Suomen viittomakielten korpusta rakentamassa

Juhana Salonen, Anna Puupponen, Ritva Takkinen & Tommi Jantunen
Jyväskylän yliopisto, kieli- ja viestintätieteiden laitos, viittomakielen keskus

Tiivistelmä
Viittomakielikorpusten rakentaminen on lisääntynyt merkittävästi 2000-luvulla: ensimmäiset korpusprojektit käynnistyivät 2000-luvun alussa Australiassa ja Hollannissa, minkä myötä laajoja, koneluettavia aineistokokoelmia on ryhdytty rakentamaan useissa Euroopan maissa 2010-luvulla. Tässä artikkelissa tarkastellaan Suomen viittomakielten, suomalaisen ja suomenruotsalaisen viittomakielen, korpuksen syntyä. Artikkeli esittelee korpuksen rakennusvaiheita eli aineiston keräämistä, käsittelyä, annotointia, pitkäaikaissäilytystä sekä julkaisua tietosuojakysymyksineen. Lisäksi artikkelissa kuvaillaan, miten korpusaineistoa on käytetty ja voidaan hyödyntää viittomakielten tutkimuksessa sekä opetuksessa.

Neljän vuoden mittainen Suomen viittomakielten korpusprojekti käynnistyi Jyväskylän yliopiston viittomakielen keskuksessa vuonna 2014. Projektin aikana kuvattiin keskusteluja ja elisitoituja kertomuksia 91 suomalaista viittomakieltä ja 12 suomenruotsalaista viittomakieltä äidinkielenään käyttävältä, eri puolilla Suomea asuvalta henkilöltä viittomakielisen kuuron projektitutkijan opastuksella. Videomateriaalia kerättiin yhteensä noin 560 tunnin edestä (seitsemästä kamerakulmasta nauhoitetut materiaalit yhteenlaskettuna).

Aineistonkeruun ja editoinnin jälkeen yhteensä 22 suomalaista viittomakieltä äidinkielenään käyttävän kielenoppaan videoaineistoihin on tehty perustason annotaatiot viittoma- ja virketasolla. Annotointivaihe eteni viittomien tunnistamisella, niiden merkitysten erottamisella ja viitotun tekstin ilmauskokonaisuuksien kääntämisellä suomen kielelle. Perusannotointi toteutettiin ELAN-ohjelmalla, jossa viittomia identifioidaan ajallisesti videoon yhteydessä olevien glossien avulla. Annotoinnissa käytettiin lisäksi Suomen Signbank -leksikkotietokantaa, johon ELAN-ohjelman glossit yhdistyvät verkkoyhteyden avulla. Laaja multimodaalinen aineistokokonaisuus täydennettiin metatiedoilla aineiston eri osa-alueista, kuten aineistokokonaisuuden yleisluonteesta, aineistonkeruussa läsnä olleista henkilöistä, videoiden sisällöistä ja video- ja annotaatiotiedostojen muodoista IMDI (ISLE Meta Data Initiative) -standardin mukaisesti. Annotoitu aineisto säilytetään ensisijaisesti Jyväskylän yliopistossa, minkä lisäksi se siirretään maaliskuun 2019 aikana FIN-CLARIN-konsortion

Kielipankkiin pitkäaikaissäilytettäväksi sekä julkaistavaksi kielenoppaiden tutkimussuostumusten ja tietosuoja-asetusten mukaisesti. Kielipankissa julkaistava korpusaineisto sisältää noin 14 tunnin edestä kuudesta kamerakulmasta kuvattua videomateriaalia 21 kielenoppaalta sekä videoihin linkitetyt annotaatiotiedostot ja IMDI-kuvaukset.

Suomen viittomakielten korpuksen luonti kehittää molempien viittomakielten kielellisten ja kulttuuristen piirteiden tutkimusta sekä opetusta. Jyväskylän yliopiston viittomakielen keskuksessa korpusaineiston pohjalta on tehty tähän mennessä useita suomalaiseen viittomakieleen keskittyviä tutkimuksia, minkä lisäksi aineistoa on käytetty myös viittomakieliä vertailevassa tutkimuksessa. Kerätty videoaineisto on ainutlaatuinen kokoelma Suomen viittomakielillä tuotettua kerrontaa ja keskusteluja: materiaali sisältää eri-ikäisten ja eri alueilta tulevien henkilöiden viittomista erilaisissa viestintätilanteissa. Systemaattisen annotoinnin myötä aineisto tulee olemaan merkittävä resurssi tutkimuksen lisäksi viittomakielten opetuksessa, viittomakieliä koskevassa koulutuksessa sekä kielisuunnittelussa.

1 Johdanto

Tässä artikkelissa kuvataan, miten CFINSL-projektissa (Corpus of Finland´s Sign Languages) on ryhdytty rakentamaan Suomen viittomakielten korpusta ja miten ensimmäinen osa korpuksesta on saatettu käytettäväksi. Suomen viittomakielten korpus on multimodaalinen aineistokokonaisuus, joka sisältää videomateriaalia sekä materiaaleihin ajallisesti sidottuja koneluettavia annotaatioita ja metatietoja. Korpusprojektin kuvaukset valmistuivat syksyllä 2017, jolloin oli kuvattu 91 suomalaista viittomakieltä ja 12 suomenruotsalaista viittomakieltä äidinkielenään käyttävän, eri puolilla Suomea asuvan henkilön viittomista. Korpuksen luonti vauhdittaa ja uudentaa molempien viittomakielten kielellisten (sanasto, rakenne, variaatio) ja kulttuuristen (kuurojenyhteisön tavat ja normit) piirteiden tutkimusta sekä kehittää molempien kielten sanakirjatyötä ja opetusta. Suomenruotsalaisen viittomakielen dokumentointi on lisäksi tärkeää kielen elvyttämisen kannalta (ks. De Meulder 2016). Korpusaineistoa taltioidaan FIN-CLARIN-konsortion

Kielipankkiin[1], josta aineistoa voidaan käyttää tutkimus- ja opetustarkoituksiin kielenoppaiden antamien lupien rajoissa.

2 Aineistonkeruu CFINSL-projektissa

Vuonna 2014 käynnistyneen CFINSL-projektin tavoitteena oli kerätä aineistoa yhteensä sadalta suomalaista ja suomenruotsalaista viittomakieltä käyttävältä kielenoppaalta ympäri Suomea. Korpusaineiston keruu tuli ajankohtaiseksi muissa maissa käynnistyneiden projektien johdosta. Ensimmäinen projekti käsitteli australialaista viittomakieltä, jossa materiaalin keruu alkoi vuonna 2004 sadan kielenoppaan voimin (Johnston 2010). Sittemmin ovat seuranneet projektit hollanti- laisesta viittomakielestä (92 kielenopasta vuosina 2006–2008)[2], brittiläisestä viittomakielestä (249 kielenopasta vuosina 2008–2011)[3] ja ruotsalaisesta viittomakielestä (42 kielenopasta vuosina 2009–2011)[4]. Näistä projekteista, erityisesti Ruotsista, saaduista vaikutteista muodostettiin CFINSL-projektin suuntaviivat.

CFINSL-projektin alkuvaiheessa laadittiin suunnitelma siitä, miten etsiä ja tavoittaa kielenoppaita, kuinka tiedottaa viittomakieliselle yhteisölle meneillään olevasta projektista ja miten rakentaa aineiston käsittelyyn vaadittava infrastruktuuri. Korpusprojektin aikana vuosina 2014–2017 aineistonkeruu toteutettiin pääosin Jyväskylän yliopiston AV-studiossa, minkä lisäksi pieni osa aineistosta kuvattiin Helsingissä ja Oulussa. Aineistoa kerättiin seitsemältä pääalueelta, jotka määritettiin Aluehallintoviraston toimialueiden perusteella: Etelä-Suomesta, Lounais-Suomesta, Länsi- ja Sisä-Suomesta, Itä-Suomesta, Pohjois-Suomesta ja Lapista. Lopullinen videoaineisto sisältää 91 suomalaista viittomakieltä (ikäjakauma 18–84 vuotta) ja 12 suomenruotsalaista viittomakieltä (ikäjakauma 27–89 vuotta) äidinkielenään käyttävän kielenoppaan viittomista. Jotta aineisto saatiin edustamaan mahdollisimman kattavasti Suomen viittomakielisen yhteisön keskuudessa käytettyjä kieliä, tehtiin Suomessa asuvista viittojista ennakkokartoitus, jossa selvitettiin muun muassa heidän kuulostatusta (esim. kuuro, huonokuuloinen, kuuleva), syn-

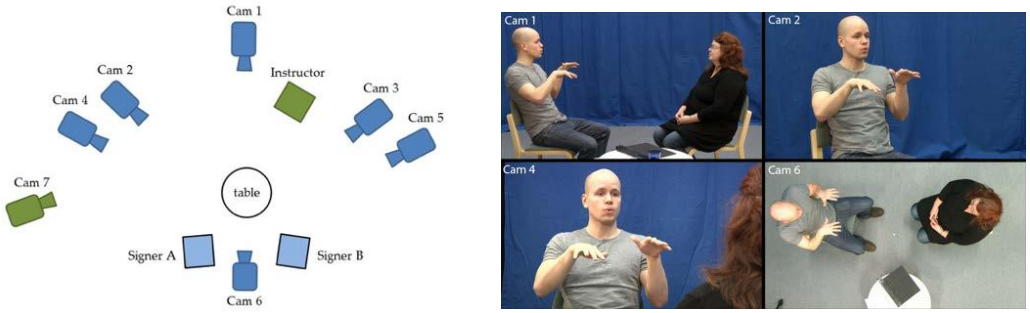
[1] https://www.kielipankki.fi/
[2] https://www.ru.nl/corpusngten/
[3] https://bslcorpusproject.org/project-information/
[4] https://www.ling.su.se/english/research/research-projects/sign-language/swedish-sign-language-corpus-project-1.59270

tymäpaikkaa ja koulutusta. Korpuksen kielenoppaat päätettiin tämän kartoitustyön avulla.

Aineistonkeruuta ja yhteydenpitoa kielenoppaisiin koordinoi viittomakielinen kuuro projektityöntekijä, joka on vastannut tiedottamisen ohella myös kielenoppaiden opastamisesta kuvaussession aikana. Kielenoppaat ovat saaneet valita itselleen oman parin, joka on läheinen tai tuttu esimerkiksi kouluajoilta. Kuvauksiin valituille pareille annettiin etukäteen tietoa yleisistä käytännön asioista, kuten korpusprojektin tavoitteista ja kuvauksille suotuisasta vaatetuksesta, paneutumatta kuitenkaan tarkemmin yksityiskohtiin, jotta voitaisiin välttää etukäteisvalmistelujen vaikuttamista viittojien kielenkäyttöön.

Korpusaineiston kuvauksissa käytettiin seitsemää Full HD-laatuista videokameraa, joista ensimmäinen taltioi yleisnäkymän molemmista viittojista ja toinen sekä neljäs kunkin viittojan yksittäisnäkymän (ks. kuvio 1). Kolmas ja viides ka- mera tallensivat kustakin viittojasta rajatumman lähikuvan ylävartalosta kasvoihin ja kuudes kamera molemmat viittojat lintuperspektiivistä niin, että viittojien pään, vartalon ja käsien syvyyssuuntaisten liikkeiden etäisyyttä on helpompi tarkastella. Seitsemäs kamera tallensi kuvauksissa toimineen viittomakielisen opastajan toimintaa. Videot tallennettiin Material eXhange Format (MXF) -formaattiin ja pakattiin H.264-koodekilla MP4-tiedostoiksi (Puupponen ym. 2014).

Korpusaineiston kuvauksessa kielenoppaat toteuttivat seitsemän eri viestinnällistä tehtävää, joiden yhteenlaskettu kokonaiskesto vaihteli puolestatoista tunnista kahteen tuntiin. Annetut tehtävät muodostuivat keskustelusta ja kerronnasta ja olivat: (1) itsensä esittely, (2) harrastuksesta/työstä kertominen, (3) sarjakuvista viittominen (Ferd´nand-sarjakuvat), (4) videotarinasta viittominen (Mr. Bean sekä Ohukainen ja Paksukainen -videot), (5) kuvakirjasta viittominen (Lumiukko ja Sammakko, missä olet? -kuvakirjat), (6) kuurojen kulttuuriin liittyvästä tapahtumasta keskustelu ja (7) vapaa keskustelu (Salonen ym. 2016). Pareille taattiin keskusteluraha siten, että opastaja vetäytyi tehtävänannon jälkeen taustalle ja AV-teknikot pysyivät erillisessä tarkkaamossa. Tehtävien aikana oli aina mahdollista pyytää tarkennusta viittomakieliseltä opastajalta ja tehtävien välissä pidettiin myös tauko positiivisen ja luonnollisen ilmapiirin saavuttamiseksi.



Kuvio 1. Kuvaustilanne kamera-asetelmineen (a); videoaineistoa eri kuvakulmista (b). (Puupponen ym. 2014; Salonen ym. 2016.)

Kuvausten jälkeen kielenoppailta kerättiin aineiston käyttöön liittyvät suostumukset sekä taustatietoja. Kullekin parille selostettiin tutkimuslupakäytänteet suoma- laisella viittomakielellä ja kielenoppaita pyydettiin täyttämään tutkimuslupa- ja taustatietolomakkeet suomeksi. Tutkimusluvassa tuli vastata joko myöntävästi tai kielteisesti viiteen videoaineiston käyttöön liittyvään kysymykseen, jotka koskivat lupaa (1) käyttää aineistoa tutkimukseen, (2) näyttää aineistosta otteita julkisissa tilaisuuksissa, (3) irrottaa aineistosta kuvia julkaisuja varten, (4) julkaista aineisto kokonaisuudessaan verkossa ja (5) mainita kielenoppaan nimi julkaisuissa. Tausta- tietolomakkeilla kerättiin viittojista tietoa, jonka avulla voidaan verrata muun muassa eri-ikäisten ja eri paikkakunnilla asuvien viittomakielisten käyttämiä kieli- muotoja. Tutkimuslupaa on täydennetty vuonna 2018 lisäkysymyksillä. Näihin sekä metatietojen jatkokäsittelyyn palataan tämän artikkelin luvussa 4.

Projektin aikana videomateriaalia kerättiin yhteensä noin 80 tunnin edestä, mikä tarkoittaa kaiken kaikkiaan noin 560 tuntia eri kamerakulmista kuvattua videoaineistoa. Kunkin parin noin puolentoista tunnin pituinen raakavideo editoitiin eri videotiedostoihin tehtävä- ja kameranumeron mukaisessa järjestyksessä.

3 Aineiston annotointi

Aineistonkeruun ja editoinnin jälkeen videoaineiston käsittely siirtyi annotointivai- heeseen. Jotta puhe- tai multimodaaliset korpukset olisivat koneluettavia, tulee vi-

deoaineistoon tehdä siihen ajallisesti sidottuja merkintöjä eli annotaatioita. Nämä merkinnät mahdollistavat aineistossa navigoinnin, tarkasteltavien kohtien rajaamisen sekä erilaiset haut. Annotaatioiden avulla aineistoon on helpompi myös palata myöhemmin uudestaan. Suurten aineistokokonaisuuksien annotoinnin on tärkeää olla systemaattista, jotta aineisto soveltuisi useiden tutkijoiden käyttöön ja sopisi erilaisiin tutkimustavoitteisiin.

CFINSL-projektissa kuvattu videomateriaali annotoitiin Max Planck -instituutissa Nijmegenissä kehitetyllä ELAN-ohjelmalla (Eudico Linguistic Annotator; Crasborn & Sloetjes 2008)[5]. Videoaineiston perusannotointia eli viittomien ja lauseiden mahdollisimman neutraalia annotointia ryhdyttiin tekemään yhteensä 22 suomalaista viittomakieltä äidinkielenään käyttävän kielenoppaan materiaaleista. Tämä aineisto on kestoltaan yhteensä noin 16 tuntia. Annotointiprosessi alkoi viitomien tunnistamisella, niiden merkitysten erottamisella ja viitotun tekstin ilmaus- kokonaisuuksien kääntämisellä suomen kielelle.

3.1 Glossaus eli viittomatason annotointi

Viittomatason annotointi voidaan toteuttaa eri tavoin. Yleisesti ottaen viittomakielten viittomien merkitsemisessä käytetään glosseja, jotka ovat yleensä suuraakkosin kirjoitettuja puhutun kielen sanoja. Glossiksi valikoituu yleensä sana, jonka merkitystä viittoma vastaa mahdollisimman hyvin (Johnston 2016). Suomalaisen viittomakielen kohdalla käytetään usein suomenkielisiä glosseja, jotka kirjoitetaan perusmuotoisena (esim. Savolainen 2000). Myös CFINSL-projektissa suomalaista viittomakieltä koskeva viittomien perusannotointi on tehty käyttäen suomenkielisiä glosseja (Salonen ym. 2019).

Toimivan korpuksen rakentamisen edellytyksinä ovat yhtenäisyys ja johdonmukaisuus. Tämä tarkoittaa, että yhteisistä periaatteista ja annotointikonventioista tulee sopia kaikkien annotoijien kesken. Annotointikonventioita on kehitelty useissa eri viittomakielten korpusprojekteissa (esim. Johnston 2016 Australia; Crasborn ym. 2015 Hollanti; Wallin & Mesch 2018 Ruotsi). CFINSL-projektissa konventioita kehiteltiin perustason annotointia tehtäessä (ks. Keränen ym. 2016). Ensimmäinen versio annotaatiokonventioista julkaistiin keväällä 2018 ja se sisälsi viittomatason annotointiin liittyvät periaatteet CFINSL-projektissa. Konventioissa

[5] https://tla.mpi.nl/tools/tla-tools/elan/

kuvaillaan viittomiston erilaisten osien, kuten esimerkiksi eriasteisesti leksikaalisten viittomien (ks. Jantunen 2018) muistiinmerkintään liittyviä periaatteita (Salonen ym. 2018). Konventioiden toisessa, helmikuussa 2019 julkaistussa versiossa on kuvattu myös virketason käännöksiin liittyvät periaatteet (Salonen ym. 2019).

Vuosina 2015 ja 2016 CFINSL-projektissa annotoitiin merkityslähtöisesti siten, että samanmuotoisille viittomaesiintymille nimettiin eri glossi (so. merkitysglossi) lausetason kontekstuaalisen merkityksen perusteella, mikä oli aikaa vievää. Annotoijien kesken oli työlästä löytää yhteistä linjaa paisuvalle merkitysglossijoukolle, saati hallita rakentumassa olevaa korpusta. Esimerkiksi yksi opiskeluun viittaava muoto saatettiin tilanteen mukaan glossata opiskelua, oppimista ja kurssia tarkoittavilla merkitysglosseilla.

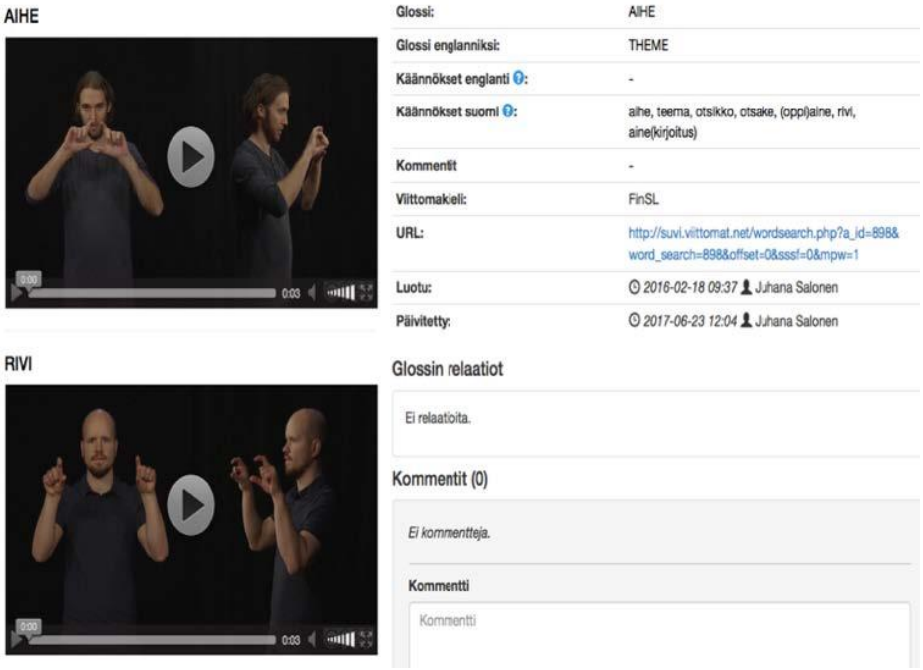
Vuonna 2017 päätettiin siirtyä merkitysglosseista ID-glosseihin, jolloin samanmuotoisia (homonyymisiä, polyseemisiä ja foneettisesti varioivia) viittomia alettiin koota yhteen samoin tunnisteglossein. ID-glossilla tarkoitetaan sellaista nimikettä, joka on valittu edustamaan merkitykseltään varioivaa, mutta muodoltaan identtistä viittomaa laajassa korpuksessa (Johnston 2008, 2010). Esimerkiksi suomalaisessa viittomakielessä esiintyy manuaaliselta (käsillä tuotetulta) artikulaatioltaan samanmuotoinen viittoma, joka voi tarkoittaa lauseyhteydestä riippuen arkea, farkkuja, maaseutua, raitista tai oranssia. Aikaisemmin merkitysglossiksi valikoitiin aina jokaisen esiintymän kohdalla kontekstin mukainen merkitys, mutta ID-glossiksi kaikille esiintymille valikoitui ARKI[6]. ID-glossi ei näin ollen ilmaise viittoman merkityskäännöstä, vaan annotoijien kesken sovittua tunnistetta. Useimmiten ID-glossiksi valikoidaan se merkitys, jonka frekvenssi aineistossa on suurin. ID-glossaus mahdollistaa merkitysglossausta tehokkaammat haut aineistosta.

Toteutimme ID-glossauksen kahdella toisiinsa yhteydessä olevalla alustalla: ELAN-ohjelmassa ajallisesti videoon yhteydessä olevat glossit yhdistyvät myös Suomen Signbank -tietokantaan verkkoyhteyden avulla. Suomen Signbank on suomalaiselle ja suomenruotsalaiselle viittomakielelle rakennettu leksikkotietokanta, jonka perustehtävänä on toimia työkaluna viittomakielisten tekstien annotoinnissa[7]. Tietokannassa olevat glossitietueet sisältävät itse glossin lisäksi suomenkieliset käännösvastineet (vrt. merkitysglossit), videon viittomasta, jota glossilla merkitään, sekä tarvittaessa muuta tietoa viittomasta (ks. kuvio 2).

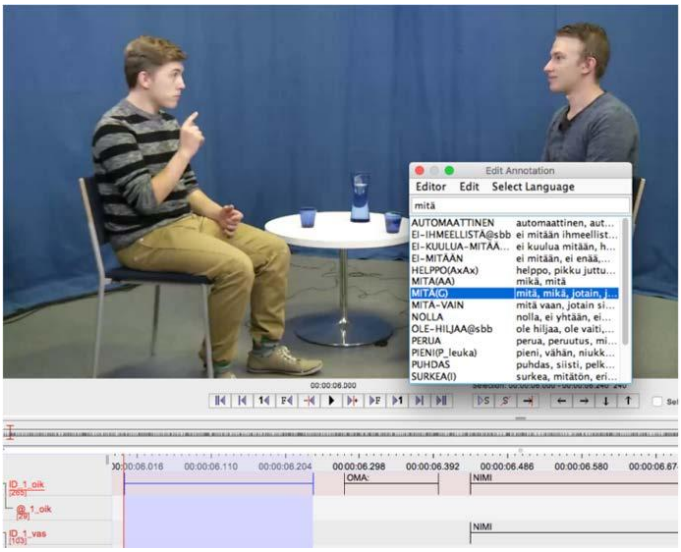
---

[6] ID-glossin ARKI tietue Signbank-tietokannassa: https://signbank.csc.fi/dictionary/gloss/3564/
[7] https://signbank.csc.fi

ELAN-ohjelma sisältää ominaisuuden, jonka myötä ohjelma voi ottaa annotointeja tehtäessä yhteyden ulkoisella verkkopalvelimella ylläpidettyyn kontrolloituun sanastoon (ECV, external controlled vocabulary). CFINSL-projektissa kontrolloitu sanasto eli lista aikaisemmin luoduista glosseista sijoitettiin Suomen Signbankiin kehittämällä Signbank-alustaa tähän soveltuvaksi. Glossilistan ansi- osta annotoija voi joka annotaatiosolun nimeämisen yhteydessä tarkistaa, löytyykö glossi jo tietokannasta. Jos glossi on jo olemassa, annotoija voi valita sen listalta ja välttää näin ei-automatisoidussa transkriptiossa herkästi syntyviä kirjoitusvirheitä. Jos glossia ei ole vielä luotu kyseiselle viittomalle, tietokantaa voi täydentää luo- malla sinne uuden glossitietueen videoineen, käännöksineen ja muine lisätietoineen. Kuviossa 3 havainnollistetaan, kuinka ELAN-ohjelmassa annotaatiosolun sisältöä luotaessa voidaan hakea Suomen Signbank -tietokannasta sopivaa glossia joko ID- glossin (laatikon vasemmanpuoleinen sarake) tai sen käännösvastineiden (laatikon oikeanpuoleinen sarake) avulla (Salonen ym. 2018). Signbankissa käsin tehdyt glossi- ja käännösvastinemuutokset päivittyvät automaattisesti kaikkiin linkitettyi- hin annotaatiosoluihin jatkuvan ECV-yhteyden myötä.



Kuvio 2. Esimerkkinäkymä Signbankin glossitietueesta.



Kuvio 3. Näkymä annotoinnista ELAN-ohjelmassa käytettäessä Signbankiin sijoitettua kontrolloitua sanastoa.

Suomen Signbankin kehittämistyöstä on vastannut CFINSL-projekti yhdessä Kuurojen Liiton tutkimus- ja sanakirjatyön kanssa. Tietokannan taustalla on alun perin australialaisessa viittomakielityössä kehitelty Auslan Signbank sekä tämän myöhempi sovellus, hollantilainen Signbank-tietokanta. Lähdekoodit kaikista Signbank-tietokannoista ovat saatavissa Github-versionhallintasivustolta. [8] Signbankin kehittämistä on ryhdytty 2010-luvulla tekemään kansainvälisessä yhteistyössä Australiassa, Hollannissa, Suomessa ja Iso-Britanniassa sijaitsevien tutkimusryhmien välillä (Cassidy ym. 2018). Suomen Signbankin kehittelyssä teknisen dokumentoinnin lisäksi tietokannan rakennetta ja ominaisuuksia on dokumentoitu käyttäjälähtöisesti Githubin FinSL-signbank wikiin. Suomen Signbank -tietokannasta löytyvä CFINSL-projektin leksikko perustuu työstettävänä olleeseen noin 16 tunnin pituiseen materiaaliin, ja se julkaistiin huhtikuussa 2018. Tämän lisäksi Suomen Signbank sisältää kesäkuussa 2017 julkaistun Kuurojen Liiton Kipo-korpuksen leksikon, joka pohjautuu noin 2,5

[8] https://github.com/Signbank

tunnin mittaiseen Suomen viittomakielten kielipoliittisen ohjelman annotoituun materiaaliin (Kuurojen Liitto ry 2015).

3.2 Virketason käännöksiä koskeva annotointi

Viittomatason annotoinnin lisäksi CFINSL-projektissa tehtiin annotointia myös virketasolla. Tämä tarkoittaa käytännössä viittomisen kääntämistä suomen kielelle. Käännöksen alkuvaiheessa viitotusta tekstivirrasta eroteltiin kääntäjien intuition perusteella mielekkäitä virkkeitä ilman tarkempaa lauseiden erottelua, mikä on varsinaisen tutkimuksen tehtävä perusannotoinnin jälkeen. Käännös on toteutettu sellaisessa muodossa, joka huomioi lähtökielen tapaa ilmaista asiat niin manuaalisesti (käsillä) kuin ei-manuaalisesti (päällä, keholla ja kasvoilla). Lisäksi käännöksiin on lisätty sulkeiden sisään osia, joita sujuva suomenkielinen teksti edellyttää, mutta jotka tulevat viittomakielisessä tekstissä ilmi edeltävästä diskurssikontekstista tai joita ei välttämättä edellytetä lainkaan (mm. lauseen tekijä, kopula, eräät konjunktiot; ks. esimerkki 1). Käännöksiä koskevat periaatteet kuvataan tarkemmin CFINSL-projektin annotointikonventioissa (ks. Salonen ym. 2019).

(1) KATSOA ULKONA SATAA-LUNTA LUMI SATAA-LUNTA
(Hän) huomaa, (että) ulkona sataa lunta.

Käännös tarjoaa kokonaisvaltaisemman kuvan viitotuista teksteistä, sillä ID-glossauksessa keskitytään pelkästään manuaaliseen artikulaatioon. Käännöksestä voidaan myös tarkistaa, mihin merkitykseen kullakin ID-glossilla on viitattu. Tähän mennessä virketason käännökset on tehty 22 suomalaista viittomakieltä käyttävän kielenoppaan materiaaleista. Kääntäjät ovat myös tehneet Suomen Signbank-tietokannan glossitietueisiin viittomien suomenkielisiä käännösvastineita sen mukaan, millaisia merkityksiä ID-glosseilla merkityillä viittomilla on ilmennyt kääntämisen yhteydessä.

4 Aineiston pitkäaikaissäilytys, julkaisu ja tietosuojakysymykset

CFINSL-projektissa käynnistyneen korpustyön tavoitteena on sekä pitkäaikaissäilyttää aineistoa että julkaista siitä erilaisin käyttöoikeuksin rajattuja osia kielenoppaiden tutkimussuostumusten ja tietosuojalainsäädännön sallimissa puitteissa. Korpusaineistoa rakennettaessa aineistoa säilytetään ensisijaisesti

Jyväskylän yliopistossa, mutta pitkäaikaissäilytystä ja julkaisua varten aineistoa siirretään myös FIN-CLARIN-konsortion Kielipankkiin. Ensimmäinen osakokonaisuus aineistosta siirrettiin Kielipankkiin maaliskuussa 2019. Aineiston nimi on Suomalaisen viittomakielen korpus (Corpus FinSL)[9], ja se sisältää yhteensä noin 14 tuntia videomateriaalia: suomalaisella viittomakielellä viitottuja tarinoita ja keskusteluja yhteensä 21 kielenoppaalta. Corpus FinSL -aineisto on jaettu Kielipankissa käyttöoikeuksien ja videoaineiston sisällön perusteella kahteen osa-aineistoon: Elisitoituihin kertomuksiin (Elicited narratives)[10] ja Keskusteluihin (Conversations)[11]. Elisitoidut kertomukset sisältävät viestinnällisten tehtävien 3–5 materiaaleja (ks. luku 2). Aineiston yhteiskesto on noin 5 tuntia, ja se on julkisesti saatavilla tutkijoiden, kouluttajien sekä laajemman yleisön käyttöön Creative Commons BY NC SA 4.0 -lisenssillä. Keskustelut sisältävät viestinnällisten tehtävien 1, 2, 6 ja 7 aineistoa, joiden yhteis- kesto on noin 9 tuntia. Keskusteluaineiston käyttö edellyttää tutkimussuunnitelmaa sekä henkilökohtaista käyttöoikeutta Kielipankin RES-lisenssin mukaisesti.

Pitkäaikaissäilytykseen ja monipuoliseen tutkimuskäyttöön tarkoitetun laajan, multimodaalisen aineistokokonaisuuden tulee sisältää itse videoaineistot, videoihin synkronoidut annotaatiot sekä riittävän kattavat metatiedot aineiston eri osa-alueista. Metatiedot kuvaavat tässä tapauksessa aineistokokonaisuuden yleisluonnetta, aineistonkeruussa läsnä olleita henkilöitä, videoiden sisältöjä ja video- ja annotaatiotiedostojen muotoja. Kielipankkiin siirrettävän Corpus FinSL -aineiston ano- nymisoidut metatiedot on kuvattu IMDI (ISLE Meta Data Initiative) -standardin mukaisesti. IMDI on Max Planck -instituutissa Nijmegenissä kehitelty kuvausstandardi monimediaisten ja multimodaalisten kieliaineistojen yhdenmukaiseen kuvaukseen[12]. CFINSL-projektissa tuotettiin IMDI-standardien mukaisesti yleiskuvaus aineistosta (Corpus FinSL), sen taustalla olevasta projektista (CFINSL Project) sekä osa-aineistojen sisällöistä (Elicited narratives; Conversations). Osa-aineistojen osalta annettiin myös yleiskuvaukset kustakin viestinnällisestä tehtävästä (1–7, ks. luku 2). Tämän lisäksi jokaisen parin yksittäisistä viestintätehtävistä tehtiin tilannekohtainen kuvaus (Session), joka sisältää tietoja aineistonkeruutilanteen osallistujista (Actors); viestintätilanteen

[9] Suomalaisen viittomakielen korpus: http://urn.fi/urn:nbn:fi:lb-2019012321
[10] Elisitoidut kertomukset: http://urn.fi/urn:nbn:fi:lb-2019012322
[11] Keskustelut: http://urn.fi/urn:nbn:fi:lb-2019012323
[12] https://tla.mpi.nl/imdi-metadata/

laadusta, vuorovaikutuksellisuudesta ja keruu- menetelmästä (Content); videomateriaaleista (MediaFiles) ja annotaatioista (WrittenResources).

Kielenoppaisiin liittyviä taustatietoja kerättiin CFINSL-projektin aineistonkeruun aikana vuosina 2014–2017 hyvin kattavasti. Näistä Kielipankkiin rakennettuun IMDI-kuvaukseen valikoituivat lopulta vain henkilön yksilöivä anonymisoitu koodi, ikä ikäryhmittäin, sukupuoli, asuinalue sekä kätisyys (oikea/vasen). Kaiken kaikkiaan Kielipankkiin siirretty Corpus FinSL -aineisto koostuu 71 viestintätilanteesta ja sisältää yhteensä 343 videotiedostoa (kamerakulmat 1–6), 142 annotaatiotiedostoa (ELANin .eaf- ja .pfsx-tiedostot) sekä IMDI-kuvaukset (taulukko 1).

Taulukko 1. Suomalaisen viittomakielen korpus (Corpus FinSL) Kielipankissa.

Koko aineisto	14 tuntia ja 22 minuuttia
Elisitoidut kertomukset (CC-lisenssi)	5 tuntia ja 4 minuuttia
Keskustelut (RES-lisenssi)	9 tuntia ja 18 minuuttia
Videotiedostot	343 mp4-tiedostoa
Annotaatiotiedostot	142 tiedostoa (eaf + pfsx)
Kielenoppaiden määrä	21 kielenopasta

Vuoden 2018 toukokuussa voimaan tulleen EU:n tietosuoja-asetuksen myötä Corpus FinSL -aineiston lupia oli tarpeen täydentää. Kun alkuperäiset luvat käsittelivät aineiston verkkojulkaisua yhtenä yleisenä kokonaisuutena (ks. luku 2), niin lisäluvat pyytävät kielenoppaiden suostumusta aineiston pitkäaikaissäilytykseen Kielipankissa sekä kunkin viestinnällisen tehtävän vapaaseen julkaisuun samalla alustalla. Tietosuoja-asetuksen mukaisesti kielenoppailta on pyydetty nimenomaista suostumusta myös siihen, että aineisto sisältää heihin liitettäviä erityisiin henkilötietoryhmiin kuuluvia tietoja eli käytännössä kielenoppaiden itsensä kertomaa tietoa oman kuulonsa asteesta. Tarve tämän luvan pyytämiseen kumpuaa erityisesti vuorovaikutuksellisesta tehtävästä 1, jossa kielenoppaat viittomakieliselle kulttuurille luonnollisella tavalla usein identifioivat itsensä kuuroiksi itseään esitellessään. Käytännössä lisälupalomakkeella on pyydetty kielenoppaiden suostumusta myös heidän koko aineistonsa lisensointiin kaupallisen käytön kieltävällä Creative Commons BY NC SA 4.0 -lisenssillä, joskin lopulta Creative Commons -lisenssillä on lisensoitu ainoastaan vapaasti julkaistava Elisitoidut kertomukset -osa-aineisto.

Corpus FinSL -aineiston jako kahteen osa-aineistoon on niin ikään seurausta EU:n tietosuoja-asetuksen mahdollisimman tarkasta noudattamisesta. Kun

Elisitoidut kertomukset -osa-aineisto sisältää ainoastaan temaattisesti rajattuja monologeja, joissa kielenoppaat toisintavat erilaisten kuvamateriaalien tarinoita, niin Keskustelut-osa-aineisto sisältää temaattisesti rajaamattomampia dialogeja, joissa kielenoppaat saattavat ilmaista välillisesti henkilötietoja myös kolmansista osapuolista. Näiden henkilötietojen vapaaseen julkaisemiseen ei ole ollut mahdollista pyytää lupaa suostumuslomakkeella, joten Keskustelut-osa-aineiston saatavuutta on haluttu rajoittaa.

5 Loppusanat

Tässä artikkelissa on esitelty Suomen viittomakielten korpusprojektissa tehtyä viittomakielisen aineiston keruuta, annotointia, pitkäaikaissäilytystä sekä julkaisua. Laaja, elektronisessa muodossa oleva ja tietokoneluettava aineisto tarjoaa uusia mahdollisuuksia Suomessa käytettyjen viittomakielten – suomalaisen ja suomenruotsalaisen viittomakielen – määrälliseen ja laadulliseen tutkimukseen. Kattavasta, usean henkilön viittomista sisältävästä aineistosta voi tutkia esimerkiksi eri-ikäisten tai eri puolilla Suomea asuvien viittomakielisten käyttämää viittomistoa sekä eri tekstilajien välillä olevia eroja viittomistossa ja kielen rakenteessa. Laajat, osin julkisesti saatavilla olevat aineistot mahdollistavat myös aivan uudella tavalla viittomakieliä vertailevan tutkimuksen. Tätä tukee erityisesti eri viittomakielten korpusprojekteissa käytetyt samankaltaiset aineistonkeruumenetelmät.

Viittomakielikorpuksen tuomat mahdollisuudet näkyvät selvästi 2010-luvulla Jyväskylän yliopistossa tehdyssä viittomakielen tutkimuksessa. Korpusaineiston pohjalta on tähän mennessä tehty tutkimusta suomalaisen viittomakielen rakenteesta muuan muassa lauseenjäsenten järjestyksestä intransitiivi- ja transitiivilauseissa (Jantunen 2017), kuvailevista viittomista (Takkinen ym. 2018) sekä ei-manuaalisuudesta pään ja kehon liikkeiden suhteen (Puupponen 2018). Lisäksi aineistoa on käytetty viittomakieliä vertailevissa tutkimuksissa (Jantunen ym. 2016; Puupponen ym. 2016) sekä suomalaisen viittomakielen oppiaineesta valmistuneissa maisterintutkielmissa (esim. Syrjälä 2018; Puhto 2018). Suomenruotsalaisen viittomakielen aineistoa on käytetty Helsingin yliopiston, Kuurojen Liiton ja Humanistisen ammattikorkeakoulun tulkkikoulutuksen yhteisissä projekteissa koskien suomenruotsalaisten kielitietoisuuden kehittämistä ja tulkkikoulutusta. Lisäksi suomenruotsalaisesta viittomakielestä on parhaillaan tekeillä kerättyä aineistoa hyödyntävä väitöskirja Helsingin yliopistossa.

Suomen viittomakielten korpuksella tulee olemaan merkittävä vaikutus Suomen viittomakieliseen yhteisöön sekä viittomakielten yhteiskunnalliseen asemaan. Viittomakielisille se tarjoaa mahdollisuuden kehittää kielitietoisuutta omasta äidin- kielestään, jota ei monille ole opetettu perusopetuksessa. Viittomakieltä vieraana kielenä käyttäville henkilöille – kuten esimerkiksi viittomakielentulkeille – korpus tarjoaa opetusmateriaalia muun muassa kielenkäyttäjien välisten sosiolingvististen erojen tunnistamiseen. Korpusaineistoa tullaan myös hyödyntämään Jyväskylän yliopistossa käynnissä olevassa, opetus- ja kulttuuriministeriön rahoittamassa koulutushankkeessa, jonka tavoitteena on kehittää täydennyskoulutustarjontaa viittomakielen opettajille Suomessa. Opetus- ja koulutussovellusten lisäksi korpus voi tulevaisuudessa toimia myös kielenhuollon ja kielisuunnittelun työvälineenä.

Lähteet

Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E. & John- ston, T. (2018). Signbank: Software to Support Web Based Dictionaries of Sign Language. Proceedings - The 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, 2359–2364.

Crasborn, O., Bank, R., Zwitserlood, I., Kooij, E., Meijer, A. & Sáfár, A. (2015). Annotation Conventions for The Corpus NGT. Version 3. Radboud University Nijmegen: Centre for Language Studies & Department of Linguistics.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. Proceedings - The 3rd Workshop on the Representation and Processing of Sign Lan- guages: Construction and Exploitation of Sign Language Corpora, 39–43.

De Meulder, Maartje (2016) Promotion in Times of Endangerment: The Sign Language Act in Finland. Language Policy, 16(2), 189–208.

Jantunen, T. (2017). Fixed and NOT free: Revisiting the order of the main clausal constituents in Finnish Sign Language from a corpus perspective. SKY Journal of Linguistics 30, 137–149.

Jantunen, T. (2018). Viittomakielet hybridisysteemeinä: hämärärajaisuus ja epäkonventionaalisuus osana viittomakielten rakennetta. Puhe ja Kieli 38(3), 109–126.

Jantunen, T.; Mesch, J.; Puupponen, A. & Laaksonen, J. (2016). On the rhythm of head movements in Finnish and Swedish Sign Language sentences. Proceedings - Speech Prosody 2016, 850–853.
Johnston, T. (2008). Corpus linguistics and signed languages: No lemmata, no corpus. Proceedings - The 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora, 82–87.
Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics 15 (1), 106–131.
Johnston, T. (2016). Auslan Corpus Annotation Guidelines. Centre for Languge Sciences, Department of Linguistics, Macquarie University (Sydney) and La Trobe University (Melbourne), Australia.
Keränen, J., Syrjälä, H., Salonen, J. & Takkinen, R. (2016). The Usability of the Annotation. Proceedings - The 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, 111–116.
Kuurojen Liitto ry (2015). Suomen viittomakielten kielipoliittinen ohjelma 2010 - korpus, annotoitu versio. Kielipankki. Saatavilla http://urn.fi/urn:nbn:fi:lb-2014073031
Puhto, J. (2018). Päänpudistuksen käyttötavat ja frekvenssit suomalaisessa viittomakielessä. Suomalaisen viittomakielen pro gradu -tutkielma. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto.
Puupponen, A.; Jantunen, T.; Takkinen, R.; Wainio, T. & Pippuri, O. (2014). Taking non-manuality into account in collecting and analyzing Finnish Sign Language video data. Proceedings - The 6th Workshop on the Representation and Processing of Sign Languages: Beyond the manual channel, 143–148.
Puupponen, A., Jantunen, T. & Mesch, J. (2016). The Alignment of Head Nods with Syntactic Units in Finnish Sign Language and Swedish Sign Language. Proceedings - Speech Prosody 2016, 168–72.
Puupponen, A. (2018). The relationship between the movements and positions of the head and the torso in Finnish Sign Language. Sign Language Studies 18(2), 175–214.
Salonen, J., Takkinen, R., Puupponen, A., Nieminen, H. & Pippuri, O. (2016). Creating Corpora of Finland´s Sign Languages. Proceedings - The 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, 179–184.

Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2018). Suomen viittomakielten korpus -projektin (CFINSL) annotointiohjeet. 1. versio. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto. http://r.jyu.fi/ygQ
Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2019). Suomen viittomakielten korpus -projektin (CFINSL) annotointiohjeet. 2. versio. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto. http://r.jyu.fi/ygR
Savolainen, L. (2000). Viittomakielten erilaiset muistiinmerkitsemistavat. Teoksessa A. Malm (toim.) Viittomakieliset Suomessa (pp. 189–200). Helsinki: Finn Lectura.
Syrjälä, H. (2018). Hakukysymysviittoman paikka suomalaisessa viittomakielessä. Suomalaisen viittomakielen pro gradu -tutkielma. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto.
Takkinen, R., Keränen, J. & Salonen, J. (2018). Depicting Signs and Different Text Genres: Preliminary Observations in the Corpus of Finnish Sign Language. Proceedings – The 8th Workshop on the Representation and Processing of Sign Languages: Involving the Lan- guage Community, 189–194.
Wallin, L. & Mesch, J. (2018). Annoteringskonventioner för teckenspråkstexter. Version 7. Avdelningen för teckenspråk, Institutionen för lingvistik, Stockholms universitet.