VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

MITEN VIITTOMAKIELEN KORPUSTA LUODAAN JA MIHIN SITÄ TARVITAAN? VIITTOMAKIELTEN KORPUKSET JA NIIDEN TEHTÄVÄT

Ritva Takkinen, Jyväskylän yliopisto,
kieli- ja viestintätieteiden laitos

Juhana Salonen, Jyväskylän yliopisto,
kieli- ja viestintätieteiden laitos

Anna Puupponen, Jyväskylän yliopisto,
kieli- ja viestintätieteiden laitos

Henri Nieminen, Jyväskylän yliopisto,
kieli- ja viestintätieteiden laitos

Artikkeli käsittelee suomalaisen ja suomenruotsalaisen viittomakielen korpusten luontia CFINSL-projektissa (Corpus project of Finland's sign languages, Suomen viittomakielten korpusprojekti). Viittomakielillä ei ole kirjoitettua muotoa, joten korpusten laatiminen vaatii erilaista lähestymistä kuin korpusten luonti sellaisille puhutuille kielille, joilla on kirjoitettu muoto. Artikkelissa kuvataan ne menetelmät, joilla Jyväskylän yliopiston viittomakielen keskuksessa on koottu aineistoa suomalaisen ja suomenruotsalaisen viittomakielen korpukseen. Lisäksi kuvataan korpusaineiston teknistä käsittelyä, annotointia, metatietojen keruuta ja käsittelyä sekä aineiston säilytystä ja tutkijoiden käyttöön saattamista. Korpuksen lisäksi ja sen käyttöön luotiin myös leksikkotietokanta, Signbank, joka hyödyttää sekä itse annotointiprosessia että korpusten käyttöä niin tutkimuksessa kuin opetuksessakin. Korpukset tallentavat Suomessa käytettyjä viittomakieliä niin tutkijoiden kuin kummankin kieliyhteisön tämän päivän jäsenille ja tulevien sukupolvien saataville.

Avainsanat: annotaatio, korpus, leksikkotietokanta, Signbank, suomalainen viittomakieli, suomenruotsalainen viittomakieli, viittomakielen korpus

Kirjoittajien yhteystiedot:

Ritva Takkinen	Juhana Salonen	Anna Puupponen	Henri Nieminen
ritva.takkinen@jyu.fi	juhana.salonen@jyu.fi	anna.m.puupponen@jyu.fi	henri.nieminen@gmail.com

1 JOHDANTO

1.1 Suomen viittomakielet ja niiden käyttäjät

Suomessa on tällä hetkellä kaksi kotimaista viittomakieltä, suomalainen viittomakieli (SVK) ja suomenruotsalainen viittomakieli (SRVK). Niistä kummankin juuret ovat ruotsalaisessa viittomakielessä, jonka toi Suomeen maamme ensimmäinen kuurojen opettaja Carl Oscar Malm (1826–1863) opiskeltuaan itse Tukholmassa kuurojen Manilla-koulussa 1800-luvun puolivälissä. Palattuaan Suomeen hän käytti opetuksessaan oppimaansa ruotsalaista viittomakieltä (svenskt teckenspråk). Suurimmassa osassa myöhemmin Suomeen perustetuista kuurojen kouluista (esimerkiksi Turussa ja Kuopiossa) käytettiin kirjoituksessa suomen kieltä, mutta Pietarsaaren ja Porvoon kouluissa kirjotuskielenä oli ruotsi. Tämän vuoksi näihin kouluihin menivät opiskelemaan juuri ruotsinkielisten perheiden kuurot lapset (Salmi & Laakso, 2005; Wallvik, 1997).

Opetus kuurojen kouluissa muuttui 1800-luvun lopulla lähes yksinomaan suomen- tai ruotsinkieliseksi, ja viittomakielen käytöstä tuli kouluissa yleisesti kiellettyä. Tästä huolimatta oppilaat viittoivat keskenään, ja viittomakieli säilyi, mutta suomen- ja ruotsinkielisten koulujen oppilaiden viittomakielet alkoivat kehittyä erilaisiksi: esimerkiksi viittomien yhteydessä käytettyjen huulioiden muodot alkoivat erota näissä kielivarianteissa. Ajan kuluessa myös osa viittomista on muovautunut erilaiseksi (Hoyer, 2000). Samalla Suomessa käytettävät viittomakielet ovat irtautuneet ruotsalaisesta viittomakielestä siinä määrin, että voidaan puhua erillisistä kielistä (Jantunen, 2000; Mesch, 2006).

Suomalaista viittomakieltä käyttävät lähinnä suomenkielisissä tai suomalaista viittomakieltä (SVK) käyttäneissä perheissä kasvaneet kuurot, jotka ovat käyneet suomenkielistä kuurojen koulua. Heitä on Kuurojen Liiton (2018) mukaan arvioitu olevan noin 4000–5000, Rainòn tutkimuksen mukaan määrä on lähempänä 3000:a (Posti, 2008). Kuulevia, äidinkielenään suomalaista viittomakieltä käyttäviä (lähinnä kuurojen vanhempien kuulevia lapsia, engl. Coda) ja toisena tai vieraana kielenä käyttäviä (esimerkiksi kuurojen henkilöiden kuulevat perheenjäsenet sekä työssään säännöllisesti viittomakieltä käyttävät kuten tulkit ja viittomakielisten opettajat) on arvion mukaan noin 6000–9000 (Kuurojen Liitto, 2018). Suomenruotsalaista viittomakieltä käyttävät lähinnä Suomen rannikkoalueilla asuvat suomenruotsalaisissa tai suomenruotsalaista viittomakieltä (SRVK) käyttävissä perheissä kasvaneet kuurot, jotka ovat käyneet ruotsinkielisiä, nyt jo suljettuja kouluja. Kuurojen lisäksi tätä kieltä käyttävät jonkin verran myös kuulevat, lähinnä kuurojen vanhempien lapset. Suomenruotsalaista viittomakieltä käyttäviä kuuroja on tällä hetkellä vain noin 90, joista suurin osa on yli 55-vuotiaita (Soininen, 2016), mikä tekee suomenruotsalaisesta viittomakielestä hyvin uhanalaisen kielen (ks. myös Hoyer, 2004; 2012).

1.2 Kielikorpukset

Korpuksella tarkoitetaan sellaista elektronisessa muodossa olevaa kirjoitetun, puhutun tai viitotun kielen editoitua ja annotoitua kokoelmaa, jonka avulla voidaan tutkia kielen sanastoa, kieliopillisia rakenteita ja käyttöä (esim. Johnston, 2010; 2012; Lüdeling & Kytö, 2008; Sinclair, 2005; Wichmann, 2008). Moderniin korpukseen liittyy erilaisten kielellisten piirteiden annotaatio eli merkitseminen konehakua varten. Lisäksi siihen liittyy metadata eli sosiolingvistinen ja aineistoon liittyvä taustatieto, joka kuvaa kielenkäyttäjää, kielenkäyttötilannetta ja aineiston sisältöä. Korpuksia on hyvin erilaisia, ja niiden kokoamisessa käytetään erilaisia kriteereitä sen mukaan, millaiseen tarkoitukseen

ja tutkimukseen ne luodaan (esim. Hunston, 2008; Sinclair, 2005).

Korpukset mahdollistavat laajempien aineistojen käytön tutkimuksessa ja antavat siten luotettavamman kuvan kielen rakenteesta ja käytöstä kuin introspektioon ja pieniin aineistoihin perustuva tutkimus. Elektroniset aineistot nopeuttavat ja tehostavat aineiston analyysiä. Korpuksia hyödyntämällä voi myös tehdä eri kieliä vertailevaa tutkimusta. Lisäksi korpuksia käytetään myös opetuksessa. Nykyisin on kerätty myös kielenoppijoiden aineistoja, joista voi tutkia muun muassa kielen oppimisen etenemistä ja prosesseja[1]. Tällainen on esimerkiksi Kansainvälinen oppijansuomen korpus (Jantunen, 2011; Kielipankki[2]).

Tässä artikkelissa esitellään niitä vaiheita, joita liittyy viittomakielen korpuksen luomiseen. Aluksi tutustutaan eri maissa viittomakielistä tehtyihin korpuksiin (luku 2), minkä jälkeen perehdytään Jyväskylän yliopiston viittomakielen keskuksessa olevaan Suomen viittomakielten CFINSL-korpusprojektiin (luku 3). Tässä yhteydessä käydään läpi projektia edeltänyt pilottihanke, jonka kokemuksia hyödyntämällä kehittyivät käytänteet aineiston keruuseen, videoiden editointiin, metatietojen keruuseen ja käsittelyyn, videoaineiston annotointiin sekä aineiston säilytykseen ja julkaisuun. Päätäntöluvussa pohditaan korpusten merkitystä Suomen viittomakielisille yhteisöille sekä viittomakielille ja niiden tutkimukselle ja opetukselle.

2 VIITTOMAKIELTEN KORPUKSET

Viittomakielten lingvistisessä tutkimuksessa korpusaineistoilla voidaan ajatella olevan erityisen tärkeä rooli. Viittomakielten heikko asema vähemmistökielenä, pitkälle kehittyneiden, yhteisöllisten standardien puute sekä keskeytynyt periytyminen sukupolvelta toiselle (kuuroille vanhemmille syntyy useimmiten kuulevia lapsia, jolloin viittomakieli ei välttämättä jää käyttöön) ovat muun muassa niitä tekijöitä, joiden myötä laajojen kieliaineistojen tarkastelu on tärkeää tehtäessä viittomakielten kuvauksia ja kielioppeja (ks. Johnston, 2010). Pieniin aineistoihin tai muutaman kielenoppaan intuitioon pohjautuvat tutkimukset ovat alttiita väärintulkinnoille ja voivat osaltaan vääristää tutkittavana olevasta viittomakielestä muodostuvaa kokonaiskuvaa.

Viittomakielten korpusten kokoamiseen on vaikuttanut merkittävästi teknologian kehitys, joka on taannut aikaisempaa paremman videoiden laadun ja tallentamisen. Tietotekniikan edistyminen multimedia-annotointiohjelmineen antaa mahdollisuuden luoda myös viittomakielestä elektronisia aineistoja. Ensimmäinen viittomakielen korpus luotiin australialaisesta viittomakielestä (Auslan). Sen koonti aloitettiin 2000-luvun alussa, ja valmis korpus[3] julkaistiin 2008 ELAR-arkistossa (Endangered Language Archive).

Hollantilaisen (NGT)[4], brittiläisen (BSL)[5] ja ruotsalaisen viittomakielen (STS)[6] korpukset ovat australialaisen viittomakielen korpuksen ohella ensimmäisiä viittomakielten korpusprojekteja. Tällä hetkellä viittomakielten korpusten laatiminen on meneillään

[1] UCL, Centre for English Corpus Linguistics: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

[2] FIN-CLARINin Kielipankki: https://www.kielipankki.fi

[3] Auslan-korpus: https://researchdata.ands.org.au/auslan-australian-sign-language-corpus/125009

[4] NGT-korpus: https://www.ru.nl/corpusngtuk/

[5] BSL-korpus: http://bslcorpusproject.org

[6] STS-korpus: https://www.ling.su.se/teckenspråksresurser/teckenspråkskorpusar/svensk-teckenspråkskorpus

muun muassa USA:ssa ja Japanissa. Euroopassa korpuksia kootaan esimerkiksi Unkarissa, Puolassa, Italiassa, Ranskassa, Belgiassa, Tanskassa ja Norjassa. Yhtä laajimmista korpusaineistoista ollaan parhaillaan työstämässä saksalaisesta viittomakielestä (DGS)[7]. Taulukoon 1 on suomalaisten viittomakielten korpuksen lisäksi koottu vertailevaa tietoa australialaisen, hollantilaisen, brittiläisen, saksalaisen ja ruotsalaisen viittomakielen korpusten luonnista, koosta, sisällöstä ja niissä käytetyistä ohjelmista.

Korpusten koot ovat luonnollisesti verrannollisia maan tai alueen väkilukuun ja kuurojen määrään. Saksalaisen ja brittiläisen viittomakielen korpukset ovat suurimpia sekä informanttimäärältään että aineiston laajuudeltaan. Kaikissa korpuksissa on koottu aineistoa äidin- tai ensikieleltään viittomakielisiltä henkilöiltä. Tasaisen laadun takaamiseksi aineistot on kerätty studio-oloissa useammalla korkealaatuisella videokameralla samanaikaisesti. Viittomakielikorpusten videoaineistot sisältävät yleensä korpusta varten kehitettyjen tehtävien avulla kirvoitettua materiaalia. Australialaisessa korpuksessa on osakorpuksena myös sosiolingvistisen variaation tutkimukseen kerättyä aineistoa. Tekstityyppeinä ovat vapaa ja osittain elisitoitu keskustelu, kuvista tai videosta kertominen ja vapaa kertominen. Joissakin viittomakielikorpuksissa on lisäksi haastatteluja ja viittomistoelisitaation myötä kerättyä materiaalia. Eri viittomakielten korpusaineistoja kuvaillaan tarkemmin taulukossa 1.

Koska informantit esiintyvät korpuksessa omilla kasvoillaan ja aineistot sisältävät henkilöitä suoraan identifioivaa materiaalia, viittomakielten korpustyöhön liittyvät aina henkilökohtaiset suostumukset, joissa informantit määrittelevät, mihin tarkoituksiin aineistoa saa käyttää. Tutkimuslupiin liittyviä asioita käsitellään CFINSL-projektin osalta tarkemmin tämän artikkelin luvussa 3.2.2.

Annotaatio eli erilaisten kielenaineksen luokittelumerkintöjen liittäminen aineistoon on olennainen osa korpuksia. Viittomakielikorpuksen annotointi tarkoittaa yksinkertaisimmillaan käsien artikulaation luokittelua sekä kielenaineksen kääntämistä. Vaikka aineistoa on annotoitu runsaasti esimerkiksi Australiassa, Hollannissa, Iso-Britanniassa ja Ruotsissa, yksikään korpus ei ole tältä osin valmis edes yksinkertaisimmalla tasolla. Hidas, käsityönä tehtävä annotointi etenee usein tutkimusintressien mukaan. Sitä varten luodaan systemaattiset konventiot, mikä on edellytys tutkittavien kielellisten piirteiden hakemiselle aineistosta. (Ks. esimerkiksi Auslan Corpus Annotation Guidelines 2019)[8]. Annotointiin on käytetty yleensä Max Planck -instituutissa kehitettyä ELAN-ohjelmaa[9] (EUDICO Linguistic Annotator), joka sopii multimedia-aineiston käsittelyyn. Poikkeuksena tästä saksalaisessa viittomakielen korpuksessa käytetään erityisesti viittomakielen annotointiin kehitettyä saksalaista iLex-ohjelmaa[10], joka sisältää myös leksikkotietokannan. Useissa muissa korpuksissa leksikkotietokannan hallintajärjestelmänä on australialainen Signbank, josta on kussakin korpusprojektissa luotu omalle viittomakielelle soveltuva versio. Lisäksi aineistoista on tehty metatietokuvaukset IMDI-metatietostandardin[11] mukaisesti.

[7] DGS-korpus: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html

[8] Auslan Corpus Annotation Guidelines 2019: https://www.academia.edu/40088269/Auslan_Corpus_Annotation_Guidelines_August_2019_version_

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

[10] iLex; http://www.sign-lang.uni-hamburg.de/ilex/

[11] IMDI: https://archive.mpi.nl/forums/t/imdi-metadata-information/2639

TAULUKKO 1. Suomalaisen (SVK), australialaisen (Auslan), hollantilaisen (NGT), brittiläisen (BSL), saksalaisen (DGS) ja ruotsalaisen (STS) viittomakielen korpusten vertailua *

	CFINSL (SVK)	**Auslan**	**NGT**	**BSL**	**DGS**	**STS**
Koko	91 informanttia, 5 alueelta 67 tuntia	100 informanttia 5 alueelta 300 tuntia	92 informanttia eri puolilta maata	249 informanttia 8 alueelta	330 informanttia 12 paikkakunnalta n. 500 tuntia	42 informanttia 3 alueelta 24 tuntia
Koostaminen	2013–2017	2004–2007	2006–2008	2008–2011	2009–	2009–2011
Informantit	kuurot ja kuulevat äidinkieleltään viittomakieliset	kuurot ja kuulevat äidinkieleltään viittomakieliset	kuurot äidinkieleltään viittomakieliset	kuurot äidinkieleltään viittomakieliset	kuurot äidinkieleltään viittomakieliset	kuurot ja kuulevat äidinkieleltään viittomakieliset
Keräystapa	videointi pareittain 6 kameraa	videointi pareittain	videointi pareittain, 4 kameraa	videointi pareittain 3–4 kameraa	videointi pareittain 3 kameraa	videointi pareittain 5 kameraa
Tekstilajit	vapaa ja elisitoitu kerronta, keskustelu	vapaa ja elisitoitu kerronta, keskustelu, haastattelu	vapaa ja elisitoitu kerronta, keskustelu	vapaa ja elisitoitu kerronta, keskustelu, haastattelu, elisitoitu kieliopillinen aineisto & viittomisto (102 peruskäsitettä)	vapaa ja elisitoitu kerronta, keskustelu, elisitoitu viittomisto	vapaa ja elisitoitu kerronta, keskustelu
Tehtävänanto	7 tehtävää	11 tehtävää	8 tehtävää	tehtäviä 4:stä eri tehtävätyypistä	20 tehtävää	7 tehtävää
Tehtävän suoritus	korpusta varten kerätty, spontaani, elisitoitu	luonnollinen aineisto, elisitoitu	korpusta varten kerätty, spontaani, elisitoitu	korpusta varten kerätty, elisitoitu, spontaani	korpusta varten kerätty, elisitoitu, spontaani	korpusta varten kerätty, elisitoitu, spontaani
Annotointiohjelma	ELAN	ELAN	ELAN	ELAN	iLex	ELAN
Annotointi	manuaalinen, viittomien ID-glossit, osittaista fonologista ja kieliopillista tietoa, käännös	manuaalinen, ID-glossit, fonologista, morfologis-syntaktista & semanttista tietoa, käännös	manuaalinen, ID-glossit, fonologista tietoa, osittaista kieliopillista tietoa, käännös	manuaalinen, ID-glossit, käännös, fonologista tietoa osittaista kieliopillista annotointia	manuaalinen, glossit, käännös leksikaalista tietoa, viittomien rakenne annotoitu HamNoSys -transkriptiolla, osittaista fonologista & non-manuaalista tietoa	manuaalinen, glossit, fonologista tietoa, kieliopillista tietoa non-manuaalista tietoa käännös
Leksikkotietokanta	Suomen Signbank	Auslan Signbank	NGT Signbank	BSL Signbank	iLex (Lexikalische Datenbank)	ELAN-ohjelman kontrolloitu sanasto
Metadatakriteeristö	IMDI	IMDI	IMDI	IMDI	iLex/IMDI	IMDI

• Taulukon pohjaidea Jantusen (2011) artikkelin vertailutaulukosta.

Korpukset julkaistaan useimmiten tutkijoiden saataville. Kielipankissa on esillä tällä hetkellä CFINSL-projektin ensimmäisen julkaistun osan lisäksi myös kaksi muuta pientä suomalaisen viittomakielen korpusta, Kipo-korpus ja Snowfrog-korpus. Kipo-korpus on Kuurojen Liiton ja Kotimaisten kielten tutkimuskeskuksen julkaiseman Suomen viittomakielten kielipoliittisen ohjelman (2010) annotoitu käännösversio, joka on tehty Kuurojen Liitossa. Snowfrog on Jyväskylän yliopistossa vuosina 2013–2018 toteutetussa ProGram-projektissa käytetty aineisto, joka sisältää Snowman ja Frog, where are you? -kuvakirjojen pohjalta viitottuja tarinoita sekä annotaatioita muun muassa suomalaisen viittomakielen syntaktisista piirteistä. Snowfrog-aineiston videomateriaali on osa CFINSL-projektissa kerättyä aineistoa (Jantunen & Pippuri, 2016)[12].

Suomalaisen viittomakielen tutkimus on pohjautunut 2010-luvulle asti varsin pieniin aineistoihin. Siksi laajemman aineiston keruu ja annotoidun korpuksen koostaminen on modernille viittomakielen tutkimukselle välttämätöntä. Korpukset antavat paremmat mahdollisuudet myös viittomakielten väliseen vertailevaan tutkimukseen. Ne antavat pienaineistoja oikeamman kuvan kielistä ja niiden ilmaisuvoimasta. Korpukset tarjoavat aidoista kielenkäyttötilanteista koostuvan, tilastollisen analyysin mahdollistavan ja kielenkäyttöä edustavasti kuvaavan aineistokokonaisuuden.

[12] Snowfrog-aineisto Kielipankissa: http://urn.fi/urn:nbn:fi:lb-1001100113005

3 SUOMEN VIITTOMAKIELTEN KORPUSTEN LUOMINEN

3.1 Pilotista systemaattiseen korpustyöskentelyyn

Suomalaisten viittomakielten korpustyö alkoi pilottivaiheella vuonna 2013, jolloin aineistoa kerättiin Jyväskylän yliopistossa suomalaisen viittomakielen syventäviin opintoihin kuuluvan korpuskurssin yhteydessä neljältätoista suomalaista viittomakieltä äidinkielenään käyttävältä henkilöltä. Kirjallisuuden pohjalta tutustuttiin myös aineiston käsittelyyn. Tämän pilottivaiheen kautta alettiin suunnitella suomalaisen viittomakielen korpuksen keruuta ja käsittelyä.

Kurssilla tehtyä pilottia varten valittiin kerättävät tekstilajit sekä keinot, joilla kirvoitettaisiin informantit tuottamaan näitä tekstilajeja mahdollisimman vapautuneesti. Aikaisempien korpusprojektien käyttämien tehtävien ja elisitaatiomateriaalien pohjalta laadittiin tehtäväsarja kirvoittamaan keskustelua, vapaata kerrontaa sekä visuaalisen materiaalin avulla elisitoitua kerrontaa. Muunkielisissä korpuksissa on aineistonkeruussa käytetty yleensä 7–10 tehtävää. Pilottihankkeessamme käyttöön otettiin kuusi tehtävää: 1) kaksi informanttia kerrallaan kertovat itsestään toisilleen, 2) kumpikin informantti kertoo opiskelustaan (yliopistokonteksti), 3) kumpikin osallistuja valitsee neljästä sarjakuvasta kolme, joista hetken niihin tutustuttuaan kertovat toiselleen, 4) kumpikin informantti tutustuu hetken tekstittömään kuvakirjaan ja kertoo sitten kirjan tarinan toiselle, 5) informantit valitsevat yhdessä kuurojen kulttuuriin liittyvän aiheen, josta keskustelevat, 6) informantit valitsevat yhdessä esimerkiksi urheiluun tai elokuviin liittyvän tai muun kiinnostavan aiheen, josta keskustelevat.

Tutkimuslupaa varten laadittiin suostumuslomake, jossa informanteilla oli mahdollisuus valita, miten laajasti kunkin omaa aineistoa saa käyttää tutkimuksessa ja tallentaa julkiseksi

materiaaliksi verkkoon. Tutkimusluvan lisäksi informanteilta kerätään erillisellä taustatietolomakkeella tietoa viittomakielen omaksumisesta, koulutuksesta ja työstä sekä osallistumisesta kuurojen yhteisön toimintaan. Taustatietolomakkeen tiedot kootaan tulevaisuudessa metadatan hallintaan soveltuvan standardin mukaisesti muiden metatietojen yhteyteen. Taustatieto- ja lupalomakkeisiin saatiin perusta erityisesti ruotsalaisen viittomakielen korpusprojektista.

Pilottivaiheessa suunniteltiin aineistonkeruutilanteen studioasetelma ja kameroiden lukumäärä. Tässä auttoivat yliopistomme AV-keskuksen tekninen henkilökunta sekä muiden viittomakielten korpusprojektien aineistonkeruukokemukset. Näin luotujen studio-olojen toimivuutta keskustelijoiden ja kuuden kameran asemointeineen testattiin pilottitutkimuksessa.

Kuvausten jälkeen videot editoitiin ja tallennettiin ELAN-ohjelmassa toimiviksi mp4-tiedostoiksi. Editoinnissa eri kuvakulmista tallennetut videot synkronoitiin ja jaettiin tehtäväkohtaisiin episodeihin, metatiedot yhdistettiin niihin ja editoidut videoleikkeet nimettiin systemaattisesti pilotissa luotujen konventioiden mukaisesti. Tällä tavoin mahdollistetaan taustamuuttujiltaan erityyppisten aineistojen haku ja keskinäinen vertailu. Editoitu materiaali tallennettiin sekä yliopiston palvelimelle että CSC:n[13] IDA-tallennuspalveluun.

Pilottivaiheen jälkeen keväällä 2014 käynnistyi systemaattinen aineiston keruu ja käsittely sekä annotointikonventioiden kehittäminen ja metatietojen käsittely nelivuotisessa CFINSL-projektissa. Tällöin tarkasteltiin myös pilottivaiheen kokemuksia ja palautteita, joiden pohjalta työskentelyä kehitettiin. Tähdellistä oli myös selvittää, missä aineiston raakamateriaali sekä editoidut ja annotoidut videot metatietoineen säilytetään.

Pilottivaiheen arvioinnissa havaittiin, että aineistonkeruusession ohjeistus toimi melko hyvin. Kuitenkin informanttien tutustuttaminen kuvaustilanteeseen ja korpuksen tarkoitukseen on tehtävä huolellisemmin. On myös tarpeellista selittää informanteille viittomakielellä suomeksi kirjoitettujen lupa- ja taustatietolomakkeiden sisältö. Lisäksi aineistokeruusession ohjaajan pitää olla viittomakielinen, eikä kuvausstudiossa saa olla läsnä hänen lisäkseen muita henkilöitä.

Pilottitutkimus osoitti myös, että tehtävät ja elisitaatiomateriaalit toimivat hyvin, mutta monissa korpusprojekteissa käytetty videon uudelleen kertominen olisi hyvä täydentävä tehtävä. Lisätehtävästä huolimatta aineistonkeruutilanne saisi kuitenkin kestää korkeintaan noin tunnin. Lisäksi on tarpeen pitää taukoja tehtävien välillä ja tarkistaa, että informantit ovat ymmärtäneet tehtävät. Heillä pitää olla myös mahdollisuus välillä kysyä tai tarkentaa epäselväksi jäänyttä asiaa. Pilotissa korostui myös se, että rauhallisen ja kiireettömän tunnelman aikaansaaminen on tärkeää, jotta informantit voivat viittoa mahdollisimman vapautuneesti.

Annotaatiokonventioiden luominen on aikaa vievää ja vaatii systemaattista työtä aineiston parissa, joten niiden perusteellinen kehittäminen jäi varsinaisen korpusprojektin tehtäväksi. Tässä ongelmia aiheutti se, ettei Suomessa käytetyistä viittomakielistä ole olemassa laajoja sanakirjoja, joihin glossien valinnan voisi perustaa annotointikonventioita rakennettaessa.

3.2 Aineistonkeruu

3.2.1 Informantit ja kielet

Jo pilottivaiheessa tehtiin varsinaisen korpuksen kielivalintaa ja kattavuutta koskeva aineistonkeruusuunnitelma. Koska sekä suomalainen että suomenruotsalainen viittomakieli ovat vähemmistökieliä ja varsinkin

[13] CSC (Tieteen ja tietotekniikan keskus): https://www.csc.fi

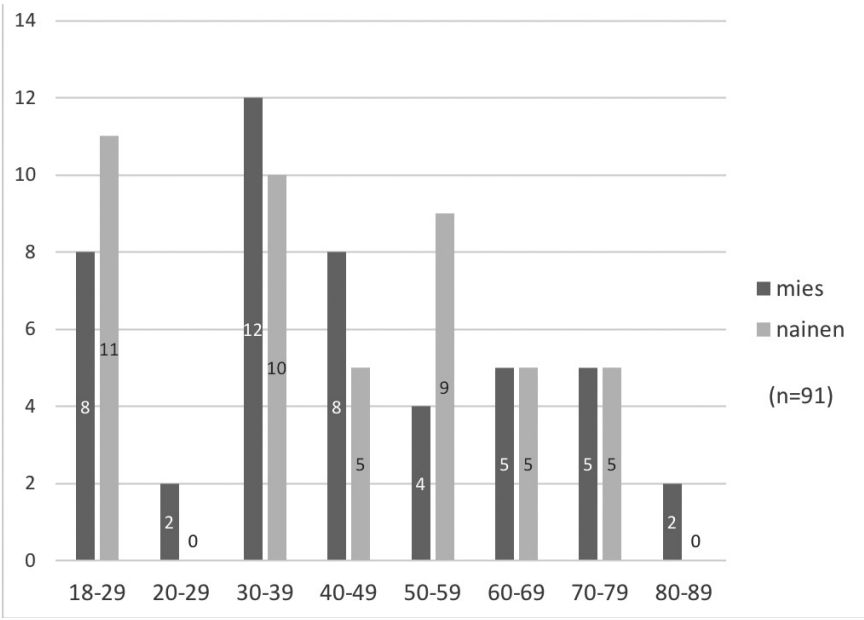
suomenruotsalainen viittomakieli erittäin uhanalainen kieli, katsottiin kummankin kielen dokumentointi ja aineiston keruu välttämättömäksi sekä kielitieteen että kieliyhteisön näkökulmasta. Suunnitelmaan kirjattiin myös, miltä alueilta aineistoa kerätään, jotta saataisiin kattava kuva kielistä ja kielenkäytöstä. Laadun ja kattavuuden kannalta on myös tarpeellista, että aineistoa kootaan eri-ikäisiltä naisilta ja miehiltä.

Suomenruotsalaisen viittomakielen käyttäjiä asuu Helsingin seudulla sekä Pohjanmaan alueella. Suomalaisen viittomakielen käyttäjiä haluttiin saada eri puolilta Suomea, joten pääalueiksi muodostuivat suurimpien kuurojenyhdistysten alueet: Helsingin seutu, Turun seutu, Keski-Suomi (Jyväskylän ja Tampereen alueet), Pohjois-Suomi (Oulun seutu, Kainuu ja Lappi) sekä Itä-Suomi. Lähes koko 1900-luvun ajan Suomessa oli viisi suomenkielistä ja yksi ruotsinkielinen kuurojen valtionkoulu, joissa kuurot viittoivat keskenään, vaikkei viittomakieliä käytetty opetuksessa. Tällöin eri kouluissa ja kuurojenyhdistysten alueilla kehittyi osin erilaisia, paikallisia viittomistoja. Tästä syystä on olennaista saada aineistoa eri puolilta Suomea asuvilta viittomakielisiltä. Korpusprojektin myöhempänä tavoitteena on kerätä aineistoa myös muutamilta taktiilia viittomakieltä käyttäviltä henkilöiltä. Taktiili viittomakieli on kuurosokeiden käyttämä kielimuoto, jossa keskustelu tapahtuu viittomalla kädestä käteen.

Laajaa viittomakieliaineistoa kerättäessä yhteys kuurojen yhteisöön on tärkeä, jotta yhteisön jäsenillä on mahdollisuus saada tietoa hankkeesta. Sen perusteella kukin voi miettiä, haluaako lähteä mukaan informantiksi. Tällä tavoin kielenkäyttäjät kokevat myös arvostusta kieltään kohtaan. Aineistonkeruun käynnistyessä korpusprojektin työntekijä kävi kunkin alueen kuurojenyhdistyksessä kertomassa viittomakielten korpustyöstä, ja hän myös rekrytoi kultakin alueelta yhdyshenkilön informanttien löytämiseksi. Informanttien rekrytoinnissa käytettiin myös sosiaalisen median kanavia.

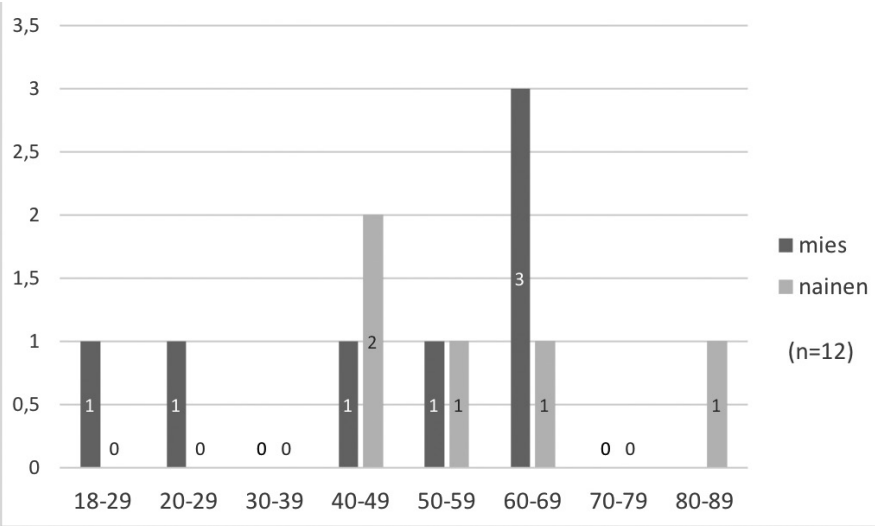
Alueellisen tasapainon lisäksi pyrittiin myös löytämään miehiä ja naisia sekä eri ikäluokkia mahdollisimman tasapuolisesti. Eri-ikäisten viittomisesta voidaan tutkia, millaista lingvististä vaihtelua kielen eri tasoilla esiintyy. Toisaalta narratiivisen perimätiedon näkökulmasta on relevanttia taltioida vanhemman sukupolven viittominen tulevia sukupolvia varten. Nuoremmissa ikäluokissa viittomakielisiä on yhä vähemmän, ja he hyötyvät vanhempien sukupolvien viittomakielen dokumentoinnista. Korpusprojektin alkuperäiseksi tavoitteeksi asetettiin kerätä aineistoa 80:ltä suomalaista viittomakieltä ja 20:ltä suomenruotsalaista viittomakieltä käyttävältä henkilöltä. Aineistonkeruun päätyttyä kuvattuja viittojia oli yhteensä 103, joista 91 oli suomalaisen ja 12 suomenruotsalaisen viittomakielen käyttäjää. Iältään he olivat 18–89 -vuotiaita (kuvio).

Kuvatuista 91:stä suomalaista viittomakieltä käyttävästä viittojasta 46 oli miehiä ja 45 naisia. Suomenruotsalaista viittomakieltä käyttävästä 12 informantista seitsemän oli miehiä ja viisi naisia. Kuviossa 1 esitellään suomalaista viittomakieltä käyttävien informanttien ikäjakauma sukupuolen mukaan eriteltynä ja kuviossa 2 samat tiedot suomenruotsalaista viittomakieltä käyttävistä informanteista. Kuviosta 3 käy ilmi, missä päin Suomea suomalaista (tummanharmaa) ja suomenruotsalaista (vaaleanharmaa) viittomakieltä käyttävät informantit olivat syntyneet. Suomalaista viittomakieltä käyttävistä henkilöistä valtaosa oli syntynyt Länsi- ja Sisä-Suomessa (31). Lisäksi mukana oli 24 Itä-Suomessa, 13 Etelä-Suomessa ja 13 Pohjois-Suomessa syntynyttä viittojaa. Lounais-Suomessa syntyneitä oli vähiten (8). Suomenruotsalaista viittomakieltä käyttävistä kuusi oli syntynyt Etelä-Suomessa, neljä Länsi- ja Sisä-Suomessa ja yksi Lounais-Suomessa.

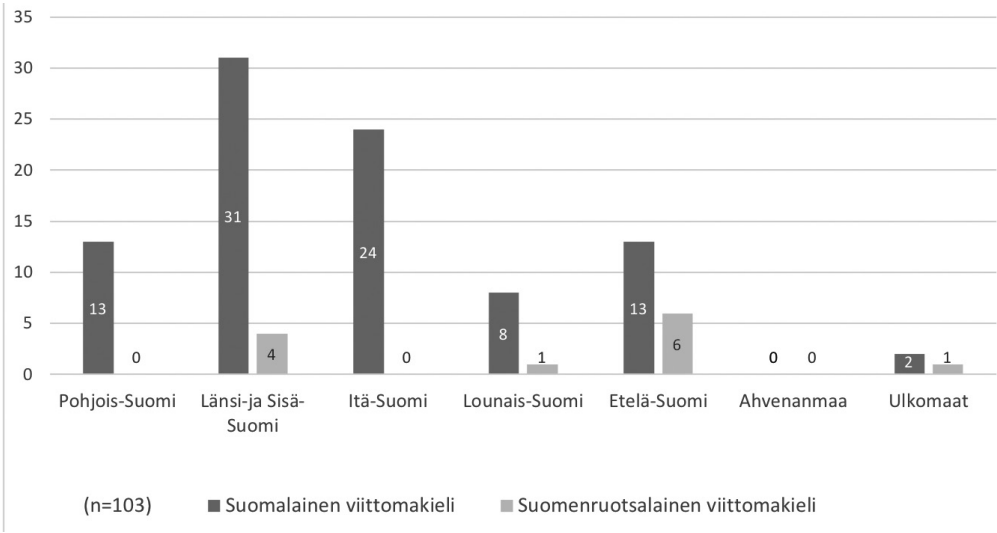


KUVIO 1. Suomalaista viittomakieltä käyttävien informanttien ikäjakauma sukupuoliryhmittäin.

Ahvenanmaalta ei ollut yhtään informanttia. Ulkomailla syntyneitä informantteja koko korpusaineistossa oli kolme henkilöä. Syntymäpaikka yhdistää viittojat eri koulualueisiin ja siten eri kielivariantteihin. Alueelliset kuurojen koulut ovat vaikuttaneet suomalaisen viittomakielen kehitykseen sekä variaatioon.



KUVIO 2. Suomenruotsalaista viittomakieltä käyttävien informanttien ikäjakauma sukupuoliryhmittäin.



KUVIO 3. Suomalaista ja suomenruotsalaista viittomakieltä käyttävien informanttien jakauma syntymäpaikkojen perusteella.

3.2.2 Keruumenetelmä

Informantit tulivat kuvaustilanteeseen pareittain. Heti ilmoittauduttuaan mukaan projektiin he olivat saaneet valita tutun parin, jonka kanssa olisi luontevaa keskustella videointitilanteessa. Aluksi informanteille esiteltiin kuvausstudio ja videotyöskentelyn kulku. Ilmapiiri pyrittiin luomaan mahdollisimman vapautuneeksi, jotta tilanne ei jännittäisi osallistujia. Kuvaustilanteen aikana informantit saivat yhden keskustelu- tai kerrontatehtävän kerrallaan toimintaohjeineen. Informanttien lisäksi studiossa oli ohjaaja, joka tarvittaessa neuvoi ja opasti kuvattavia. Hän ei kuitenkaan ollut suoraan informanttien näkökentässä, jotta ei läsnäololollaan häiritsisi keskustelua.

Seitsemän erilaisen tehtävän avulla elisitoitiin keskusteluja ja kerrontaa. Ensimmäisessä esittelytehtävässä keskustelijat kertoivat vuorotellen esimerkiksi omasta lapsuudestaan, viittomakielen oppimisesta, koulunkäynnistään ja perheensä kommunikaatiosta. Keskustelukumppanit saivat myös esittää toisilleen kysymyksiä. Toisessa tehtävässä osallistujat keskustelivat työstään tai jostain mieleisestä harrastuksestaan. He kertoivat kokemuksiaan eri kuurojen urheilutapahtumista ja järjestötoiminnasta sekä kokemuksistaan työuralla esimerkiksi, miten olivat onnistuneet saamaan jonkin työpaikan ja miten kommunikointi kuulevien työtovereiden kanssa oli luonnistunut.

Kolme seuraavaa tehtävää olivat kerrontatehtäviä. Kolmannessa tehtävässä kumpikin informantti sai valita kahdeksasta tekstittömästä Ferd'nand-sarjakuvasta neljä, tutustua niihin hetken ja sitten kertoa niistä vuorotellen. Neljännessä tehtävässä kumpikin katsoi lyhyen videon (Ohukainen ja Paksukainen tai Mr. Bean) ja sen jälkeen kertoi videon tapahtumat keskustelukumppanilleen. Viidennessä tehtävässä kumpikin osanottaja tutustui ensin tekstittömään kuvakirjaan (Lumiukko tai Frog, where are you?) ja kertoi sen jälkeen kirjansa tarinan parilleen.

Viimeiset kaksi tehtävää olivat keskustelutehtäviä. Kuudennessa tehtävässä informantit valitsivat kuurojen maailmaan liittyvän keskusteluaiheen, jonka kokivat itselleen tär-

keäksi. Tehtävän aikana keskusteltiin muun muassa yhdistystoiminnasta, kuurojen olympialaisista, kongresseista, kuurojen kulttuuripäivistä, sekä kuurojen kouluajoista ja oralismiin eli puhetta ja huuliltalukua korostavaan opetusmetodiin liittyvistä kokemuksista. Seitsemännessä tehtävässä informantit saivat keskustella vapaavalintaisesta aiheesta, kuten esimerkiksi matkailusta, TV-ohjelmasta tai urheilusta.

Monia näistä tehtävätyypeistä on käytetty useiden eri viittomakielten korpusmateriaalin keruussa, joten aineisto sallii myös viittomakielten välisen vertailevan tutkimuksen eri näkökulmista. Tämä antaa mielenkiintoisen mahdollisuuden myös suomalaisen ja suomenruotsalaisen viittomakielen vertailevaan tutkimukseen.

Kuvaustilanteen jälkeen jokainen informantti täytti suostumuslomakkeen, jossa hän määritteli oman kielellisen tuotoksensa käyttöoikeudet. Ymmärtämisen varmistamiseksi suomen- tai ruotsinkielisen lomakkeen sisältö viitottiin informanteille ennen täyttämistä. Jokaisen informantin oli mahdollista rajata oman aineistonsa käyttöä vain tutkimukseen ilman, että siitä esitetään edes kuvia artikkeleissa tai opetuksessa. Mahdollista oli myös sallia valokuvien käyttö mutta kieltää videomateriaalin käyttö muuhun kuin tutkimustarkoituksiin. Väljemmän rajauksen valinneet informantit antoivat luvan käyttää videoaineistoa tutkimuksessa ja tutkimusjulkaisuissa, mutta ei avoimesti. Laajin lupa antoi suostumuksen julkaista materiaali verkossa. Informantilla on halutessaan myöhemmin oikeus myös pyytää korpuksen hallinnoijaa poistamaan oman videoaineistonsa aineistokokoelmasta.

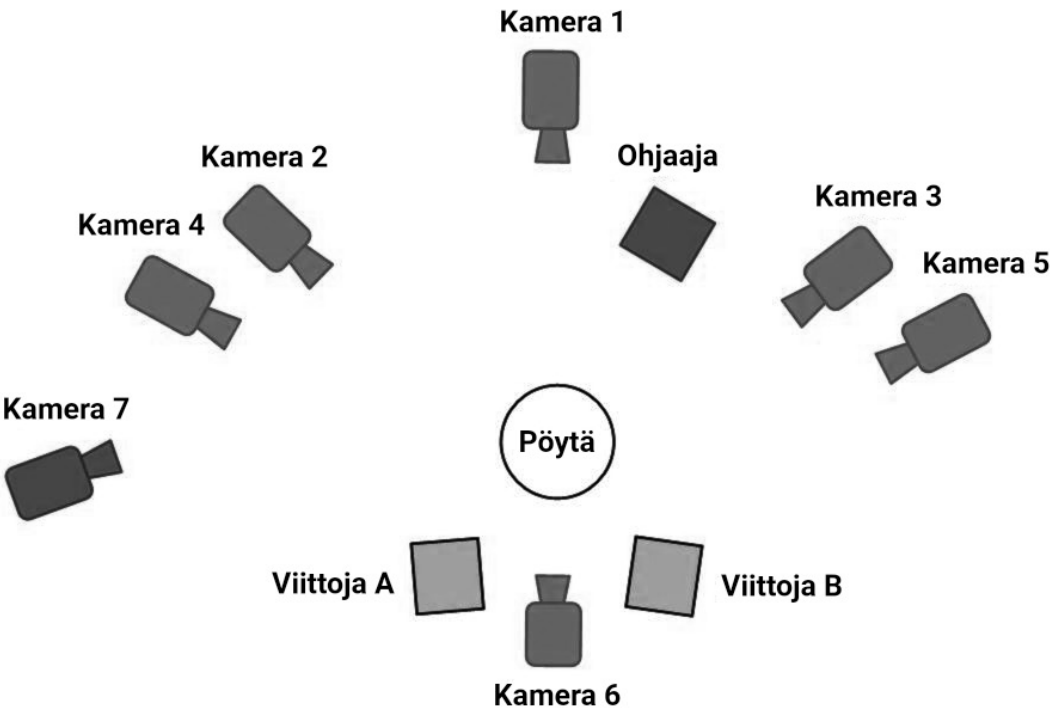
3.2.3 Aineiston kuvaus, säilytys ja saatavuus

Suurin osa aineistosta on kuvattu Jyväskylän yliopiston audiovisuaalisessa studiossa (ks. Kuva 1). Pieni osa aineistosta on kuvattu Oulun yliopiston tiloissa ja Kuurojen Liiton studiossa Valkeassa talossa Helsingissä. Kaikissa kuvauspaikoissa olosuhteet on tehty samanlaisiksi. Informantit matkustivat kotipaikkakunniltaan kuvauspaikoille, ja matkakustannukset korvattiin heille päivärahoineen.

KUVA 1. Studioasetelma kuvauksissa.

Kuvauksissa käytettiin seitsemää korkealuokkaista kameraa (Panasonic-videokamerat 3 x AG-HPX371E, 1 x AW-HE120KE, 3 x AG-HPX171E). Kamera 1 kuvasi molemmista viittojista yleiskuvaa, kamerat 2 ja 3 yleiskuvaa viittojista erikseen, kamerat 4 ja 5 lähikuvaa viittojista erikseen ja kamera 6 molempia ylhäältä päin (ks. Kuva 2). Useat kamerakulmat mahdollistavat viitottujen vuorovaikutustilanteiden moniulotteisen tarkastelun. Esimerkiksi kattoon kiinnitetyllä kameralla kuvatusta materiaalista viittojien käsien, pään ja ylävartalon syvyyssuuntaisten liikkeiden tarkastelu helpottuu huomattavasti. Seitsemännellä kameralla tallennettiin aineistonkeruun ohjaajan ja informanttien väliset keskustelut, mikäli informanteilla oli jotakin kysyttävää tehtävän aikana. Kukin tehtävä kesti keskimäärin 10–15 minuuttia lukuun ottamatta sarjakuvista kerrontaa, joka vei vain noin 5 minuuttia. Kokonaisuudessaan kuvaustilanteeseen kului aikaa 1–1 ½ tuntia.

Kuvausten jälkeen videoaineisto editoitiin tehtäväkohtaisiksi videoleikkeiksi niin, että eri kameroilla kuvattu materiaali synkronoitiin ajallisesti. Kuvatut HD-videot tallennetiin P2-kovalevyille (50 fps) MXF-formaattiin. Editoitu videoaineisto pakattiin lopulta sekä matala- että korkearesoluutioisiksi MP4-tiedostoiksi. Jokaisen informanttiparin kuvauksesta editoitiin ja tallennettiin keskimäärin 42 videoleikettä (7 tehtävää x 6 kameraa). Editoitua materiaalia kertyi suomalaisesta viittomakielestä 67 tuntia 15 minuuttia ja suomenruotsalaisesta viittomakielestä 7 tuntia 15 minuuttia. Korpuksen yhteenlaskettu aineistomäärä on siis 74 tuntia 30 minuuttia.



KUVA 2. Kameroiden asemat kuvaustilanteissa.

Aineisto (raakamateriaali, editoitu ja annotoitu materiaali sekä editointitiedot) säilytetään työstövaiheessa neljässä eri paikassa: ulkoisilla kovalevyillä kahdessa eri tilassa, Jyväskylän yliopiston humanistis-yhteiskuntatieteellisen tiedekunnan palvelimella sekä CSC:n IDA-tallennuspalvelussa[14]. Sen lisäksi annotoidut tiedostot viedään informanttien antamien lupien niin salliessa FIN-CLARINin hallinnoimaan Kielipankkiin, joka on osa kansainvälistä CLARIN-infrastruktuuria. Suomessa käytettävien viittomakielten aineisto tulee olemaan saatavilla tutkimus- ja opetuskäyttöön kunkin informantin antaman suostumuksen rajoissa. Ensimmäinen osa aineistosta on julkaistu keväällä 2019[15].

3.2.4 Metatiedot

Metatieto on Burnardin (2014) mukaan ”tietoa tiedosta”, ja se on tärkeä osa korpusta. Metatieto liitetään jokaiseen video- ja annotaatiotiedostoon, ja hyvin koottuna ja dokumentoituna se mahdollistaa pääsyn korpukseen. Metatietoja tarvitaan, jotta aineiston myöhempi käyttö olisi mahdollista, aineisto olisi ymmärrettävää ja siitä voisi tehdä monipuolisia hakuja. Viittomakielten korpuksen metatiedot dokumentoitiin ensin Excel-tiedostoon, jonka jälkeen aineistosta tuotettiin IMDI-metatietostandardien mukaiset kuvaukset. Myöhemmin IMDI-kuvauksista voidaan tarvittaessa luoda CMDI-metadatainfrastuktuurin (Component MetaData Infrastructure) mukaisia kuvailutietueita. CMDI on CLARINin kehittämä viitekehys, jota käytetään laajasti esimerkiksi kieliaineistojen metatietojen kuvaukseen ja niiden uudelleenkäyttöön[16].

[14] IDA-tallennusalvelu: https://www.fairdata.fi/en/
[15] CFINSL-aineiston julkaistu osa: http://hdl.handle.net/11113/00-0000-0000-0000-4F9F-A@view
[16] Clarin: https://www.clarin.eu/content/component-metadata

CFINSL-korpusprojektissa koottu ja dokumentoitu metatieto sisältää tietoa itse korpuksesta (korpuksen nimi, kieli, korpuksen koko, jakaja jne.), sisällöstä (kielelliset tehtävät, elisitaatiomateriaali), videoista ja annotaatiotiedostoista (muoto ja tyyppi), korpustyön taustalla olevasta projektista (nimi, kieli, tavoitteet) ja kuvaussessioista (tehtävän nimi, osanottajat, tekstilajin ja kommunikaatiotilanteen piirteet jne.). Informanteista kerättiin taustatietoja varsin kattavasti (ikä, sukupuoli, syntymä- ja asuinpaikka, tietoa vanhempien viittomakielen taidosta, kieliympäristö lapsena, koulukieli, koulutus, ammatti, yhdistystoiminta, kätisyys ym.), mutta Kielipankissa julkaistun aineiston IMDI-kuvauksen mukaisissa tiedoissa on vain henkilön yksilöivä anonymisoitu koodi, ikä ikäryhmittäin, sukupuoli, asuinalue sekä kätisyys (ks. Salonen ym. 2019).

3.3 Aineiston annotointi ja leksikkotietokanta Signbank

Annotoinnilla tarkoitetaan kirjoitetun, puhutun tai viitotun aineiston kuvaamista, luokittelua ja jäsentelyä systemaattisella tavalla. Samalla viitottu tai puhuttu aineisto muutetaan koneluettavaan muotoon (Johnston, 2010; 2016). Annotaatioon voidaan liittää esimerkiksi fonologista, morfologista ja syntaktista tietoa ilmaisujen rakenteesta, minkä ansiosta aineistohakuja voidaan tehdä erilaisilla kriteereillä. CFINSL-korpuksen aineiston annotoinnissa käytetään ELAN-ohjelmaa, joka soveltuu multimedia-aineiston monipuoliseen annotointiin (ks. Kuva 3). Ohjelman avulla voi tehdä kielellisten piirteiden hakuja myös monesta tiedostosta samanaikaisesti (Crasborn & Sloetjes, 2008).

Annotaatiokonventioiden (Salonen ym., 2018; 2019) luominen on pitkällinen prosessi. Konventioiden tulee olla systemaattisia ja johdonmukaisia, koska hakujen tekeminen korpuksesta perustuu niihin. Kaikkien

annotoijien on noudatettava samoja yhteisesti sovittuja periaatteita, jotta korpuksesta tulee tasalaatuinen, sen tutkimuksellinen käyttö on mahdollista ja sen avulla saadut tulokset luotettavia.

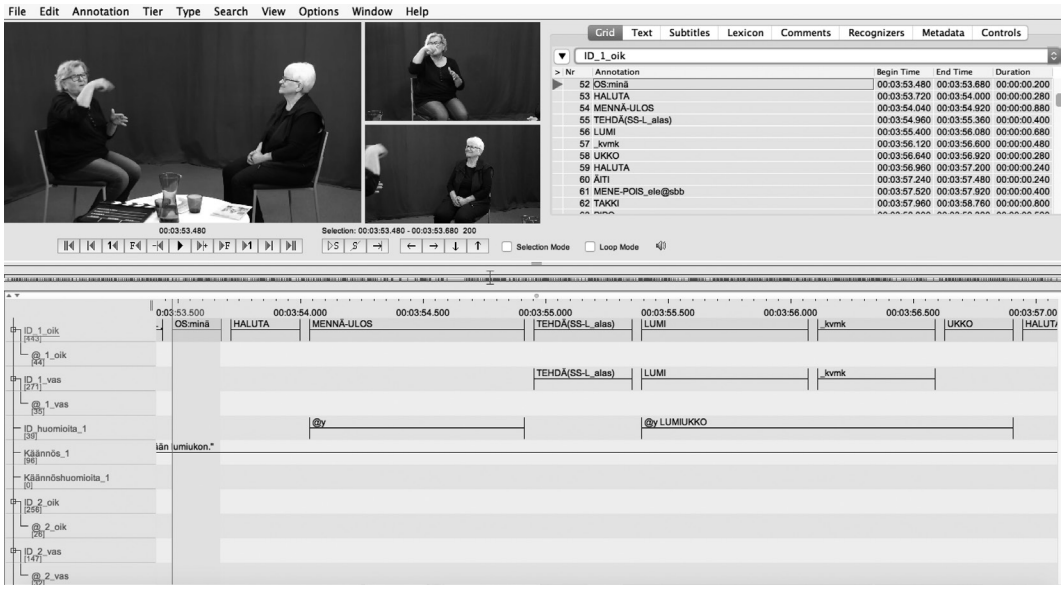
Viitotuilla kielillä ei ole yleisesti käytettyä kirjoitettua muotoa. Jonkin verran käytetään SignWriting-järjestelmää[17] erityisesti opetuksessa, mutta korpustyöhön se ei ole levinnyt. Foneettis-fonologisen tason kirjoittamiseen käytetään toisinaan HamNoSys-järjestelmää[18], joka on luotu Saksassa Hampurin yliopistossa, missä sitä käytetään myös korpustyöskentelyssä. Kun viittomakielen kirjoittamiseen ei ole omaa järjestelmää, useimmissa korpusprojekteissa käytetään puhutusta kielestä lainattuja sanoja eli glosseja, jotka valitaan sen mukaan, mikä sana kuvaa parhaiten kunkin viittoman keskeistä merkitystä. CFINSL-korpuksessa viittomien merkitsemiseen siis käytetään suomen- tai ruotsinkielistä glossia, joka kirjoitetaan suuraakkosin ja perusmuotoisena. Esimerkkinä tällaisesta on seuraava lause:

(1) OS:MINÄ HALUTA MENNÄ-ULOS TEHDÄ LUMI vartalo(B)_kvmk UKKO.

'Minä haluan mennä ulos tekemään lumiukkoa.'

(kv=muotoa kuvaileva viittoma, vartalo(B)=viitotaan käsimuodolla B)

Korpustyön tärkeä osa ja apuväline on leksikkotietokanta, johon tallennettuun aineistoon annotaatio pohjautuu. CFINSL-projektissa on luotu annotaatiotyön ohessa verkkopohjaista leksikkotietokantaa, jonka työkaluna on alkuaan Australiassa kehitetty Signbank[19]. Suomalaiseen korpustyöhön saatiin Hollannista Signbank-tietokantaversio, jota on edelleen kehitetty Suomen viittomakielten kon-



KUVA 3. Ruutukaappaus ELAN-ohjelman näkymästä annotaatioriveineen.

[17] SignWriting: http://www.signwriting.org/about/what/what02.html

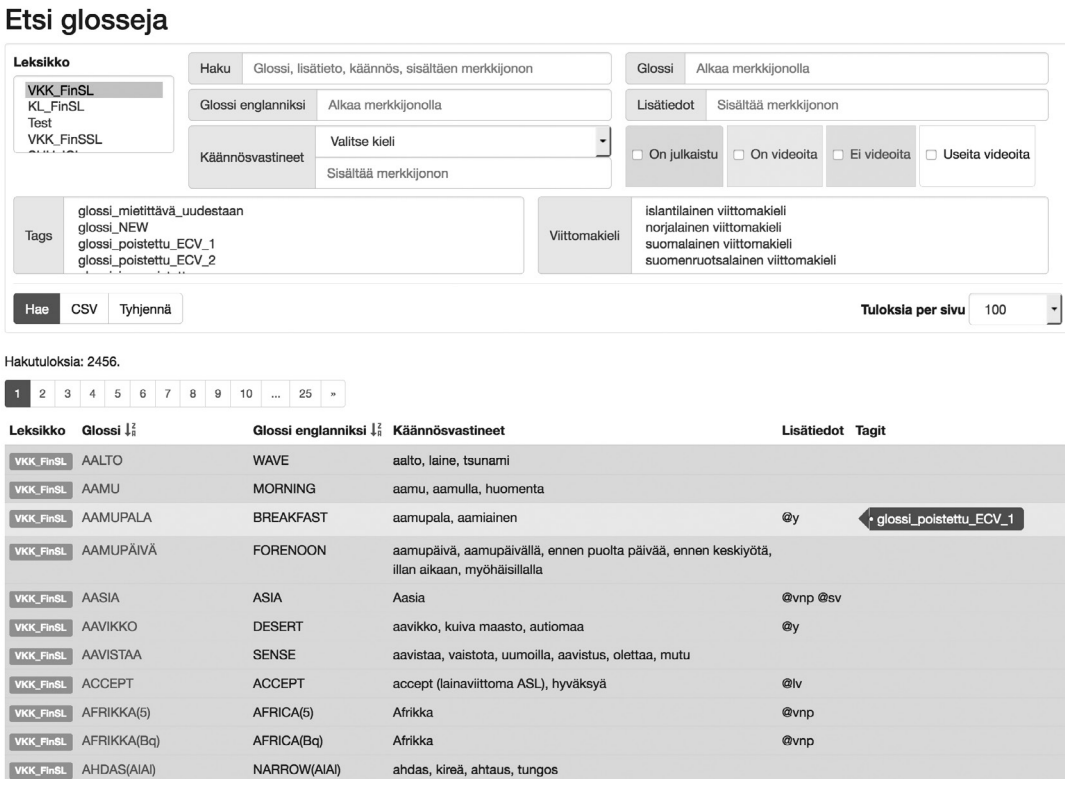
[18] HamNoSys https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/hamnosys-97.html

[19] http://www.auslan.org.au

teksteihin. Tästä syntyi FinSL-signbank[20], ohjelmisto, joka toimii Suomen Signbankin[21] alustana. Sitä on kehitetty CFINSL-projektin ja Kuurojen Liiton korpus- ja sanakirjatyön yhteistyönä. FinSL-signbank on avoimen lähdekoodin sovellus, joka on kenen tahansa vapaasti käytettävissä.

Koska annotoinnin pohjaksi ei alussa ollut riittävän laajaa viittomakielen sanakirjaa, aloitettiin työ merkitysjohtoisesta annotoinnista ja koottiin viittomia kuvaavia glosseja Excel-taulukkoon. Myöhemmin suomalaisen Signbank-version kehittyessä glossit siirrettiin Signbank-tietokantaan, jossa niitä voitiin hallita. Jokaiseen glossitietueeseen liitettiin myös video viittoman muodosta. Verkkoyhteyden välityksellä glosseja voidaan käyttää ELAN-ohjelmassa olevan ECV-ominaisuuden (external controlled vocabulary) ansiosta. Annotoitaessa ECV hakee glosseja Signbankissa olevasta aineistosta. Jos viittomalle ole vielä glossia, se voidaan lisätä Signbankiin ECV:n välityksellä (Kuva 4). Tämä työkalu nopeuttaa ja johdonmukaistaa annotointia sekä minimoi manuaalisessa annotoinnissa mahdollisesti tapahtuvia virheitä. Annotointityön edetessä leksikkotietokanta täydentyy koko ajan.

Kun viittomiston määrä leksikkotietokannassa kasvaa, voidaan siirtyä käyttämään yleiseli tunnisteglosseja, joita nimitetään myös ID-glosseiksi eli viittomia identifioiviksi glosseiksi (Johnston, 2016). Silloin viittoma, jolla on useita toisiinsa läheisesti tai etäisem-



KUVA 4. Kuvakaappaus Signbankin glossilistasta

[20] FinSL-signbank https://github.com/Signbank/FinSL-signbank
[21] Suomen Signbank https://signbank.csc.fi/

min liittyviä merkityksiä, merkitään annotaatiossa samalla glossilla, tunnisteglossilla. Esimerkkinä tästä on viittoma YRITTÄÄ (Kuva 5), jolla on monia merkityksiä mutta jonka tunnisteglossi on annotaatiossa aina sama.

KUVA 5. Viittoma YRITTÄÄ ja käännösvastineet: 'harrastaa', 'harrastus', 'yrittää', 'ahkera', 'ahkeroida', 'uuras', 'uurastaa', 'uuttera', 'innokas', 'kiire'.[22]

[22] https://signbank.csc.fi/dictionary/?gloss=yrittää&keyword=&dataset=1

Korpuksen perusannotaatiossa käytetään viittomia identifioivia tunnisteglosseja sekä SVK-aineistossa vähintään suomenkielistä ja SRVK-aineistossa ruotsinkielistä käännöstä. Kommentteja varten ELAN-ohjelmassa on oma rivinsä. Kukin tutkija voi myöhemmin lisätä perusannotaatioon oman tutkimusteemansa mukaisia rivejä.

Leksikaalistuneet viittomat merkitään glossilla (esim. HALUTA, PALLO, PUNAINEN). Luokkatunnisteen leksikaalistuneista viittomista saavat ainoastaan numeraaliviittomat (_num), esimerkiksi SATA-VIISI_num. Tämän lisäksi luokkatunnisteita käytetään kuvailevista viittomista (_kv) (ks. esimerkki (1) edellä), elemäisistä viittomista (_ele) (KÄMMEN-ALAS_ele) ja sormiaakkosviittomista (_sa) (k-a-l-l-e_sa). Viittomiin voidaan liittää lisätietoja esimerkiksi viittoman käsimuodosta, liikkeestä, käden orientaatiosta tai artikulaatiopaikasta.

Tunnisteglossit merkitään ELAN-annotaatiossa emo- eli pääriveille, kun taas kieliopillista tietoa voidaan merkitä emoriviin yhteydessä olevalle tytärriville (alenevalle riville), jonka tunnisteena käytetään @-merkkiä. Kieliopillista tietoa merkitään kiellosta, monikosta ja toistosta, listapoijuiksi kutsutuista diskurssinmerkitsimistä (ks. Liddell 2003) ja yhdysviittomista, lainaviittomista, epäselvistä viittomista sekä päännyökkäyksestä ja -pudistuksesta. Kuva 6 havainnollistaa ELAN-annotointia.

KUVA 6. Kuvakaappaus ELAN-annotaatiosta.

Metakielen eli toisen kielen (CFINSL-korpuksessa suomi ja ruotsi) käyttö viittomakielisen aineiston annotoinnissa aiheuttaa monia haasteita. Esimerkiksi niin sanotut visuospatiaaliset kuvailevat viittomat, joilla ei ole kiinteää muotoa, kääntyvät puhutulle kielelle useampana sanana tai jopa lauseena. Tällaiset viittomat identifioidaan aineistossa kuvailevan viittoman luokkatunnisteella *_kv*. Tätä havainnollistaa edellä olevan esimerkin (1) ilmaus ja Kuva 3, jossa lumesta tehtyä hahmoa kuvaillaan sen muotoa jäljittelevällä viittomalla, joka on annotoitu glossilla *vartalo(B)_kvmk*. Glossissa lyhenne *_kv* identifioi kuvailevan viittoman ja *mk*-tarkenne luokittelee viittoman nimenomaan muotoa ja kokoa kuvaavaksi (ks. tarkemmin Takkinen, 2008; Takkinen, Keränen & Salonen, 2018).

Kääntäminen suomen ja ruotsin kielelle aloitetaan erottelemalla viitotusta tekstistä mielekkäitä ilmauskokonaisuuksia. Käännöksissä pyritään ilmauksiin, jotka kertovat viitotun asiasisällön lähtökielen tapaa noudatellen. Ne sisältävät viittomin ilmaistun sisällön lisäksi myös ei-manuaalisen sisällön. Tällöin käännöksistä on tukea tulkittaessa glossirivejä.

Viittomakielen annotointi eroaa siis radikaalisti puheen annotoinnista kielissä, joilla on kirjoitusjärjestelmä. Vaatii runsaasti aikaa ja annotointikokeiluja, ennen kuin käyttökelpoisimmat konventiot muotoutuvat. Niiden luonti eteneekin pienin askelin ja usein aiempia tapoja paremmiksi muokaten. Lisäksi eri metakielet voivat aiheuttaa erilaisia haasteita (vrt. englanti vs. suomi vs. japani). Jokainen

konteksti edellyttää omanlaistaan hienosäätöä, jotta annotaatio tukee saumattomasti aineiston hakuprosesseja.

3.4 Typologinen näkökulma Suomen viittomakielten korpukseen

Jantunen tuo esiin oppijansuomen korpusta esittelevässä artikkelissaan (2011: 90–92) muutamia korpusten typologisia dimensioita. Niitä ovat esimerkiksi

genredimensio
(yleistekstilajinen vs. monitekstilajinen)

teemadimensio
(yleiskorpus vs. terminologinen korpus)

rekisteridimensio
(kirjoitetun vs. puhutun kielen korpus)

kielidimensio
(yksikielinen vs. kaksikielinen (rinnakkais-) vs. monikielinen)

varianttidimensio
(yksivarianttinen vs. verrannollinen, joka sisältää useita variantteja tai osakorpuksia)

käännösdimensio
(ei-käännöskorpus vs. käännöskorpus)

aikadimensio
(synkroninen vs. diakroninen)

otantadimensio
(kokotekstikorpus vs. otekorpus)

mediumdimensio
(sähköisenä vs. käsinkirjoitettuna kerätyt tekstit)

annotaatiodimensio
(raakatekstikorpus vs. annotoitu korpus)

Näiden dimensioiden valossa tarkasteltuna muun muassa tässä esitellyt viittomakielten korpukset ovat *useita tekstilajeja* sisältäviä *yleiskorpuksia*. Ne ovat *kokotekstikorpuksia*, jotka on *kerätty videoimalla* ja sitten siirretty *annotoimalla* koneluettavaan tekstimuotoon. Tällaiset viittomakielikorpukset ovat lähinnä *puhutun kielen korpuksien* kaltaisia, *yksikielisiä ei-käännöskorpuksia*[23], jotka voivat sisältää myös *osakorpuksia*. Tässä artikkelissa esitellyt viittomakielikorpukset ovat myös *synkronisia* (ei siis ajallisesti peräkkäistä aineistoa). Esimerkiksi viittomakielen omaksumiseen liittyvät korpukset voivat kuitenkin olla diakronisia eli kerätty samoilta henkilöiltä eri aikoina (esim. Takkinen, 2003; 2013).

4 PÄÄTÄNTÖ

CFINSL-korpuksella on tärkeä merkitys sekä suomalaisen että suomenruotsalaisen viittomakielen aseman vahvistamisessa yhteiskunnallisesti ja kielenhuollon näkökulmasta (alueelliset variaatiot huomioiden). Korpuksen avulla viittomakieliämme voidaan dokumentoida ja tallentaa nykyisille ja tuleville sukupolville. Viittomakielten korpusten luonti ja tallentaminen on kulttuurisesti merkittävä työ. Se on kieliyhteisöjä ja niiden kulttuuria arvostavaa toimintaa, joka vahvistaa viittomakielten tunnettuutta ja kielenkäyttäjien kielellistä identiteettiä. Korpukset lisäävät mahdollisuuksia avoimeen keskusteluun viittomakieltemme merkityksestä nykypäivänä. Nämä näkökulmat ovat tärkeitä kummallekin maassamme käytettävälle viittomakielelle, mutta erityisen tärkeitä ne ovat suomenruotsalaiselle viittomakielelle, joka Unescon kriteerien mukaan on vakavasti uhanalainen kieli[24].

CFINSL-korpusaineisto sisältää eri-ikäisten kielenkäyttäjien viittomista. Koska korpusaineisto on synkronista, siitä ei voi kuitenkaan tehdä samanlaisia päätelmiä kielen

[23] Poikkeuksena myöhemmin tekstissä mainittu Kuurojen liiton Kipo-korpus, joka on käännöskorpus.
[24] https://www.kotus.fi/kielitieto/kielet/suomen_viittomakielet

muutoksesta kuin diakronisesta aineistosta. Siitä voi kuitenkin tarkastella eri-ikäisten henkilöiden välisiä eroja viittomistavoissa, jotka taas heijastavat kielen muuttumista ajan ja maailman muutoksen mukana. Toisaalta korpus on poikkileikkaus 2010-luvulla Suomessa käytetyistä viittomakielistä, ja se mahdollistaa tulevaisuudessa aineistojen vertailun nuorempien polvien viittomakieliin, joihin vaikuttavat esimerkiksi valtakunnalliset viestintäkanavat ja uudet mediat. Lisäksi suomalaisesta ja suomenruotsalaisesta viittomakielestä samalla tavalla kerätty aineisto tarjoaa oivan mahdollisuuden näiden kielten vertailevaan tutkimukseen niin viittomiston kuin rakenteenkin näkökulmasta. Viittomakielten korpusten avulla voidaan myös vertailla eri maiden viittomakieliä, mikä on tärkeää muun muassa kielitypologisen tutkimuksen näkökulmasta.

Korpukset mahdollistavat siis luotettavan ja systemaattisen pohjan viittomakielten tutkimukselle ja opetuksen kehittämiselle. Kun tutkimus perustuu isompiin aineistoihin, ovat tulokset luotettavampia. Näin voidaan myös testata pienemmillä aineistoilla saatuja tuloksia. Annotoitu kieliaineisto myös nopeuttaa tutkimusta ja edesauttaa kielten kuvausta ja kieliopin laadintaa. Se on lisäksi avain kielen variaation hahmottamiseen. Viittomakielen opetukselle korpusaineisto tarjoaa autenttista kielimateriaalia, ja toisaalta korpusaineistoon nojautuen on aikaisempaa helpompi laatia varsinaista oppimateriaalia. Kielenoppijan ja muiden korpusten käyttäjien on tärkeä oppia hyödyntämään korpusaineistoa siinä käytettyjen ohjelmien ja hakutoimintojen avulla. Tämä edellyttää viittomakielisen yhteisön jäsenten ja viittomakielen opettajien koulutusta korpuksen käytössä.

Korpuksen luominen kaikkine vaiheineen on hyvin työlästä ja aikaa vievää. Siksi on tärkeää, että mahdollisimman moni pääsee hyötymään julkaistusta aineistosta tutkijana, opettajana tai oppijana – ylipäätään kielestä kiinnostuneena ihmisenä. Tästä syystä tavoitteena on julkaista suomalaisten viittomakielten korpukset Kielipankissa muiden kielten korpusten joukossa.

CFINSL-projektia ovat tukeneet Opetus- ja kulttuuriministeriö, Bovalliuksen säätiö ja Svenska kulturfonden, mistä heille kiitokset. Osoitamme lämpimät kiitokset kielenoppaille, jotka ovat osallistuneet aineiston tuottamiseen. Parhain kiitos myös artikkelin anonyymeille arvioitsijoille rakentavista kommenteista ja parannusehdotuksista.

LÄHTEET

Burnard, L. (2014). *Metadata for corpus work*. https://www.academia.edu/3234836/Metadata_for_corpus_work

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. Teoksessa *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages* (s. 39–43). Paris: ELRA.

Hoyer, K. (2000). *Variation i teckenspråk: en studie av släktskapsterminologi i teckenspråk.* Helsingfors: Finlands dövas förbund.

Hoyer, K. (2004). The sociolinguistic situation of Finland-Swedish Deaf people and their language, Finland-Swedish Sign Language. Teoksessa M. Herreweghe & M. Vermeerbergen (toim.), *To the lexicon and beyond: Sociolinguistics in European Deaf Communities* (s. 3–23). Washington DC: Gallaudet University press.

Hoyer, K. (2012). *Dokumentation och beskrivning som språkplanering : perspektiv från arbete med tre tecknade minoritetsspråk*. Akademisk avhandling. Nordica Helsingensia 29. Finska, finskugriska och nordiska institutionen, Helsingfors universitet. URN:ISBN:978-952-10-7612-1

Hunston, S. (2008). Collecting strategies and design decisions. Teoksessa A. Lüdeling & M. Kytö (toim.), *Corpus Linguistics. An International Handbook. Volume 1* (s. 154–167). Berlin: De Gruyter.

Jantunen, J. H. (2011). Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. Teoksessa A. Kaivapalu, P. Muikku-Werner, J. Laakso & M-M. Sepper (toim.), *Lähivõrdlusi. Lähivertailuja No* 21 *(s.* 86–105). Tallinn: Eesti Rakenduslingvistika Ühing. doi:10.5128/LV21.04

Jantunen, T. (2000). *Suomalaisen viittomakielen synnystä, vakiintumisesta ja kuvaamisen periaatteista*. Yleisen kielitieteen pro gradu -tutkielma. Helsingin yliopisto

Jantunen, T. & Pippuri, O. (2016). Snowfrog loikkasi LATiin – ProGramin tarina-aineisto Kielipankissa. *Kielisilta, 3*, 13–16.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, *15,* 106–131.

Johnston, T. (2012). Lexical frequency in signed languages. *The Journal of Deaf Studies and Deaf Education, 17*, 163–193.

Johnston, Trevor (2016). *Auslan Corpus Annotation Guidelines. February 2016 version.* Centre for Language Sciences, Department of Linguistics, Macquarie University, Sydney, Australia.

Kuurojen liitto (2018). *Viittomakielet ja viittomakieliset*. Haettu 31.1.2019 osoitteesta www.kuurojenliitto.fi/fi/viittomakielet/viittomakielet-ja-viittomakieliset

Liddell, S. (2003). *Grammar, gesture, and meaning in American sign language*. Cambridge: Cambridge University Press.

Lüdeling, A. & Kytö, M. (toim.) (2008). *Corpus Linguistics. An International Handbook. Volume 1*. Berlin: De Gruyter.

Mesch, J. (2006). Påminner nationella teckenspråk om varandra? Teoksessa K. Hoyer, M. Londen & J-O. Östman (toim.), *Teckenspråk: Sociala och historiska perspektiv* (s. 71–95). Teckenspråkstudier 2. Helsingfors: Helsingfors universitet, Institutionen för nordiska språk och nordisk litteratur.

Posti, A. (2008). Onko viittomakielemme uhattuna? *Kuurojen Lehti, 113(4),* 12–13.

Salmi, E. & Laakso, M. (2005). *Maahan lämpimään. Suomen viittomakielisten historia.* Helsinki: Kuurojen Liitto.

Salonen, J., Puupponen, A., Takkinen, R. & Jantunen, T. (2019). Suomen viittomakielten korpusta rakentamassa. Teoksessa J. H. Jantunen, S. Brunni, N. Kunnas, S. Palviainen & K. Västi (toim.), *Proceedings of the Research data and humanities (RDHum) 2019 conference: data, methods and tools* (s. 83–98). Studia Humaniora Ouluensia 17. Oulu: Oulun yliopisto, Humanistinen tiedekunta

Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2018). *Suomen viittomakielten korpusprojektin (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 korpusprojektin (CFINSL) annotointiohjeet.* 2. versio. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto. http://r.jyu.fi/ygR

Sinclair, J. (2005). Corpus and Text — Basic Principles. Teoksessa M. Wynne (toim.), *Developing Linguistic Corpora: a Guide to Good Practice.* Haettu 18.2.2019 osoitteesta http://ota.ox.ac.uk/documents/creating/dlc/

Soininen, M. (2016). Selvitys suomenruotsalaisen viittomakielen kokonaistilanteesta. Selvityksiä ja ohjeita 2/2016. Oikeusministeriö. http://urn.fi/URN:ISBN:978-952-259-490-7

Suomen viittomakielten kielipoliittinen ohjelma (2010). Helsinki: Kuurojen Liitto & Kotimaisten kielten tutkimuskeskus. http://scripta.kotus.fi/www/verkkojulkaisut/julk15/

Takkinen, R. (2003). Viittomakielen omaksuminen äidinkielisessä ja kuulevassa viittomakieltä käyttävässä ympäristössä. *Puhe ja kieli, 23*, 151–164.

Takkinen, R. (2008). Kuvailevat verbit suomalaisessa viittomakielessä. *Puhe ja kieli, 28*, 17–40.

Takkinen, R. (2013). Sisäkorvaistutetta käyttävien lasten viittomakielen ja puhutun kielen omaksuminen. Teoksessa A. Kaivapalu, P. Muikku-Werner, J. Laakso, K. Õim & M-M. Sepper (toim.), Lähivõrdlusi. *Lähivertailuja* No 23 (s. 371–402). Tallinn: Eesti Rakenduslingvistika Ühing. DOI: http://dx.doi.org/10.5128/LV23.15

Takkinen, R., Keränen, J. & Salonen, J. (2018). Depicting Signs and Different Text Genres: Prelimary Observations in the Corpus of Finnish Sign Language. In M. Bono, E. Efthimiou, F. Stavroula-Evita, T. Hanke, J. Hochgesang, J. Kristoffersen, J. Mesch & Y. Osugi (toim.), *Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* [organized as a part of LREC'18 at Miyazaki, Japan, May 12, 2018] (s. 189–194). Paris: European Language Resources Association (ELRA). http://lrec-conf.org/workshops/lrec2018/W1/pdf/18038_W1.pdf

Wallvik, B. (1997). *...ett folk utan land....* Borgå: Döva och hörselskadade barns stödförening.

Wichman, A. (2008). Speech corpora and spoken corpora. Teoksessa A. Lüdeling & M. Kytö (toim.), *Corpus Linguistics. An International Handbook. Volume 1* (s. 187–206). Berlin: De Gruyter.

HOW IS A SIGN LANGUAGE CORPUS CREATED AND FOR WHAT?

Ritva Takkinen, University of Jyväskylä, Department of Language and Communication Studies
Juhana Salonen, University of Jyväskylä, Department of Language and Communication Studies
Anna Puupponen, University of Jyväskylä, Department of Language and Communication Studies
Henri Nieminen, University of Jyväskylä, Department of Language and Communication Studies

This article deals with the construction of the corpora of Finnish sign language and Finland-Swedish sign language in the CFINSL project (Corpus project of Finland's Sign Languages). Sign languages do not have a written form, thus the construction of corpora demands a different approach compared to the spoken languages which have a written form. This article presents the corpora constructed in the Sign Language Centre in the University of Jyväskylä: the collection of the material; the technical processing of the videos; the collection and the processing of metadata; the annotation of the recorded material; and the storage and the publication of the material. In addition to the corpora, a lexical database, Signbank, has been created. It facilitates the annotation process and helps the use of the corpora in research and instruction. The corpora also document the sign languages used in Finland for the language societies today and for future generations.

Keywords: annotation, corpus, Finnish sign language, Finland-Swedish sign language, lexical database, Signbank, sign language corpus