toisaalta tekstin merkityksen täydellinen ymmärtäminen edellyttää lauseiden syntaksin dekoodausta. Lauseita muodostavat komponentit (lauseenjäsenet) ja jäsenten väliset suhteet ovat edellytys korpusten hyödyntämiselle mm. automaattisessa kääntämisessä ja puheentunnistamisessa (Leech & Eyes 1997: 34). Vastaavalla tavalla syntaktinen koodaus voi auttaa myös oppijankielentutkimusta laajentamalla tutkimuskohteet kielellisten ilmiöiden välisiin suhteisiin. Syntaktiseen annotointiin kehitetyt ohjelmat perustuvat kuitenkin pääosin englannin kieleen eivätkä siten sellaisenaan sovi malliksi morfologisesti monipuolisempien kielten syntaktiseen käsittelyyn (Leech & Eyes 1997: 47). The Helsinki Constraint Grammar -parseria on kuitenkin käytetty myös englantia morfologisesti monimutkaisempien kielten syntaktiseen annotointiin (Karlsson, Voutilainen, Heikkilä & Anttila 1995). Lisäksi käytössä olevat sanaluokkajaottelut ja koodauksessa käytetyt ohjelmat on yleensä kehitetty natiivikielten pohjalta tai niitä var-

ten, mikä aiheuttaa ongelmia erityisesti oppijankielen annotoinnissa (ks. Diaz-Negrillo, Meuers, Valer & Wunsch 2010; Rastelli 2009). Rastellin (2009) mukaan muun muassa liian tiukka kohdekielen mukainen sanaluokkakoodaus ei sovi SLA-tutkimukseen (*Second Language Acquisition*), koska siinä kiinnostuksen kohteena on oppijoiden tuottama kieli, johon kuuluvat olennaisesti sekä korrektit että epäkorrektit ilmaukset, eikä sanaluokkajaottelu aina ole itsestään selvää. Sama pätee myös muihin oppijankielen ilmiöihin.

Koska annotoinnin perimmäisenä tarkoituksena on antaa aineistoa selittävää lisäinformaatiota, pitää se välttämättä sisällään myös jonkinlaisen inhimillisen käsityksen kielestä, eikä mikään kokonaan automaattinen järjestelmä voi täysin virheettömästi käsitellä kielen kompleksisuutta (Leech 1997a: 2; Heikkinen ym. 2012: 373–374). Olemassa olevissa syntaktisissa annotointijärjestelmissä automaation ja inhimillisen työn osuudet vaihtelevat suuresti, mutta aina niihin kuitenkin kuuluu vähintään manuaalinen annotoinnin tarkastus (ks. Bateman, Forrest & Willis 1997: 167). Myös sanaluokkakoodauksessa oleellinen kysymys on, missä määrin manuaalista työtä tarvitaan automaattisten koodausten editoimisessa (Leech 1997b: 20).

Suomen kaltainen morfosyntaktisesti kompleksinen kieli vaatii omat lähestymistapansa. Pelkkä POS- tai lauseenjäsenkoodaus eivät anna riittävää lingvististä informaatiota tutkijan käyttöön, koska oppijoiden ongelmiksi koituvat usein juuri morfosyntaktiset sijavalinnat. Annotointiin tulee siis sisältyä vielä sanaluokka- ja lauseenjäsenkoodaustakin enemmän lingvististä tietoa, jotta tutkijat voivat kohdistaa huomionsa juuri haluttuun kielen ilmiöön (ks. Ragheb & Dickinson 2012). ICLFI-korpuksessa on päädytty morfosyntaktiseen kieliopilliseen annotointiin, johon kuuluvat 1) lemmatisointi sekä 2) sanaluokan, 3) sijamuodon ja 4) lauseenjäsenaseman koodaaminen. Annotoinnissa pyritään aina maksimoimaan automaation osuus, mutta jokainen vaihe vaatii myös manuaalista työtä, näin on myös tässä esiteltävän aineiston kohdalla.

2.2 Oppijankielikorpusten virheannotointi

Oppijankielen korpuksia on koottu useiden eri äidinkielten puhujilta, ja ne edustavat eri kohdekieliä. Korpukset eroavat keskenään muun muassa siinä, miten paljon aineistoa on käsitelty: onko aineistoa annotoitu kieliopillisesti tai virheiden analysoinnin näkökulmasta? Oppijanenglannin korpuksia on useita, ja niistä esimerkiksi *International Corpus of Learner English* (ICLE) ja *Corpus of Japanese Learner English* (NICT JLE) ovat ainakin osittain virheannotoituja (korpuksista ks. esim. Tono 2003: 802–803; Diaz-Negrillo & Fernandes-Dominguez 2006: 87; ICLE, Granger, Dagneaux & Meunier 2002; NICT JLE, Izumi, Uchimoto & Isahara 2005). Myös oppijansaksan (esim. FALKO) ja -ranskan (FRIDA) korpuksia on virheannotoitu (ks. esim. Diaz-Negrillo & Fernandes-Dominguez 2006: 87).

(Kattava luettelo oppijankieliaineistoista löytyy osoitteesta http://www.uclouvain.be/en-cecl-lcworld.html.)

Oppijankielen korpusten virheanalyysia on perusteltu muun muassa sillä, että oppijan tekemien virheiden analysointi on yksi tehokkaimmista keinoista kuvailla oppijoiden tuottaman kielimuodon ominaispiirteitä ja kehitysvaiheita. Tätä tietoa voidaan jatkossa käyttää hyväksi toisen kielen omaksumisen tutkimuksessa ja kielen opetuksessa. (Izumi ym. 2005: 71; Granger 2002: 14.) Virheiden merkitseminen on tullut keskeiseksi osaksi oppijankielen analyysia, joka tunnetaan nimellä *computer-aided error analysis* (Dagneaux, Dennes & Granger 1998: 163). Toisaalta virheiden analysointi on myös ollut kritiikin kohteena. Sitä on pidetty epätieteellisenä ja sekavana oppijankielen negatiivisiin piirteisiin keskittyvänä lähestymistapana. (Granger 2003: 466, 2002: 14.) Grangerin (2003: 466) mukaan virheet ovat kuitenkin erottamaton osa oppijankieltä ja siinä mielessä yhtä lailla analyysin arvoisia kuin mitkä tahansa muutkin oppijankielen piirteet.

Virheiden koodaamisella on useita etuja, joista suurimmat liittyvät virheiden hakemiseen aineistosta. Kokonaan virhekoodattu korpus paljastaa epätyypilliset muodot ja mahdollistaa virheiden tehokkaan hakemisen virhetyypin tai tietyn kielenoppijaryhmän mukaan. Virhekoodatun korpuksen avulla voidaan kuvata esimerkiksi sitä, mitä virheitä jokin kielenoppijaryhmä tekee eniten ja miten virheiden määrä ja laatu muuttuvat kielitaidon kehittyessä. Materiaalista voidaan löytää sekä odotuksenmukaisia että täysin ennakoimattomia virheitä. Lisäksi aineistosta voidaan koodauksen avulla löytää ns. nollaesiintymät, kun kielenoppija on jättänyt merkitsemättä esimerkiksi sanan, artikkelin tai konjunktion. (Dagneaux ym. 1998: 72.) Seuraavaksi esittelemme ICLFI-korpuksen ja käsittelemme sen morfologista annotointia ja virheannotointia sekä näissä ilmeneviä ongelmia ja käytettyjä ratkaisumalleja.

3 *Kansainvälinen oppijansuomen korpus* (ICLFI) ja sen annotointiprosessit

Kansainvälinen oppijansuomen korpus (ICLFI, *International Corpus of Learner Finnish*), jota on kerätty Oulussa vuodesta 2007 lähtien, on yksi kuudesta Suomessa kootusta suomi toisena ja vieraana kielenä -tekstimateriaaleja sisältävästä sähköisestä aineistosta. Muita ovat Yleisten kielitutkintojen *YKI-korpus*, *Cefling-korpus*, *Topling-korpus, Dialuki-korpus* (Jyväskylä) ja *Edistyneiden suomenoppijoiden korpus* (Turku). ICLFI on ns. suomi vieraana kielenä -korpus, sillä tekstit ovat peräisin opiskelijoilta, jotka opiskelevat suomea ulkomaisissa yliopistoissa – joko pää- tai sivuaineena tai yksittäisinä kursseina. Taulukkoon 1 on kuvattu korpuksen tämänhetkinen tilanne erilaisina tunnuslukuina ja -piirteinä.

TAULUKKO 1. ICLFI tunnuslukuina ja -piirteinä (8/2014).

Koko - saneita - tekstejä	 1 miljoonaa sanetta n. 6 000
Annotointi - kieliopillinen annotointi - virheannotointi	 92 % 5 %
Taitotasoarviointi (EVK) - A1 - A2 - B1 - B2 - C1 - C2	 0,1 % 7,3 % 43,2 % 36,1 % 11,9 % 2,0 %
Lemmatisointi	92 %
Äidinkieliä	22
Aineiston kerääminen	Sekä käsin että tekstinkäsittelyohjelmalla kirjoitettuja tekstejä
Tekstilajit	Fiktiivisiä ja asiatekstejä
Tehtävän suoritus	Opetuksen yhteydessä tehty harjoitustyö tai koe

Jotta korpusaineistoja on mahdollista hyödyntää monipuolisesti, on niiden suunnitelmalliseen keräämiseen ja taustamuuttujien tallentamiseen kiinnitettävä erityistä huomiota. ICLFI-aineisto sisältää runsaasti metatietoa tekstintuottajista, tilanteesta, jossa tekstit on kirjoitettu, ja itse tekstistä. Taustamuuttujat on kuvattu seuraavassa asetelmassa:

Tekstintuottaja
- henkilötiedot: ikä, syntymäpaikka, sukupuoli, asuinpaikka
- kielitaito: äidinkieli ja muut osatut kielet
- taitotaso: opiskeluajan mukaan

Oppimiskonteksti
- opittavalle kielelle altistuminen: vanhempien äidinkielet, suomen käyttö kotikielenä, sukulaisten antama opetus, oleskelu Suomessa, opettajan äidinkieli
- käytetyt oppikirjat

Teksti
- tekstilaji ja kirjoituksen tehtävänanto
- ajankäyttö: rajattu vai rajaamaton
- testiluonteisuus: harjoitustehtävä vai koe
- apuvälineiden käyttö: sanakirjat ym.
- kirjoituspaikka: kotona, luokassa, muualla
- taitotaso: EVK:n mukaan

Muut
- keräyspaikka ja -aika
- medium: käsin tai tekstinkäsittelyohjelmalla kirjoitettu

Ehdottomasti taajimmin tutkimuksissa huomioon otettuja taustamuuttujia ovat opiskelijan äidinkieli ja taitotaso. Tällä hetkellä oppijoiden äidinkielen mukaan koostetuista osakorpuksista kahdeksan (viron-, venäjän-, saksan-, puolan-, ruotsin-, kiinan-, tsekin- ja hollanninkieliset) muodostaa niin laajan osakorpuksen, että niiden avulla voidaan tehdä esimerkiksi transferia koskevaa tutkimusta. Muut osakorpukset ovat sen verran pieniä, etteivät ne sovellu yksinään tämäntyyppiseen tutkimukseen. Toki niitä voi käyttää mukana sellaisessa tutkimuksessa, jossa äidinkielellä ei ole roolia vaan tarvitaan vain suurta oppijankielen tekstimassaa, tai tutkimuksessa, jossa tarvitaan yhden äidinkielen osakorpuksen lisäksi laaja verrannollinen yleinen oppijankieliaineisto (ks. esim. Jantunen & Brunni 2012). Kuten taulukosta 1 havaitsee, suurin osa aineistosta sijoittuu Eurooppalaisen viitekehyksen (EVK) taitotasoille A2–C1. Tämä johtuu jo pelkästään kielen oppimisesta ja opetusmenetelmistä, sillä aivan alkeistasolla oppijat eivät vielä kirjoita sellaisia tekstejä, joita aineistoon on kerätty. On kuitenkin huomattava, että EVK:n mukaan annettu taitotaso on nimenomaan tekstin – ei opiskelijan – tasoa kuvaava: jokainen teksti on saanut vähintään kaksi erillistä arvioita, ja saman kirjoittajan tekstit voivat myös poiketa tasoltaan toisistaan. (Jos teksti on saanut kaksi eri arviota, on pyydetty lisäksi vielä kolmas arvio.) EVK:n mukaisen taitotasoarvion lisäksi metadatassa on myös tieto opiskelijan saamasta opetuksen määrästä, joka voidaan haluttaessa ottaa huomioon taitotasoa kuvatessa. Kaikki edellä kuvattu metadata on kerrottu aineiston nykyisessä formaatissa jokaisen tekstin metadataosiossa ennen varsinaista kirjoitussuoritusta.

Korpusaineistoja kuvataan usein erilaisten luokittelupiirteiden avulla. Myös oppijankielen aineistoja voidaan luokitella erilaisten dimensioiden mukaan; kattavin oppijankieliaineistoja koskeva dimensioluokittelu löytynee toistaiseksi Jantusen (2011) kuvauksesta. Dimensioiden mukaan (luettelossa vasemmalla) ICLFI voidaan luokitella seuraavasti:

genre:	monitekstilajinen korpus
teema:	yleiskorpus
rekisteri:	kirjoitetun kielen korpus
kieli:	yksikielinen korpus
variantti (verrannollisuus):	ei-verrannollinen korpus (ei natiiviaineistoa)
kääntäminen:	ei-käännöskorpus
aika:	synkroninen (osittain diakroninen) korpus
otanta:	kokotekstikorpus
medium:	käsin ja tekstinkäsittelyohjelmalla kirjoitettuja tekstejä

annotointi:	raakateksti- ja annotoitu versio
äidinkieli:	moniäidinkielinen korpus
taitotaso:	monitaitotasoinen korpus (A1–C2)
oppimiskonteksti:	vieraan kielen korpus
oppimismenetelmä:	formaali oppiminen

3.1 Morfologinen annotointiprosessi

ICLFI-aineiston morfologinen annotointi on monesta eri vaiheesta koostuva prosessi, jossa raakateksti lemmatisoidaan ja siihen lisätään merkitsimillä kieliopillista informaatiota. Kuten kieliopillisessa annotoinnissa yleensäkin (ks. Leech 1997a: 8), prosessi on osin automaattinen, mutta aineiston käsittely sisältää myös manuaalisen tarkastuksen. (Annotoinnin yleisistä prosesseista ks. esim. Heikkinen ym. 2012.) Oppijansuomen morfologinen annotointi on aikaa vievää siitä syystä, että tietokonesovelluksella tehty automaattinen analyysi ei anna yhtä hyvää tulosta oppijoiden tuottamasta materiaalista kuin natiiviaineistosta (ks. De Haan 2000: 71). Lisäksi oppijoille vaikea suomen kielen morfologia johtaa virheellisiin leksikaalisten ja kieliopillisten morfeemien yhdistelmiin, jotka taas automaattinen analysaattori tulkitsee helposti väärin. Täysin manuaalisesti annotointi on aivan liian työläs toteuttaa (ks. myös Jelínek, Štindlová, Rosen & Hana 1999: 132–133), joten ICLFI on koodattu puoliautomaattisesti.

Oppijankielen annotoinnissa erityisen ongelmallisia ovat oppijoiden tuottamat virheelliset muodot. ICLFI:n annotoinnissa tämä ongelma on ratkaistu siirtämällä tekstitiedosto Microsoft Word -tekstinkäsittelyohjelmaan jo ennen automaattista koodausta. Tässä vaiheessa tekstistä poistetaan oikeinkirjoitusvirheet ja taivutusmuodoissa esiintyvät ongelmat. Microsoft Word auttaa tässä merkitsemällä kyseiset kohdat automaattisesti virheellisiksi (esimerkiksi kvantiteetti- ja astevaihteluvirheet), jotka sijoitetaan kulmasulkeiden sisään, jolloin annotointiohjelma jättää sulkeiden sisäpuolisen aineiston huomiotta. Tämän vaiheen tarkoituksena on muokata tuotettua tekstiä juuri sen verran, että annotointiohjelma osaa lukea sanan ja jäsentää sen (esimerkit 1 ja 2). (Jantunen 2011: 98.)

(1) *Minun <kodussa> kodissa monet kirjat.*
(2) *<Sängi> Sänky on iso ja mukava.*

Tarkoitus ei ole korjata virhettä muuttamalla sanaa tai taivuttamalla sitä kontekstiin paremmin soveltuvaksi. Vaikka taivutus olisi kontekstiin sopimaton ja Microsoft Word todennäköisesti merkitsee ongelmakohdat, ei virhettä kuitenkaan korjata annotointiohjelmaa varten (esimerkki 3). Virheitä korjataan siis mahdollisimman vähän.

(3) *Menen*
ostamaan
valkosipuliin (**illatiivi**, ei genetiivi)

Virhekorjauksen jälkeen tiedosto kopioidaan Connexorin Fi-fdg-jäsentimeen (Järvinen, Laari, Lahtinen, Paajanen, Paljakka, Soininen & Tapanainen 2004; ks. myös Heikkinen ym. 2012), minkä jälkeen tiedosto tuodaan takaisin ICLFI-kiintolevylle annotoituna. Tällöin jokainen annotoitu tiedosto erotetaan vastaavasta raakatekstitiedostosta ja sijoitetaan oikeaan kansioon. Tämän vaiheen tulos on lemmatisoitu ja morfosyntaktisesti koodattu teksti. Automaattisen koodauksen jälkeen täytyy tekstintuottajan kirjoittamat virheet palauttaa tekstitiedostoon alkuperäisessä asussaan.

Koska automaattisen koodauksen jäljiltä tekstiin jää jonkin verran virheitä, viimeisenä työvaiheena on morfosyntaktisen koodauksen tarkistaminen manuaalisesti. Tämä onkin annotoinnin hitaimmin etenevä vaihe. Jäsennin tarjoaa monin paikoin useita vaihtoehtoisia koodauksia yhdelle muodolle, jolloin annotoijan tehtävänä on poimia oikea vaihtoehto manuaalisesti. Vaihtoehtoisia morfologisia tulkintoja voi myös lisätä ongelmakohtiin. Esimerkiksi seuraavassa tapauksessa (4) tulkintana on 1. infinitiivi ja persoonapääte tai yksikön 1. persoonan taivutusmuoto (annotointimerkinnät ja -selitykset tarkemmin liitteessä 1). Annotoija voi myös valmista analyysia tarkastaessaan lisätä useita morfologisia tulkintavaihtoehtoja (esimerkki 5), joiden perusteella tutkija voi myöhemmin tehdä erilaisia hakuja valmiista aineistosta. (Lehto, Brunni & Jantunen 2013.)

(4)	*Minä*	@NH PRON SG P1 NOM	
	katsoan	@MAIN V ACT INF F1 SG P1	
		@MAIN V ACT IND PRES SG P1	
	televisiota	@NH N SG PTV	
(5)	*tulevana*	@PREMOD N SG ESS	
	vuonna	@NH N SG ESS	
	touhikuussa	touhikuussa	@Heur
		toukokuu	@NH N SG INE

Tarkastusprosessin tulos on alkuperäinen teksti lemmatisoituna ja kieliopillisesti koodattuna. Lemmaksi merkitään se, joka on tekstistä ilmeisimmin nähtävissä, ei sitä, joka mahdollisesti sopii paremmin kontekstiin (6):

(6)	*Kotini*	koti	@NH N SG NOM CLI POSS P1
	sijoittaa	**sijoittaa** (ei sijaitsee)	@MAIN V ACT IND PRES SG P3
	Tartossa	Tartto	@NH N SG INE Prop

Pienet taivutus- tai kirjoitusvirheet (esim. astevaihtelu- ja kvantiteettivirheet) eivät muuta lemmaa, jos siitä ei ole epäselvyyttä kontekstin perusteella (7).

(7)	*Jouluna*	joulu	@NH N SG ESS
	me	me	@NH PRON PL P1 NOM
	onneksi	onneksi	@ADVL ADV
	tapamme	**tavata** (ei tappaa)	@MAIN V ACT IND PRES PL P1
	kaikki	kaikki	@NH PRON NOM

Vieraskielisyydet sekä tunnistamattomat sanat merkitään Heur-koodilla (8). Tämä ei kuitenkaan koske lainasanoja, vaikka ne joissain tapauksissa olisivatkin kielenoppijoiden itsensä virheellisesti tuottamia.

(8)	*On*	@MAIN V ACT IND PRES SG P3
	muodostunut	@MAIN V ACT PCP PAST
	kielibarjääri	@NH Heur N SG NOM

Ohjelma ei välttämättä tunnista kirjoja, tv-sarjoja tai elokuvia erisnimiksi, joten tarvittaessa niihin lisätään manuaalisesti erisnimeä merkitsevä Prop-koodi. Tervehdykset ja huudahdukset (*hei, huomenta, moi*) merkitään interjektioiksi INTERJ-koodilla. Sen sijaan puhekielisille sanoille ei ole olemassa erillistä merkintätapaa.

Connexor-annotointiohjelma tekee joitain toistuvia virheitä. Sovellus tulkitsee esimerkiksi virkkeenalkuisen *minä*-pronominin *mikä*-pronominin essiiviksi. Homonymiatapauksista löytyy usein Connexorin tekemiä virheitä tai kaksi erillistä tulkintaa, jotka pitää korjata tai disambiguoida. Virkkeenalkuisia sanoja ohjelma jäsentää joissain tapauksissa propreiksi ja merkitsee teonnimien lemmat paikoin automaattisesti kantaverbien perusmuodoiksi, joten ne on korjattava substantiiveiksi. Ohjelma ei myöskään tunnista adverbien komparaatioita eikä merkitse automaattisesti A-infinitiivin translatiivia. *Nen*-päätteiset substantiivit, kuten erisnimet ja kansallisuudet, jäsentyvät usein adjektiiveiksi. On tärkeää myös tarkistaa mahdollinen sananmuodon leksikaalistuminen tai partisiipin adjektiivistuminen, joita sovellus ei ota huomioon. Joissain tilanteissa Connexor jäsentää vain genetiivin kanssa esiintyvän adposition oikein. Sekä adpositioiden että adverbien suhteen on tärkeää huomioida konteksti, sillä kyseessä saattaa olla tilanteesta riippuen kumpi tahansa. (Ks. lisää annotointiohjelmien tekemistä virheistä Heikkinen ym. 2012.)

Morfosyntaktisessa annotoinnissa pyritään noudattamaan Ison suomen kieliopin (ISK 2004) esittämää luokittelutapaa. Joissakin tapauksissa siitä on kuitenkin poikettu. Esimerkiksi partikkeleille ei ole omaa erillistä merkintätapaa (paitsi aiemmin mainituille interjektioille), ja järjestyslukujen katsotaan kuuluvan numeraaleihin adjektiivien sijasta. Myös pre- tai postpositioiden täydennykset merkitään edussanoiksi koodauksen helpottamiseksi. Kieliopillisen annotoinnin yleiset periaatteet sekä mahdolliset poikkea-

mat ISK:n esittämästä luokituksesta on kirjattu ICLFI-projektin annotointimanuaaliin (ei julkaistu).

3.2 Virheannotointi

ICLFI:n virheannotointisysteemin luominen on aloitettu vuoden 2013 alussa, ja tähän mennessä on saatu aikaan toimiva luokitus sekä virhekoodisto. Virheannotoituja tekstejä on tällä hetkellä nelisensataa kappaletta, noin 48 000 sanetta, mikä on noin viisi prosenttia ICLFI:n kokonaissanemäärästä. Virheannotoidut tekstit ovat äidinkieleltään ruotsin-, hollannin- ja vironkielisten opiskelijoiden kirjoittamia. Virheiden merkitseminen ja korjaaminen on aikaa vievää työtä (Dagneaux ym. 1998), ja virheiden koodaus ICLFI-aineistoon on tällä hetkellä täysin manuaalista: virheet merkitään suoraan tekstitiedostoon eikä virheiden merkitsemisen ja korjaamisen apuna ole toistaiseksi tätä tarkoitusta varten luotuja työkaluja (vrt. esim. ICLE:n virhe-editori, Granger 2002: 19–20).

ICLFI:n näkökulmasta tärkeitä ovat olleet suomen ja sen sukukielten oppijankielen korpukset: niiden virheluokitukset soveltuvat samankaltaisen rikkaan morfologiansa takia paremmiksi vertailukohdiksi kuin esimerkiksi englannin pohjalta tehdyt. EVKK:n (*Eesti vahekeele korpus*, Eslon & Metslang 2007), Indianan yliopiston opiskelijoiden parissa kerätyn oppijanunkarin korpuksen (Dickinson & Ledbetter 2012) sekä Edistyneiden suomenoppijoiden korpuksen (LAS2, Ivaska & Siitonen 2009) virheannotointisysteemit ja -luokitukset poikkeavat toisistaan, mutta niitä on kuitenkin voitu hyödyntää ICLFI:n virheannotointia suunniteltaessa.

3.2.1 Virheiden luokittelu

ICLFI:n virheluokitus perustuu virheiden luonteeseen eli siihen, ovatko ne esimerkiksi sanastollisia vai syntaktisia virheitä (ks. lisää virhetyypeistä esim. Granger 2002: 19). Luokituksen suunnittelussa ja luomisessa on otettu huomioon suomen kielen morfologinen erityislaatu. ICLFI:n virheluokitus on hierarkkinen ja kattaa kaikki kielen tasot fonologiasta syntaksiin, sanastoon ja fraseologiaan asti. Yksi virheiden yläkategoria on esimerkiksi morfosyntaktiset virheet, ja sen alla ovat muun muassa objektin luku- ja sijavirheet. (Ks. muiden korpusten virhekategorioista esim. Granger 2003: 467.)

Muissa korpuksissa (esim. LAS2 ja ICLE) käytettyjen virheannotointisysteemien tarkastelun jälkeen on tehty virheannotointipilotointeja pienillä aineistoilla, minkä jälkeen on alettu suunnitella luokitusta ja itse virhekoodistoa. Tämänhetkinen virheluokitus on esitetty taulukossa 2.

TAULUKKO 2. ICLFI:n virheluokitus.

1 ORTOGRAFISET	1A oikeinkirjoitus 1B välimerkit 1C yhdistäminen
2 FONOLOGISET	2A kvantiteetti 2B diftongi
3 MORFOFONOLOGISET	3A astevaihtelu 3B vokaalisointu
4 MORFOLOGISET	4A nominintaivutus, muoto 4B nominintaivutus, käyttö 4C verbintaivutus, muoto 4D verbintaivutus, käyttö 4E vaillinaisesti taipuvat, muoto 4F vaillinaisesti taipuvat, käyttö
5 MORFOSYNTAKTISET	5A viittaussuhde, possessiivisuffiksi 5B kongruenssi 5C subjektin sija ja luku 5D objektin sija ja luku 5E predikatiivi sija ja luku 5F adverbiaalin sija ja luku 5G rektio
6 SYNTAKTISET	6A sanajärjestys 6B lauseenvastikkeet 6C lauseke 6D lausetyyppi 6E ylimääräinen sana
7 LEKSIKAALISET	7A nominin muodostus 7B verbin muodostus 7C sananvalinta 7D uudismuodoste 7E tyyli ja rekisteri 7F vierassana 7G sana puuttuu
8 FRASEOLOGIA	8A fraseologia
9 SELITTÄMÄTÖN	9A selittämätön

Virhekategorioiden heikkoutena on pidetty muun muassa sitä, että ne ovat huonosti määriteltyjä, subjektiivisia ja perustuvat sekalaisiin kriteereihin (Dagneaux ym. 1998: 164). Yksi toimivan virheannotointisysteemin perusteista onkin järjestelmän sisäinen yhtenäisyys. Virheiden tarkka kuvaus ja virhekategorioiden määrittely sekä koodaamisen periaatteet tulisi ilmaista virhekoodausmanuaalissa, jotta koodaajasta johtuvaa subjektiivisuutta saadaan häivytettyä. (Granger 2003: 467.) Näin on tehty myös ICLFI:n virheannotointisysteemissä. Lisäksi aineiston luotettavuutta on lisätty päättämällä virheannotoinnin ratkaisuista usean tutkijan kesken. ICLFI:n virheannotointimanuaali on

koottu siinä vaiheessa, kun virheannotointia on jo jonkin verran tehty. Näin on saatu kuvaus siitä, mitä virheluokat sisältävät, ja samalla luokitusta on voitu tarkentaa ja parantaa.

ICLFI:n virhekoodi on pyritty tekemään universaaliin muotoon, sillä Grangerin (2003: 467) mukaan virhekategorioiden tulisi olla tarpeeksi yleisiä, jotta niitä voisi hyödyntää useiden eri kielten tarpeisiin. Oppijankielen korpukset ovat kuitenkin rajoittuneet tiettyihin kieliin ja morfologisesti rikkaiden kielten osalta virheannotointia ei ole juuri tehty (Dickinson & Ledbetter 2012). Esimerkiksi indoeurooppalaisia kieliä varten luodut systeemit (ks. esim. ICLE, Granger ym. 2002) eivät ole sellaisenaan soveltuneet ICLFI:n virheluokituksen pohjaksi. Tämä johtuu siitä, että suomen kaltaisten agglutinoivien kielten virheet ja virheiden kokonaisuus ovat erilaisia esimerkiksi fuusiokieliin verrattuna (Dickinson & Ledbetter 2012). Oppija on esimerkiksi sanoja taivuttaessaan ja morfeemeja yhdistellessään voinut tuottaa muotoja, jotka eivät ole virhetulkinnaltaan yksiselitteisiä tai läpinäkyviä.

ICLFI:ssä esiintyvät virheet voi koodauksen vaativuuden näkökulmasta jakaa karkeasti kolmeen luokkaan (Lehto ym. 2013):

1. Selkeät ja helposti luokiteltavat virheet
   Esimerkiksi vokaalisoinnussa tai possessiivisuffiksin käytössä tehdyt virheet kuuluvat tähän luokkaan.

*Ensimmäisessä **kerroksessä***	vokaalisointuvirhe
*Minun **huoneella** on yksi ikkuna*	väärä adverbiaalin sijavalinta ja puuttuva possessiivisuffiksi

2. Virheet, joissa on useita tulkintavaihtoehtoja
   Näissä tapauksissa on usein vaikea tietää varmasti, mihin luokkaan virhe tulisi sijoittaa.

*Mielipiteet **jakaavat***	Kyseessä voi olla joko kvantiteettivirhe tai verbin taivutusvirhe, jossa kielenoppija on liittänyt persoonapäätteen infinitiivimuotoon.
*Minun syö **makkara***	Joko objektin sija- tai kvantiteettivirhe.

3. Tapaukset, joissa ei varmuutta, mistä virheestä on kysymys
   Lause on usein niin epäselvä, että virhettä ei voi luokitella, ja usein kontekstistakaan ei ole apua. Vaarana on liika tulkitseminen oletetun tavoitemuodon pohjalta.

Kaupungin takaisin lähtööni maanantaina.	Onko kyseessä lause vai lauseke?
Mutta myös minä on työssä italialaiselta, suomalaiselta ja historialta.	Konteksti ei paljasta, mistä on kysymys.

Kuten edellisistä esimerkeistä huomaa, ICLFI:n virheannotointisysteemissä on mahdollista huomioida virheiden päällekkäisyys. Miltonin ja Chowdhuryn (1994) mukaan virheiden luokittelu on epävarmaa, koska aina ei ole mahdollista nimetä virhettä selvästi yhteen virheluokkaan kuuluvaksi. Heidän mukaansa koodaus pitäisi luoda sellaiseksi, että se mahdollistaa useampien tulkintamahdollisuuksien lisäämisen (ks. myös Granger 2003: 467). Tätä periaatetta on noudatettu ICLFI:n virheannotoinnissa: sen virheluokituksessa yksi virhe voikin kuulua useampaan eri luokkaan riippuen esimerkiksi siitä, millä kielen tasolla virhettä tarkastellaan (erityisesti 2. luokkaan kuuluvat virheet). Virheiden koodauksessa tämä on huomioitu antamalla virhekoodissa vaihtoehtoja. Epäselvissä tapauksissa siis lopulta jää aineistoa käyttävän tutkijan päätettäväksi, mikä virhe on kyseessä. Osa vaihtoehtoisista tulkinnoista voi näyttää turhilta, mutta todellisuudessa koodaaja ei voi kuitenkaan edes tuotetun muodon perusteella varmasti tietää, minkä virheen kielenoppija on tehnyt. Milton ja Chowdhury (1994: 129) toteavat, että vaikka kaikki oleelliset tulkinnat yritettäisiin lisätä, analyysissa päästään tuskin koskaan kaikki mahdolliset vaihtoehdot kattavaan virheannotointiin.

Virheannotointi tapahtuu käytännössä siten, että tekstitiedostoon lisätään virhekoodi morfologisten koodien perään. Virhekoodi sisältää tavoitemuodon eli korjauksen (jos mahdollista), virheen luokan ja tavoitemuodon morfologisen tulkinnan (9):

(9)	*Minä*	minä	<@subj_PRON_SG_P1_NOM>
	en	ei	<@pred_Aux_V_ACT_SG_P1>
	tarvitse	tarvita	<@V_ACT_PRES_NEG>
	kengät	kenkä	<@obj_N_PL_NOM>
			<err=U 'kenkiä'_MSYN_OBJ_PL_PTV>

Esimerkistä 9 näkee virhekoodin rakenteen: ensin on ilmoitettu, onko kyseessä muotoon vai käyttöön liittyvä virhe (F=form ,U=use). Seuraavaksi tulee tavoitemuoto *kenkiä*, koska kieltolauseen objektin tulee olla partitiivissa. Koodin osa MSYN OBJ kertoo virheen luokan: yläluokka morfosyntaksi (MSYN) ja sen alaluokka objektin sija- ja/tai lukuvirhe (OBJ). Lopuksi käy vielä ilmi, että virheannotoijan kirjaama tavoitemuoto on monikon partitiivi (PL PTV). ICLFI:n kieliopillinen ja virheannotointi tarjoavat siten monipuolisia hakumahdollisuuksia: tietoa voi hakea morfologisella tai virhekoodilla erikseen tai koodit voi yhdistää hakukomentoon. Edellinen esimerkki havainnollistaa, kuinka vaikkapa objektivirheitä voi hakea yhdistämällä haussa tiedon opiskelijan tuottamasta muodosta

(kieliopillinen annotaatio, PL NOM) ja tiedon siitä, mikä muodon olisi pitänyt olla (virhekoodi, PL PTV).

3.2.2 Virheannotoinnin ongelmia ja ratkaisuja

Grangerin (2003: 467) mukaan virheannotoinnin tulisi olla edellä esitettyjen kriteerien lisäksi myös informatiivista eli tarpeeksi yksityiskohtaista, jotta se voi tarjota tarpeeksi tietoa oppijan tekemistä virheistä. Informatiivisuuden täytyisi kuitenkin pysyä sellaisissa rajoissa, että annotoijan on helppo käyttää virheannotointiluokittelua virheanalyysin pohjana. ICLFI:n virheannotointisysteemi on pyritty muokkaamaan sellaiseksi, että se tarjoaisi tutkijalle tarpeeksi vaihtoehtoja, mutta samalla karsisi pois turhat ja epäolennaiset virhetulkinnat. Kun virheannotointiprosessin edetessä esiin nousseita ongelmia on tarkasteltu ja niihin on harkittu ratkaisuja, on saatu muodostettua joitakin suuntaviivoja, joiden mukaisesti annotointia tehdään. Nämä suuntaviivat ovat: kontekstin huomioiminen, yksinkertaisuus virheiden tulkinnassa, virheiden kertautumisen välttäminen ja täsmällisyys.

Kontekstin huomioiminen tarkoittaa kontekstin käyttämistä tulkinnan apuna: koodataan virhe sen mukaan, mikä se todennäköisimmin on. Kontekstin mukainen tulkinta on ensisijainen, mutta ylitulkitsemista tulee välttää. **Yksinkertaisuus** taas tähtää virheen mahdollisimman helpon tulkinnan löytämiseen. Jos virheestä näkee suoraan, mihin luokkaan se kuuluu, ei tulkintoja pidä lähteä tekemään liian monimutkaisin perustein.

ICLFI:n virhekoodauksessa ei huomioida enempää virheitä kuin on pakko, eli siinä vältetään kielenoppijan tekemien virheiden **kertautumista**. Jos esimerkiksi määrite kongruoi edussanansa kanssa, mutta edussanan sijamuoto on väärin, merkitään virhe ainoastaan edussanaan eikä enää määritteeseen. **Täsmällisyys** taas tarkoittaa virheannotoinnissa sitä, että toiset virheluokat kertovat virheen laadusta enemmän kuin toiset. Esimerkiksi sananmuodostusvirhe on diftongivirheen luokkaa epämääräisempi. Täsmällisyyden periaate ei kuitenkaan poista sitä, että virheissä on päällekkäisyyttä, jolloin yksi tulkinta ei ole toista parempi.

Kuten aiemmin on tullut ilmi, ICLFI:n virheannotointisysteemi mahdollistaa useiden vaihtoehtoisten tulkintojen lisäämisen tekstiin. Samoin systeemi huomioi yhdessä sanassa esiintyvät useat virheet: ne kaikki koodataan näkyviin (ks. aiheesta myös Granger ym. 2002). ICLFI:n virheitä ei ole koodattu morfeemi morfeemilta, kuten ei ole koodattu kieliopillista tietoakaan (vrt. Dickinson & Ledbetter 2012), joten tutkijan on lähes välttämätöntä osata suomen kieltä, jotta hän voi havaita virheen sijainnin sanassa.

4 Lopuksi

Edellä esitetystä oppijankieliaineiston annotointiprosessin kuvauksesta käy eittämättä ilmi, että annotointi ei ole yksinkertaista. Oppijankielimateriaalin annotointi on niin sanottua natiiviaineistoa ongelmallisempaa erityisesti siksi, että kielenoppijoiden tuottamat muodot poikkeavat usein suurestikin kohdekielen tavoitemuodoista. Koska kieliopillinen annotaatio riittää harvoin oppijankieliaineiston kuvaukseen, on rinnalle tehtävä virheannotaatio. Huolimatta manuaalisen osuuden työläydestä oppijankieliaineistolle kannattaa tehdä sekä kieliopillinen annotaatio että virhekoodaus, koska aineistojen käytettävyys paranee prosessien tuloksena huomattavasti. ICLFI:n annotointiprosesseissa oleellista onkin aineiston muokkaaminen siihen muotoon, että tutkijat saavat siitä mahdollisimman paljon ja täsmällistä tietoa. Morfologinen ja virheannotointi täydentävät toisiaan: niiden sisältämää tietoa voi esim. yhdistää aineistoon kohdistuvissa hakulausekkeissa. Näin aineistoihin voidaan kohdistaa hyvinkin erilaisia tutkimuskysymyksiä erilaisten metodien avulla ja niistä voidaan saada oppijankieltä laadullisesti ja määrällisesti kuvaavaa tietoa.

Kirjallisuus

Bateman, J., J. Forrest & T. Willis 1997. The use of syntactic annotation tools: partial and full parsing. Teoksessa R. Garside, G. Leech & A. McEnery (toim.) *Corpus annotation. Linguistic information from computer text corpora.* New York: Longman, 1–18.

Dagneaux, E., S. Dennes & S. Granger 1998. Computer-aided error analysis. *System*, *26*, 163–174.

Dickinson, M. & S. Ledbetter 2012. Annotating errors in Hungarian learner corpus. Teoksessa N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (toim.) *Proceedings of the 8th language resources and evaluation conference (LREC 2012),* Istanbul, Turkey. European Language Resources Association (ELRA), 1659–1664. Saatavissa: http://www.lrec-conf.org/proceedings/lrec2012/pdf/758_Paper.pdf.

Díaz-Negrillo, A. & J. Fernandes-Dominguez 2006. Error tagging systems for learner corpora. *Resla*, *19*, 83–102.

Díaz-Negrillo, A., D. Meurers, S. Valer & H. Wunsch 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, *36* (1–2) (Special issue: corpus linguistics for teaching and learning. In honour of John Sinclair), 139–154. Saatavissa: http://www.sfs.uni-tuebingen.de/~dm/papers/diaz-negrillo-et- al-09.html.

Ellis, R. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

Eslon, P. & H. Metslang 2007. Õppijakeel ja eesti vahekeele korpus. Teoksessa H. Metslang, M. Langemets & M.-M. Sepper (toim.) *Eesti rakenduslingvistiika ühingu aastaraamat 3 - Estonian papers in applied linguistics 3*. Tallinn: Eesti Keele Sihtasutus, 99–116.

Garside, R., G. Leech & A. McEnery (toim.) 1997. *Corpus annotation. Linguistic information from computer text corpora.* New York: Longman.

Granger, S. 2002. A bird's-eye view of learner corpus research. Teoksessa S. Granger, J. Hung & S. Petch-Tyson (toim.) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins, 3–33.

Granger, S. 2003. Error-tagged learner corpora and CALL: a promising synergy. *CALICO Journal*, *20* (3), 465–480.

Granger, S., E. Dagneaux & F. Meunier 2002. *International corpus of learner English*. Version 1.1. Université catholique de Louvain: Centre for English Corpus Linguistics.

de Haan, P. 2000. Tagging non-native English with the TOSCA-ICLE tagger. Teoksessa C. Mair & M. Hundt (toim.) *Corpus linguistics and linguistic theory. Papers from the twentieth international conference on English language research on computerized corpora (ICAME 20)*. Amsterdam: Rodopi, 69–79.

Heikkinen, V., M. Lounela & E. Voutilainen 2012. Automaattinen analysaattori tekstilajitutkimuksessa. Teoksessa V. Heikkinen, E. Voutilainen, P. Lauerma, U. Tiililä & M. Lounela (toim.) *Genreanalyysi – tekstilajitutkimuksen käsikirja*. Kotimaisten kielten keskuksen julkaisuja 169. Helsinki: Gaudeamus.

ISK = Hakulinen, A., M. Vilkuna, R. Korhonen, V. Koivisto, T. Heinonen & I. Alho 2004. *Iso suomen kielioppi*. Helsinki: SKS.

Ivaska, I. & K. Siitonen 2009. Syntaktisesti koodattu oppijankielen korpus: mahdollisuuksia ja kysymyksiä. Teoksessa P. Eslon & K. Õim (toim.) *Korpusuuringute metodoloogia ja märgendamise probleemid. Tallinna Ülikooli eesti keele ja kulttuuri instituudi toimetised 11*. Tallinn: Tallinna Ülikooli, 54–71.

Izumi, E., K. Uchimoto & H. Isahara 2005. Error annotation for corpus of Japanese learner English. Teoksessa *Proceedings of 6th International Workshop on Linguistically Interpreted Corpora (LINC-2005)*. Jeju Island, 15 October 2005 (South Korea), 71–80.

Jantunen, J. H. 2011. Kansainvälisen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. Teoksessa A. Kaivapalu, J. Laakso, P. Muikku-Werner & M.-M. Sepper (toim.) *Lähivõrdlusi. Lähivertailuja 21*. Tallinn: Eesti Rakenduslingvistiika Ühing, 86–105.

Jantunen, J. H. & S. Brunni 2012. Morfologinen priming ja fraseologia vieraan kielen oppimisessa: korpustutkimus oppijansuomesta. Teoksessa A. Kaivapalu, P. Muikku-Werner, J. H. Jantunen & M.-M. Sepper (toim.) *Lähivõrdlusi. Lähivertailuja 22*. Tallinn: Eesti Rakenduslingvistiika Ühing, 71–100.

Jelínek, T., B. Štindlová, A. Rosen & J. Hana 1999. Combining manual and automatic annotation of a learner corpus. Teoksessa V. Matousek, P. Mautner, J. Ocelíková & P. Sojka (toim.) *Text, speech and dialogue: second international workshop, TSD'99 Plzen, Czech Republic September 13–17, 1999 Proceedings*. Berlin: Springer, 126–134.

Järvinen, T., M. Laari, T. Lahtinen, S. Paajanen, P. Paljakka, M. Soininen & P. Tapanainen 2004. Robust language analysis components for practical applications. Teoksessa B. Gambäck & K. Jokinen (toim.) *Coling 2004. Proceedings of the workshop 'Robust and adaptive information processing for mobile speech interfaces'*. Riga: The Baltic Perspectives, 53–56. Saatavissa: https://www.academia.edu/3146747/Robust_and_Adaptive_Information_Processing_for_Mobile_Speech_Interfaces.

Karlsson, F., A. Voutilainen, J. Heikkilä & A. Anttila (toim.) 1995. *Constraint grammar: a language-independent system for parsing unrestricted text*. Berlin: Mouton de Gruyter.

Leech, G. 1997a. Introducing corpus annotation. Teoksessa R. Garside, G. Leech & A. McEnery (toim.) *Corpus annotation. Linguistic information from computer text corpora*. New York: Longman, 1–18.

Leech, G. 1997b. Grammatical tagging. Teoksessa R. Garside, G. Leech & A. McEnery (toim.) *Corpus annotation. Linguistic information from computer text corpora*. New York: Longman, 20–33.

Leech, G. 2004. Adding linguistic annotation. Teoksessa M. Wynne (toim.) *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books, 17–29. Saatavissa: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm.

Leech, G. & E. Eyes 1997. Syntactic annotations: treebanks. Teoksessa R. Garside, G. Leech & A. McEnery (toim.) *Corpus annotation. Linguistic information from computer text corpora*. New York: Longman, 34–52.

Lehto, L.-M., S. Brunni & J. H. Jantunen 2013. How to annotate morphologically rich language? Problems and solutions. *Poster presentation in Learner Corpus Research 2013 Conference September 27–29*. Bergen: University of Bergen.

Milton, J. & N. Chowdhury 1994. Tagging the interlanguage of Chinese learners of English. Teoksessa L. Flowerdew & A. Tong (toim.) *Entering text*. Hong Kong: The Hong Kong University of Science and Technology, 127–143.

Pienemann, M. 1998. *Language processing and second language development: processability theory.* Amsterdam: John Benjamins.

Ragheb, M. & M. Dickinson 2012. Defining syntax for learner language annotation. Teoksessa *Proceedings of the 24th international conference on computational linguistics (COLING 2012), Poster Session.* Mumbai, India, 965–974. Saatavissa: http://cl.indiana.edu/~md7/papers/ragheb-dickinson12.pdf.

Rastelli, S. 2009. Learner corpora without error tagging. *Linguistik Online, 38* (2), 57–66. Saatavissa: http://www.linguistik-online.de/38_09/rastelli.pdf.

van Rooy, B. & L. Schäfer 2003. *An evaluation of three POS taggers for the tagging of the Tswana learner English corpus* [luettu 28.11.2013]. Saatavissa: http://www.corpus4u.org/forum/upload/forum/2005092023174960.pdf.

Schmidt, H. 1994. Probabilistic part of speech tagging using decision trees. Teoksessa *Proceedings of the international conference on new methods in language processing*, Manchester, UK. Saatavissa: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf.

Tono, Y. 2003. Learner corpora: design, development and applications. Teoksessa D. Archer, P. Rayson, A. Wilson & T. McEnery (toim.) *Proceedings of the corpus linguistics 2003 conference*. UCREL technical paper number 16. UCREL: Lancaster University, 800–809.

LIITE 1.

Annotointimerkinnät ja -selitykset

#	yhdyssana
(INF F4)	*-minen*-deverbaalisubstantiivi
@ADVL	adverbiaali
@CC	rinnastuskonjunktio
@MAIN	verbi
@NH	edussana
@PREMARK	pre- tai postpositio tai konjunktio
@PREMOD	määrite
<p>	kappale vaihtuu
<s>	virke loppuu
A	adjektiivi
Abbr	abbreviation = lyhenne esim. EUR
ABE	abessiivi
ABL	ablatiivi
ACC	akkusatiivi
ACT / PASS	aktiivi / passiivi
ADE	adessiivi
ADV	adverbi
ALL	allatiivi
Aux	apuverbi (kieltoverbi, *olla*-verbi perfektissä ja pluskvamperfektissä)
CARD / ORD	perusluku / järjestysluku
CLI	liite(partikkeli)
CMP / SUP	komparatiivi / superlatiivi
COM	komitatiivi
CS	alistuskonjunktio
ELA	elatiivi
ESS	essiivi
GEN	genetiivi
Heur	tuntematon sana
ILL	illatiivi
IND/IMP/CND/SUB	indikatiivi/imperatiivi/konditionaali/potentiaali
INE	inessiivi
INF F1	A-infinitiivi
INF F2	E-infinitiivi
INF F3	MA-infinitiivi
INF F5	*-maisillaan*-muoto

INS	instruktiivi
INTERJ	interjektio
KAAN / -KIN / -S / -PA /-HAN / -KO / -KA	
N	substantiivi
NEG	kieltoverbiä seuraava verbi
NOM	nominatiivi
NOM / GEN / PTV...	sijamuoto
NUM	numeraali
P1 / P2 / P3	1.persoona / 2.persoona / 3.persoona
PCP AGT	agenttipartisiippi
PCP PAST	NUT-partisiippi
PCP PRES	VA-partisiippi
POSS P1 / P2 / P3	1. / 2. / 3. persoonan possessiivisuffiksi
POST	postpositio
PREP	prepositio
PRES / PAST	preesens / imperfekti
PRON	pronomini
Prop	erisnimi eli propri
PTV	partitiivi
SG / PL	yksikkö / monikko
TRA	translatiivi
V	verbi