SYNTAKTISESTI KOODATTU OPPIJANKIELEN KORPUS: MAHDOLLISUUKSIA JA KYSYMYKSIÄ

Ilmari Ivaska, Kirsti Siitonen

Abstrakti

Oppijankieleen kohdistuva korpuslingvistiikka on kasvava tutkimusala, joka voi kertoa oppijankielestä paljon sellaisia asioita, jotka ovat aiemmin olleet tutkijoiden ulottumattomissa. Syntaktisesti koodattu korpus laajentaa näitä tutkimusmahdollisuuksia entisestään. Syntaktisesti koodatun korpuksen kehittäminen on kuitenkin pitkä prosessi, jonka laatiminen herättää runsaasti kysymyksiä kaikissa sen työstämisen vaiheissa. Tässä artikkelissa esittelemme Turun yliopiston Lauseopin arkiston osaksi tulevaa Edistyneiden suomenoppijoiden korpuksen (LAS2) toteutusta sekä sen kehittämisessä ilmi tulleita kysymyksiä ja valittuja ratkaisuja. Korpuksen ensimmäinen osa on hakusanoitettu ja se on koodattu niin morfologisesti kuin syntaktisestikin, minkä lisäksi korpukseen liitetään virhekoodaus. Korpus mahdollistaa myös informanttikohtaisen pitkittäistutkimuksen.

Avainsanat: kielen koodaus, korpuslingvistiikka, suomi toisena kielenä, syntaksi

1. Johdanto

Turun yliopiston edistyneiden suomenoppijoiden kielestä koostuvan korpuksen (LAS2) koodaus on aloitettu syksyllä 2008. Koodaus toteutetaan Turun yliopiston Lauseopin arkiston mallin mukaisesti ja aineisto muokataan XML-muotoon (”Extensible Markup Language”, ks. alempana). Korpus tullaan liittämään osaksi Turun yliopiston Lauseopin arkistoa, jossa se on tutkijoiden käytettävissä verkkoselaimen välityksellä.

Koodaaminen tehdään vaiheittain kumuloituvasti ja siitä pyritään tekemään mahdollisimman joustava niin, että edeltävä vaihe luo pohjaa seuraavalle. Ensimmäisessä vaiheessa tarkoituksena on keskittyä morfologisen tason ja sanatason koodaukseen merkitsemällä aineistoon runsaasti morfologista informaatiota kustakin sanasta. Toisessa vaiheessa morfologinen koodaus kontekstualisoidaan, aineisto rakenteistetaan ja siihen merkitään syntaktista informaatiota. Rinnan koodauksen eri vaiheiden kanssa merkitään kaikki standardoidusta yleiskielestä poikkeavat muodot ja käyttöyhteydet kommenttikoodilla. Turun yliopiston Edistyneiden suomenoppijoiden korpus seuraa pääpiirteissään Nobufumi Inaban kehittämää syntaktisesti koodatun korpuksen tekotapaa (Ks. Inaba 2007: 147–161).

2. Koodatun oppijakorpuksen laatiminen

2.1. Aineisto ja tekstilaji

Aineistona on Turun yliopiston suomen ja sen sukukielten maisteriohjelman opiskelijoiden kirjoittamaa opiskeluun liittyvää asiaproosaa. Seuranta-aika on 2–3 vuotta. Opiskelijoiden suomen kielen taito on jo alussa vahva. Lisäksi korpuksen osaksi kootaan ensikielisten suomen kielen opiskelijoiden tenttivastauksista koostuva vertailuaineisto.

Tutkittavana on monipuolinen tekstilajivalikoima, joka sisältää tenttivastauksia, esseitä ja tutkielmia sekä katsauksia ja raportteja. Korpuksen laatimisen ensimmäisessä vaiheessa pääpaino on tenttivastauksissa. Aineisto karttuu jatkuvasti. 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. Katsauksia ja raportteja tulee vain muutama vuosittain, noin 10 000 saneen luokkaa.

Tekstilajilla on luonnollisesti suuri merkitys aineiston laatuun. Esimerkiksi tenttivastauksista voidaan olettaa, että niiden muoto on osittain tenttikysymyksen muodon inspiroima. Ne lienevät myös jossakin määrin stereotyyppisiä niin, että esimerkiksi eri vastausten alut voivat olla hyvinkin samankaltaisia. Kuten Jyrki Kalliokoski huomauttaa, tekstilaji on ennen kaikkea sosiokulttuurinen käsite (Kalliokoski 2006: 240). Tekstilajin hallinta on siis niin ikään kielen omaksumiseen kuuluvaa stereotyyppistä tietoa siitä, miten kielenkäyttäjän on tapana toimia kussakin tilanteessa. Se on osa kielellistä sujuvuutta, jossa keskeistä on variaatio (Mt., 248). Sylvaine Grangerin mukaan korpusten avulla onkin mahdollista tarkastella kielen variaation ja eri tekstilajien yhteyksiä (Granger 2002: 4–5). Vertailuaineiston avulla tällaisten tekstilajikohtaisten ilmiöiden tarkasteleminen on mahdollista.

2.2. Morfologinen koodaus ja sanakirja

Aineisto muokataan, kuten jo johdannossa todettiinkin, XML-muotoon. Teknisesti tämä tarkoittaa html-kielen kaltaisten tagien käyttöä, joiden avulla aineisto rakenteistetaan hierarkkisesti. Koodaus on Kotimaisten kielten tutkimuskeskuksessa Kotuksessa kehitetyn TEI-ohjeistuksen (*The Text Encoding Initiative*) muunnoksen mukainen (Inaba 2007: 151).

Morfologisessa koodauksessa aineistoon merkitään seuraavat tiedot: oikeakielinen hakusanamuoto silloin, kun se on mahdollista (<lemma>[1]); sanaluokka (<pos>); muoto-opilliset metatiedot silloin, kun niiden tulkitseminen on mahdollista (<mrp>); virhekommentit silloin, kun niiden tunnistaminen on mahdollista (<com>). Morfologisista metatiedoista koodataan sanaluokasta riippuen seuraavat seikat: sijamuoto, komparaatiomuoto, luku, omistusliitteet, pääluokka, tempus, modus, persoona ja finiittisyys. Virhekommentoinnissa voidaan tässä vaiheessa huomioida seuraavat ilmiöt: vartalovirheet, astevaihteluvirheet, morfologiset taivutusvirheet ja sanojen sekoittuminen. Esimerkiksi virheellisesti muodostetun *hyvä*-adjektiivin vertailumuoto *hyvempi* koodataan sanakirjavaiheessa seuraavasti:

<w lemma=”hyvä” pos=”a” cmp sg nom" com="virheellinen cmp">hyvempi</w>[2]

Tätä työskentelyvaihetta kutsutaan sanakirjan tekemiseksi. Sanakirjaa tehtäessä koko aineistoa käsitellään sanamuodoittain aakkosjärjestyksessä. Kukin sanamuoto koodataan vain kerran, mistä syystä homonymiatapaukset koodataan kaikin mahdollisin tulkinnoin. Tällöin todennäköisempi tulkinta koodataan sanan morfologiseksi koodaukseksi ja epätodennäköisempi sanan kommentiksi. Tarpeeton koodaus poistetaan sitten, kun koodaus kontekstualisoidaan. Kun esimerkiksi *tulla*-verbin konnegaatiomuoto ja imperatiivimuoto *tule* lankeavat yhteen, on sanan koodaus sanakirjavaiheessa seuraavanlainen:

---

[1] Sulkeissa oleva merkintä kertoo kunkin seikan kohdalla korpuksessa käytettävän XML-tagin muodon.

[2] w = word, hyvä = hakusana, a = adjektiivi, cmp = komparaatio, sg = yksikkö, nom = nominatiivi, virheellinen cmp = virheellisesti muodostettu komparaatio.

<w lemma="tulla" pos="v" mrp="conneg ind pres" com="fin impv pres sg2">tule</w>[3]

Sanakirjan tekeminen on osa koodauksen automaattistamista. Kun kukin sanamuoto on koodattu kerran, samaa morfologista koodausta voidaan käyttää korpuksen karttuessa myös jatkossa. Kun korpusta kasvatetaan lisäämällä siihen uutta digitalisoitua aineistoa, ajetaan uusi aineisto ensimmäisessä vaiheessa sanakirjan kautta. Tämän jälkeen sanakirjaan koodataan ainoastaan ne sanamuodot, joita se ei ennestään sisältänyt. Korpuksen kasvaessa ja sanamuotojen lisääntyessä morfologinen raakakoodaus hioutuu siis lähes automaattiseksi.

2.3. Syntaktinen koodaus

Syntaktisessa koodauksessa morfologinen koodaus kontekstualisoidaan ja aineisto rakenteistetaan informanteittain tekstintuottamisajankohdan, tekstikokonaisuuksien, kappaleiden, virkkeiden ja lauseiden osalta. Samalla aineistoon lisätään syntaktisten funktioiden koodaus ja virhekommentointiin lisätään syntaktisia seikkoja koskevat huomautukset.

Syntaktisen koodauksen aluksi valmis sanakirja syötetään takaisin aineistoon eli morfologinen koodaus kontekstualisoidaan. Koska aineisto on digitalisoitu virkkeittäin ja siihen on merkitty kappaleet ja tekstikokonaisuuksien alkukohdat, voidaan virkkeet (<s>), kappaleet (<p>), tekstikokonaisuudet (<teksti>) sekä tekstintuottoajankohdat (<tentti> ja <paivamaara>) rakenteistaa automaattisesti skriptien eli komentosarjojen avulla. Tämä koodaus ei ole virheetön, mutta se no-

[3] tulla = hakusana, v = verbi, conneg = konnegaatiomuoto, ind = indikatiivi, pres = preesens, fin = finiittimuoto, impv = imperatiivi, sg2 = yksikön 2. persoona.

peuttaa työskentelyä, sillä virheet voidaan korjata käsin kontekstisidonnaista koodausta tehtäessä.

Kontekstisidonnainen koodaus tarkoittaa aineiston jakamista lauseisiin (<cl>) ja sanojen syntaktisen roolin (<fun>) koodausta sekä lauseiden tyypittelemistä myönteisiin ja kielteisiin väite- ja kysymyslauseisiin (<cl type>). Samalla aineistosta poistetaan sanakirjan aiemmin mainitut päällekkäiset morfologiset koodaukset sekä muut vastaan tulevat koodausvirheet. Asetelmissa 1 ja 2 on virke *Merkitys on myös sellainen, että sana ei voi olla kovin vanha, lisäksi ovat vastineet indoeurooppalaisissa kielissä.* ennen kontekstisidonnaista koodausta ja sen jälkeen.

Asetelma 1. Korpuksen virke *Merkitys on myös sellainen, että sana ei voi olla kovin vanha, lisäksi ovat vastineet indoeurooppalaisissa kielissä.* ennen kontekstuaalista koodausta.

```
<s num="9">
<cl type="" fun="" com="">
<w lemma="merkitys" pos="n" mrp="sg nom" fun="" com="">Merkitys</w>
<w lemma="olla" pos="v" mrp="fin ind pres sg3" fun="" com="">on</w>
<w lemma="myös" pos="adv" mrp="" fun="" com="">myös</w>
<w lemma="sellainen" pos="a" mrp="sg nom" fun="" com="">sellainen</w>
<w lemma="" pos="" mrp="" fun="" com="">,</w>
<w lemma="että" pos="cnj" mrp="" fun="" com="">että</w>
<w lemma="sana" pos="n" mrp="sg nom" fun="" com="">sana</w>
```

```
<w lemma="ei" pos="neg" mrp="sg3" fun="" com="">ei</w>

<w lemma="voida" pos="v" mrp="conneg ind pres " fun=""
com="fin ind pres sg3">voi</w>

<w lemma="olla" pos="v" mrp="inf1" fun=""
com="">olla</w>

<w lemma="kovin" pos="adv" mrp="" fun=""
com="">kovin</w>

<w lemma="vanha" pos="a" mrp="sg nom" fun=""
com="">vanha</w>

<w lemma="" pos="" mrp="" fun="" com="">,</w>

<w lemma="lisäksi" pos="p:post" mrp="" fun=""
com="">lisäksi</w>

<w lemma="olla" pos="v" mrp="fin ind pres pl3" fun=""
com="">ovat</w>

<w lemma="vastine" pos="n" mrp="pl nom" fun=""
com="">vastineet</w>

<w lemma="indoeurooppalainen" pos="a" mrp="pl ine"
fun="" com="">indoeurooppalaisissa</w>

<w lemma="kieli" pos="n" mrp="pl ine" fun=""
com="">kielissä</w>

<w lemma="" pos="" mrp="" fun="" com="">.</w>

</cl>

</s>
```

Virke on kontekstuaalisesti koodattaessa jaettu kolmeen lauseeseen ja kunkin lauseen alkuun on merkitty lauseen tyyppi. Kuhunkin sanaan on koodattu sen syntaktinen funktio, toisen

lauseen konnegaatiomuotoisesta verbistä *voi* on poistettu vaihtoehtoinen koodaus ja kolmannen lauseen predikaattiin ja subjektiin on lisätty kommentti niiden odotuksenvastaisesta muodosta. (Ks. asetelma 2).

Asetelma 2. Korpuksen virke *Merkitys on myös sellainen, että sana ei voi olla kovin vanha, lisäksi ovat vastineet indoeurooppalaisissa kielissä.* kontekstuaalisen koodaamisen jälkeen. Koodauksessa tehdyt muutokset on lihavoitu.

<s num="9">

<cl type="**affdecl**" fun="" com="">

<w lemma="merkitys" pos="n" mrp="sg nom" fun="**npsubj**" com="">Merkitys</w>

<w lemma="olla" pos="v" mrp="fin ind pres sg3" fun="**pred**" com="">on</w>

<w lemma="myös" pos="adv" mrp="" fun="**advl**" com="">myös</w>

<w lemma="sellainen" pos="a" mrp="sg nom" fun="**compl:s**" com="">sellainen</w>

<w lemma="" pos="" mrp="" fun="" com="">,</w>

</cl>

<cl type="negdecl" fun="" com="">

<w lemma="että" pos="cnj" mrp="" fun="**lauseyhd**" com="">että</w>

<w lemma="sana" pos="n" mrp="sg nom" fun="**npsubj**" com="">sana</w>

<w lemma="ei" pos="neg" mrp="sg3" fun="**pred**" com="">ei</w>

<w lemma="voida" pos="v" mrp="conneg ind pres" fun="**pred2**" com=**POISTO[""]**>voi</w>

<w lemma="olla" pos="v" mrp="inf1" fun="**pred3**" com="">olla</w>

<w lemma="kovin" pos="adv" mrp="" fun="**nmod**" com="">kovin</w>

<w lemma="vanha" pos="a" mrp="sg nom" fun="**compl:s**" com="">vanha</w>

<w lemma="" pos="" mrp="" fun="" com="">,</w>

</cl>

<cl type="affdecl" fun="" com="">

<w lemma="lisäksi" pos="p:post" mrp="" fun="**advl**" com="">lisäksi</w>

<w lemma="olla" pos="v" mrp="fin ind pres pl3" fun="**pred**" com="**kongr_e**">ovat</w>

<w lemma="vastine" pos="n" mrp="pl nom" fun="**npsubj**" com="**sija_e_subj**">vastineet</w>

<w lemma="indoeurooppalainen" pos="a" mrp="pl ine" fun="**nmod**" com="">indoeurooppalaisissa</w>

<w lemma="kieli" pos="n" mrp="pl ine" fun="**advl**" com="">kielissä</w>

<w lemma="" pos="" mrp="" fun="" com="">.</w>

</cl>

</s>

Kuten aiemmin mainittiin, kieliopillisen informaation lisäksi aineisto identifioidaan informanttikohtaisesti tekstintuottamisajankohdan perusteella. Tämä metodi mahdollistaa pitkittäistarkastelun tekemisen.

Syntaktinen koodaus on huomattavasti vaativampaa kuin morfologinen koodaus. Lauseoppia käsittelevässä tutkimuksessa esitetään hyvin paljon keskenään toisistaan poikkeavia näkemyksiä. Syntaktisessa koodauksessa onkin tehtävä paljon enemmän itsenäisiä päätöksiä ja valintoja kuin morfologisessa koodauksessa. Näin ollen korpuksen syntaktista koodausta aloitettaessa eri ratkaisuista on keskusteltava tarkoin ja työn edetessä kaikki ratkaisut on huolellisesti kirjattava. Tärkein tavoite on korpuksen yhdenmukaisuus ja tehtyjen ratkaisujen transparenttius. Esimerkiksi *haluan tehdä* -tyyppiset verbiketjut on syntaktiselta funktioltaan koodattu Ison suomen kieliopin tulkinnan mukaisesti verbiketjuiksi (koodeina pred, pred2 jne) (ISK: 493–495). Samoin kieltomuodot, kuten *ei voi olla*, koodataan syntaktisesti verbiketjuiksi (pred, pred2, pred3). Muita tulkinnanvaraisia lauseopillisia ilmiöitä ovat esimerkiksi: *meistä tulee opettajia* kaltaisen tuloslauseen subjektipaikkaisen jäsenen *opettajia* syntaktinen funktio, mikä on korpuksessa koodattu subjektinpredikatiiviksi (compl:s); lauseensisäisen infinitiivilausekkeen jäsentäminen, missä on päädytty tulkitsemaan lausekkeen pääsanan edustavan kulloinkin kyseessä olevaa lauseenjäsentä ja muiden jäsentyvän osana infinitiivilauseketta, esim. *Sen* (nmod) *tehtävänä* (advl:p) *on* (pred) *analysoida* (infsubj) *virheitä* (npobj)[4].

Korpukseen tullaan jatkossa soveltamaan Mikael Agricolan teosten tieteellinen editio ja morfosyntaktinen tietokanta -tut-

[4] nmod = substantiivin määrite, advl:p = predikatiiviadverbiaali, pred = predikaatti, infsubj = infinitiivisubjekti, npobj = substantiiviobjekti.

kimushankkeessa kehitettyä koodausjärjestelmää. Inaban mukaan syntaksin puoliautomaattinen koodaus on näillä metodeilla sanaluokasta riippuen parhaimmillaan 80–90 prosentin luokkaa. Adverbien ja adpositioiden osalta koodauksen tarkkuus on yli 90 prosenttia ja adjektiivien osalta noin 80 prosenttia. Substantiivien kohdalla tarkkuus laskee alle 60 prosentin (Inaba 2009). Osittaisesta automatisoimisesta huolimatta koko aineisto tullaan jatkossakin käymään läpi myös manuaalisesti koodauksen tarkistamiseksi ja rakenteistamiseksi, virheiden korjaamiseksi ja ylimääräisten koodirivien karsimiseksi (ks. esim. Inaba 2007). Automatisoitavien prosessien avulla koodaustyö kuitenkin kevenee huomattavasti, mikä mahdollistanee jatkossa koodatun aineiston nopeamman kasvattamisen ja korpuksen luotettavuuden parantamisen.

2.4. Virhetyypittely

Korpukseen merkitään myös virhekoodaus. Tämä koodaus ja siihen liittyvä virhetyypittely tehdään aiemmin kuvattujen koodausvaiheiden aikana tehdyn virhekommentoinnin pohjalta. Näin varmistetaan seuraavassa kappaleessa tarkemmin esiteltävä aineiston laadun kannalta relevantti luokittelu.

Morfologisen ja kontekstuaalisen koodauksen jälkeen niissä merkityt kommentit aineiston normipoikkeamista ja kohdekielen kannalta epätyypillisistä kielellisistä ratkaisuista kerätään yhteen. Näiden kommenttien pohjalta pystytään hahmottamaan aineiston kannalta tarkoituksenmukaisia virheryhmiä niin niiden laadun kuin esiintymien määränkin puolesta. Tämän tyypittelyn perusteella Turun yliopiston Lauseopin arkiston Edistyneiden suomenoppijoiden korpukseen koodataan viisiportainen hierarkkinen virhetyyppiluokitus.

Luokituksen ensimmäinen taso on kunkin tekijän norminmukaisuus tai normipoikkeama. Tämän jälkeen tulee tyypittelyn kuusi pääluokkaa, jotka ovat seuraavat:

1) sanastolliset virheet
2) sanastollis-johto-opilliset virheet
3) sanastollis-morfologiset virheet
4) morfologiset virheet
5) syntaktiset virheet
6) lauserakenteelliset virheet.

Luokittelu on väistämättä joiltakin osin ristikkäinen, sillä monet aineiston normipoikkeamat ovat joiltakin osiltaan monitasoisia. Korpuksessa käytettävässä tyypittelyssä on keskeistä se, että luokitus on toimiva ainoastaan yhdessä korpuksen muun koodauksen kanssa. Näin esimerkiksi objektin sijavalinnassa esiintyvä normipoikkeama merkitään virhekoodauksessa ainoastaan sijavirheenä, mutta koska sanaan on koodattu sen syntaktinen funktio, korpuksesta pystytään tarpeen mukaan tarkastelemaan niin yleistä objektien sijavariaatiota, objektien herkkyyttä normipoikkeamille kuin objektien sijassa esiintyviä normipoikkeamiakin. Virhetyypittelyä ja sen teknistä toteutusta ollaan parhaillaan kehittämässä.

2.5. Subjektiivisen tulkinnan ongelma

Oppijankieltä käsittelevää, syntaktisesti koodattavaa korpusta tehtäessä kohdataan väistämättä lukuisia koodaajien subjektiiviseen tulkintaan liittyviä ongelmia, joista osaa on sivuttu tässä artikkelissa jo aiemmin. Kuten sanojen syntaktisten funktioiden koodaamista käsiteltäessä todettiin, tärkeintä korpuksen käytettävyyden kannalta on tehtyjen ratkaisujen yhdenmukaisuus ja niiden tarkka ja transparentti dokumentointi ja perustelu.

Suurin ja monitulkintaisin ongelma korpuksessa on aiemmin esitelty virhetyypittely ja sen liittäminen osaksi koodausta. Virheen käsite on hyvin monisyinen. Hankkeessa työskentelevät ovat erittäin tietoisia siitä, että välikielen poikkeamia ei pidetä virheinä ja että sellaisessakin kielitaitonäkemyksessä, jossa jokin ilmiö luokitellaan virheeksi, ilmiö ei välttämättä ole virhe kaikissa rekistereissä. Korpuksen myöhempää käyttöä varten on kuitenkin pidetty järkevänä, että äidinkielisen koodaajan kieli-intuition vastaiset ilmaukset merkitään. Myöhemmin korpusta tutkimuksessaan käyttävä ratkaisee suhteensa merkintöihin tapauskohtaisesti. Ennen korpuksen virhekoodauksen julkistamista on välttämätöntä, että sen tarkistaa koodaajan lisäksi vähintään yksi ensikielenään suomea puhuva hankkeen työntekijä.

3. Korpuksen soveltaminen suomi toisena kielenä -tutkimukseen

Syntaktisesti koodattu oppijankorpus pystyy tarjoamaan uutta tietoa sellaisista ilmiöistä, joita on aiemmin ollut työlästä tai vaikeaa tutkia kvantitatiivisin metodein. Syntaktinen koodaus mahdollistaa esimerkiksi lauserakenteiden esiintymisfrekvenssien tarkastelun suhteessa ensikielisten suomenpuhujien lauserakenteisiin sekä tällaisten ilmiöiden esiintymisehtojen ja niiden valintapreferenssien tarkastelun niin itse oppija-aineistossa kuin korpuksen osaksi liitettävässä vertailuaineistossakin (vrt. esim. Jantunen 2004: 15, 29). Keskeistä syntaktisessa koodauksessa on lisäksi se, että sen avulla pystytään luontevasti nostamaan tarkastelun kohteeksi nimenomaan lauseet pelkkien muotojen asemesta. Kuten Granger kirjoittaa, korpusten teho on nimenomaan kielen frekvenssejä tarkasteltaessa, sillä se on ala, josta kielenpuhujilla on hyvin vähän intuitiivista tietoa (Granger 2002: 4).

Hakusanoitettu eli lemmattu aineisto mahdollistaa periaatteessa myös monipuolisen sanastontutkimuksen. Toistaiseksi korpuksen koodatun aineiston laatu kuitenkin rajoittaa tätä tutkimusta, sillä kielitieteeseen painottuvat tenttivastaukset eivät anna todenmukaista kuvaa kielenoppijoiden sanastosta. Kun korpusta laajennetaan myös muihin tekstilajeihin, tämä on mahdollista.

Jarmo Jantunen kirjoittaa sellaisista oppijankielen universaaleista, joiden tutkimiseen korpustutkimus tarjoaa uusia keinoja ja näkökulmia. Näitä ovat hänen mukaansa kielenainesten epätyypilliset frekvenssit, kieltenvälinen vaikutus, yksinkertaisuus ja epäkonventionaalisuus (Jantunen 2008: 70–81). Voidaankin laajasti ottaen ajatella, että syntaktisesti koodattu oppijankielenkorpus mahdollistaa näiden seikkojen tutkimuksen myös lauseopillisten seikkojen osalta. Suurista aineistoista on mahdollista myös nähdä, minkälaisia ensikielen puhujille tyypillisiä piirteitä tai ilmiöitä ei ollenkaan esiinny oppijoiden teksteissä. Näin mahdollistuu siis myös hyvin hankalasti havaittavan välttämisen (*avoidance*) tutkiminen.

Edellä esitetyn lisäksi voidaan korpuksen teksteistä tutkia ylipäänsä kaikkea sellaista, mitä muustakin S2-aineistosta voidaan tutkia. Tutkimusta vain helpottaa se, että kaikki materiaali on valmiiksi järjestettyä ja digitaalisessa muodossa.

Tärkeä seikka korpusta käytettäessä on sen tulevan virhekoodauksen soveltaminen. Virhekoodausta on syytä tulkita kriittisesti eikä sitä voi missään tapauksessa käyttää tutkimushypoteesien suorana toteennäyttäjänä. Virhekoodauksen perusteella ei esimerkiksi ole syytä laskea suoraan virhefrekvenssejä. On muistettava, että virhekoodaus on ennen kaikkea tutkimuksen lähtökohta, ei sen tulos. Sen avulla pystytään osoittamaan laajasta aineistosta edistyneiden oppijoiden suomelle

ominaisia piirteitä, jotka tuntuvat ensikielisen suomenpuhujan intuition vastaisilta. Tämä on kuitenkin vasta lähtökohta, jolle tutkimus voidaan perustaa.

4. Lopuksi

Koodauksen tässä vaiheessa haasteena on hahmottaa kokonaisuus kussakin vaiheessa siten, että tehty työ ei rajoita tulevia, toistaiseksi osin huomaamattomia oivalluksia. Parhaimmassa tapauksessa nyt tehtävä työ helpottaa tulevaa ilman, että siitä koituu varsinaista lisävaivaa missään vaiheessa. Aineisto koodataan lauseopillisesti ja sitä kasvatetaan jatkuvasti niin määrällisesti kuin laadullisestikin. Monipuolinen pitkittäisaineisto tarjoaa ajallisen perspektiivin kunkin informantin tuotoksiin. Sen lisäksi arkisto sisältää lukuisia eri tekstilajeja samoilta informanteilta. Kun nämä seikat saadaan mukaan ja näkyviin korpuksen kokonaisuuteen, korpuksen käytettävyys ja monipuolisuus paranee entisestään.

Arkistoon pyritään tallentamaan myös saman informantin saman tekstin eri versioita, joiden avulla voitaneen tavoittaa esimerkiksi pedagogisesti kiinnostavaa tietoa korjaamisesta sekä fossilisoituneista rakenteista.

Kirjallisuus

Granger, Sylviane 2002. A Bird's-eye view of learner corpus research. – Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching / Toim. S. Granger & J. Hung & S. Petch-Tyson. Philadelphia: John Benjamins, 3–37.

Inaba, Nobufumi 2007. Mikael Agricolan teokset tietokannan muodossa. – Agricolan aika / Toim. K. Häkkinen & T. Vaittinen. Helsinki: BTJ, 147–161.

Inaba, Nobufumi 2009. Re: Koodaamisen automatisoiminen. [Sähköpostiviesti 11.3.2009]

ISK = Hakulinen, Auli & Vilkuna, Maria & Korhonen, Riitta & Koivisto, Vesa & Heinonen, Tarja Riitta & Alho, Irja 2004. Iso suomen kielioppi. SKST 950. Helsinki: Suomalaisen Kirjallisuuden Seura.

Jantunen, Jarmo 2004. Synonymia ja käännössuomi: korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin. Joensuu: Joensuun yliopiston humanistisia julkaisuja.

Jantunen, Jarmo 2008. Haasteita oppijankielen analyysille: oppijankielen universaalit. – Õppijakeele analüüs: Võimalused, probleemid, vajadused / Toim. P. Eslon. Eesti filoloogia osakonna toimetised 10. Tallinn: TLÜ Kirjastus, 67–91.

Kalliokoski, Jyrki 2006. Tekstilajin taju ja toisella kielellä kirjoittaminen. – Genre – tekstilaji / Toim. A. Mäntynen & S. Shore & A. Solin. Tietolipas 213. Helsinki: Suomalaisen Kirjallisuuden Seura, 240–265.

Syntactically encoded corpus of learner language: Opportunities and challenges

Ilmari Ivaska, Kirsti Siitonen

Summary

The encoding of the corpus of advanced Finnish began in autumn 2008 at the University of Turku. The corpus is being modified into the XML-format (”Extensible Markup Language) and it follows the TEI-directions (The Text Encoding Initiative) given by the Research Institute for the Languages of Finland.

The encoding is done cumulatively in order to retain the flexibility of the corpus throughout the progress of the project. The first stage is to focus on the morphological and lexical level. All word types are encoded alphabetically with various morphological information. The lemmatisation of the material is executed synchronously. All the word types are encoded only once and the encoding is, then, extended to every token. This method is called the compiling of the corpus' dictionary.

The second stage is the syntactical encoding, in which the morphological encoding is contextualised separately for each informant. In this phase, the corpus is organised by the date of the texts, text entities, paragraphs, sentences and clauses. Simultaneously, each token is encoded by its syntactical function.

Alongside these steps, the material is also commented in the respect of errors and unidiomatic formal and lexical solutions. In the third step the comments are collected and sorted to generate a corpus-based error tagging system. In these manners the corpus will be labelled with a hierarchical error tagging. However, the concept of an error is highly complex and, thus, the error tagging should be considered solely as the basis of every particular study.

Syntactically encoded learner language corpus can provide new knowledge about the phenomena that have so far remained too demanding to reach with non-computer based quantitative methods. It enables for example the studies of the frequencies of syntactical structures in advanced learner's Finnish.

Keywords: encoding, corpus linguistics, Finnish as a second language, syntax