Creating Corpora of Finland's Sign Languages

Juhana Salonen, Ritva Takkinen, Anna Puupponen, Henri Nieminen, Outi Pippuri
Department of Languages (Sign Language Centre), University of Jyväskylä, Finland
P.O. Box 35, FI-40014 University of Jyväskylä, Finland
E-mail: {juhana.salonen, ritva.a.takkinen, anna.puupponen, outi.pippuri}@jyu.fi & henri.i.nieminen@student.jyu.fi

Abstract

This paper discusses the process of creating corpora of the sign languages used in Finland, Finnish Sign Language (FinSL) and Finland-Swedish Sign Language (FinSSL). It describes the process of getting informants and data, editing and storing the data, the general principles of annotation, and the creation of a web-based lexical database, the FinSL Signbank, developed on the basis of the NGT Signbank, which is a branch of the Auslan Signbank. The corpus project of Finland's Sign Languages (CFINSL) started in 2014 at the Sign Language Centre of the University of Jyväskylä. Its aim is to collect conversations and narrations from 80 FinSL users and 20 FinSSL users who are living in different parts of Finland. The participants are filmed in signing sessions led by a native signer in the Audio-visual Research Centre at the University of Jyväskylä. The edited material is stored in the storage service provided by the CSC – IT Center for Science, and the metadata will be saved into CMDI metadata. Every informant is asked to sign a consent form where they state for what kinds of purposes their signing can be used. The corpus data are annotated using the ELAN tool. At the moment, annotations are created on the levels of glosses and translation.

Keywords: sign language corpus, Finnish Sign Language, Finland-Swedish Sign Language, annotation, metadata, Signbank

1. Background

In Finland there are two official sign languages, Finnish Sign Language (FinSL) and Finland-Swedish Sign language (FinSSL). FinSL is used mainly by deaf people who come from Finnish-speaking families and have attended Finnish deaf schools. The estimated number of deaf FinSL users is 4000–5000 and of hearing native signers (mainly codas) and second language users approximately 6000–9000[1]. FinSSL, on the other hand, is used mainly in the coastal areas of Finland among those deaf people whose family background is Swedish speaking. The number of deaf FinSSL users is now estimated at approximately 90, most of them over 55 years of age (Soininen, 2016). The creation of corpora will enable us to conduct wider, deeper, more diverse and more reliable research, on which we will be able to construct a comprehensive dictionary and a descriptive grammar of these two languages. Creating the corpus especially for FinSSL is crucial as the number of users is very small and includes mainly elderly people. It is essential that the documentation of the language takes place at once.

The corpus project was piloted at the Sign Language Centre of the University of Jyväskylä in 2013. In spring 2014 the four-year (2014–2018) CFINSL[2] project began, its aim to document both FinSL and FinSSL. The documentation will serve both linguistic (vocabulary, structure, language use, variation) and cultural (topics related to the deaf community) purposes as well as teaching. We aim to collect conversations and narrations from 80 FinSL users and 20 FinSSL users who are living in different parts of Finland.

[1] *http://www.kuurojenliitto.fi/fi/viittomakielet/viittomakielet-ja-viittomakieliset#.VrBp5E1f3L8*

[2] *https://www.jyu.fi/hum/laitokset/kielet/oppiaineet_kls/viittomakieli/tutkimus/menossa-olevat-projektit/suomen-viittomakielten-korpusprojekti*

2. Procedure

2.1 Collecting the data

In the project we collect data from participants in different parts of Finland with the help of contact persons in the deaf clubs. The material is recorded in the Audio-visual Research Centre at the University of Jyväskylä. The material is recorded in a professional setting in order to produce high-quality video material, for example for the quantitative phonetic analysis of FinSL and FinSSL (with e.g. computer-vision based technologies). We have tried to ensure a wide range of regional variation by recruiting participants from seven different parts of Finland (see the map in Figure 1).

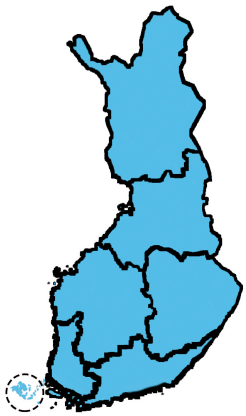


Figure 1: A map showing the areas where participants are recruited.

Normally both of the participants in a dialogue setting are from the same area in order to preserve and make clear any regional variation. Also the age variety is taken into account: we aim to get as even a distribution as possible across a range of ages: 18–29, 30–39, 40–54, 55–69 and 70–.

Participants are invited to a signing session led by a native signer. They are asked to perform seven language tasks, all of which are carried out in a dialogue setting. Tasks 1–2 and 6–7 are discussions, while tasks 3–5 are semi-interactive monologues. The tasks involve 1) introductions, 2) a discussion of work or hobbies, 3) narrating about cartoon strips (Ferd'nand), 4) narrating about a video, 5) narrating a story from a picture book (The Snowman, and Frog, where are you?), 6) discussing a topic related to the deaf world, and 7) free discussion (e.g. on travelling, TV-programmes, sports). Since some of the elicitation materials have also been used when collecting corpus material in other sign languages (e.g. Nishio et al., 2010; Mesch, 2015), the data will allow cross-linguistic comparison.

The video recording takes place in a studio of the Audio-visual Research Centre (see Figure 2). Before the recording session the instructor, a native signer, has a discussion with the two participants and explains what will happen in the signing situation. During the recording the instructor and the participants are present in the studio and the technicians are in a separate control room (see Figure 3). The instructor gives the participants instructions before each task. During the tasks he is available if more information is needed but otherwise he leaves the participants to discuss freely.

In the first task the participants take it in turn to introduce themselves. The other participant can ask for more information if he/she wants to. Task 2, telling about work or hobbies, is also signed by each signer in turn but discussion is free during each turn. Narrations about cartoons, videos and picture books (tasks 3, 4 and 5) are individual narrations, and discussion may take place afterwards. Tasks 6 and 7 are free dialogues and include a discussion about the deaf world and a free discussion. The length of the sessions is between one and a half and two hours.

Figure 2: The studio

Figure 3: The control room

The video recording takes place at the studio with seven Panasonic video cameras (3 x AG-HPX371E, 1 x AW-HE120KE, 3 x AG-HPX171E). Camera 1 records a general view of the situation, camera 2 records a complete picture of Signer B and camera 3 a complete picture of Signer A. Cameras 4 and 5 are angled towards the torso and face of Signers B and A, respectively. Camera 6 is angled towards the signers from directly above in order to get exact information of the movements of the head, body and hands on the sagittal plane. Camera 7 is directed towards the instructor in order to record the instructions given before and possibly during the tasks (see Figure 4). The HD films are saved in P2-disks (25–50 fps), stored in MXF format and compressed into low and high resolution MP4 files.

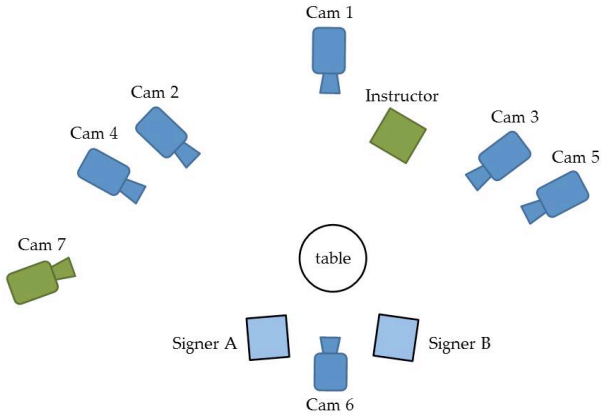


Figure 4: Recording in the studio: Camera setting

The edited material is stored in the storage service provided by the CSC – IT Center for Science[3], which is a state-owned company administered by the Finnish Ministry of Education and Culture. In addition, the annotated files will be stored in the Language Bank of Finland administered by FIN-CLARIN[4], which is part of the international CLARIN infrastructure. The data will be available for research and teaching purposes when permitted by the language informants.

2.2 Metadata

Metadata, in other words "data about data" (e.g. Burnard, 2014), are a crucial part of a corpus. Relevant metadata make the data accessible, and are appended to all media and annotation files. The metadata documented in the CFINSL project include information about the corpus itself (its name, language, the size of the corpus, distributor etc.) as well as about the participants (region, sex, age and education etc.), the content (the various language tasks and elicitation materials used), media (format and type), project (name, language, methodology) and

[3] CSC - IT Center for Science Ltd. maintains and develops the state-owned centralised IT infrastructure and uses it to provide nationwide IT services for research, libraries, archives, museums and culture as well as information, education and research. *https://www.csc.fi/csc*

[4] *https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/*

session (task name, participants, etc.). The metadata are currently documented in Excel, from which they will be converted into CMDI metadata (Component MetaData Infrastructure), a framework initiated and developed by CLARIN for the description and use of metadata. Searching the data can be done in ELAN[5] (Crasborn & Sloetjes, 2008), which is also the tool used for annotating the material.

2.3 Consent

The establishment of a corpus of sign language with open access is a sensitive issue because visual material is used. It is important to show the face of the language informant because the facial area carries a lot of grammatical and lexical information. Thus the informant cannot be made anonymous. It is therefore essential to carefully explain to the informant in both written form and in sign language that her/his signing will be available for research and later will be partly publicly available on the Internet. Every informant is required to sign a consent form where consent for different kinds of uses of her/his signing is sought separately, allowing every informant to decide for what purpose(s) he/she will permit his/her signing material to be used. On the consent form there are five different parts and participants must choose either the yes or the no option for each of them:

- Video material can be used for research purposes in the CFINSL project but publishing video clips or still images is prohibited
- Video material can be presented in public events (e.g. academic presentations and teaching)
- Still images can be taken from the video material for publications (electronic or paper)
- The whole video material can be published electronically e.g. in the Internet
- The name of the participant can be mentioned in publications

The informant will have the right to check her/his material, before its presentation or publication. Moreover, she/he can ask the administrator to remove her/his recorded video material from the corpus if she/he so wishes.

In addition, we will comply with the Personal Data Act (523/1999) as well as with the regulations set out by Office of the Data Protection Ombudsman concerning a Personal Data File, by creating a Description of File.

3. Annotation

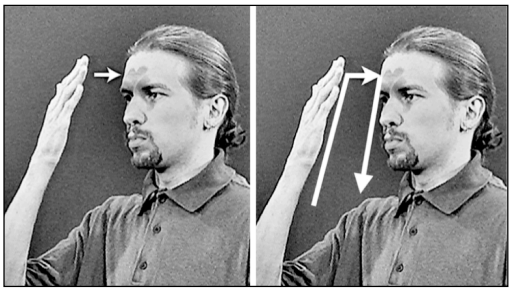
The ongoing process of annotating CFINSL data began during 2015. The corpus data are annotated using the ELAN tool, which enables time-aligned annotations to video media. The work started with the raw annotation of the narrative and discourse data of altogether 22 participants and approximately four hours of material. We started the annotation with a small amount of data with the aim of drawing up guidelines for creating annotation conventions for the corpus annotation. The work group in the annotation process consists of several annotators (most of them native signers) and translators (both native signers and native speakers). The annotation process was divided into three rounds:

1. The first round (raw annotation throughout 2015) was based on annotation with two tiers (a gloss tier and its comments)
2. The second round (during January-August 2016) is based on annotation with five tiers (a gloss tier of left/right hand, their comments, translation and comments on it)
3. The third round (starting in September 2016) will be a systematization of the annotation of the second round.

Our annotation work in the CFINSL project is based on four principles:

1. The length of the annotation cells is based on a view of the sign as a relatively long unit
2. Structural information is used in glosses to distinguish between forms with the same meaning (both phonetic and lexical variation)
3. Glosses are created for form-meaning pairs according to the contextual meaning of signs in discourse
4. Annotation is seen as a tool for future research and teaching

The first principle is related to Jantunen's (2013, 2015) understanding of a sign as a relatively long unit (Figure 5, see also the concept of broad segmentation in Hanke et al., 2012). This specification of the sign influences our annotation in that the annotation cells are presumed to be longer than the annotations done on the basis of the present mainstream view of a sign's length.



BLACK (short) BLACK (longer)

Figure 5: The length of the sign according to the mainstream view (short) and according to the view on which CFINSL annotation is based (longer). (Images from Jantunen, 2015: 117.)

In the present work we do not focus on exploring the borders (on- and off-sets) of the sign but rather annotate long cells, which are sure to capture the whole sign for

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

further investigation in the future. In practice this means that the annotated cell starts in the frame in which one of the parameters (usually the handshape, the orientation or the non-manual elements) of the sign is noticeable for the first time, and ends in the frame in which one of those parameters is noticeable for the last time.

The second principle emphasises the sign's phonological parameters. We use information about the four parameters of a sign (the handshape, location, movement and palm orientation) to code structural differences between signs with the same meaning. These differences may be free variation of only one parameter (phonetic variation) or differences between several parameters (lexical variation). With relation to the first option, at this stage of the annotation process we bring out equally all possible phonetic variants of a sign without combining them in the same 'family' as is done in the ID-gloss system (see Johnston & Schembri, 1999; Johnston, 2010; Cormier et al., 2012; Schembri et al., 2013). For example, we append information concerning the handshape, location, movement or orientation of the palm to a gloss, which helps us to distinguish the variants from each other (see Table 1 and 2).

HANDSHAPE	**LOCATION**
RUN(BB)	SKIN(cheek)
RUN(SS)	SKIN(back of a hand)

Table 1: Phonological parameters handshape and location differentiating between glosses for signs which differ in one parameter.

MOVEMENT	**ORIENTATION**
ARRANGE(sliding)	FINISHED(palms_down)
ARRANGE(bouncing)	FINISHED(palms_forward)
	FINISHED(palms_backward)

Table 2: Phonological parameters movement and orientation differentiating between glosses for signs which differ in one parameter.

We use information concerning the parameters of signs also when glossing different signs which have the same meaning (lexical variation). In this case we typically choose the most salient parameter for the gloss. E.g., Australia can be signed in at least three different ways in FinSL (see Figures 6–8)[6]. We have chosen the handshape of the signs as the most salient parameter to separate these signs from each other; otherwise the glosses are similar.

[6] Images for three signs meaning 'Australia' taken from KOTUS (2003).

Figure 6: AUSTRALIA(B)

Figure 7: AUSTRALIA(3)

Figure 8: AUSTRALIA(middle-finger)

The third principle concerns the semantics and grammatical phenomena of a sign. Every sign in the data is annotated (given a gloss) according to its meaning in context, although the same form might be used in another meaning in some other context. In this aspect our annotation is at this stage different from traditional ID-glosses (see e.g. Johnston, 2010). For example, the form demonstrated in Figure 9 can refer to the meanings 'everyday', 'jeans', 'countryside', 'sober' and 'redneck'. We annotate such form-meaning pairs in separate glosses, according to the contextual meaning of the occurrences.

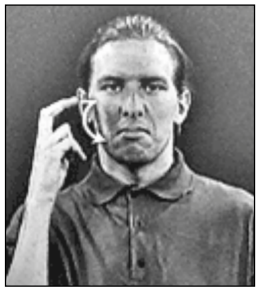


Figure 9: The manual form for form-meaning pairs glossed as EVERYDAY, JEANS, COUNTRYSIDE, SOBER and REDNECK (image from Suvi, the on-line dictionary of Finland's Sign Languages).

On the grammatical level, after the gloss we add codes indicating how the sign is modified morphologically. Grammatical codes (e.g. negation or descriptive utterances) are appended so that linguistic structures can be researched efficiently in the future. The annotation conventions for grammatical phenomena are currently being processed. We are considering using the symbol @ before coding different grammatical features after a gloss (see Wallin & Mesch, 2014). For instance:

- GLOSS@gesture
- GLOSS@depicting-sign
- GLOSS@repetition

During the annotation we have often used a comment tier in ELAN, into which we put different remarks about the modifications of a sign for subsequent work in coding these features in the glosses.

The last principle concerns an annotation as a tool. The aims of our glossing system are efficient search functions and machine readability, which are the same goals as for the ID-glosses used in the corpus work of several other sign languages. However, we are first creating glosses systematically with the help of the three previously mentioned principles, which will help us to build glosses as tools with different purposes in the future. It is important to remember that the glosses which are used in annotations have long-term effects on the research, teaching and learning of a sign language. We need first to test how well the glosses we have used serve different search processes; they must be as logical and usable as possible. We plan to arrange tests of the use of the corpus material in the contexts of teaching (pedagogical view) and research (linguistic view). We thereby hope to achieve a logical and usable glossing system which will serve as many aims as possible.

It is important to strive for consistency, usability and compatibility in the order of the glosses (see Keränen et al., 2016). The examples in Figure 10a demonstrate how a gloss can be a tool for the research and teaching of a sign language. It is much easier for researchers, teachers and students to search for a gloss (e.g. verbal KNOW) and its different structural and grammatical features by appending these features after a verbal of the same form (i.e. in this case the verbal KNOW). For instance, students can learn how the verbal KNOW can be modified in different ways (phonetic variation, prosody, negation), or how two completely different signs can have the meaning 'know' (lexical variation, TIETÄÄ-EI(55) and TIETÄÄ(repetition)). This aim corresponds with the ID-gloss system. If a gloss and its features were not in systematic order, according to a basic form, it would be much more difficult and messy to search for and find a certain gloss from a gloss list (see Figure 10b).

TIETÄÄ-EI(55) 'not know'
TIETÄÄ-EI(BB) 'not know'
TIETÄÄ-PALJON 'to know a lot'
TIETÄÄ(loiva) 'to know (gentle)'
TIETÄÄ(toisto) 'to know (repetition)'

a

EI-TIETÄÄ(55)
EI-TIETÄÄ(BB)
PALJON-TIETÄÄ
TIETÄÄ(loiva)
TIETÄÄ(toisto)

b

Figure 10: Examples of (a) an efficient search according to systematic annotation, (b) how unsystematic annotation may affect the search.

4. Creating a lexical database: Signbank

Our team has processed the glosses from the basic annotation work with the help of two lexical databases. Firstly, during the first round of annotation, we collected all the lexical glosses in Excel, as we did not yet have the FinSL Signbank in use. In this file we all commented on the existing glosses and then modified them as necessary, systematizing and confirming the glosses on the basis of the comments and the above-mentioned four principles of our linguistic concept.

Secondly, since May 2015 we have been working on the FinSL Signbank[7], a web-based lexical database used by researchers to store videos and relevant information about glosses. During the second round of annotation we are gradually transferring the lexicon from Excel to the FinSL Signbank and ultimately we plan to use only the FinSL Signbank. The FinSL Signbank has been developed on the basis of the NGT Signbank[8], which is a branch of the Auslan Signbank[9]. The source codes for these three versions of Signbank are all available on Github (https://github.com/Signbank). Some features of the NGT Signbank that are not necessary for our work at the moment were modified, hidden or deleted from the FinSL Signbank in order to match the current and future needs of the CFINSL project.

Our three main objectives for the use of Signbank in the CFINSL project are to allow two different research teams to upload their data sets, make the user interface translatable into multiple languages, and to be able to export glosses from Signbank to ELAN. With regard to the first objective, our current aim in the CFINSL project is to include the annotated glosses of both FinSL and FinSSL in two different dictionaries inside Signbank. In addition, co-operation between the CFINSL project and the corpus project of the Finnish Association of the Deaf[10] may result in three separate corpus lexicons within the FinSL Signbank. Work on this feature started in January 2016 and is currently in progress.

With regard to the second objective, the interface of the FinSL Signbank is now translatable into multiple languages. The process began during June 2015 and it was done with internationalization and localization features of Django[11], the web framework with which the

[7] *http://signbank.csc.fi*
[8] *http://signbank.science.ru.nl*
[9] *http://www.auslan.org.au*
[10] *http://www.kuurojenliitto.fi/en*
[11] *https://docs.djangoproject.com/en/1.8/topics/i18n/*

Auslan Signbank was built. These internationalization and localization features are needed in order to provide the interface in at least three languages: Finnish, Swedish and English.

Finally, in relation to the third objective, exporting glosses from the FinSL Signbank to ELAN works, but the feature needs further testing so that we can avoid possible problems in the future. In addition, we have added some new functions to the FinSL Signbank interface in order to help annotators' work in the lexical database. One of the new functions is the creation of colour codes for the glosses listed on the search page in Signbank. The listed glosses are automatically given a colour code according to whether the gloss entries include videos, are under evaluation, or have been approved by the administrators.

Signbank is an important tool for annotating new material efficiently and for observing coherent annotation conventions for the FinSL and FinSSL corpora. The use of the FinSL Signbank for annotation purposes began with multiple tests during autumn 2015. At the time of writing, we have begun transferring all the confirmed (i.e. our commonly accepted) glosses from the Excel lexicon into the FinSL Signbank. For the moment, the process of creating lexical entries in the FinSL Signbank follows the annotation conventions and principles described in Section 3 of the current paper. The description of the lexical entries begins with glosses and translation equivalents.

5. Conclusion

In this paper we have described the work of creating corpora of Finland's Sign Languages. This work, which is being carried out at the University of Jyväskylä, is still in its early stages. We have described the process of collecting the data, consents and metadata; the process of annotating the data and developing conventions for the annotation; and the process of building a web-based lexical database for the corpus lexicon. The CFINSL project will document and store both sign languages for present and future generations: the annotation conventions and lexical database will work as a tool for the research, teaching and learning of FinSL and FinSSL.

6. Bibliographical References

Burnard, L. (2005). Metadata for Corpus Work. In M. Wynne (Ed.) *Developing Linguistic Corpora: A guide to good practice. AHDS Guides to Good Practice.* Oxford: Oxbow Books, pp. 30–46.

Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R., & Woll, B. (2012). From corpus to lexical database to online dictionary: Issues in annotation of the BSL Corpus and the development of BSL SignBank. In *Proceedings of the 5th Workshop on the representation and processing of sign languages: Interactions between corpus and lexicon.* Paris: ELRA, pp. 7–12.

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

Hanke, T., Matthes, S., Regen, A. & Worseck, S. (2012). Where does a sign start and end? Segmentation of continuous signing. In *Proceedings of the 5th LREC Workshop on the Representation and Processing of Sign Languages.* Paris: ELRA, pp. 69–74.

Jantunen, T. (2013). Signs and transitions: Do they differ phonetically and does it matter? *Sign Language Studies* 13(2), pp. 211–237.

Jantunen, Tommi (2015). How long is the sign? *Linguistics* 53(1), pp. 93–124.

Johnston, T. & Schembri, A. (1999). On defining lexeme in a signed language. *Sign language & linguistics*, 2(2), pp. 115–185.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15(1), pp. 106–131.

Keränen, J., Syrjälä, H., Salonen, J. & Takkinen, R. (2016). The Usability of the Annotation. To appear in *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*. Paris: ELRA.

KOTUS (2003). The minutes of a meeting of the Board of Sign Languages of the Institute for the Languages of Finland, 9 October 2003. http://www.kotus.fi/files/509/kokous23-090603.pdf

Mesch, J. (2015). Svensk teckenspråkskorpus – dess tillkomst och uppbyggnad [Building the corpus for Swedish Sign Language]. In *Forskning om teckenspråk XXIV* [Research on sign language XXIV]. Stockholm: Stockholm University, Department of Linguistics, pp. 1–25.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., Rathmann, C. (2010). Elicitation methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies.* Paris: ELRA, pp. 178–185.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S. & Cormier, K. (2013). *Building the British Sign Language Corpus*. Language Documentation and Conservation 7, 136–154.

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

Wallin, L. & Mesch, J. (2014). *Annoteringskonventioner för teckenspråkstexter. Version 5.* Avdelningen för teckenspråk, Institutionen för lingvistik, Stockholms universitet.

7. Language Resource References

SUVI. The online dictionary of Finland's Sign Languages. The Finnish Association of the Deaf. http://suvi.viittomat.net