SignWiki – an Experiment in Creating a User-based Corpus

Sonja Erlenkamp[1], Olle Eriksen[2]
[1]University-College of Sør-Trøndelag
Department of Teacher and Interpreter Education
7004 Trondheim, Norway

[2]Møller Resource Center
PO box 175 Heimdal
7473 Trondheim, Norway
E-mail: sonja.erlenkamp@hist.no, olle.eriksen@statped.no

Abstract
In comparison to other signed languages, Norwegian Sign Language (NTS) is not well researched and documented while at the same time the need for documentation of NTS in a corpus based dictionary has been apparent to the field for quite a while. Despite some high quality applications to raise funding for corpus work, the field in Norway has not succeeded to gain enough understanding in governmental research funding institutions for the need of a corpus based dictionary, mainly because of the rather small population of NTS users. As a result, a new approach is used by involving the NTS community to create a database of signs, including their use, distribution and as far as possible other metadata. Tegnwiki (=Signwiki) is a first attempt at creating a user-based database of NTS by allowing users to contribute videos and information on isolated signs on a Wiki platform. Like Wikipedia, the Signwiki will be open accessible, but administered by a group of experts. Obviously a Signwiki cannot replace a scientific corpus. But if this experiment is successful it might be a good starting point for countries with no or little funding for corpus projects where involvement of users is the key factor.

Keywords: Norwegian Sign Language, Wiki, Sign Database

1. Introduction

Norwegian Sign Language (henceforth NTS for Norsk TegnSpråk) is one of the known yet little described signed languages in Europe. It is one of the genuine Norwegian languages used by a language minority of several thousand deaf and hearing users. Since 1825 it has been school language in Norway at schools for the deaf. As early as 1875 Norwegian Sign Language was labelled as a language (Skavlan, 1875), but as in many other western countries this attitude towards a signed language did not survive the period of oralism first in the late 1970s and early 1980s, the idea of NTS as a natural full-fledged grammatical language evolved again. Through the past 3 decades, several official documents and articles (e.g. Bergh, 2004; Erlenkamp, 2007; Erlenkamp et al. 2007) have operated with a number of about 4000 to 5000 deaf Norwegians and an unknown number of hearing Norwegians using NTS as one of their first languages. It is estimated that about 15.000 of the 5 million Norwegians use this language as a first or second language.

Thus, the language community is rather small. By now, NTS has gained a relatively wide acceptance in the Norwegian Society; on April 28th 2009, a proposition was passed by the Norwegian Parliament that NTS should become one of several official languages (GP35 2008).

Sign language studies and interpreting studies have been offered at several Universities and University Colleges since the mid-1990s. Moreover, in the 1990s the government established a 40 weeks free course in NTS for hearing parents of deaf children to help closing the gap between the hearing parent's signed language knowledge and skills and the practical skills of their deaf children in NTS.

NTS is however, in comparison to other signed languages, not well researched and documented. Basic aspects about NTS grammar, acquisition and sign variation (e.g. dialects, sociolects) have not been documented yet, and the documentation of Norwegian Sign Language has so far only been conducted by a handful of researchers (e.g. Greftegreff, 1991; Vogt-Svendsen, 1990, 2001; Selvik, 2006; Erlenkamp, 2009, 2011 a and b). Furthermore, reliability of many research projects depend on the availability of large amounts of annotated language data. Thus, the need for documentation of NTS in a corpus based dictionary has been apparent to the field for quite a while and for that reason scientific groups in a number of European countries are currently developing signed language data corpus. In Norway the establishment of any collection of mentionable size has been prevented by the lack of a standard written representation which results in an extremely time consuming process where the visual data must be annotated manually. Despite some high quality applications to raise funding for corpus work, the field has not succeeded to gain enough understanding in governmental research funding institutions for the need of a language corpus and a corpus based dictionary, mainly because of the rather small population of NTS users. The existing sign language dictionary project "Norsk tegnordbok" developed by the Møller Resource Centre has due to economic limitations and limited

availability of expertise reached a state where single signs are searchable online, but crucial information on linguistic categories of the signs, their usage and affiliation to regional or social varieties is not available. Thus the glossary consists of videos of simple isolated signs (in total about 6500) and corresponding Norwegian translations (http:tegnordbok.no). The translation lemmata are the basis for a search in this glossary.
As a result, the field is trying out a new approach by involving the NTS community to create a larger database of signs, including their use, distribution and as far as possible other metadata. Tegnwiki (=Signwiki) is a first attempt at creating a user-based database of NTS. If successful, this project could contribute in a large scale to increased accessibility of data on isolated signs in NTS. In that case, Tegnwiki will be functioning as an aid for interpreters in need of vocabulary, as well as part of signed language learning. Furthermore, users can suggest and discuss new signs on the Signwiki and thus the Signwiki will open up for a more democratic process on the development of new signs in underdeveloped domains of NTS.

2. Project Goals

The project has the following goals:

1) developing a user interface based on a common wiki user interface, modified to allow easy integration of videos even for users with little experience in using video tools
2) developing a standard for each article (each article will be linked to one sign), including slots for metadata about the usage of the sign, as well as opportunities for user discussions about the sign
3) informing and encouraging the NTS community to participate in this project

Above all, this project is an attempt to involve signers in a project about their own language and gather some information of signs based on the user's knowledge. As a consequence, the expectations on what can be collected and the level of quality of each article have to be kept at a reasonable level. It is, however, planned to make the signs from the already existing glossary available on the wiki as well, in an attempt to obtain more information about these signs and thus hopefully to create a synergy effect between the Signwiki and the dictionary project.

3. Technological, Economical and Scientific Requirements

Research on any sign language during the past decades has been shaped by two main limitations due to the visual modality the language data are based on. Both the lack of a standard written representation of signed languages which also could serve as a base for an annotation system, and the technological limitations regarding storage and access of large amounts of signed language data, have until recently made it impossible to work on large amounts of sign language data. Technological developments have already improved the latter, through streaming and other methods of accessing films. The former limitation is currently undergoing a major change due to the development of software tools for annotations of visual data like syncWRITER (Hanke & Prillwitz 1995), ELAN and iLex (Hanke & Storz, 2008).
These developments do, however, not change the fact that the process of developing, annotating and storing a large signed language data corpus depends on several external factors:

1. availability of staff trained in signed language and annotation of visual data (or the opportunity to train the staff)
2. access to research based knowledge about signed linguistic categories for the respective signed language used as background for tagging
3. access to (and funding of) the necessary technical equipment, e.g. internet access, servers, cameras
4. funding of the time consuming process of creating a data base

Any signed language corpus project is highly dependent on all of these four factors. Thus, so far only countries with enough expertise, financial background and technical infrastructure have implemented such a project. Countries that for whatever reasons do not manage to meet these factors will have to develop other strategies for building accessible information about signs/signed language. As Norway definitely has the sufficient means for every factor, willingness of prioritizing this kind of project in research politics is a crucial part of the funding process.
In distinction to large data corpus project, a Signwiki has to deal with only two of the four factors:

1. access to research based knowledge about signed linguistic categories
2. access to (and funding of) the necessary technical equipment, e.g. internet access, servers, cameras

The first requirement will be met with a small group of experts to develop a standard for linguistic data and eventually work with securing the quality of information on the Wiki.
The second factor is not an issue in Norway, since Norway has one of the highest internet access per capita rates in the world. Furthermore, other needed equipment like web cameras are affordable for most people in Norway and many deaf and hearing Norwegians already possess one.
In order to achieve the highest possible accessrate for users in the Wiki, the project aims nevertheless towards mainstream technical solutions, i.e. standard hardware, in spite of possible decrease in video-quality. The user interface depends heavily

on the use case and must be easily accessible with a low learning demand. Thus, the SignWiki aims at using a common Wiki user interface, with integrated functions for use of videos in order to record and play videos.
The user does not have to store data on her/his own computer. All data are manipulated on the central server. The user needs an internet connection with "Fast broadband connection".
Ideal for video recording would be at least 25 frames per second and a standard PAL resolution (720x576), which however cannot be expected from most home-used webcameras. Thus, we can distinguish between contributions from the project group (with data from the sign dictionary) and contributions from the non-expert users. The former group will deliver data meeting the requirements for a data corpus, while the latter might not. This may result in a number of videos less clear, sharp and smooth than sought for in other signed language data corpus projects and will in consequence lead to less adoptability of the Signwiki's data to other projects.

4. Availability of Data and Privacy

Like Wikipedia, the Signwiki will be open accessible, but administered by a group of experts. Access to the infrastructure will be through an internet portal. Most data will be publicly accessible, while there might be some data, like experts discussions, that are only accessible to the project group, in particular throughout the establishment of the project.
Norwegian rules about privacy in connection with research projects are very strict and would limit access to a research data base: Access to the raw data would be limited to researchers and application designers; before granted access, users would need to sign agreements about confidential use of the data. By developing an open Signwiki where every user can control her/his own data input, privacy will not be an issue.

5. Linguistic Data and Metadata

Each article in the Wiki will cover one single sign with the opportunity to comment on and contribute to meta-information about the sign. This form of data collection leads to some advantages in terms of being user based, but also comprises a number of risks regarding the reliability and quality of the data.
Research on NTS has been performed by only a few researchers since the early 1980s and has thus focused mainly on certain areas of grammatical description, like sign classes (Erlenkamp, 2000), different parameters of signs like mouthing (Vogt-Svendsen, 2001) and hand shapes (Greftegreff, 1991), time expressions (Selvik, 2006) and sentence types (Vogt-Svendsen, 1990). At present, a language model for NTS grammar and iconicity is under development, including grammatical descriptions like word order and grammatical relations (Erlenkamp, 2009; 2011 a and b). Since the number of researchers on NTS is small, most projects on NTS follow international developments on theory and methods and contribute to these developments. The results of earlier research will serve as background for the development of a list of desired metadata about a sign. This list will be matched up against a realistic expectation of what native signers without expertise in signed language linguistics can contribute with, e.g.:

List of desirable linguistic data:	**Realistic expectations towards user provided data:**
Phonological data:	Most of these data can be obtained at any time by analyzing the video data by experts, independent from the Wiki users)
Morphological/syntactic categories:	Unlikely to be provided by users
Data about the distribution of the sign:	Likely to be able to get some indication from users
Examples of usage:	Very likely to be able to get reliable data from the users

One of the major questions is how the data on the Signwiki will be searchable. One obvious solution is a search function based on the Norwegian translation word(s) for each sign. In addition we are going to look into possible solutions for a search based on sign configurations. For that purpose, a Ph.D.-candidate will try to develop and apply her model for searching on the Wiki as part of her project.
Since the Signwiki-project is not meant as a scientific database, metadata common in scientific signed language corpora cannot be expected to be obtained through Signwiki. As part of a Wiki, some metadata are collected automatically, like when the sign was put online and by whom. This kind of metadata, however, are not the most interesting for signed language data comparison.
Obviously a Signwiki cannot replace a scientific corpus. But if this experiment is successful it might be a good starting point for countries with no or little funding for corpus projects were the involvement of users is the key factor.
Technical solutions for video presentation of signed language data might also be of interest for other publishers of websites on NTS. In a best case scenario, Signwiki will serve as a contribution for a sign language dictionary platform and as a supplier of examples as well as a democratic platform for the development of new signs in NTS.

6. Acknowledgements

This project has been financially supported by the Norwegian ExtraFoundation for Health and Rehabilitation through EXTRA funds.

7. References

Bergh, G. (2004) (ed.). Norsk tegnspråk som offisielt språk. [Norwegian Sign Language as an Official Language]. ABM. ABM-skrift nr. 10. Report.

ELAN. EUDICO Linguistic Annotator. http://www.lat-mpi.eu/tools/elan/

Erlenkamp, S. (2007). Et tospråklig liv med norsk tegnspråk. Språknytt 4:2007, pp. 16-19.

Erlenkamp, S., Gjøen, S., Haualand, H., Kvitvær, H.B.; Petterson P.R.; Schrøder, O.-I. & Vonen, A.M. (2007). Begrunnelser for å gjøre norsk tegnspråk til offisielt språk. Notat frå utval oppretta av styret for Norges Døveforbund. [Arguments for why Norwegian Sign Language Should Become Official Language. Report of the Committee Established by the Norwegian Deaf Association]. Oslo: Norges Døveforbund.

Erlenkamp, S. (2009). "Gesture Verbs" - Cognitive-visual Mechanisms of "Classifier Verbs" in Norwegian Sign Language. CogniTextes 3. http://cognitextes.revues.org/250

Erlenkamp, S. (2011). Grunntegnstilling i norsk tegnspråk. Erlenkamp, S.; Amundsen, G.; Halvorsen, R. P. & Raanes E. (Eds.), Norsk Lingvistisk Tidskrift 29(1), Oslo: Novus, pp. 87-116.

Erlenkamp, S. (2011). Norsk tegnspråk: helt norsk og veldig annerledes. Skisse av en ny beskrivelsesmodell for norsk tegnspråk. Erlenkamp, S.; Amundsen, G.; Halvorsen, R. P. & Raanes E. (Eds.), Norsk Lingvistisk Tidskrift 29(1), Oslo: Novus, pp. 26-37.

GP35 (2008). Governmentproposition no. 35 (2007-2008) "Mål og meining. Ein heilskapleg norsk språkpolitikk". http://www.regjeringen.no/pages/2090873/PDFS/STM200720080035000DDDPDFS.pdf

Greftegreff, I. (1991). Håndformer og håndformendringer i norsk tegnspråk: En innledende undersøkelse av foneminventaret. M.A.-thesis in linguistics, Trondheim: University of Trondheim.

Hanke, T. & Prillwitz, S. (1995). syncWRITER: Integrating Video into the Transcription and Analysis of Sign Language. In Bos, H. & Schermer, G. (Eds.), Sign Language Research 1994: Proceedings of the Fourth European Congress on Sign Language Research, Munich, September 1-3, 1994. Hamburg: Signum, pp. 303-311. http://www.signum-verlag.de/BTitel/pdf/3-927731-57-9b.pdf.

Hanke, T. & Storz. J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguisticsand Sign Language Lexicography. Proceedings of 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora at LREC 2008.

Selvik, K.-A. (2006). Spatial Paths Representing Time. A Cognitive Analysis of Temporal Expressions in Norwegian Sign Language; Oslo: Department of Linguistic and Nordic Studies, University of Oslo. Ph.D.-thesis.

Skavlan, S. (1875). Throndhjems Døvstumme-Institut Program, udgivet i Anledning af Institutets 50-aarige Bestaaen. Trondheim: Statped (original: Lie & Sundts Bogtrykkeri).

Vogt-Svendsen, M. (1990). Interrogative strukturer i norsk tegnspråk. En analyse av nonmanuelle komponenter i 86 spørsmål. Trondheim: Faculty of History and Philosophy, University of Trondheim. Ph.D. thesis.

Vogt-Svendsen, M. (2001). A Comparison of Mouth Gestures and Mouthings in Norwegian Sign Language (NTS). In Boyes Braem, P. & Sutton-Spence, R. (Eds.), The Hands are the Head of the Mouth: The Mouth as Articulator in Sign Language. Hamburg: Signum, pp. 9-39.