Sharing sign language corpora online: proposals for transcription and metadata categories

Onno Crasborn[*], Els van der Kooij[*], Daan Broeder[†], Hennie Brugman[†]

[*] Department of Linguistics, University of Nijmegen
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
(o.crasborn, e.van.der.kooij}@let.kun.nl
[†] Technical group, Max Planck Institute for Psycholinguistics
PO Box 310, 6500 AH Nijmegen, The Netherlands
{daan.broeder, hennie.brugman}@mpi.nl

Abstract

This paper presents the results of a European project called ECHO, which included an effort to publish sign language corpora online. The aim of the ECHO project was to explore the intricacies of sharing data using the internet in all areas of the humanities. For sign language, this involved adding a specific profile to the IMDI metadata set for characterizing spoken language corpora, and developing a set of transcription conventions that are useful for a broad audience of linguists. In addition to presenting these results, we outline some options for future technological developments, and bring forward some ethical problems relating to publishing video data on internet.

1. The ECHO project

Within the EU project 'European Cultural Heritage Online' (ECHO)[1], one of the five case studies is devoted to the field of language studies. The case study is titled 'Language as cultural heritage: a pilot project with sign languages'[2]. New data have been recorded from three sign languages of different Deaf communities in Europe: Sign Language of the Netherlands (abbreviated SLN), British Sign Language (BSL) and Swedish Sign Language (SSL). By having people retell written fable stories, comparable data resulted that can be used for cross-linguistic research. In addition to these semi-spontaneous data, we have elicited basic word lists and included some sign language poetry (some newly recorded, some already published).

The first aim of this paper is to characterize the conventions that were used and to explain why these can be considered as useful to a larger audience of linguists. The ELAN and IMDI software tools that were used to enter the transcriptions and metadata store their data in XML files whose format is described by open schemata and which can be accessed by other software tools as well. Using these open-standard tools, we developed a set of transcription conventions that are likely to be useable by a large group of researchers with diverse interests.

The second aim of this paper is to outline some desired functionalities of these tools that will make it more attractive to actually use existing corpora. Finally, we will outline some ethical challenges that have not yet received much discussion in the sign language field

2. The need for standardization

For actual cross-linguistic studies to take place, it is necessary that not only the same stimulus material is used, or otherwise comparable data are used, but also that the same conventions for annotating these data are used, both in terms of linguistic transcription and in terms of metadata description. The availability of a small corpus of video recordings from different languages, as published for the ECHO project, hopefully promotes standardization.

2.1 Metadata standards

In our case, metadata descriptions of language corpora characterize the documents and data files that make up the corpus in terms of descriptors that pertain to the whole unit of media and transcription files, rather than to individual sections within the files. For example, information about the subjects, the identity of the researchers involved in the collection and the register used by the speakers or signers typically belongs to the metadata domain. Users can then search within and across large corpora for all transcribed video material with male signers older than 45 years, for example. However, for such searches to be possible, it is essential that users obey the same conventions for labeling corpora. A proposal for such a standard is presented in section 3[3]. This is a specialization of the IMDI set of metadata descriptors for language resources[4].

2.2 Transcription standards

Several tools are currently available for annotating video data. Both SyncWriter (Hanke & Prillwitz 1995) and SignStream[5] have developed especially for sign language data, whereas ELAN started its life in the domain of gesture research (former versions were called MediaTagger)[6].

These new technologies for presenting sign language data and transcriptions pose the following question: to what extent should we use standard transcription conventions? If all the raw material (the video sources)

[1] http://echo.mpiwg-berlin.mpg.de/

[2] http://www.let.kun.nl/sign-lang/echo/; project partners were the University of Nijmegen, City University London, and Stockholm University.

[3] Further information on the proposed standard can be found at http://www.let.kun.nl/sign-lang/IMDI/.

[4] http://www.mpi.nl/IMDI

[5] http://www.bu.edu/asllrp/SignStream/

[6] http://www.mpi.nl/tools/elan.html

is available, do we need full transcriptions? In principle, one can look at the video source for all kinds of information that are traditionally included in various transcription system, such as eye gaze, head nods, etc. On the other hand, the great strength of computer tools such as ELAN is that it allows for complex searches in large data domains and for the immediate inspection of the video fragments relating to the search results; this is typically very time consuming when using paper transcription forms or even digitized transcription forms that are not directly linked to the original data.

Within the ECHO project, we therefore wanted to establish an annotation system that could be useful for any researcher, with a focus on the syntactic and discourse domains. We tried to be careful not to impose too much analysis on any tier by saying that a specific phonetic form is an instance of 'person agreement', for example. On the other hand, analytical decisions are constantly being made in any transcription process. For example, even adding multiple tiers with translations in various written languages (in the case of the ECHO project: Dutch, English and Swedish) implies taking (implicit or explicit) decisions about where sentence boundaries are located.

While every research project will have its own research questions and require special transcription categories, it should be possible to define a standard set of transcription tiers and values that are useful to large groups of researchers, regardless of their specific interests. For example, a translation at sentence level to a written language is always useful, if only for exploring a video recording. Working with three teams of linguists from different countries, each with their own research interests, the ECHO project formed a good start for developing such a standard set of transcription conventions. This ECHO set is described in section 4.

The relatively small set of transcription tiers allows for the coding of a relatively large data set, which can be further expanded by researchers according to their specific needs. ELAN will see several updates in the near future; one of the future functions will be the possibility to expand a publicly available transcription file with ones own additions, including extra tiers, and storing these additions in a local file while maintaining the link to the original transcription that will be stored on a remote server.

3. Metadata description of sign language corpora: expanding the IMDI standard

3.1 The IMDI standard and profiles

The set of IMDI metadata descriptors that was developed for spoken language corpora distinguishes 7 categories for each session unit:

1. *Session*. The session concept bundles all information about the circumstances and conditions of the linguistic event, groups the resources (for example, video files and annotation files) belonging to this event, and records any administrative information for the event.
2. *Project*. Information about the project for which the sessions were originally created.
3. *Collector*. Name and contact information for the person who collected the session.
4. *Content*. A set of categories describing the intellectual content of the session.
5. *Actors*. Names, roles and further information about the people involved in the session, including the signers and addressees, but also, for example, the researchers who collected and transcribed the material.
6. *Resources*. Information about the media files, such as URL, size, etc.
7. *References*. Citations and URLs to relevant publications and other archive resources.

Each of these seven categories allow for extension by users, in the form of 'key–value pairs'. A key specifies an extra category, an extra field, for which a value can be specified. For example, one might specify a key called *Backup Copy* to quickly specify whether a back-up copy of the original tape has already been made (yes vs. no).

In a workshop for the ECHO project, held at the University of Nijmegen in May 2003, a group of sign linguists from various countries and with varying research interests sat together to see how these categories could be applied to sign language data. The outcome of that workshop was a set of key fields to describe sign language corpora. These extra categories have now been bundled in an extension to the standard IMDI metadata specification, called 'sign language profile'. Profiles in the IMDI Editor tool offer sets of extra fields that apply to specific types of data, in this case communication in a specific modality.

3.2 The sign language profile

The sign language profile adds key fields in two areas in the IMDI set: content and actors. All of the fields can be specified or left empty.

In content, *Language Variety* describes the specific form of communication used in the session, and *Elicitation Method* specifies the specific prompt used to elicit the data at hand. A set of four keys describes the communication situation with respect to interpreting: who was the interpreter (if any) interpreting for (*Interpreting.Audience*), what were the source and target modalities (*Interpreting.Source* and *Interpreting. Target*), and is the interpreter visible in the video recording (*Interpreting.Visibility*)?

Secondly, four sets of keys are defined that can be used to describe various properties of each actor who is related to the session: properties pertaining to deafness, the amount of sign language experience, the family members, and the (deaf) education of the actor.

Deafness.Status describes the hearing status of the actor (deaf, hard-of-hearing, hearing), and *Deafness.AidType* describes the kind of hearing aid the actor is using (if any).

The amount of *Sign Language Experience* is expressed by specifying the *Exposure Age*, *Acquisition Location* and experience with *Sign Teaching*.

The family of the actor can be described by specifying *Deafness* and *Primary Communication Form* for *Mother*, *Father* and *Partner*.

Finally, the *Education* history of the actor can be specified in a series of keys: *Age* (the start and end age

of the actor during his education), the *School Type* (primary school, university, etc.), the *Class Kind* (deaf, hearing, etc.), the *Education Model*, the *Location*, and whether the school was a *Boarding School* or not.
A more complete definition of the whole sign language profile is given in Crasborn & Hanke (2003).

4. A standard set of linguistic transcription conventions for sign language data

4.1 An introduction to ELAN and the 'tier' concept

Below we describe the different tiers used for the ECHO project[7]. A tier is a set of annotations that share the same characteristics, e.g. one tier containing all the glosses for the right hand and another tier containing the Dutch translations. ELAN distinguishes between two types of tiers: "parent tiers" and "child tiers". Parent tiers are independent tiers, which contain annotations that are linked directly to a time interval of the media frame. Child tiers or referring tiers contain annotations that are linked to annotations on another tier (the parent tier)[8]. ELAN provides the opportunity to select one or more video frames and assign a specific value to this selected time span. For example, when the eye brows are first up and then down (neutral) in the same sign, one would only select the time interval in the video in which the eyebrows are up for the brows tier, and mark that time-domain with a specific code (for instance 'up'). This is possible for all tiers that one creates.
It is important to emphasize that, unlike in the IMDI software, there is no standard set of tiers for any document. Tiers have to be set up by the user for every annotation file that is created to annotate a media file. The set that we propose is just that: a proposal for a set of tiers that cover elementary transcription categories that can be useful for many different kinds of research. The use of this set of tiers is exemplified by the data transcribed for the ECHO project[9]. Any user can add both additional tiers and additional annotations on existing tiers to the documents that have been published in the context of the ECHO project.

4.2 Tiers with general information

General information that can be supplied for every fragment of a video file includes *Translation* tiers for English, Swedish and Dutch. Each of these tiers target a translation at sentence level. An annotation on the *Role* tier indicates that the signer takes on the role of a specific discourse participant, as commonly happens in sign language discourse. Finally, the *Comments/notes* tier can be used to add any kind of comment by the user.

4.3 Tiers with manual information

Manual behavior is systematically described separately for the two hands. For both the left and the right hand, there is a *Gloss* tier. Child tiers for each of these two articulators specify whether there is *Repetition* in the movement of the glossed unit, and what the *Direction & Location* of each of the hands is.

4.4 Tiers with non-manual information

A set of non-manual tiers allow for the specification of some of the relevant properties of the face, head, and body of the signer. Movement of the *Head* and *Eye Brows* can be specified, as well as the amount of *Eye Aperture* (including the notation of eye blinks) and the direction of *Eye Gaze*.
A new system was devised to specify the behavior of the *Mouth*, including the tongue, which in previous systems was often treated in a rather fragmentary manner (Nonhebel, Crasborn & van der Kooij 2004b).

4.4 Properties of the transcription conventions

The transcription system outlined in the sections above had two central goals. First of all, it should be easy and relatively quick to use for encoders, so that users can transcribe considerable amounts of data within a reasonable time frame. This inevitably goes at the expense of detail. For example, for facial expression, the FACS system (Ekman, Friesen & Hager 2002) is the most detailed and accurate transcription method that is known, but it is extremely time-intensive to learn to master and use, and offers far more detail than is necessary for the large majority of research projects. The tiers for non-manual activity that we propose aim to form an optimal compromise between the amount of detail available to the user and the time investment made by the transcriber.
Secondly, we tried to systematically separate form from function for all tiers. Since the function of a given linguistic form can vary from language to language, it is crucial to emphasize the coding of the form of linguistic behavior.

5. Specifications for future tools

Most importantly in the context of this paper, searching across both data and metadata domains will need to be an important target of further development. In the present state of the tools, one needs to first search within the set of metadata categories, and in the resulting set of transcription files search for data categories one-by-one. Finding all cases of weak hand spreading by people younger than 20 thus becomes a very time-consuming task, whereas corpora are particularly useful for those kinds of complex queries.
In the sign language research community, working with corpus data is still very uncommon, presumably in part because there are no commonly used written forms of sign languages until now that have allowed to create text corpora. Now the computer technology is available to build up corpora of digitized video recordings and annotate these, in addition to the search facilities, software is needed to provide basic statistical functions in ELAN, including frequencies of annotation values on

[7] An extensive description is available in Nonhebel, Crasborn & van der Kooij (2004a).
[8] See also ELAN manual, available at http://www.mpi.nl/tools/elan.html.
[9] These data can be freely downloaded from http://www.let.kun.nl/sign-lang/echo/data.html.

different tiers and the distribution of the durations of these annotation values. Currently, the most obvious way to perform quantitative analyses of transcription files at this moment is to export data to a spreadsheet program for further analysis.

A function that is currently being implemented is to add a visualization of kinematic recordings with the transcription of video material, similar to the display of the oscillogram of sound files in ELAN. These numerical data can then be more easily integrated with qualitative analyses based on transcription. Additionally, the software will need to provide numerical analyses appropriate to phonetic analysis of sign languages, similar to the 'Praat' software for speech analysis (Boersma & Weenink 2004). As the field of sign language phonetics is still in its infancy, the specifications of such functionality will have to develop over the years to come. Finally, a similar integration of quantitative data from eye-tracking equipment would enhance the usability of the software for some research groups.

Working together with colleagues anywhere in the world on the same annotation document at the same time is another function currently under development. Using peer-to-peer technology, it will become possible to look at the same annotation document on different computers connected to the internet, and instantly see modifications that are being made by the other party. In combination with a chat function, one can jointly look at existing annotations and create new annotations (see Brugman, Crasborn & Russel 2004 for further details on this 'collaborative annotation' concept).

6. Ethical aspects of publishing sign language data online

Needless to say, the privacy of subjects in scientific studies has to be respected. For the sign language study in the ECHO project, this gives rise to extra problems not previously encountered in the creation of spoken language corpora that just make use of sound recordings. The visual information in the video recordings contains a lot more personal information than audio recordings of voices, including not only the identity of the signer (i.e., the visual appearance of the face), but also more clues to the emotional state and age of the person, for example.

While it is common practice to ask subjects in linguistic recordings for their explicit written permission to use the recordings for various purposes, including making images for publications, discussion among sign language specialists revealed that this permission is a rather sensitive issue in the case of internet publication. Publication of data online imply that the information is available to the whole world, and not just to a limited group of people with access to specific university libraries, for example, as in the case of video tape recordings used until recently. Signers who have no problem with the inclusion of the video data at the time of recording may well regret this choice 15 years later. Can this be considered the problem of the person involved, or should researchers make more of an effort to outline the implications of sharing data to subjects?

Alternatively, data access can be restricted to linguists registered as users of the corpus by the host institution, but this comes down to restricting access to data that were intended to be public – at least within the open access concept that is central to the ECHO project.

Future projects aimed at making data accessible online should explore these issues in more depth, with assistance from both legal and ethics specialists.

7. References

Boersma, P. & D. Weenink (2004) Praat. Doing phonetics by computer. http://www.praat.org.

Brugman, H., O. Crasborn & A. Russel (2004) Collaborative annotation of sign language data with peer-to-peer technology. Paper presented at LREC 2004, Lissabon.

Crasborn, O. & T. Hanke (2003) Additions to the IMDI metadata set for sign language corpora. http://www.let.kun.nl/sign-lang/echo/docs/SignMetadata_Oct2003.doc.

Ekman, P., W. Friesen & J. Hager (2002) *Facial Action Coding System*. Salt Lake City: Research Nexus.

Hanke, T. & S. Prillwitz (1995) syncWRITER: Integrating video into the transcription and analysis of sign language. In H. Bos & T. Schermer (eds.) *Sign language research 1994*. Hamburg: Signum. Pp. 303-312.

Nonhebel, A., O. Crasborn & E. van der Kooij (2004a) Sign language transcription conventions for the ECHO project. http://www.let.kun.nl/sign-lang/echo/docs/transcr_conv.pdf.

Nonhebel, A., O. Crasborn & E. van der Kooij (2004b) Sign language transcription conventions for the ECHO project. BSL and NGT mouth annotations. http://www.let.kun.nl/sign-lang/echo/docs/transcr_mouth.pdf.