

SignSpeak

Deliverable 1.2

Nature of available NGT corpora (ECHO and CNGT)

30 July 2009

Dr. O.A. Crasborn
Centre for Language Studies
Radboud University Nijmegen



1. Introduction

The goal of this report is to sketch the nature of available annotated data. As the DOW mentions in the description of Task 1.2, only for NGT there are open access corpora that are annotated to some degree. These two corpora (ECHO and Corpus NGT) will be characterised in section 2. A short characterisation of data collections for other signed languages (not available) is given in section 3. Finally, section 4 will sketch the current developments (Task 1.3) that are undertaken to enhance the existing annotations of the Corpus NGT.

2. Corpora for Sign Language of the Netherlands

2.1 General description

The ECHO corpus was part of a large European project (European Cultural Heritage Online) that aimed to integrate data from all different disciplines in the humanities. One of the four pilot studies concerned the creation of a cross-linguistic sign language corpus. Similar data were collected from British Sign Language (BSL), Swedish Sign Language (SSL), and Sign Language of the Netherlands (SLN/NGT). The corpus can be accessed at http://corpus1.mpi.nl/ds/indi_browser?openpath=MPI84302%23. General information is published at <http://www.let.ru.nl/sign-lang/echo>.

A rough description of the corpus is presented in Table 1. Annotations include not only glosses and translations, but also non-manuals. The conventions for glossing distinguish the left from the right hand, but also show some deviations from the Corpus NGT conventions.

| Language | Type | No. of people | No. of minutes | Annotated |
|----------|---------------|---------------|----------------|---------------|
| BSL | Interview | 2 | 10 | no |
| | Fable stories | 2 | 15 | yes |
| | Lexicon | 1 | 24 | no |
| | Poetry | 2 | 27 | in part |
| NGT | Interview | 3 | 20 | no |
| | Fable stories | 4 | 30 | 3 of 4 people |
| | Poetry | 1 | 30 | yes |
| | Lexicon | 4 | 50 | 2 of 4 people |
| SSL | Interview | 2 | 10 | no |
| | Fable stories | 2 | 15 | yes |
| | Lexicon | 2 | 10 | no |
| | Poetry | 1 | 1.5 | yes |

Table 1: overview of the ECHO corpus.

The Corpus NGT built on the experiences gained with the ECHO project. It increased the number of signers (92) and the types of conversation that were elicited (see tables below). Moreover, an effort was made to include all five regional variants.

| Duration of sessions | | S1 region | | | | | | |
|------------------------|-------------------------|-----------------|-----------------|----------------|----------------|--------------------|-----------------|-----------------|
| Interactie | InhoudType | Amsterdam | Groningen | Overige | Rotterdam | St. Michielsgestel | Voorburg | Grand Total |
| dialogoog | Conversatie | 0:14:35 | 0:15:46 | 0:42:38 | 0:00:32 | 0:01:14 | 0:12:06 | 1:26:51 |
| | Discussie (doof) | 5:44:13 | 6:56:25 | 1:45:51 | 1:28:38 | 1:17:06 | 4:12:34 | 21:24:47 |
| | Discussie (gebarentaal) | 2:13:29 | 3:36:02 | 0:30:25 | 0:41:59 | 0:07:52 | 1:41:33 | 8:51:20 |
| | Voorstellen | 0:39:33 | 1:13:06 | 0:12:39 | 0:14:42 | 0:10:59 | 0:33:42 | 3:04:41 |
| | Zoek de 10 verschillen | 1:26:06 | 2:15:33 | 0:19:43 | 0:25:01 | 0:21:38 | 1:28:15 | 6:16:16 |
| dialogoog Total | | 10:17:56 | 14:16:52 | 3:31:16 | 2:50:52 | 1:58:49 | 8:08:10 | 41:03:55 |
| monoloog | Eigen ervaring | 1:27:39 | 2:18:54 | 0:38:58 | 0:38:57 | 0:13:29 | 1:15:52 | 6:33:49 |
| | Fabel | 1:25:39 | 2:26:33 | 0:22:49 | 0:30:55 | 0:16:31 | 1:16:08 | 6:18:35 |
| | Funniest Home Video | 1:47:09 | 3:00:11 | 0:28:12 | 0:32:43 | 0:16:21 | 1:25:52 | 7:30:28 |
| | Kikkerverhaal | 0:24:52 | 0:35:32 | 0:07:09 | 0:06:03 | 0:05:51 | 0:24:24 | 1:43:51 |
| | Strip Boerke | 0:23:53 | 0:43:13 | 0:07:49 | 0:07:45 | 0:04:58 | 0:29:21 | 1:56:59 |
| | Tellen (1-100) | | 0:03:49 | | | | | 0:03:49 |
| | Tweety & Sylvester | 1:25:04 | 2:37:51 | 0:24:43 | 0:26:54 | 0:13:55 | 1:11:10 | 6:19:37 |
| monoloog Total | | 6:54:16 | 11:46:03 | 2:09:40 | 2:23:17 | 1:11:05 | 6:02:47 | 30:27:08 |
| Grand Total | | 17:12:12 | 26:02:55 | 5:40:56 | 5:14:09 | 3:09:54 | 14:10:57 | 71:31:03 |

Table 2: overview of the size of the Corpus NGT by interaction type, content type and region.

2.2 Lexical statistics

Sign language annotation is extremely time-consuming. A recent estimate for DGS (German Sign Language) was that a first parse of the video for glosses takes about 200 times realtime (R. Konrad, *Corpus Linguistics* 2009). This is taking into account the large lexicon that is linked to the DGS gloss annotations, which is not available in the Netherlands. The problems with glossing reside largely in the fact that the glosses do not consist of transcriptions in the language itself, but in translations of the sign language items as spoken language items. There is thus also translation involved, and one-to-one translation is simply impossible – certainly with such structurally extremely different languages. Multiple rounds of checking by different signers (subsequent parses) would therefore be desirable for accuracy, yet this is not realistic. For the Corpus NGT, all cases of doubt will be double-checked, and several measures will be taken to ensure consistency even without the presence of a usable lexicon. In the statistics below, these measures have not yet been implemented.

Lexical statistics are difficult to derive from the current annotation documents because the information is distributed over tiers for the left hand and the right hand. In no corpus it is annotated explicitly whether the annotated sign is one-handed or two-handed. While it is safe to assume that the vast majority of overlapping identical glosses on the left and right hand tiers are in fact lexically two-handed signs, there may be instances of one-handed signs that are doubled. Should one only take into consideration the tier for the right hand, one would leave out signs that are made on the left hand of either predominantly left-handed signers, or of lexical items realised by the ‘weak’ hand – whether or not the strong (right) hand was realising another sign at the same time. Below, these intricacies are not solved; we have simply calculated gloss frequencies for the left and right hand separately. The tables below show only the type-token information; the whole frequencies can be found in the accompanying Excel files.

| Project & tier | Tokens | Types |
|------------------|--------|-------|
| ECHO Left Hand | 974 | 396 |
| ECHO Right Hand | 2.261 | 694 |
| CNGT LH signer 1 | 11.061 | 2.346 |
| CNGT RH signer 1 | 27.694 | 3.802 |
| CNGT LH signer 2 | 10.534 | 2.149 |
| CNGT RH signer 2 | 21.064 | 3.056 |

Table 3: number of gloss annotations for the different tiers of the ECHO corpus and the Corpus NGT, as of July 2009.

3. Other languages

A large number of data collections have been made in the past decades for signed

languages. In most cases, these consist of collections of analogue video tapes that have often not been catalogued, are often transcribed or otherwise annotated on paper (if at all), and at best have received a transcription in some kind of text document. The consent obtained from informants typically applied to the specific research goal at hand, and metadata for the data are often sparse.

Over the past ten years, more and more digital collections of linguistic research data have been created, but again these often arose in response to specific research questions, and are far more often short pieces of elicited signing than spontaneous interactive discourse. Their annotation is sometimes done in annotation tools such as SignStream or ELAN, but often still in other standard office software. None of this material is published online, or otherwise made public. A recent survey of the composition of the data collections and the data elicitation methods was made as part of a dissertation on German Sign Language.¹

Only recently systematic larger corpora have been set up. The data collections for Auslan (the sign language of Australian) and NGT are the first to be completed. The former will only be made public in 2012. At the moment of writing this report, similar corpora were being collected in the USA, UK, Italy, Ireland and Germany. Plans for similar data collections are set up in Sweden and Norway. All these projects share the overall approach of recording many dozens of signers in one-to-one conversation with multiple cameras, aiming to elicit both narratives and more interactive materials, and a basic annotation of glosses and translations for at least part of the video recordings. In many cases open access is foreseen for (some of) the data. Little information on the intended distribution is currently available.

4. Conclusion

While it is clear that Sign Language of the Netherlands is the only signed language for which currently an open access annotated corpus exists, it is also clear that it is by no means a fully annotated corpus. A good start had been made with glossing at the time of the publication of the materials in 2008: about 8 of the 70 hours had been glossed. Since the start of SignSpeak, about 2 hours has been added. From a review of all glosses so far, it is clear that several steps need to be taken to make the gloss annotations useable for automatic sign recognition in SignSpeak:

1. General cleaning of spelling and typing errors, and inconsistencies in the application of the annotation guidelines.
2. Systematic labelling of (regional) variants.
3. An attempt to categorise productive sign productions that are the result of productive morphology (incl. classifier use), pantomimic expressions, and manipulation of iconic signs.

¹ Konrad, Reiner. 2009: Die lexikalische Struktur der DGS im Spiegel empirischer Fachgebärdenlexikographie. Zur Integration der Ikonizität in ein korpusbasiertes Lexikonmodell. Universität Hamburg. [Unveröffentl. Dissertation]. [The lexical structure of German Sign Language (DGS) as mirrored by empirical sign language lexicography of technical terms. A corpus-based lexicon model of DGS taking into account the iconicity of signs. University of Hamburg. [Unpublished dissertation.]]

This process has started in June 2009, and will be continuing until at least October 2009. The annotation guidelines will be adapted for new annotation documents. In addition to this attempt to maximally systematise the annotations, an improved type-token ratio can be obtained by selecting a subset of the Corpus NGT; least variation among signers is expected among younger signers. As the Groningen region is most strongly represented in the Corpus NGT, it appears that it would be best to select Groningen signers from under 50 for further annotation. In addition, the signing style of each signer will be qualified in terms of the use of Sign Supported Dutch, the importance of mouthing in the communication, and the overall fluency. The results of this evaluation will be used in further refining the selection of segments within the above group.