The *Corpus NGT*: an online corpus for professionals and laymen

Onno Crasborn, Inge Zwitserlood

Department of Linguistics, Radboud University Nijmegen

PO Box 9103, NL-6500 HD Nijmegen, The Netherlands

E-mail: o.crasborn@let.ru.nl, i.zwitserlood@let.ru.nl

Abstract

The *Corpus NGT* is an ambitious effort to record and archive video data from Sign Language of the Netherlands (Nederlandse Gebarentaal: NGT), guaranteeing online access to all interested parties and long-term availability. Data are collected from 100 native signers of NGT of different ages and from various regions in the country. Parts of these data are annotated and/or translated; the annotations and translations are part of the corpus. The *Corpus NGT* is accommodated in the Browsable Corpus based at the Max Planck Institute for Psycholinguistics. In this paper we share our experiences in data collection, video processing, annotation/translation and licensing involved in building the corpus.

1. Introduction

As for most sign languages, NGT resources are scant. Still, such resources are direly needed for several purposes, sign language research not the least. The aim of the *Corpus NGT* is to provide a large resource for NGT research in the shape of movies of native NGT signers. The signed texts include several different genres, and the signers form a diverse group in age and regional background. Besides the movies, crude annotations and translations form (a small) part of the corpus, so as to ease access to the data content. The corpus is made publicly available to answer the need for NGT data (e.g. by NGT teachers and learners and interpreters).

2. Data collection

2.1 Participants

The initial aim was to record 24 native signers, divided over two regions where two different variants of NGT are reported to be used. The plan was changed in its early stages so as to include a much larger number of participants, spread over all five reported variant regions.. Moreover, by including participants from different ages, it was possible to record older stages of NGT, even male and female variants in these older stages. Altogether, this ensures a good sample of the current state of the language

The participants were invited to take part in the recordings by announcements on Deaf websites, flyers and talks at Deaf clubs, and by 'sign of hand'. Interestingly, when the project became familiar in the Deaf community, many older people wanted to participate, in order to preserve their own variant of NGT. Because most signers are familiar with the use of contact varieties combining signs with spoken Dutch and because the variation in the form of such contact varieties is very large, participants were selected who are deaf from birth or soon after and who started to use NGT at a very early age (preferably before the age of 4). Also, we tried to eliminate standardised NGT (an artificial variant of NGT, recently constructed on request of the Dutch government; Schermer 2003).

2.2 Tasks and materials

In building the corpus, we followed the project design developed by the constructors of the Auslan corpus project[1], although adaptations were made to match the situation in the Netherlands. This means that a subset of the tasks given to the participants of the Auslan project were used, using the same or similar stimuli. These included narratives based on cartoons (the Canary Row cartoon of Tweety & Sylvester), fable stories presented to the signer in NGT, comic stories (including the Frog Story children book), and TV clips (e.g. funniest home videos). Besides elicitation of such monologue data, (semi-)spontaneous conversation and discussion forms a substantial part of the *Corpus NGT*. Using the advice from the Auslan experience, the elicitation materials that were used contained as little written text as possible. The participants were all asked to briefly introduce themselves and to tell about one or more life events they experienced. Most importantly (in terms of quantity and content), they were asked to discuss a series of topics introduced to them in NGT movies concerning Deaf and sign language issues. Finally, they engaged in a task where they had to spot the differences between two pictures they had in front of them. In addition to these tasks, occasional free conversation was also recorded.

2.3 Recording situation

The participants were recorded in pairs, to encourage 'natural' signing as much as possible.

[1] http://www.hrelp.org/grants/projects/index.php?lang=9

Beforehand, the purpose of the corpus and the tasks and proceedings were explained to them by a native Deaf signer, who also led the recordings. Explanation and recording took approximately 4 hours, and resulted in ± 1.5 hours of useable signed data per pair. Some recordings were made at the Radboud University and the Max Planck Institute for Psycholinguistics, both in Nijmegen. However, most recordings were made in Deaf schools, Deaf clubs or other places that were familiar to the Deaf participants. All recordings from the northern region (Groningen) were made at the Guyot institute for the Deaf in Haren.[2] As a result of the different sizes and light circumstances of the rooms, there is some variation in the recordings. All recordings were made with consumer quality cameras; no additional lighting equipment was used.

In a recording session, the participants were seated opposite each other, preferably in chairs without armrests as these might hamper their signing. An upper body view and a top view of each signer were recorded. This situation is illustrated in Figure 1. In combination, these front and top views approximate a three-dimensional view of the signing. Previous research has shown that such a view can give valuable information on the use of space and even on the shape of signs, if these are not completely clear from the front view (Zwitserlood, 2003). The top views were recorded with two Sony DV cameras on mini-DV tapes. The cameras were attached with bolts to metal bookstands that could be easily attached to the ceiling above the seated participants. The front views were recorded using two Sony High Definition Video (HDV) cameras on mini-DV tapes; these were mounted on tripods. The upper body view was recorded slightly from the side. This had the advantage of a better view of the signing (since a recording straight from the front does not always give reliable or clear information on the location and handshape(s) in particular signs). Also, when one looks at the front view recordings of both participants in a session, the slight side view gives a better impression of two people engaged in conversation, rather than two people signing to cameras.

We chose to use HDV recordings for the front views because of the high resolution (the full HD recording includes 1920x1080 pixels in contrast to normal digital video, with a format of 720x568 pixels for the European PAL format), resulting in recordings that are very detailed in comparison to standard PAL video. Furthermore, we wanted to provide detailed information on facial expressions; the HDV resolution allowed cutting out a view on the face, rather than having to use two additional cameras that could be zoomed in on the face.

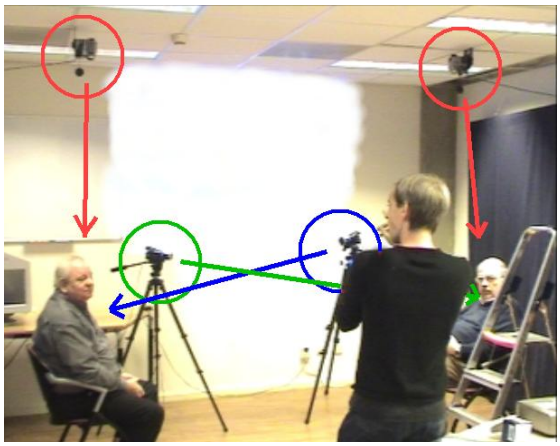

Figure 1: Recording situation

The recording sessions were lead by a Deaf native signer, who would explain the aims of the project and the procedure beforehand to the participants, allowing ample time for questions, and who stressed the fact that we were especially interested in normal signing, viz. they should try not to sign “neater” or “more correct” than usual. Every new task was explained in detail, and if necessary, the session leader would give examples or extra information during the execution of a task.

For each pair of participants, there were three one-hour recording sessions. In between there were breaks in which the participants could rest and chat, and the tapes were replaced by new ones. Since the cameras were not connected to each other electronically and since switching the four cameras into recording mode by a remote control proved unreliable, each camera was switched on by hand. When all four cameras were running, there would be three loud hand claps, that would show in all the recordings and could thus be used to synchronise the four video streams afterwards.

3. Data processing

We took the following steps in processing the recorded data: data capturing, editing, and compression. These are explained in the following sections.

3.1 Capturing and editing

For capturing and editing of the recorded tapes, the video processing programme Final Cut Pro (version 5.3.1, later version 6.0.2) was used. This is a professional video editing programme and the only one that, at the time, was able to handle HDV format video as well as normal DV video. The content of the videotapes was captured in Apple computers (using OS X version 10.4, later 10.5). A Final Cut project contains the four tapes of a

[2] We thank Annemieke van Kampen for her work in finding participants and in leading all the recording sessions in Groningen.

recording session, that are then synchronised on the basis of the clap signal. Subsequently, as many fragments as possible were selected for further use (even those where signers were grunting about a particular task), and all other bits in between were cut out (where a participant was looking at the stimuli or de session leader was explaining something). The selected fragments were assigned a specific "session code" (e.g. CNGT0018) with a postscript indicating the signer (S001 to S100) and the viewpoint of the camera ('t' for top view, 'f' for face and 'b' for body) exported to Quicktime movies in DV and HDV format, respectively. These 'raw DV' files were too large to be included in the corpus or to be used productively in applications such as ELAN; for that reason, all movies were compressed to different MPEG formats.

3.2 Compression

The project aimed at providing movies that can be used for different purposes and in different applications; moreover, the video should still be accessible in a few decades from now. For this reason, we followed the policy of the data archive at the Max Planck Institute for Psycholinguistics (which is also the location of the DOBES archive) to use MPEG-1 and MPEG-2 video files.. The latter keeps the original PAL image size, while the former reduces the size to about one quarter, often 352x288 pixels. The various MPEG standards are a publicly defined and accessible standard, and are not a commercial format promoted and protected by a company (such as the Flash video standard is owned by Adobe).

The resulting movie clips can be easily used in various software applications such as the annotation tool ELAN (see section 5.1). The combination of the MPEG-1 format and the segmenting of video recordings into smaller clips ensures that the movies can be readily downloaded or viewed online. The MPEG-2 version of the top view movies are also included in the corpus for those who need a higher quality image; and also as a relatively unreduced original that can be converted to future video standards in the future. The hosting of the whole corpus at MPI ensures that the material in the corpus will be converted to future standards along with the many other corpora in the corpus in the future.

For the body and face views, a different procedure was followed. In the first stages of the project (late 2006), we were not able to find a compression technique that was able to maintain the full resolution of the HDV recordings. Although the H.264 compression method that is part of the MPEG-4 standard should in principle be able to maintain the full spatial and temporal resolution at highly reduced data rates, we were not able to produce such files. Since both this standard and the HDV recording techniques were only just appearing on the (consumer) market, we decided to wait with a decision on the high-quality archive format of the HDV recordings. For now, such high-resolution recordings will not be frequently used anyway, given the infrequent use of high-resolution displays: the 1920x1200 resolution is equal to the better graphic cards and (22" and 23") monitors on the market nowadays, and few computer setups will be used with two such displays side by side (needed to play back the conversations in the corpus). At the end of the project in May 2008, we still have not yet decided what to use as a full-resolution format; the 'raw' HDV exports from Final Cut Pro will be included in the corpus for future processing. They can be played back in Quicktime-compatible video players, but are not yet de-interlaced.

To be able to use the recordings productively, we decided to create two MPEG-1 movies from every HDV file. Since the aspect ratio of MPEG-1 (4x3) does match that of HDV (16x9), cropping was necessary anyway; different cropping settings were used to create cut-outs of the face and of the whole upper body plus head; in addition, the face versions were scaled down less than the upper body versions. Thus, for every section of the recordings, we have six MPEG-1 movie files, three for each signer.

At the start of the project, Apple's Compressor (version 2) appeared to be unreliable for the compression to MPEG-2 format. Therefore, the programme MPEG Encoder Mac 1.5 from MainConcept was used for this type of compression initially. This program has proved to produce good quality MPEG-1 and MPEG-2 movies. However, its disadvantage is that there is no easy way compress large numbers of movies in an easy batch mode; all settings have to be re-applied for every movie. Because of the large numbers of movies in the corpus, this was too labour-intensive. Midway the project, when Compressor version 3 proved to have a reliable MPEG-2 compression option, we switched to that programme for the production of both MPEG-1 and MPEG-2 versions.

In all parts of the corpus, even in the 'monologue' story-telling, two signers interact. For a good understanding of the signing one therefore needs the movies of both participants, and they should be played in synchrony. While this is a standard function of the ELAN annotation software (see section 5), most common movie players that are integrated in web browsers are not built to play separate movies simultaneously. Therefore, we also provide movies in which the MPEG-1 movies of the front view of both signers are combined into one wide-screen image. These combined movies also have MPEG-1 compression settings, but the aspect ratio is that of two juxtaposed MPEG-1

movies. This process was carried out by the Ffmpeg and Avisynth tools for Windows.

Finally, after the MPEG-1 and MPEG-2 movies have been published online as part of the MPI Browsable Corpus, in the near future all movies will be converted into streaming MPEG-4 clips and made accessible through MPI's streaming server. In this way, movies can be easily accessed by online tools such as ANNEX[3].

4. Access

4.1 Metadata

The *Corpus NGT* is published by the MPI for Psycholinguistics, as part of their growing set of language corpora. We follow the IMDI standard for creating metadata descriptions and corpus structuring.[4] These metadata concern information about the type of data (narrative, discussion, retelling of cartoon, etc) and about the participants. Although all data are freely accessible and the participants are clearly visible in the movies, their privacy is protected as much as possible by restricting the metadata for the participants in the corpus to their age at the time of the recording, their age of first exposure to NGT, their sex, the region where they grew up, and their handedness. Researchers who need more information (e.g. about the fact whether there are deaf family members) can request such information from the corpus manager. Names or initials are not used anywhere in the metadata description for the participants.

4.2 Access for all

Although the corpus is mainly intended for linguistic research, the data can have several other uses. Because of the need of NGT data indicated earlier, we are happy to share the data in the corpus with other people who need such data or are interested in them, providing open access to all video and annotation documents. Other interested scientists may be psychologists, educators, and those involved in constructing (sign) dictionaries. Deaf and hearing professionals in deaf schools and in the Deaf community may want to use the material, including NGT teachers, developers of teaching materials, and NGT interpreters. Many hearing learners of NGT will benefit from open access to a large set of data in their target language. Deaf people themselves may be interested in the discussion on deaf issues that forms part of every recording session.

All participants in the corpus signed a consent form that explicitly mentions the online publication and the open access policy. The forms in Dutch were explained by the Deaf person leading the recording session. Most importantly, the publication and possible use of the material was explained to the signers before they agreed to come and participate. During the actual recordings, signers were encouraged to limit the amount of personal information they might reveal in their discussions. In a few cases, we decided to leave out privacy-sensitive segments after the recordings, often in conformance with requests from the participants.

Since the construction of large sign language corpora is a recent phenomenon, we hope that our experiences will be valuable for other projects. Therefore, the project's open access policy extends beyond the video data to the annotations, workflows and guidelines for tools that have been used, which will all be published online.

Although everyone has free access to the data in the MPI Browsable Corpus that is available via the internet,[5] searching and finding interesting movies in the large corpus is not an easy or quick task. Therefore, we are currently designing a few websites for specific target groups (e.g. NGT teachers, deaf children and their parents, NGT interpreters), from which websites selected movies are easily accessible.

4.3 Licensing

The use and reuse of the data are encouraged *and* protected at the same time by Creative Commons licenses (see Crasborn, this volume, for further discussion). Creative Commons offer six types of protection, ranging from restrictive to highly accommodating. We chose the combination BY-NC-SA:

1. Attribution: when publishing (about) data of this corpus, mention the source (BY);
2. Non-commercial: no part of this corpus can be used for commercial purposes (NC);
3. Share alike: (re)publishing (parts of) data of this corpus should be done under the same licenses (SA).

The first two licenses are self-explanatory. The third license is meant to encourage other people to make use of the data and to share new data based on data from the corpus with others (while, again, protecting the new data). For example, an NGT teacher may want to use a part of a movie to point out particular grammatical phenomena to her students, or provide a movie with subtitles, and share the new movie with colleagues. Alternatively a researcher interested in a particular aspect of NGT may use an annotation file, add new annotations and share the enriched file with other researchers. The licenses are mentioned in the metadata. Also, the licenses are part of all the movies in the corpus: a short message in Dutch and

[3] http://www.lat-mpi.eu/tools/annex/
[4] http://www.mpi.nl/imdi/
[5] http://corpus1.mpi.nl

English is shown at the start and end of each movie.

5. Accessibility of the data

Not all people who may be interested in the data of the corpus are fluent in NGT. For these users, the corpus provides ways to gain (better) access to at least parts of the data, viz. annotations and translations.

5.1 Annotation

For annotation the annotation tool ELAN (Eudico Linguistic Annotator), developed at the MPI for Psycholinguistics, was used.[6] This program is currently widely in use for the annotation of various kinds of linguistic and other communicative data. This tool allows online transcription where the original data (a sound or video file) can be played and annotations can be aligned with the audio of video signal. Originally used for annotation of gesture, it has been improved substantially since it also started to be used for sign language annotation. Based on experiences in previous projects (e.g. ECHO[7]) and desired functionality in the corpus project, various new features were formulated and implemented in the software (see Crasborn & Sloetjes, this volume). The extension of ELAN as well as the integration of ELAN and IMDI (the data and metadata domains) formed a substantial part of the project.

Annotation is an enormously time-consuming process. Due to time and budget limitations (the project was funded for two years), and as we invested in more recordings than originally planned which left less time for annotation, it was only possible to provide crude gloss annotations of a small subset of the data. Four Deaf assistants were assigned this job, on a part-time basis to avoid health problems because of the intensive use of the computer. They were trained to use ELAN (showing only a front view of both participants) and to gloss the signs made by the left and right hand with a Dutch word, or a description if there was no appropriate Dutch word available. They could use a bilingual Dutch-NGT dictionary holding approximately 5000 lemmas and Dutch (picture and normal) dictionaries to check Dutch spelling, as well as a reference list with the gloss conventions to be used. These conventions were based on and adapted from the conventions used in ECHO; see Nonhebel *et al.* 2004. At the end of the project, 160 movies were annotated, totalling almost 12 hours of signing and 64.000 glosses. Unfortunately, the assistants' skills in Dutch appeared to be quite poor, resulting in a rather large amount of spelling and writing mistakes in the annotations. In addition, they did not remember conventions well enough and/or seemed to be reluctant in looking up information that they needed. Also, it appeared to be a hard task to focus solely on the manual component of signing in determining a gloss text, annotators almost automatically look at the meaning of the whole construction, including facial expression and other non-manual behaviour. Because of that, several other mistakes occur in the annotations, including misalignments between start and end of many signs and their annotations. We corrected the most salient spelling mistakes and diacritics used in the wrong places. Furthermore, some of the ELAN files were corrected by a Deaf signer experienced in the use of ELAN and in annotation. Still, the current annotations should not be blindly relied upon. We plan to do further corrections and to provide more and more detailed annotations in future research projects.

5.2 Translation

Annotations are very helpful in doing linguistic research. However, besides researchers, the data are also made available to other interested parties. In order to make as much of the data set accessible to a large audience, parts of the data are provided with a voice-over translation, done by interpreters and interpreter students. For this, empty ELAN files were created, only showing front view movies of two participants for the data to be translated. The interpreters were instructed in the navigation of ELAN and in the use of a Sony minidisc recorder with one or two microphones (depending on whether the movies to be translated involved monologues or dialogues). Their job was to look at a particular movie one or two times, if necessary to discuss difficult parts with a colleague, switching on the minidisc recorder and give a running translation while watching the movie. The audio files on the minidiscs were processed into WAV files, aligned with the movies and connected to ELAN files.

The interpretation of the (often unknown) participants in discussion turned out to be a challenging task. The option to play back the movie is almost irresistible to interpreters if they know that they may not have fully understood every detail. As sign language interpreters are rarely in such a position, typically doing simultaneous interpreting in events where they have little control over things like signing rate, the voice interpreting for the *Corpus NGT* was an attractive task, precisely because of the option to replay and discuss the interpretation with their colleague. The nature of this process can be considered a mix between interpretation and translation. On average, the interpreting process (including administrative

[6] http://www.lat-mpi.eu/tools/elan/
[7] http://www.let.ru.nl/sign-lang/echo/

and technical tasks related to the recording procedure) took ten times realtime (thus, one hour of signing took ten hours to record on minidisc). Because of the increase in recorded hours of signing with respect to the original plan, it was not possible to provide a voice-over with all video recordings.

Originally we had hoped for the possibility to transfer the speech signal of the interpreters into written Dutch using speech recognition software. However, this appeared not to be possible because of a combination of factors. First, most speech recognition programs needs to be trained to recognize the speech of the interpreters; it appeared to be impossible to set this up logistically. Second, speech recognition software that we could use does not need the auditory signal for training, but instead, uses word lists. However, the wide range of lexical items and spontaneous nature of the spoken translations appeared to be too variable for reliable transfer to written text. Taking into account the post-hoc corrections that would be necessary, it is probably cheaper and more reliable to use typists. This is clearly an option for the future.

6. Future developments

It is clear from the programme of the present workshop alone that we can expect rapid developments in the field of corpus studies in signed languages. There is an enormous increase in the data that linguists have at their disposal, which will enable deeper insights in the linguistic structure and in the variability of signing within a community. Even though the *Corpus NGT* explicitly aimed to exclude signers that only used some form of sign supported Dutch, the influence of Dutch appears to vary greatly across age groups, an observation that has not yet received any attention in the literature.

In order to carry out such linguistic studies, we need clear standards for annotation and transcription in sign language research. While there have been some efforts in the past, for example as collected in the special double issue of *Sign Language & Linguistics* (issue 4, 2001), there is very little standardisation for common phenomena such as gloss annotations. We hope that the increasing use of shared open source tools such as ELAN that use published XML file formats will increase the possibilities for exchanging data between research groups and countries, and promote standardisation among linguists.

In terms of technology, progress is slowly being made in automatic sign recognition. Having tools that enable some form of automatic annotation would constitute a next large jump in the construction and exploitation of sign language corpora. Recording and publishing video data online is now possible, but the Achilles heel in using them remains in accessing the large amounts of data: search tools need to be enhanced, but for these tools just as for the linguistic eye, annotations remain crucial, yet require an enormous investment in time and money. For the Corpus NGT, we hope that its use by various researchers in the near future will slowly increase the 15% of the data that have been glossed until now.

7. Acknowledgements

The *Corpus NGT* project is funded by the Netherlands Organisation for Scientific Research (NWO) by a two-year grant, nr. 380-70-008. We thank the participants in the recordings for their valuable contributions.

8. References

Nonhebel, A., Crasborn, O. & Van der Kooij, E. (2004) Sign language transcription conventions for the ECHO project. Version 9, 20 January 2004. Radboud University Nijmegen.

Schermer, T. (2003). From variant to standard. An overview of the standardization process of the lexicon of Sign Language of the Netherlands (SLN) over two decades. *Sign Language Studies*, 3(4), 96-113.

Zwitserlood, I. (2003) Classifying Hand Configurations in Nederlandse Gebarentaal (Sign Language of the Netherlands). Utrecht, LOT Dissertation Series 78.