Dicta-Sign – Building a Multilingual Sign Language Corpus

Silke Matthes[1], Thomas Hanke[1], Anja Regen[1], Jakob Storz[1], Satu Worseck[1], Eleni Efthimiou[2], Athanasia-Lida Dimou[2], Annelies Braffort[3], John Glauert[4], Eva Safar[4]

[1]Institute of German Sign Language and Communication of the Deaf, University of Hamburg, [2]Institute for Language and Speech Processing / “Athena” R.C., [3]LIMSI/CNRS, [4]University of East Anglia
{silke.matthes,thomas.hanke,anja.regen,jakob.storz,satu.worseck}@sign-lang.uni-hamburg.de,
{eleni_e,ndimou}@ilsp.gr, annelies.braffort@limsi.fr, {J.Glauert,E.Safar}@uea.ac.uk

Abstract

This paper presents the multilingual corpus of four European sign languages compiled in the framework of the Dicta-Sign project. Dicta-Sign researched ways to enable communication between Deaf individuals through the development of human-computer interfaces (HCI) for Deaf users, by means of sign language. Sign language resources were compiled to inform progress in the other research areas within the project, especially video recognition of signs, sign-to-sign translation, linguistic modelling, and sign generation. The aim for the corpus data collection was to achieve as high a level of naturalness as possible with semi-spontaneous utterances under lab conditions. At the same time the elicited data were supposed to be semantically close enough to be comparable both across individual informants and for all four sign languages. The sign language data were annotated using iLex and are now made available via a web portal that allows for different access options to the data.

Keywords: Sign language technologies, multilingual sign language resources, annotation

1. Introduction

Within the framework of the Dicta-Sign project (2009-2012) sign language resources were compiled for four European languages: British, German, Greek, and French Sign Language (BSL, DGS, GSL, and LSF). These resources were used to inform progress in other research areas within the project, especially sign recognition, sign-to-sign translation, linguistic modelling, and sign generation, which then in turn was used to improve sign language technology. At the same time the data are to serve as a self-contained resource for future research.

In a first step, a multilingual lexical database providing a core lexicon of approximately 1000 entries in the four project sign languages was built. The shared list of concepts chosen for the lexicon are of everyday use or specifically related to the field of Dicta-Sign's main topic, European travel. Signs were recorded for each language and annotated assigning gloss labels, form description (HamNoSys) and a rough meaning.

In a second step, a new corpus on the domain “Travel across Europe” was produced by using the same elicitation materials for all four sign languages. Prior to the project, parallel corpus collection for sign languages had only been undertaken in minimal sizes or for spoken language simultaneously interpreted into several sign languages, but not for semi-spontaneous signing by native signers. Because of the “oral” nature of sign language and the risk of influences from written majority languages the collection of parallel sign language data is a difficult task. Corpus planning therefore needs to balance between naturalness of the data to be collected on the one side and the degree of parallelisability of the data across languages on the other side. Within Dicta-Sign, the aim for the data collection was to elicit sign language data as natural as possible with semi-spontaneous utterances under lab conditions. With respect to parallelisability of the sign language data, elicitation tasks had to be designed that result in semantically close answers without predetermining the choice of vocabulary and grammar (Matthes et al. 2010).

Corpus data collection took place in each of the four countries involved in the project and the sign language data were annotated using iLex. A web portal was developed to allow access to the corpus data for research purposes.

2. Compilation of the Multilingual Corpus

A multilingual corpus on the domain “Travel across Europe” was compiled for the four sign languages involved in the project (BSL, DGS, GSL and LSF). Elicitation tasks were developed specifically for the project's purposes. After recording had taken place in all four countries, the sign language data were annotated on different levels.

2.1 Corpus Data Elicitation

With the objective of gaining sign language data as natural as possible on the one hand and comparable across languages as well as individual informants on the other hand elicitation tasks and materials were designed specifically for the Dicta-Sign corpus collection. One key point in the planning was to film Deaf informants in pairs, interacting with each other. The tasks therefore mostly required the active involvement of both conversational partners, asking them to discuss and negotiate on certain topics or to describe and explain things to the partner. The elicitation material consists of 10 different tasks, aiming at a session length of approximately two hours, and covers different interaction formats ranging from monologues to sequences of very short turns, also with different levels of predictability. It includes communication for transport by different means and contexts as well as related personal experiences (Matthes et al. 2010).

The complex studio setup that was decided to be used for Dicta-Sign's data collection consisted of seven cameras, two of them stereo cameras (Hanke et al. 2010a). The different camera perspectives (front, side and bird's eye view) were to help annotators interpret the signing. The additional stereo cameras provide footage that allows image analysis to reconstruct 3D information and help automatic processing. In each country, 16 to 18 informants were filmed in sessions lasting about two hours each. Not counting task explanations or material that needed to be excluded for certain reasons, the corpus now consists of 8 to 10 hours of signed data from 14 to 16 different signers per language.

A variety of post-processing steps were needed before annotation work could start, most importantly providing backup data, compression of the video files as well as precise frame-by-frame synchronisation.

2.2 Annotation

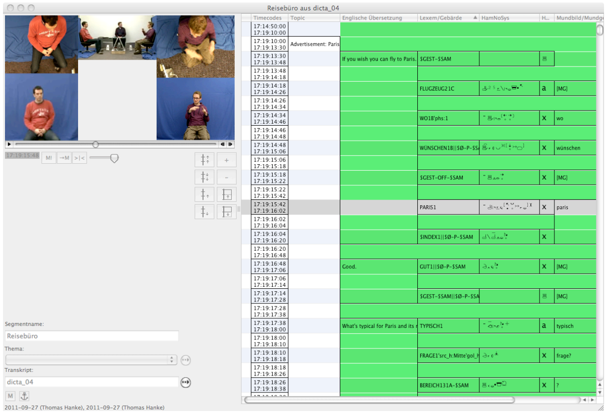
Corpus annotation work for all four sign languages within the Dicta-Sign project was carried out using iLex, an annotation environment that is linked to a lexical database (Hanke/Storz 2008). The video data were integrated into iLex and transcripts were produced for all tasks. At UHH, where Session Director had been used to run the elicitation sessions (Hanke et al. 2010a), it was possible to provide automatic tagging specifying start and end of the individual tasks and subtasks using time information from the Session Director log files. The annotation consists of a basic annotation on sign level for subsets of the sign language data as well as content tags that allow detecting comparable content across different signers and languages.

2.2.1 Sign Level Annotation

Sign level annotation of the corpus data is now available for about 40 minutes up to 5.5 hours per language, including segmentation of signs, lemmatisation, form description, as well as further details depending on the individual language. With regard to segmentation of the continuous signing it was decided to treat transitional movements between individual signs not as part of either sign (i.e. there are gaps between two signs during which the articulators move from the end of one sign to the beginning of the next). This approach, though more time consuming than segmenting once in the middle of a transition, offers advantages for subsequent processing: Firstly it results in the fact that a token tag only represents that part of the signal that is described by HamNoSys. Secondly, variation between tokens is much lower than if the transition would be part of the sign.

After segmentation the individual signs were lemmatised, i.e. unique glosses were assigned by means of type-token matching. In iLex this is done by linking tags to type entries in the database, which results in filling the transcript and a growth of the sign language database at the same time. A form description of the sign types was added using HamNoSys (Hanke 2004).

As a further step to enrich the corpus data, individual project partners conducted extra annotation work on data from their respective sign language: For the DGS data mouth patterns were annotated by assigning either written German words to represent mouthings or "MG" as a preliminary tag for mouth gestures. Furthermore, in addition to the annotation of HamNoSys on type level form deviations between types and the respective tokens were tagged as such in order to provide more reliable training data for image processing. The GSL data include a tagging of clause boundaries, and for LSF pointing, buoys and depicting signs were annotated using categories close to those proposed in the Auslan annotation guideline provided by Trevor Johnston (2011), and for pointing and depicting signs some indication on the use of the signing space were added. Furthermore, English translation is provided for parts of the German, Greek, and French subcorpora as well as a French translation for the majority of the LSF data.



Reisebüro aus dicta_04

17:19:15:48

Segmentname: Reisebüro

Thema:

Transkript: dicta_04

2011-09-27 (Thomas Hanke), 2011-09-27 (Thomas Hanke)

Timecodes	Topic	Englische Übersetzung	Lexem/Gebärde	HamNoSys	H...	Mundbild/Mundge
17:14:50:00 17:19:10:00						
17:19:10:00 17:19:13:30	Advertisement: Paris					
17:19:13:30 17:19:13:48		If you wish you can fly to Paris.	$GEST-$SAM			
17:19:13:48 17:19:14:18						
17:19:14:18 17:19:14:26			FLUGZEUG21C	[illegible]	a	[MG]
17:19:14:26 17:19:14:34						
17:19:14:34 17:19:14:46			WO1B'phs:1	[illegible]	X	wo
17:19:14:46 17:19:14:48						
17:19:14:48 17:19:15:06			WÜNSCHEN1B||$Ø-P-$SA	[illegible]	X	wünschen
17:19:15:06 17:19:15:18						
17:19:15:18 17:19:15:22			$GEST-OFF-$SAM	[illegible]	X	[MG]
17:19:15:22 17:19:15:42						
17:19:15:42 17:19:16:02			PARIS1	[illegible]	X	paris
17:19:16:02 17:19:16:04						
17:19:16:04 17:19:16:20			$INDEX1\|\|$Ø-P-$SAM	[illegible]	X	
17:19:16:20 17:19:16:48						
17:19:16:48 17:19:17:06		Good.	GUT1||$Ø-P-$SAM	[illegible]	X	[MG]
17:19:17:06 17:19:17:14						
17:19:17:14 17:19:17:28			$GEST-$SAM||$Ø-P-$SA			[MG]
17:19:17:28 17:19:17:38						
17:19:17:38 17:19:18:00		What's typical for Paris and its r	TYPISCH1	[illegible]	a	typisch
17:19:18:00 17:19:18:10						
17:19:18:10 17:19:18:18			FRAGE1'src_h:Mitte'gol_h	[illegible]	X	frage?
17:19:18:18 17:19:18:26						
17:19:18:26 17:19:18:38			BEREICH131A-$SAM	[illegible]	X	?

Figure 1: Transcript of Task "Travel Agency" (DGS informant)

2.2.2 Content Tagging

The Dicta-Sign multilingual corpus is not a parallel corpus in the classical sense as the "oral" nature of sign language as well as the risk of influences from written majority languages do not allow for such an approach. Instead, the aim within Dicta-Sign was to elicit semantically close answers without predetermining the choice of vocabulary and grammar. In order to allow identifying video sections in the corpus with comparable content across individual informants and languages, especially for those parts without sign level annotation, content tags were assigned that reflect the topics the informants signed about. The detailedness of the content tags varies across the different tasks, ranging from very broad content descriptions that mainly reflect the given structure of the subtasks to a more detailed specification of the topics covered (see examples below).

Example 1:
For the task “Public transportation” (task 1, category “Route descriptions”) the informants are asked to explain how to get from a certain place to another using public transportation. A map is provided to both of them displaying different means of public transport and stations. In five subtasks different stations are given as departure and destination points and each informant is asked to suggest one possible route per subtask.
For each subtask between nine and 12 different routes were described. While many of the 60 informants described similar routes, several routes occurred only once or twice. Mapping information was needed to compare information from the different sign languages: Route codes were agreed on and pictures were produced for each of the routes in order to ease the mapping (see pictures below). Discussion about the chosen routes was included in the route tags, but in cases were further discussion evolved (e.g. advantages of taking the bus) this was tagged separately.

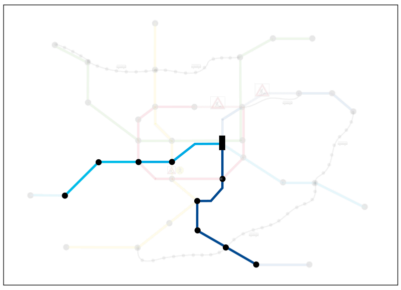


Figure 2: Route R2.2 (by 23 signers)

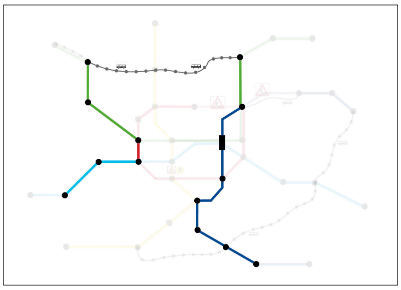


Figure 3: Route R2.6 (by 1 signer)

Example 2:
For the task “At the airport” (task 4, category “Description of Places and Activities”) one informant is asked to explain the procedures taking place at the airport as if the other has never travelled by plane before. Pictures displaying different aspects as checking in, boarding, and baggage claim are shown in chronological order.

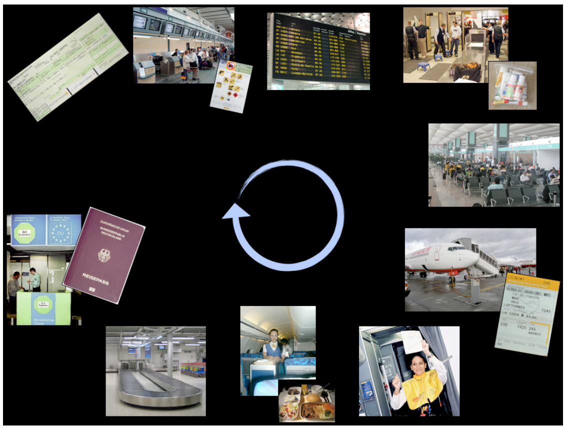


Figure 4: “At the airport” (German version)

The content tags distinguish the different steps that were described by the individual informants. Most of the topics are directly related to the elicitation material – as e.g. *check-in*, *information board*, *security check*, *food and drinks on board*, and *baggage claim* – and were covered by almost all 23 informants across three sign languages.[1] However, additional topics occurred: e.g. *Preparation of the trip* was mentioned by four informants (DGS, GSL, and LSF), *airplane fuelling* by three informants (GSL and LSF) and *Amusement activities on board* by 10 informants altogether (DGS, GSL, and LSF).

Example 3:
For the task “Expectation & Reality” (task 6, category “Narration”) the informants were asked to tell short stories based on picture cards showing somebody’s expectations of a certain situation and the actual situation. Topics of the stories were: small hotel room, cancelled flight, crowded museum, posh restaurant, rained off BBQ, and missed sunset.

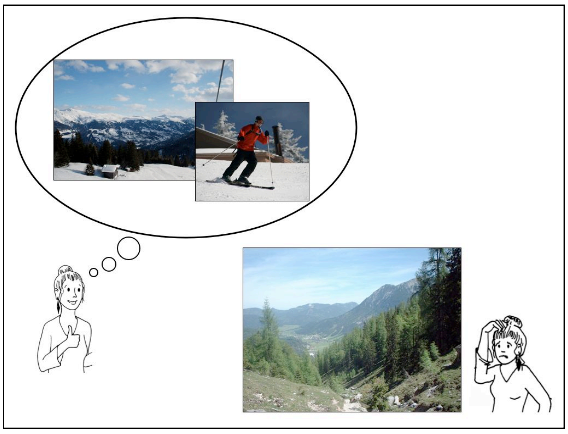


Figure 5: Example from the task explanation

[1] Only one informant per pair performed this task. For BSL tagging of this task is not available.

With respect to content tagging, the signing was first segmented into individual stories (i.e. subtasks), using Session Director log file information where available, and further divided into the ‘expectation’ and the ‘reality’ part of the story. For this task, applying only two content tags per subtask seems appropriate to mark the content, as the individual stories told by the informants are comparably short. For example, in the DGS data the content tags are – depending on the subtask – of a length ranging from 15sec up to 1min:19sec, with the ‘expectation’ part always being slightly longer than the ‘reality part’. This results in stories with an average length of about 1min:20sec. Both parts of the six stories could be detected in the data of almost all informants of the four sign languages.

2.3 Metadata

Personal metadata was collected from the informants by means of questionnaires based on the IMDI standard with sign language-specific extensions as defined in Crasborn/Hanke 2003. As that set covers a variety of purposes for metadata (e.g. to support language acquisition studies), but does not explicitly define subparts, Dicta-Sign defined a subset that seemed suitable for the kind of study conducted here and also minimised the questionnaire filling effort for the informants.

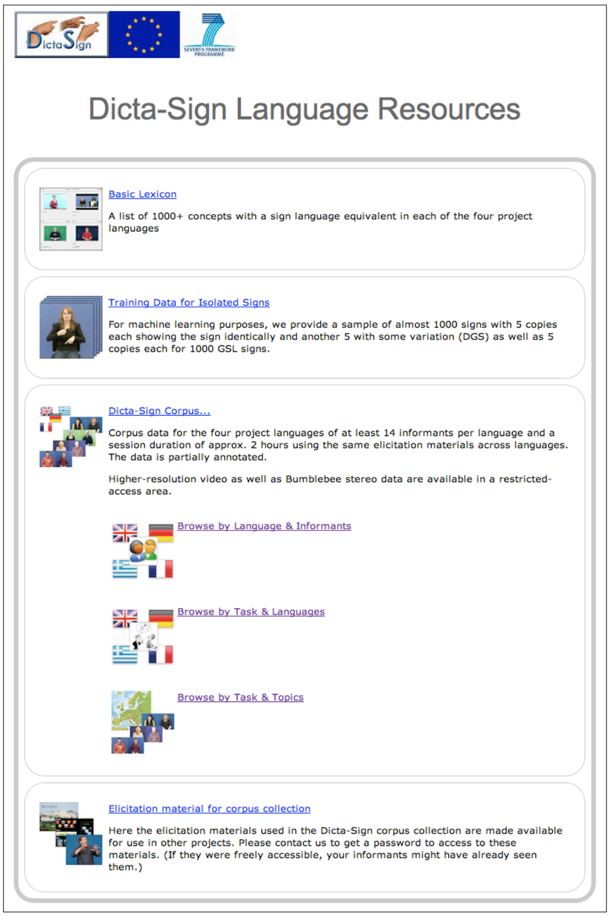
Metadata was collected in a finer granularity than appropriate for publication, however standards are not yet available that specify suitable coarsenings for such data. Therefore, two levels of coarsening were defined within Dicta-Sign for different publicity levels of informant data (see below on portal structure). For example, the informant’s date of birth is converted to the age in years for restricted access and age range (e.g. 41-50) for public access.

For the time being, data are made available in IMDI session file format. We plan, however, to convert these data into the CMDI component structure.

3. Exploitation

3.1 The Dicta-Sign Web Portal

A web portal was developed to allow access to the Dicta-Sign language resources for public use as well as research purposes. It can be accessed from the Dicta-Sign website: http://www.dictasign.eu/Main/Portal. Besides Dicta-Sign’s basic lexicon and further training data for sign recognition the portal presents the multilingual corpus, allowing for different access options to the data.



Dicta-Sign Language Resources

Basic Lexicon

A list of 1000+ concepts with a sign language equivalent in each of the four project languages

Training Data for Isolated Signs

For machine learning purposes, we provide a sample of almost 1000 signs with 5 copies each showing the sign identically and another 5 with some variation (DGS) as well as 5 copies each for 1000 GSL signs.

Dicta-Sign Corpus...

Corpus data for the four project languages of at least 14 informants per language and a session duration of approx. 2 hours using the same elicitation materials across languages. The data is partially annotated.

Higher-resolution video as well as Bumblebee stereo data are available in a restricted-access area.

Browse by Language & Informants

Browse by Task & Languages

Browse by Task & Topics

Elicitation material for corpus collection

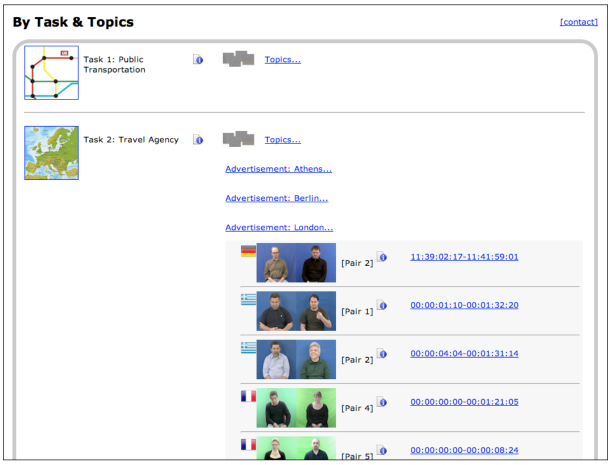
Here the elicitation materials used in the Dicta-Sign corpus collection are made available for use in other projects. Please contact us to get a password to access to these materials. (If they were freely accessible, your informants might have already seen them.)

Figure 6: The Dicta-Sign web portal

3.1.1 Layout

The Dicta-Sign web portal offers different approaches to access the corpus data:

- *By Language & Informants:* For every sign language the available recording sessions (i.e. pairs of informants) are listed. Via the sessions all tasks performed by the respective informants as well as informant metadata information can be accessed.
- *By Task & Languages:* For each task that the informants were asked to perform a short description as well as the elicitation material is provided. Grouped by languages, all data-by-task items for a certain task are listed and can be accessed. In addition, the content tags defined for each task are presented.
- *By Task & Topics:* This approach makes use of the topics identified as part of the annotation process. By listing all content tags of the individual tasks it allows access to comparable data across individual informants and languages.



By Task & Topics

[contact]

Task 1: Public Transportation

Topics...

Task 2: Travel Agency

Topics...

Advertisement: Athens...

Advertisement: Berlin...

Advertisement: London...

[Pair 2] 11:39:02:17-11:41:59:01

[Pair 1] 00:00:01:10-00:01:32:20

[Pair 2] 00:00:04:04-00:01:31:14

[Pair 4] 00:00:00:00-00:01:21:05

[Pair 5] 00:00:00:00-00:00:08:24

Figure 7: Content tags in the web portal

3.1.2 Access Levels

There are three accessibility levels to the corpus data:

- *Publicly available on the project site:* Metadata and "appetisers" to the video and transcript data. Information on elicitation materials is at the same level of detail as previously published (Matthes et al. 2010).
- *Restricted access for researchers:* Elicitation materials available for researchers' own purposes as well as to fully understand the data collection. Video and transcript data as available in a standard format (H.264 for video, ELAN and iLex export format for transcripts), more detailed metadata.
- The third level is not available online, but requires arrangements between the researcher interested and the individual partner owning the data. This includes higher-resolution less compressed video and stereo data and even more refined metadata.

The corpus video and annotation data linked to the portal via the different access options are made available as videos with and without subtitles, in iLex export format as well as in ELAN format. The elicitation material including task explanations in the respective sign language is provided as Keynote or PowerPoint documents.

Additionally the web portal includes contact forms for researchers who request higher-resolution and stereo data from individual partners or ask to contact informants to be given access to more detailed metadata or to suggest additional data collection.

3.2 Finding the most Parallel Content Tags

The content tagging, as described in chapter 2.2.2, facilitates a rough comparison of the corpus data on the semantic level. Via the "Task & Topics" approach of the portal access to individual topics is provided and allows for direct comparison across languages and individual informants. The problem remains how, for a given topic tag, to find the closest match in another language, from the set of identically tagged stretches of signing offered by the portal.

Here we report on the experiments undertaken to gain a better understanding of what can be done for sign language corpora. For written language texts, a variety of similarity measures have been suggested in the literature, often relying on probabilistic models. As the needed statistical data are not yet available for sign language lexicons, we started with a very simple measure, namely lexical overlap count relative to sample size within one language (DGS) in the "At the airport" task. Not surprisingly, this measure highly depends on lexical variation. In fact, it becomes useless if signers with different sign dialects are involved. However, computing overlap in the semantic domain (concept entries assigned as meanings to the types) and thereby eliminating the influence of lexical variation provided results coming close to the annotators' intuition.

In order to apply this approach to content tags from different languages, a common semantic basis such as compatible SignNets in the sense of WordNets would be needed. Dicta-Sign has provided a list of 1000 concepts and signs in each of the four project languages for each of these. In many cases, WordNet sense keys could be assigned to the concepts whereas in other cases the sign languages require a granularity not provided by a WordNet for English.

Now we used the same measure as before, but only the instances of types with meanings in the 1000 concepts list could be taken into account. In our test case – DGS-BSL – this meant a reduction of the counts to one third, to sample sizes of 10-80 concepts. Overlap measures were no longer comparable, but seemed to provide tendencies nevertheless.

In order to provide more reliable measures, larger cross-language resources would be needed, ideally a "EuroSignNet".

4. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135.

5. References

Crasborn, O., Hanke, T. (2003). Metadata for sign language corpora. Online available at http://www.let.ru.nl/sign-lang/echo/docs/ECHO_Metadata_SL.pdf (last seen 30/03/2012).

Hanke, T. (2004). HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In: Streiter, O., Vettori, C. (Eds). LREC 2004, Workshop proceedings: Representation and processing of sign languages. Paris: ELRA, pp. 1-6.

Hanke, T., Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In: Crasborn, O., Hanke, T., Efthimiou, E., Zwitserlood, I., Thoutenhoofd, E. (Eds.): Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages. Paris: ELRA, pp. 64-67.

Hanke, T., König, L., Wagner, S., Matthes, S. (2010a). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In: Dreuw, P. et al. (Eds.): LREC 2010. 7th International Conference on Language Resources and Evaluation. Workshop Proceedings. W13. 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. May 22/23 2010. Valetta – Malta. Paris: ELRA, pp. 106-109.

Hanke, T., Storz, J., Wagner, S. (2010b). iLex: Handling Multi-Camera Recordings. In: Dreuw, P. et al. (Eds.): LREC 2010. 7th International Conference on Language Resources and Evaluation. Workshop Proceedings. W13. 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. May 22/23 2010. Valetta – Malta. Paris: ELRA, pp. 110-111.

Johnston, T. (2011). Auslan Corpus Annotation Guidelines. Online available at http://www.auslan.org.au/video/upload/attachments/AuslanCorpusAnnotationGuidelines30November2011.pdf (last seen 30/03/2012).

Matthes, S., Hanke, T., Storz, J., Efthimiou, E., Dimou, N., Karioris, P., Braffort, A., Choisier, A., Pelhate, J., Safar, E. (2010). Elicitation Tasks and Materials designed for Dicta-Sign's Multi-lingual Corpus. In: Dreuw, P. et al. (Eds.): LREC 2010. 7th International Conference on Language Resources and Evaluation. Workshop Proceedings. W13. 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. May 22/23 2010. Valetta – Malta. Paris: ELRA, pp. 158-163.