The Sign Linguistics Corpora Network: towards standards for signed language resources

Onno Crasborn
Centre for Language Studies, Radboud University Nijmegen
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
E-mail: o.crasborn@let.ru.nl

Abstract

The Sign Linguistics Corpora Network is a three-year network initiative that aims to collect existing knowledge and practices on the creation and use of signed language resources. The concrete goals are to organise a series of four workshops in 2009 and 2010, create a stable Internet location for such knowledge, and generate new ideas for employing the most recent technologies for the study of signed languages. The network covers a wide range of subjects: data collection, metadata, annotation, and exploitation; these are the topics of the four workshops. The outcomes of the first two workshops are summarised in this paper; both workshops demonstrated that the need for dedicated knowledge on sign language corpora is especially salient in countries where researchers work alone or in small groups, which is still quite common in many places in Europe. While the original goal of the network was primarily to focus on corpus linguistics and language documentation, human language technology has gradually been incorporated as a user group of signed language resources.

1. Introduction

This paper characterises a new project in the field of language resources, focused on signed languages. Signed languages have only been seriously studied by linguists since the 1970s, after the ground-breaking work of Tervoort (1953) and Stokoe (1960). In the last decade, research on signed languages has seen two major developments: the broadening of the range of languages under study beyond the US and western Europe, and the increasing use of digital resources with concomitant increases in the size of the data sets that are used for analyses, in turn giving a stronger empirical foundation to the linguistic study of signed languages. Electronic resources had played a limited role in linguistic study of signed languages, mostly consisting of electronic dictionaries targeted at a general audience (e.g. Armstrong 2003). Most recently, systematic creation of signed language corpora has announced a new phase in sign language research (Crasborn et al. 2007; Crasborn 2008; Johnston 2008ab, 2010; Crasborn & Zwitserlood 2008; Johnston & Schembri to appear). The online publication of sign language resources facilitates research by individual researchers, and it also provides shared reference points for the languages in question. The absence of writing systems in common use by deaf communities has lead to an extreme scarcity of historical data for not only individual languages, but also for signed languages as a communication type distinct from spoken languages. For this very reason, signed language corpora as they are currently being developed also have an enormous potential for the language community itself. They can serve not only as basic resources for language teaching and learning, but also form a first major documentation of language use in deaf communities, and can potentially collect a form of "oral history" that has so far not been available in most deaf communities, largely through technological restrictions related to publishing film and video. Collection and publication of signed language (if any) has been restricted to lexical resources for applied purposes (cf. the articles in Wilbur 2001).

The EU-funded 'ECHO' (European Cultural Heritage Online) project established an "open source culture of the public and scholarly exploitation of cultural heritage on the internet" ([ECHO02]). One of the four pilot projects was targeted at sign language, and included the specification of metadata descriptors for sign language datasets (Crasborn et al. 2007); in addition, ECHO supported the development of ELAN and IMDI software that is widely used to annotate video corpora and create associated metadata. Two partners contributed to this European precursor project (Nijmegen and Stockholm).

A number of sign language corpora are now available, underway of planned, addressing the sign language of deaf communities in the Netherlands, the UK, Ireland, Sweden, Greece, Australia and the US but also e.g. Mali. Increasingly, sign language corpora are planned or opened up for use in sign language teaching and learning. Finally, there are also widespread efforts to develop forms of automation, including sign and gesture recognition, movement tracking systems, and synthesisers animating avatars. These examples are all taken from the programme of the 3rd Workshop on the Representation and Processing of Sign Languages at the 2008 LREC conference ([LREC08]). In July 2009, there was a first workshop on the linguistic research based on such video corpora and related data collections [SIGN09].

The various aims and objectives of these cutting edge developments also need a clear humanities-oriented framework, because what unites all efforts to collect, describe, analyse, model and synthesise sign language data is a common interest in nurturing the native sign languages of deaf communities

around the globe This deeper motivation implies a humanities perspective, by which we mean that the language that is used among deaf people cannot be properly understood without, firstly, taking into account the status of sign language in society, as signed languages have long been oppressed (Ladd 2003). Secondly, the language community has long been associated with a strong internal cultural capital but a weak social capital; this is attributable to persistent exclusion from full social participation, which is still visible in educational and employment statistics. And thirdly, the deaf community is marked by extreme exposure to social technologies, including not only hearing aids and cochlear implantation but also health technology assessments and educational monitoring instruments. While sign language corpora can be used to analyse the first two aspects, they are part of the third aspect (Thoutenhoofd 2007, Crasborn et al. 2007, Crasborn 2010). A discussion of ethical issues is therefore an explicit part of the proposed network. Similarly, its humanities orientation will encourage historical, socio-political, and culture and media interest in the outcome of work on sign language corpus linguistics.

The Sign Linguistics Corpora Network (SLCN) is a Dutch initiative funded by the Netherlands Organisation for Scientific Research (NWO, grant no. 236-89-002) with contributions from six international partners. This network will bring together wide-ranging expertise and project experience relating to a rapidly expanding area of work: the construction and exploitation of sign language corpora. The coordinating partner (Radboud University Nijmegen) and foreign research groups unite experience from language technology (Greece, Germany), linguistics (Netherlands, Sweden, Germany), second language acquisition (UK, Germany) and sign language interpreting (UK, Germany). The following partners are involved:

- Centre for Language Studies, Radboud University Nijmegen (dr. O. Crasborn, coordinator)
- Stockholm University (dr. J. Mesch)
- Magdeburg University of Applied Sciences (prof. J. Heßmann)
- Heriot Watt University, Edinburgh (prof. G. Turner)
- Institute for Language and Speech Processing, Athens (dr. E. Efthimiou)
- Virtual Knowledge Studio, KNAW, Amsterdam (dr. E. Thoutenhoofd)
- Institute for German Sign Language, Hamburg University (dr. T. Hanke)
- Deafness, Cognition and Language research centre, University College London (dr. A. Schembri).

2. Aims of the project

The general aims of the project are as follows:

- Sharing academic work on the large-scale capture, coding and analysis of sign language data.
- Developing a stable network to support collaboration and stimulate growth and innovation in sign language corpus linguistics and technology.
- Uniting technical, computational, linguistic and educationally oriented sign language corpus projects in a collective humanities interest.

Three concrete objectives are addressed:

- Enhance existing collaboration through four open workshops, and better outward visibility by creating a web portal for anyone working with sign language corpora.
- Develop new forms of collaboration, including targeted work on specifying standards, coding, grammars and lexicons in sign linguistics and the development of specifications for software and infrastructures.
- Write and submit an international grant proposal to encourage the implemention of the above standards and software, and to work towards the new goals that will be defined in the workshops.

3. Mission statement

At the start of the project in late 2008, the following mission statement has been created:

"Documentation of the sign languages of deaf people poses specific problems due to the lack of generally accepted writing systems, the brief history of sign language linguistics, and the minority status of these languages. Technological advances have facilitated the development of digital tools for the collection of sign language data in recent years. However, there are still very few research groups that have the resources and skills to employ such tools for the creation of large scale sign language corpora that might inform linguistic research and aid in the development of language technology, be of use in teaching and training contexts, and benefit a wider deaf and hearing public.

The SLCN brings together relevant European initiatives in order to combine and share expertise in sign language corpus development and promote international cooperation. It will organize a series of international workshops to discuss data collection, technological formats, organisation of metadata, annotation processes as well as questions of accessibility, dissemination and use of sign language data. Discussion within the network aims at encouraging wider European initiatives for the preservation of sign languages as part of our cultural heritage for future generations.

The SLCN partners recognise that social transformations are changing sign language communities. Deaf communities have always faced categorically worse social conditions, and they are currently facing new challenges that arise from an unprecedented impact of science and technology on society. Since SLCN will itself be a new science and technology actor, it recognises that the construction of corpora as e-research technology and sign linguistics as a scientific community entail social responsibilities and

commitments. SLCN is therefore dedicated to promoting linguistic rights, nurturing native sign languages and developing corpus linguistic research as an engaged form of social action."

4. Workshops

The three-year programme consists of four international workshops of two days each. The workshops will also provide a context for the organisers to meet before and after to develop the SLCN. Each workshop will invite scholars who are external to the network, because a) they would be a potential partner for a European grant application, b) they have expertise from which the SLCN could benefit, or c) they can help with critical reflection on the various network aims. The concrete and focused subjects of the workshops ensure positive outcomes that will be secured for the international sign language research community by the creation of a web portal. In addition to the invited participants, the workshops will also be publicly announced, as participation of researchers from highly diverse research and education centres is an important goal of SLCN. Sign language research is often carried out by individual researchers in larger linguistic departments, and also at many non-academic institutes: deaf associations, deaf schools, interpreting programmes, lexicographic institutes for national signed languages, et cetera.

The workshops will focus on the following four subjects:

1. creating an overview of existing corpus efforts and linguistic and technological use of corpora, data collection strategies (July 2009; London)
2. exchanging experiences with metadata descriptions for signed language data, evaluating an earlier proposal for sign-specific categories (November 2009; Nijmegen)
3. exchanging proposals for annotation strategies for signed languages (June 2010; Stockholm)
4. exploitation of sign corpora; specifying the steps towards a larger European research effort for a specific grant application (end 2010; Berlin)

The first workshop in London brought together nearly 40 researchers. Ongoing and planned signed language corpus creation projects were presented, and a series of invited speakers talked about the use of spoken language corpora for studies in applied linguistics and language technology. Also, recent developments in the area of digital video were sketched.

The most prominent outcome of the workshop was the fact that due to intensive international contacts in the recent decade, the data collections that have been set up so far have very much followed the same procedure. Dialogues of many dozens of signers were recorded, carefully balancing narrative and more interactive registers; dialectal variation was explicitly targeted by all projects. Elicitation materials for narratives were often highly similar or even identical. While these shared procedures facilitate cross-linguistic studies, it is also clear that all corpora have the same restrictions that would ideally be addressed in future projects:

- Only dialogues were recorded, not multilogues
- A relatively large amount of time (±50%) was devoted to elicited narratives
- All people were only recorded on one day in one continuous session; this limits a good view on interpersonal variation
- Similarly, all signers were only recorded signing to one and the same interlocutor, likewise reducing the amount of interpersonal variation.

At the workshop on Metadata in November 2009, recent metadata developments within CLARIN were presented, pointing towards a more flexible set of metadata fields for any language resource depending on the needs of the user. This enables the flexible addition of metadata fields specific to sign language resources. The fixed 'sign language profile' that had been created for the IMDI metadata standard (Crasborn & Hanke 2003) can thus be more easily adapted in the future. At the same time, researchers at the meeting agreed that it would be profitable to standardise as many fields as possible. These specifically relate to signer properties, such as the hearing status and educational background of signers and their immediate family and their experience in acquiring or learning sign language, but also include whether or not a video recording is voice interpreted, for example.

Although there are cases of sign language corpora where even the videos are published as open content for layman usage, the discussions at the workshop did highlight the ethical problems in providing open access to the metadata themselves, which is the standard policy within CLARIN. As sign language communities can be very small, a lot of the actor information which is crucial to store for research purposes can be privacy sensitive information for outside users of the internet.

The following two workshops will take place in 2010. The first will focus on annotation and transcription of sign language corpora, exchanging experiences with annotation of sign corpora of both children and adults. The first steps towards future standardisation of some aspects of annotation will be made, within the framework of the technical standards that are already under development in CLARIN (including the EAF, ELAN Annotation Format, currently used by many linguists). The fourth and final workshop will investigate various ways in which sign language corpora as they are now being constructed can be exploited for various goals. These will involve research in linguistics and language technology, but also use as a 'deaf digital library' for non-researchers.

5. The web portal

To publish the results of the workshops, a web portal has been set up that aims to incorporate both concrete outcomes (workflow descriptions, ethical procedures and considerations, and possible linguistic approaches to corpus collection) as well as RFC documents aiming to elicit feedback from the wider research community on possible linguistic and technological standards:

www.signlanguagecorpora.org. This wiki style web portal includes an agenda with relevant events in the area of sign linguistics and language resources. For the duration of the project, practical information will be available through the project web site, www.ru.nl/slcn; relevant information for future reference will be integrated in the portal at the end of the project. The portal will serve as a reference point for contact with other relevant digital humanities projects, such as ECHO. Where relevant, the project will liaise with DANS in the Netherlands and with CLARIN in Europe in planning long-term duration of network resources or integrating them with existing infrastructure ([DANS08], [CLARIN08]).

6. Acknowledgment

The Sign Linguistics Corpora Network and the creation of this paper was supported by the Netherlands Organisation for Scientific Research (NWO) grant no. 236-89-002

7. References

Armstrong, D., Ed. (2003) Special issue on dictionaries and lexicography. *Sign Language Studies* 3(-44).

[CLARIN08] Home page, Common Language Resources and Technology Infrastructure, http://www.clarin.eu.

Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., van der Kooij, E, Woll, B., Bergman, B. (2007) Sharing sign language data online. Experiences from the ECHO project. *International Journal of Corpus Linguistics* 12-4: 535-562.

Crasborn, O. (2008) Interpreting sign language corpora. Research Seminar on Translation studies & Intercultural Communication, Heriot-Watt University, Edinburgh UK, Febr. 2008.

Crasborn, O. (2010) Ethical questions on the publication of sign language data on internet. *Sign Language Studies* 10-2: 276-290.

Crasborn, O., Hanke, T (2003). Metadata for sign language corpora. Available online at: http://www.let.ru.nl/sign-lang/echo/docs/ ECHO_Metadata_SL.pdf.

Crasborn, O., Zwitserlood, I. (2008) The Corpus NGT: an online corpus for professionals and laymen. In: O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris. Pp 44-49.

[DANS08] Home page, Data Archiving and Networked Services, February 2008. http://www.dans.knaw.nl/en/.

[ECHO02] Home page, European Cultural Heritage Online, 2002-present. http://echo.mpiwg-berlin.mpg.de/home

Johnston, T. (2008a) Corpus linguistics & signed languages: no lemmata, no corpus. In: O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris. Pp . 82-87.

Johnston, T. (2008b) Corpus of grammar and discourse strategies of deaf native users of Auslan (Australian Sign Language), Endangered Languages Archive, SOAS, University of London. http://elar.soas.ac.uk/node/3. Public access from 2012.

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

Johnston, T., Schembri, A. (to appear) Corpus Analysis of Deaf Sign Language. In Carol Chapelle (General Editor) *Encyclopedia of Applied Linguistics*. London: Wiley-Blackwell.

Ladd, P. (2003) *In Search of Deafhood*. Multilingual Matters: Clevedon.

[LREC08] Home page of the LREC workshop web site, http://www.sign-lang.uni-hamburg.de/lrec2008/cfp.html

[SIGN09] Workshop "Sign Language Corpora: Linguistic Issues", University College London, July 2009. URL: http://www.bslcorpusproject.org/sign-language-corpora-linguistic-issues

Stokoe, W. (1960) *Sign language structure. An outline of the visual communication systems of the American Deaf*. Silver Spring, MD: Linstok Press.

Tervoort, B. 1953. *Structurele analyse van visueel taalgebruik binnen een groep dove kinderen.* Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

Thoutenhoofd, E. (2007) Corpus linguistics as multimedia laboratory: Material culture and experimental practice in the social sciences. Paper presented at the 'Science and Technology Studies views of e-social science' panel, Third International Conference on e-Social Science, Ann Arbor (US), 7-9 October.

Wilbur, R., ed. (2001) Sign transcription and database storage of sign information. Double theme issue of the journal *Sign Language & Linguistics*, vol. 4:1/2.