3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora

June 1st, 2008

Onno Crasborn, Radboud University Nijmegen, The Netherlands

Eleni Efthimiou, Institute for Language and Speech Processing, Greece

Thomas Hanke, University of Hamburg, Germany

Ernst D. Thoutenhoofd, Virtual Knowledge Studio for the Humanities & Social Sciences, The Netherlands

Inge Zwitserlood, Radboud University Nijmegen, The Netherlands

ABSTRACTS

Workshop Programme

3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora

08:45 – 09:00	**Workshop opening & welcome**
09:00 – 09:30	Diane Lillo-Martin, Deborah Chen Pichler: Development of sign language acquisition corpora
09:30 – 10:00	Onno Crasborn, Inge Zwitserlood: The Corpus NGT: an online corpus for professionals and laymen
10:00 – 10:30	Trevor Johnston: Corpus linguistics & signed languages: no lemmata, no corpus.
10:30 – 11:00	**Coffee break**
11:00 – 11:30	Lorraine Leeson, Brian Nolan: Digital Deployment of the Signs of Ireland Corpus in Elearning
11:30 – 12:00	Johanna Mesch, Lars Wallin: Use of sign language materials in teaching
12:00 – 13:30	**Poster session 1**
13:30 – 14:30	**Lunch**
14:30 – 16:00	**Poster session 2**
16:00 – 16:30	**Coffee break**
16:30 – 17:00	Onno Crasborn: Open Access to Sign Language Corpora
17:00 – 17:30	Adam Schembri: British Sign Language Corpus Project: Open Access Archives and the Observer's Paradox
17:30 – 18:00	Cat Fung H-M, Scholastica Lam, Felix Sze, Gladys Tang: Simultaneity vs. Sequentiality: Developing a transcription system of Hong Kong Sign Language acquisition data
18:00 – 18:45	**General discussion**
18:45 – 19:00	**Workshop closing**

Foreword

This workshop is the third in a series on "the representation and processing of sign languages". The first took place in 2004 (Lisbon, Portugal), the second in 2006 (Genova, Italy). All workshops have been tied to Language Resources and Evaluation Conferences (LREC), the present one taking place in Marrakech, Morocco. While there has been occasional attention for signed languages in the main LREC conference, the main focus there is on written and spoken forms of spoken languages. The wide field of language technology has been the focus of the LREC conferences, where academic and commercial research and applications meet. It will be clear to every researcher that there is a wide gap between our knowledge of spoken versus signed languages. This holds not only for language technology, where difference in modality and the absence of commonly used writing systems for signed languages obviously pose new challenges, but also for the linguistic knowledge that can be used in language technologies.

The domains addressed in the two previous sign language workshops have thus been fairly wide, and we see the same variety in the present proceedings volume. However, where the first and the second workshop had a strong focus on sign synthesis and automatic recognition, the theme of this third workshop concerns construction and exploitation of sign language corpora.

Recent technological developments allow sign language researchers to create relatively large video corpora of sign language use that were unimaginable ten years ago. Several national projects are currently underway, and more are planned. In the present volume, sign language linguistics researchers and researchers from the area of sign language technologies share their experiences from completed and ongoing efforts: what are the technical problems that were encountered and the solutions created, what are the linguistic decisions that were taken?

At the same time, the contributions also look into the future. How can we establish standards for linguistic tagging and metadata, and how can we add sign language specifics to well-established or emerging best practices from the general language resource community? How can we work towards (semi-) automatic annotation by computer recognition from video? These are all questions of interest to both linguists and language technology experts: the sign language corpora that are being created are needed for more reliable linguistic analyses, for studies on sociolinguistic variation, and for building tools that can recognize sign language use from video or generate animations of sign language use.

We would like to thank the programme committee that helped us reviewing the abstracts for the workshop:

Penny Boyes Braem; Annelies Braffort; Patrice Dalle; Evita Fotinea; Jens Heßmann; Trevor Johnston; Lorraine Leeson; Adam Schembri; Graham Turner; Meike Vaupel; Chiara Vettori

We would like to point workshop participants to the proceedings of the previous two workshops, which form important resources in a growing field of research; both works were made available as PDF files for participants of the workshop.

O. Streiter & C. Vettori (2004, Eds.) *From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*. [Proceedings of the Workshop on the Representation and Processing of Sign Languages. 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisbon.] Paris: ELRA.

C. Vettori (2006, Ed.) Lexicographic Matters and Didactic Scenarios. [Proceedings of the 2nd Workshop on the Representation and Processing of Sign Languages. 5th International Conference on Language Resources and Evaluation, LREC 2006, Genova.] Paris: ELRA.

The organisers,

Onno Crasborn, Radboud University Nijmegen (NL)

Eleni Efthimiou, Institute for Language and Speech Processing (GR)

Thomas Hanke, University of Hamburg (DE)

Ernst Thoutenhoofd, KNAW Virtual Knowledge Studio (NL)

Inge Zwitserlood, Radboud University Nijmegen (NL)

Linguistic, sociological and technical difficulties in the development of a Spanish Sign Language (LSE) corpus

Patricia Álvarez Sánchez, Inmaculada C. Báez Montero, Ana Mª Fernández Soneira
Universidad de Vigo, Spain

The creation of a Spanish Sign Language corpus has been, since 1995 until 2000, one of the main aims of our Sign Languages Research Group at the University of Vigo. As a result of this attempt, these are some of our publications:

Báez Montero, I. C. & M. C. Cabeza Pereiro (1995): "Diseño de un corpus de lengua de señas española – Design of a LSE corpus", XXV Simposium de la Sociedad Española de Lingüística (Zaragoza, 11-14 de diciembre de 1995).
Báez Montero, I. C. & M. C. Cabeza Pereiro (1999): "Spanish Sign Language Project at the University of Vigo" (póster), Gesture Workshop 1999 (Gif-sur-Yvette, Francia, 17-19 de marzo de 1999).
Báez Montero, I. C. & M. C. Cabeza Pereiro (1999): "Elaboración del corpus de lengua de signos española de la Universidad de Vigo – Development of the Spanish Sign Language corpus of the University of Vigo", Taller de Lingüística y Psicolingüística de las lenguas de signos (A Coruña, 20-21 de septiembre de 1999).

At this stage, with renewed energy, we have taken up again our initial aims, crossing the technical, linguistic and sociological obstacles that had hindered our proposal to reach its end.

In our communication we will present, apart from the difficulties that we have encountered, the new proposals for solving and overcoming them, thus, finally reaching our initial aim: to develop a public Spanish Sign Language corpus that can be consulted online.

We will go into details with the criteria of versatility and representativity which condition the technical aspects. Technological advances have made possible to adapt the size of the corpus and the criteria for labelling to the interests of the final users.

The labels for marking the corpus have demanded the revision of the linguistic criteria and the grammatical bases used for describing the language samples that compose the corpus.

The revision of the sociolinguistic criteria has been caused by the selection of both, the type of discourse (interviews, dialogues, oral speech,…) and the informants chosen for a wider and better representativity in the corpus.

Finally, we will advance the utilities that we pretend to give the corpus, not only centered in the use of linguistic data for the quantitative and qualitative research of the LSE, but also centered in the use for teaching. The creation of teleteaching platforms allows us to offer the pupil real language samples which complete the process of learning started inside the classroom.

Pointing and verb modification: the expression of semantic roles in the Auslan corpus

Louise de Beuzeville
Macquarie University, Australia

As part of a larger project investigating the grammatical use of space in Auslan, 62 texts from the Auslan Corpus were annotated and analysed for the spatial modification of verbs to show semantic roles. Data were taken from two groups of native and near-native Auslan signers. Spontaneous narratives were sourced from a sociolinguistic variation corpus collected from 211 participants all over Australia. The second set of texts was elicited from 100 adult native signers of Auslan from the Auslan Corpus Project. Participants retold to a deaf interlocutor a prepared Aesop's fable and a spontaneous personal recount of a memorable event, as well as answering a series of questions on their attitudes to various factors influencing the deaf community (such as genetic testing and cochlear implants). The texts from both corpora were recorded on digital videotape and then annotated using ELAN software. Here we report on 62 texts that have been annotated (approximately 9,000 signs from 50 narrative texts and 9,000 from 10 attitude surveys). Each sign or meaningful gesture was identified, with points being categorised as pronouns or other. These signs were then classified into word class and the nouns and verbs tagged for whether they could be modified spatially. Next, the indicating nouns and verbs were annotated as to whether or not their spatial modification was realised. In this paper, we discuss the use of the ELAN search functions across multiple files in order to identify the proportion of sign types in the texts and the frequency with which indicating verbs are actually modified for space. We then searched all files again to identify all instances where pointing signs occurred directly before or after an indicating verb, in order to calculate whether the collocation of a point (pronoun, other or either) and a non-modified indicating verb was statistically significant. Despite the claim that indicating verbs in signed languages are obligatorily modified ('inflected') with respect to loci in the signing space in order to show person 'agreement', we found that these verbs are actually only spatially modified about on third of the time (Johnston et al and de B et al., forthcoming) and this study showed that to be partly as a result of presence of points. The results help determine where and when the spatial modification of indicating verbs is used in

natural Auslan texts (and potentially other signed languages). Based on this data, we suggest that 1) the degree of grammaticalization of indicating verbs may not be as great as once thought and 2) the apparent non-obligatory or variable use of spatial modifications may be partly accounted for by the presence of pointing signs—very frequent in signed texts—before or directly after the verb.

References

Johnston, T. A., de Beuzeville, L., Schembri, A., & Goswell, D. (2007) On not missing the point: indicating verbs in Auslan. Paper presented at the 10th International Cognitive Linguistics Conference, Krakow, Poland, July 15th – 20th, 2007

de Beuzeville, L., Johnston, T. A., Schembri, A., & Goswell, D. (forthcoming) The use of space with lexical verbs in Auslan.

Establishment of a corpus of Hong Kong Sign Language acquisition data: from ELAN to CLAN

Cat Fung H-M, Scholastica Lam, Joe Mak, Gladys Tang

The Chinese University of Hong Kong, China

This paper introduces the Hong Kong Sign Language Child Language Corpus currently developed by the Centre for Sign Linguistics and Deaf Studies, the Chinese University of Hong Kong. When completed, the corpus will include both longitudinal and cross-sectional data of deaf children acquiring Hong Kong Sign Language. Our research team has decided to establish a meaning-based transcription system compatible with both the ELAN and CLAN programs in order to facilitate future linguistic analysis. The ELAN program, which allows multiple-tier data entries and synchronization of video data with glosses, is an ideal tool for transcribing and viewing sign language data. The CLAN program, on the other hand, has a wide range of well-developed functions such as auto-tagging and the 'kwal' function for data search and they are extremely useful for conducting quantitative analyses. With add-on programs developed by our research team and additional functions in CLAN developed by the CHILDES research team, the transcribed data are transferable from the ELAN format to CLAN format, thus allowing researchers to optimize the use of both programs in conducting different types of linguistic analysis on the acquisition data.

Simultaneity vs Sequentiality: Developing a transcription system of Hong Kong Sign Language acquisition data

Cat Fung H-M, Felix Sze, Scholastica Lam, Gladys Tang

The Chinese University of Hong Kong, China

It is a well-known fact that sign languages are characterized with a wide range of simultaneous constructions, e.g. complex polymorphemic constructions, maintenance of list buoys in space while another hand continues signing, overlaying of various types of non-manuals with manual signing, etc. In transcribing these simultaneous constructions, decisions have to be made as to whether they should be given a single gloss or be glossed separately in two different tiers. This presentation discusses the transcription system of Hong Kong Sign Language acquisition data, with particular focus on how simultaneous constructions are analyzed and glossed, and the difficulties we encountered in the transcription process.

We are currently developing a Hong Kong Sign Language acquisition Corpus (Tang et al.) with transcriptions done with ELAN. One major advantage of ELAN is that it allows us to represent different pieces of linguistic information simultaneously on separate tiers. However, it is not always easy to decide whether two different signs produced by two hands should be glossed as a single sign or be teased apart and glossed separately on two different tiers. For example, in a typical classifier predicate such as 'a cup on a table' in example one below, the signs can either be glossed as a single entry 'CL-cup-on-table', or marked separately by 'CL-cup' and 'CL-flat surface' on two different tiers:

Example (1): 'a cup on a table'

Left hand: CL-cup

Right hand: CL-flat-surface

The advantage of having a single gloss is that it reflects the native intuition that the two classifiers form a single syntactic unit. Yet it fails to reflect the morphological complexity of the construction, leading to a potential underestimation of the morphological development of the deaf child.

On the other hand, having two separate glosses can clearly show that two classifiers are involved in the construction, reflecting its morphological complexities to some extent. From a theoretical point of view, however, once this method is adopted, the glosses are being used as 'analyzable units' to represent separate handshape morphemes. A question that arises logically is, why do we want to represent handshape morphemes separately in the transcription, but not morphemes of other phonological parameters, such as movements and locations?

Another equally thorny issue is how to gloss classifiers or signs (i.e. list buoy) that are held in space. In example (2), the signer expresses two propositions: 'A man stands here' and 'a woman shot him with a gun':

Example (2):

Left hand: MAN CL-stand ---------->

Right hand: FEMALE SHOOT-WITH-A-GUN

In terms of articulation, the classifier for 'MAN' is held in space while the second clause is signed. Syntactically, the classifier for MAN becomes the internal argument of the transitive verb SHOOT-WITH-A-GUN in the second clause. In the literature, if a sign is held in space, a broken line is usually used to represent the duration of which the sign is held. If the same method is used in the transcription, however, the fact that the classifier is the internal argument of the second clause cannot be captured. This may potentially lead to an under-estimation of the deaf child's syntactic complexity, if statistics are based on figures generated by the search functions of ELAN. In this presentation, an attempt will be made to provide solutions to the above issues.

Annotation of Non Manual Gestures: Eyebrow movement description

Emilie Chételat-Pelé, Annelies Braffort, Jean Véronis

LIMSI-CNRS, Orsay, France

This paper deals with non manual gestures annotation involved in Sign Language (SL) within the context of automatic generation of SL. Movements of the eyes, eyebrows, mouth, cheeks and head involved in SL are defined as non manual gestures or NMG. Many researches in SL emphasize the importance of NMG at different language levels and recognize that NMG are essential for the message comprehension. However these researches canít explain and define enough the way that NMG operate. A specific NMG study should allow us to know when and how NMG are involved in meaning transmission and information comprehension, in order to design a formal description usable by automatic generation system. Our purpose is to have an objective and precise description of all NMG involved in French Sign Language (LSF). At present, non manual descriptions do not allow us to deal with and to observe the movement intensity and dynamics. Therefore, we propose a new annotation methodology of NMG.

We position several 2D points on each frame of the video and export their coordinates x,y. These coordinates are used to obtain precise position of all NMGs frame by frame. Then, we use these data to evaluate the annotation by means of a synthetic face, for numerical analysis (by using curve), and, finally, to obtain numerical definition of each symbol of our set of annotation symbols based on arrows. A first annotation on the LS-COLIN corpus showed that this methodology is an answer to correctly address our purpose: All NMG can be described, with precision. Moreover, the movement dynamics can be analyzed, and each movement phase. All these results must be refined and confirmed by extending the study on the whole corpus. In a second step, our annotation will be used to produce analyses in order to define rules and a formal definition of NGM that will be evaluated in LIMSIís automatic LSF generation system.

Open access to sign language corpora

Onno Crasborn
Radboud University Nijmegen, The Netherlands

One of the ongoing developments on internet is the increasing attention for open content: data of all kinds, whether text, images or video, are made publicly available. While there may be restrictions on the type of use that s allowed, selling content and strictly protecting it under copyright laws appears not desirable necessary for some types of content. This development is sometimes characterised as a change from copyright to 'copyleft': rather than stating that "all rights are prohibited", people are encouraged to use materials for their own benefit. This presentations sketches this development and explores how it can apply to sign language corpora. As a case study, the Corpus NGT project is characterised, which publishes a large systematic collection of sign language data online. A total of 100 signers is being recorded, leading to over 75 hours of material in 2,000 video segments. The wish to publish this material not only for research purposes (cf. the Dutch Science Foundation's funding) stems from its large possible value for various parties in the Netherlands: deaf signers themselves, second language learners of sign language, interpreting students, etc.

One of the problems in publishing sign language data online is privacy protection. As sign language movies inevitable contain visual information about the identity of the signer, together with the actual content of the language production signers reveal more of themselves than uni-modal speech or text corpora. In the Corpus NGT, we try to protect the privacy of the informants in several ways: we urge people to not reveal too much personal information about themselves or about others in their stories and discussions, we limit the amount of metadata that we publish online (leaving out many of the standard fields from the IMDI metadata standard), and nowhere mention or refer to the name of the signers.

The way we aim to protect the use of the material is by publishing all materials under a Creative Commons license. Creative Commons is an international organisation that was set up especially as a bridge between national copyright laws and open content material on internet. Of the different types of licenses that are available, we chose to apply the 'BY-NC-SA' license. This license states that people may re-use the material provided they refer to the authors, that no commercial use be made, and that (modifications of) the material are distributed under the same conditions. The Creative Commons licenses are attractive because they are made available in various forms: a plain language statement (as in the previous sentence), a formal legal text, and a machine-readable version for use by software. The plain language version is attached to every movie in the Corpus NGT by a short text preceding and following every movie file, thus allowing relatively easy replacement should future changes in policy require so.

Finally, a few ethical questions are raised in relation to publishing sign language materials as open access data: although the permission for open access publication is requested of the signers in the corpus, to what extent can they foresee the consequences at that point in time? Will future technologies allow easy face recognition on the basis of movies and obliterate the privacy protection measures that have been taken? What will the (normative) effect of publishing signing of a group of 100 signers from a small community be? There is a clear risk in the publication of sign language data without an answer to these questions. The solution taken in the Corpus NGT project is to invest substantial time and energy in publicity within the deaf community, to explain the goal and nature of the corpus and to encourage use by deaf people.

Enhanced ELAN functionality for sign language corpora

Onno Crasborn[1]*, Han Sloetjes*[2]
[1]Radboud University Nijmegen, The Netherlands; [2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

The annotation tool ELAN was enhanced within the *Corpus NGT* project by a number of new and improved functions. Most of these functions were not specific for working with sign language video data, and can readily be used for other annotation purposes as well. Their direct utility for working with large amounts of annotation files during the development and use of the Corpus NGT project is what unites the various functions. The following functions appeared in a series of releases between versions 2.6 and 3.4:

- The 'duplicate annotation' function was created to facilitate the glossing of two-handed signs in cases where there are separate tiers for the left and the right hand: copying an annotation to another tier saves annotators quite some time, and prevents misspellings.
- A 'multiple file search' was implemented: structured searches combining search criteria on different tiers can be carried out in a subset of files that can be created by the user.
- The segmentation function was further developed so that annotations with a fixed, user definable duration can be created by a single key stroke while the media files are playing. The key stroke can either mark the beginning of an annotation or the end.

- A function has been added to flexibly generate annotation content based on a user definable prefix and an index number.
- A panel can be displayed that lists basic statistics for all tiers in an annotation document: the number of annotations, the minimum, maximum, average, median and total annotation duration per tier. This helps the user getting a better grip on the content in an annotation document and can be helpful in data analysis.
- The annotation density viewer can now also be set to only show the distribution of annotations of a single, selectable tier. The label of a tier in the timeline viewer can optionally show the current number of annotations on that tier.
- The property 'annotator' has been added in the specification of tiers, allowing groups of researchers to separate which tier has been filled by whom.
- Export a list of unique annotation values or a list of unique words from multiple annotation documents.
- Easy, interactive hiding and showing of any of the associated video files, without having to remove the media file association altogether.

In addition, a large number of user interface improvements have been implemented, including the following:

- Improved, more intuitive layout of the main menu bar
- Additional keyboard shortcuts; the list of shortcuts can be printed
- A recent files list has been added
- Easy keyboard navigation through the opened documents/windows
- A subtle change in the background of the timeline viewer, facilitating the perception of the distinction between the different tiers
- With the use of a new preferences system in version 3, users can now set the colour of tier labels in the timeline viewer, allowing the visual grouping of related tiers in documents containing many tiers.

Although enhanced search functionalities and templates facilitate working with multiple ELAN documents, it is not yet possible to 'manage' a set of ELAN files systematically in any way. Perl scripts were developed in order to add tiers and linguistic types to a set of documents, to change annotation values in multiple documents, and to generate ELAN and preferences files on the basis of a set of media files and existent annotation and preferences files.

Future collaboration between the ELAN developers at the Max Planck Institute for Psycholinguistics and the sign language researchers at Radboud University will be targeted at enhancing search facilities and facilitating team work between researchers using large language corpora containing ELAN documents.

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

Onno Crasborn, Inge Zwitserlood
Radboud University Nijmegen, The Netherlands

The *Corpus NGT* is an ambitious effort to record and archive video data from Sign Language of the Netherlands (NGT), guaranteeing online access and long-term availability. In this presentation, we share our experiences in building this corpus, viz. preparing for comparable data, both elicited and (semi)spontaneous, the recording set-up and procedure, processing of the data, annotation, metadata, licenses and publishing.

Initially aiming to record 24 native signers using two variants of NGT, and providing annotations of a large amount of the data, the plan changed into recording many more signers (100) using all five reported variants of NGT. This much larger collection of data ensures a good sample of the current state of the language, and, since participants are from various ages, we can also include its older stages (facilitating the study of language change). The consequence is that there is less time for making annotations. However, it will be easier to add annotations later than to make new recordings that are comparable in every respect to the initial recordings.

The project strives towards a completely open access policy: not only the video data and annotations will be available to everyone, but also the workflows and manuals for tools that have been used. Use and reuse of the data are protected by Creative Commons licenses. For now, the corpus will be published by the Max Planck Institute for Psycholinguistics, as part of their growing set of language corpora. We follow their IMDI standard for creating metadata descriptions and corpus structuring. The extension of their annotation tool ELAN as well as the integration of ELAN and IMDI (the data and metadata domains) formed a substantial part of the project.

The Corpus NGT project is funded by the Dutch Science Foundation to facilitate linguistic research. However, since there is a dire need for NGT data among several groups of people, we now are happy to include *everyone* in our target audience. 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 are interested, including teachers

of NGT, developers of teaching materials, and 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.

Participants were recorded in pairs. They performed several language tasks (producing narratives, prompted discussions, but also non-elicited signing), resulting in ±1.5 hours of useable signed data per pair. Both upper body and a top view were recorded of each signer. In combination, these recordings approximate a three-dimensional view of the signing. For extra information of the facial expressions, MPEG-1 movies showing only the face are extracted from the recordings of the body (shot in full HD resolution).

Due to time and budget limitations, it was only possible to make crude gloss annotations in ELAN of a small subset of the data. In order to make as much of the data set accessible to a large audience, a voice-over done by interpreters is provided with most of the data.

Towards Automatic Sign Language Annotation for the ELAN Tool

Philippe Dreuw and Hermann Ney
Aachen University, Germany

A new interface to the ELAN annotation software that can handle automatically generated annotations by a sign language recognition and translation framework is described. For evaluation and benchmarking of automatic sign language recognition, large corpora with rich annotation are needed. Such databases have generally only small vocabularies and are created for linguistic purposes, because the annotation process of sign language videos is time consuming and requires expert knowledge of bilingual speakers (signers). The proposed framework provides easy access to the output of an automatic sign language recognition and translation framework. Furthermore, new annotations and metadata information can be added and imported into the ELAN annotation software. Preliminary results show that the performance of a statistical machine translation improves using automatically generated annotations.

Automatic sign language recognition is a problem that is being solved by many research institutes in the world. Up to now there is a deficiency of corpora with good properties such as high resolution and frame rate, several views of the scene, detailed annotation etc. In this paper we take a closer look at the annotation of available data.

Annotating Real-Space Depiction

Paul Dudis, Kristin Mulrooney, Clifton Langdon, Cecily Whitworth
Gallaudet University, USA

“Shifted referential space” (SRS) and “fixed referential space” (FRS) (Morgan 2005) are two major types of referential space known to signed language researchers (see Perniss 2007 for a discussion of alternative labels used in the literature). An example of SRS has thesigner’s body representing an event participant. An example of FRS involves the use of “classifier predicates” to demonstrate spatial relationships of entities within a situation being described. A number of challenges in signed language text transcriptions identified in Morgan (2005) pertains to the use of SRS and FRS. As suggested in this poster presentation, a step towards resolving some of these challenges involves greater explicitness in the descriptionof the conceptual make-up of SRS and FRS. Such explicitness is possible when more than just the signer’s body, hands, and space are considered in the analysis. Dudis (2007) identifies the following as components within Real-Space (Liddell 1995) that are used to depict events, settings and objects: the setting/empty physical space, the signer’s vantage point, the subject of conception (or, the self), temporal progression, and the body and its partitionable zones. We considered these components in a project designed to assist videocoders to identify and annotate types of depiction in signed language texts. Our preliminary finding is that if we also consider the conceptual compression of space—which results in a diagrammatic space (Emmorey and Falgier 1999)—there are approximately fourteen types of depiction, excluding the more abstract ones, e.g. tokens (Liddell 1995).

Included in this poster presentation is a prototype of a flowchart to be used by video coders as part of depiction identification procedures. This flowchart is intended to reduce the effort of identifying depictions by creating binary (yes or no) decisions for each step of the flowchart. The research team is currently using ELAN (EUDICO Linguistic Annotator, www.lat-mpi.eu/tools/elan/) to code the depictions focusing on the relationship of genre and depiction type by looking at the depictions’ length, frequency, and place of occurrence in 4 different genres: narrative of personal experience, academic, poetry, conversation. We also have been mindful that a good transcription system should be

accessible in an electronic form and be searchable (Morgan 2005). In tiered transcription systems like ELAN the depiction annotation can simply be a tier of its own when it is not the emphasis of the research, or it can occupy several tiers when it is the forefront. In linear ASCII- style transcriptions the annotation can mark the type and beginning then end of the depiction. Our poster does not bring a complete bank of suggested annotation symbols, but rather the idea that greater explicitness as to the type of depiction in question may be beneficial to corpus work.

Annotation and Management of the Greek Sign Language Corpus (GSLC)

Eleni Efthimiou, Stavroula-Evita Fotinea

Institute for Language and Speech Processing (ILSP) / R.C. Athena, Athens, Greece

This paper presents the design and development of a representative language corpus for the Greek Sign Language (GSL). Focus is put on the annotation methodology adopted to provide for linguistic information and annotated corpus exploitation for the extraction of a linguistic model intended to support HCI applications based on sign recognition.

The existence of an annotated corpus is a prerequisite for the creation of linguistic resources and for the development of NLP applications for any natural language articulated either orally or through signing. In the case of a sign language corpus, annotation performed on video sequences, is intended to support exploitation of linguistic information conveyed through various combinations of spatial-temporal parameters around the signer's body.

The Greek Sign Language Corpus (GSLC) is been developed in the framework of the national project DIANOEMA (GSRT, M3.3, id 35) that aims at optical analysis and recognition of both static and dynamic signs, incorporating a GSL linguistic model in controlling robot motion. Since no previous GSL corpus is available to meet the requirements of multipurpose use in an HCI environment, the design of GSLC has taken into account annotation requirements as well as linguistic adequacy controls to ensure both corpus-based linguistic analysis and corpus re-usability. Linguistic analysis is a sufficient component for the development of NLP tools that, in the case of signed languages, support deaf accessibility to IT content and services. To effectively support this kind of language intensive operations, linguistic analysis has to derive from safe language data and also provide for an amount of linguistic phenomena, which allow for an adequate description of the language structure. In this context, safe data are defined as data commonly accepted by a specific language community. The design of GSLC content has made a distinction between three parts on the basis of the articulation units to be considered in respect to both linguistic analysis and the sign recognition process.

The first part comprises a list of lemmata which are representative of the use of handshapes as a primary sign formation component. This part of the corpus is developed on the basis of measurements of handshape frequency of use in sign morpheme formation but it has also taken into account the complete set of sign formation parameters. In this sense, in order to provide data for all sign articulation features of GSL, the corpus also includes characteristic lemmata with respect to all manual and non-manual features of the language. The second part of GSLC is composed of sets of controlled utterances, which form paradigms capable to expose the mechanisms GSL uses to expresses specific core grammar phenomena. The grammar coverage that corresponds to this part of the corpus is representative enough to allow for a formal description of the main structural-semantic mechanisms of the language. Finally, the third part of GSLC contains free narration sequences, which are intended to provide data of spontaneous language production and be used for machine learning purposes as regards sign recognition. With respect to data collection, all parts of the corpus have been performed by native signers under controlled conditions that guarantee absence of language interference from the part of the spoken language of the signers' environment. Finally, quality control mechanisms have been applied to ensure data integrity.

In the framework of the current research target, annotation on the GSLC involves, on the one hand, descriptions of the phonological structure of morphemes and, on the other hand, sentence level markers. Sign phonology involves manual and non-manual features of sign formation. For the description of the phonological composition of sign morphemes the HamNoSys coding set is being used along with GSL specific feature coding. Sentence level annotation, except for sentence boundaries, involves phrase boundary marking and grammar information marking related to multi-layer indicators, as is the case of e.g. topicalisation, nominal phrase formation, temporal indicators and sentential negation. Sentence level annotation makes use of the ELAN annotator. Annotation integrity is subject to quality controls that involve both peer and external review by expert annotators.

The GSLC current implementation has foreseen extensibility on all content levels as well as on annotation features, thus, allowing for corpus re-usability in GSL research and HCI applications beyond the scope of a specific research project.

Indicative bibliography

Bowden, R., Windridge, D., Kadir, T., Zisserman, A. & Brady, M. (2004). «A Linguistic Feature Vector for the Visual Interpretation of Sign Language», In Tomas Pajdla, Jiri Matas (Eds), *Proc. 8th European Conference on Computer Vision, ECCV04*. LNCS3022, Springer-Verlag, Volume 1, pp391- 401.

Bellugi, U. & Fischer, S. (1972). «A comparison of Sign language and spoken language: rate and grammatical mechanisms», *Cognition: International Journal of Cognitive Psychology*, 1, 173-200.

Efthimiou, E., Sapountzaki, G., Karpouzis, C. & Fotinea, S-E. (2004). «Developing an e-Learning platform for the Greek Sign Language». *Lecture Notes in Computer Science* 3118: 1107-1113. Springer.
Efthimiou, E., Fotinea, S-E. & Sapountzaki, G. (2006). «Processing linguistic data for GSL structure representation», *Proc. of the Workshop on the Representation and Processing of Sign Languages: Lexicographic matters and didactic scenarios*, Satellite Workshop to LREC-2006 Conference, May 28, pp.49-54.
ELAN annotator, Max Planck Institute for Psycholinguistics, available at: http://www.mpi.nl/tools/elan.html
Fotinea, S-E., Efthimiou, E., Karpouzis, K. & Caridakis, G. (2005). "Dynamic GSL synthesis to support access to e-content", *Proc. of the 3rd International Conference on Universal Access in Human-Computer Interaction (UAHCI 2005),* 22-27 July 2005, Las Vegas, Nevada, USA.
HamNoSys Sign Language Notation System: www.sign-lang.uni-hamburg.de/projects/HamNoSys.html
Karpouzis, K. Caridakis, G., Fotinea, S-E. & Efthimiou, E. (2005). "Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture", *Computers and Education International Journal*, Elsevier, in print, electronically available since Sept 05.
Kraiss, K.-F. (Ed.), (2006). *Advanced Man-Machine Interaction - Fundamentals and Implementation*. Series: Signals and Communication Technology, Springer.
Stokoe, W. 1978. *Sign Language Structure* (revised ed.). Silver Spring, MD: Linstok.

iLex – A database tool integrating sign language corpus linguistics and sign language lexicography

Thomas Hanke, Jakob Storz
University of Hamburg, Germany

This poster presents iLex, a software tool targeted at both corpus linguistics and lexicography. It is now a shared belief in the LR community that lexicographic work on any language should be based on a corpus. Conversely, lemmatisation of a sign language corpus requires a lexicon to be built up in parallel.

For languages with a written form and orthography, lemmatisation is a more or less straight-forward process. For sign languages, however, type-token matching is a major task by itself. Glossing or form- based transcription, e.g. with HamNoSys, may be sufficient for small single-transcriber projects. Consistency, however, cannot be guaranteed over multiple transcribers, large quantities, or longer periods of time.

iLex is therefore designed as a relational database linking tokens with their types. That means that the transcription process does not consist of assigning text tags to time intervals of the source video, but of tagging intervals with a reference to a type. The database then allows the user to review all tokens of a type at any point of time in order to verify that the intended type-token pair really fits with the type's definition and extension. Revisions of earlier decisions in the light of new data are as easy as dragging instances from one type to the other. Beyond the support in the initial type-token matching, iLex gives its users views onto the transcribed data orthogonal to the transcription itself, and thereby helps to improve transcription quality. With its ability to support users working on different projects in one database, iLex allows synergies between projects as each project immediately profits from data entered by others. The cost for these benefits is the necessity of a solid infrastructure: A database server needs to be installed, and ideally every user should have access to all videos, often requiring specialised video servers. For larger corpus projects, however, this should be taken for granted anyway. For data exchange with other research groups, iLex supports a number of file formats, such as ELAN, SignStream, and syncWRITER for transcription data and IMDI for metadata. While exporting data from iLex into these formats as well as a couple of presentation formats such as HTML with thumbnails is done with a simple menu command, importing data from other sources requires some additional steps to be done by the researcher. As other data formats consist of text tags only, some matching operations are necessary to convert from text to tokens. The newest release of iLex supports the user in this procedure: By learning a mapping from imported glosses to iLex types from user actions, it can partially automate future imports from the same source. In addition to data exchange with other transcription tools and export to presentation formats, iLex integrates with a number of tools for rapid production of sign language teaching materials and for virtual signing by means of avatars.

On the lexicography side, iLex can host all the data necessary for the production of dictionaries. With its scripting language support, iLex is able to almost completely automate the production of a variety of formats including print, DVD, online websites for computers, and online websites for iPods/iPhones.

Sign language corpora and the problems with ELAN and the ECHO annotation conventions

Annika Herrmann

University of Frankfurt am Main, Germany

Large corpus projects require logistic, technical and personal expertise and most importantly a conventionalized annotation system. In addition, relatively small projects with a definite set of data can also be an invaluable contribution to linguistic sign language research and therefore should use the same technical methods and annotation conventions for comparative reasons. The poster will present the process of building a corpus that is needed for a cross-linguistic study currently undertaken and focuses on the problems that arise with regard to annotation. The respective solutions shall be suggestions towards a unified convention.

In this project, elicited data from three European sign languages and altogether 20 informants provide a set of approx. 900 sentences and short dialogues. Metadata information about participants and the recording situation will be edited in the IMDI metadata set. ELAN provides the most adequate annotation system for my purposes as the main interest of the study lies in the use of nonmanuals. The tool is widely used for sign language annotation and I try to guarantee for comparability by mainly adopting the ECHO annotation system with a few necessary adaptations.

Problems listed below include repeatedly asked questions that are still not defined clearly yet:

a) How are the on- and offsets of signs determined? Shall we annotate the separate signs or the signing stream integrating the transition period?

b) How should pointing signs or constructions with many meaning components be transcribed?

c) Despite more or less clear definitions of what each tier should be used for, the GLOSS-tier is sometimes intertwined with external information not fitting the tier. How can these problems be avoided?

d) What kinds of disadvantages occur, if the eye gaze and eye blink annotations are not accurate?

Possible Solutions:

a) Even though the on- and offsets of signs can be defined more precisely than for words, the sign syllable not always has clear boundaries. Signing should be annotated as a streaming process that is interrupted when there is a hold or a significant pause. The transition from one sign to the other is often clearly visible through handshape change, which seems to be the more adequate marker for annotation. (The only problem left being sign duration, which cannot entirely be solved by the vague separate sign annotation either.)

b) Proposal for a more detailed distinction of pointing signs without being theoretical (at least IX-1 for signer, IX-dual (excl., incl.) e.g.) and poly-meaning constructions (e.g. BE-LOCATED-CL:vehicle instead of (p-)vehicle-be-located; BLEAK instead of (p-)bleaking sheep when SHEEP is already introduced, decision between HOLD-CL:potato and HOLD-CL:round object).

c) The GLOSS tier should only be used for manual signs or gestures, nonmanuals should not be included (*WALK-PURPOSEFUL). An additional tier is useful: other NMFs/look/other facial expressions

d) Continuous eye gaze and eye aperture annotation is necessary to exactly determine eye gaze change with or without an eye blink and the duration and timing of blinks. This can especially be relevant for prosodic analysis.

Building up digital video resources for sign language interpreter training

Jens Heßmann[1] and Meike Vaupel[2],

[1] University of Applied Sciences Magdeburg-Stendal, Germany; [2] University of Applied Sciences Zwickau, Germany

Sign language interpreter training has been offered at the universities of applied sciences in Magdeburg and Zwickau since 1997 and 1998, respectively. Both training programs are set in the institutional context of East German universities that experienced a major reorganization after the unification of Germany. The training programs share an applied perspective in research and teaching as well as many of the features typical for small scale academic ventures in a developing field. Thus, provision of teaching materials and, more particularly, sign language video resources, adequate in content, format and technical quality, has been a constant concern. Of necessity, a hands-on approach had to be chosen for the last ten years, and both programs have amassed a diversity of analogue and digital video films. In most cases, the only way of accessing this material consists in picking the brains of those colleagues who may have worked with some video clip or exercise suitable for one's own didactic or research purposes.

As it happens, Magdeburg as well as Zwickau have installed the same type of digital training facilities ('video lab') in 2007. These video labs consist of individual workstations linked to a central video server that hosts all the resources in a

unified digital format. For both institutions, a major challenge consists in organizing a process that will transform and complement existing sign language materials so as to create an accessible library of video resources for research and training purposes. Our presentation will report on our joint efforts to do the first steps in this direction. The following aspects will be discussed:

- Legal and ethical issues: Up to now, questions of ownership and property rights have often been dealt with somewhat casually. Building up a digital library of video resources implies that such questions have been formally clarified. However, just what the conditions for using video materials gathered informally, passed on from one colleague to the next or published on the internet are, may be hard to decide.
- Administrative and technical prerequisites: In order to create a solid basis for the desired cooperation and be able to access university funds, the two universities concerned will enter into formal agreements on the mutual use of video resources. This in turn, demands that there are clearly defined ways of synchronizing, complementing and accessing the respective collections of resources.
- Criteria for annotating and archiving video resources: While the process of digitizing and storing existing video materials can be dealt with somewhat mechanically, the development of systematic ways of annotating and organising sign language materials is crucial in order to make digital resources accessible. Clearly, this is an area where progress has been made in recent years, e.g. in the context of the ECHO project ('European Cultural Heritage Online,' cf. http://www.let.ru.nl/sign-lang/echo/index.html). We will add to this discussion by considering the more specific demands of sign language interpreter training and research.

Semi-automatic Annotation of Sign Language Corpora

Marek Hrúz, Pavel Campr, Miloš Železný

University of West Bohemia, Czech Republic

The first step of automatic sign language recognition is feature extraction. It has been shown which features are sufficient for a successful classification of a sign. It is the hand shape, orientation of the hand in space, trajectory of the hands and the non-manual component of the speech (facial expression, articulation). Usually the efficiency of the feature extracting algorithm is evaluated by the rate of recognition of the whole system. This approach can be confusing since the researcher cannot be always sure which part of the system is failing. However if the corpora would be available with a detailed annotation of these features the evaluation would be more precise. A manual creation of the annotation data can be very time consuming. We propose a semi-automatic tool for annotating trajectory of head and hands and the shape of the hands.

For the purpose of extracting the trajectory of hands a tracker is developed. In our case the tracker is based on similarity of the scalar description of objects. We describe the objects by seven Hu moments of the contour, a gray scale image (template), position, velocity, perimeter of the contour, area of the bounding box and area of the contour. For every new frame all objects in the image are detected and filtered. Every tracker computes the similarity of the tracked object and the evaluated object. As long as the tracker's certainty is above a threshold it is considered as ground truth. At this point all available data are collected from the object and saved as annotation. If the level of uncertainty is high, the user is asked to verify the tracking.

If a perfect tracker was available all the annotation could be created automatically. But the trackers usually fail when an occlusion of objects occurs. Because of this problem the system must be able to detect occlusions of objects and have the user verify the resulting tracking. In our system we assume that the bounding box of an overlapped object becomes relatively bigger in the first frame of occlusion and relatively smaller in the first frame after occlusion. We consider the area of the bounding box as a feature which determines the occlusion.

Up to now the annotation through tracker allows us to semi-automatically obtain the trajectory of head and hands and the shape of the hands. In the future we will extend the system to be able to determine the orientation of hands and combine it with a lip-reading system which we have ready for use. The obtained parameters can be then used as ground truth data for evaluation of feature extracting algorithm.

Corpus linguistics and signed languages: no lemmata, no corpus

Trevor Johnston
Macquarie University, Sydney, Australia

A fundamental problem in the creation of signed language corpora is lemmatisation. Lemmatisation—the classification or identification of related word forms under a single label or lemma (the equivalent of headwords or headsigns in a dictionary)—is central to the process of corpus creation. The reason is that signed language corpora—as with all modern linguistic corpora—need to be machine-readable and this means that sign annotations should not only be informed by linguistic theory but also that tags appended to these annotations should be used consistently and systematically. In addition, a corpus must also be well documented (i.e., with accurate and relevant metadata) and representative of the language community (i.e., of relevant registers and sociolinguistic). All this requires dedicated technology (e.g., ELAN), standards and protocols (e.g., IMDI metadata descriptors), and transparent and agreed grammatical tags (e.g., grammatical class labels). However, it also requires the identification of lemmata and this presupposes the unique identification of sign forms. In other words, a successful corpus project presupposes the availability of a reference dictionary or lexical database to facilitate lemma identification and consistency in lemmatisation. Without lemmatisation a collection of recordings with various related appended annotation files will not be able to be used as a true linguistic corpus as the counting, sorting, tagging. etc. of types and tokens is rendered virtually impossible. This presentation draws on the Australian experience of corpus creation to show how a dictionary in the form of a computerized lexical database needs to be created and integrated into any signed language corpus project. Plans for the creation of new signed language corpora will be seriously flawed if they do not take this into account.

Interactive HamNoSys Notation Editor for Signed Speech Annotation

Jakub Kanis, Pavel Campr, Marek Hrúz, Zdeněk Krňoul, Miloš Železný
University of West Bohemia, Czech Republic

The goal of sign language synthesis is to create an avatar which uses sign languge as main communication form. In order to emulate human behaviour during signing the avatar has to express manual components (hand position, hand shape) and non-manual components (face expression, lip articulation) of the performed signs. The task of sign language synthesis is implemented in several steps. Since the sign language has different grammar than the spoken language, the source sentence has to be translated into corresponding sequence of isolated signs. Those signs are synthesized in sequence and create output sentence in sign language. Non-manual components are synthesized by already developed Czech talking head which is able to articulate words and sentences in Czech language. Face expressions can be manually set. The synthesis process of manual movements is based on HamNoSys 3.0 notation. This notation is used for deterministic and suitable processing of the sign speech. The methodology of the notation allows precise and also extensible expression of the sign description.

Firstly, our synthesis system automatically carries out the syntactic analysis of symbolic string (in HamNoSys notation) and generates a tree structure. The tree structure is suitable for conversion of the symbols to tra jectories with application parse rules. The parsing rules were manually formed to cover all HamNoSys notation variants. There are 39 rule actions forming complete animation tra jectories. For this purpose 138 HamNoSys symbols are currently adopted. The processing of the tree is carried out by several tree walks whilst the size of the tree is reduced. The final animation tra jectories in the root node are transformed by an inverse kinematics technique to control the joints of avatar animation model. The analysis of HamNoSys symbols allows us to animate hands and the upper half-body. Thus a single sign is encoded by corresponding sequence of HamNoSys symbols.

We have developed an interactive tool which purpose is to extend our database of signs. The main application window contains list of symbols which can be clicked and added into the sequence. This sequence can be immediately converted into the movement of the avatar which is shown in the second window. This allows fast production of symbol sequences for new signs and easy modification of existing signs since the changes are directly visible. In addition it allows people who have no high experince with HamNoSys to learn it faster. At present our database contains about 300 signs which are encoded as sequeces of HamNoSys symbols. This first database is targeted to the information system for train connections. Further expansion of the database will add new areas where the avatar can be used.

Corpus-based Sign Dictionaries of Technical Terms – Dictionary Projects at the IDGS in Hamburg

Lutz König, Susanne König, Reiner Konrad, Gabriele Langer
University of Hamburg, Germany

At the Institute of German Sign Language (IDGS), six dictionary projects in such diverse technical fields as computer technology, psychology, joinery, domestic science, social work as well as health and nursing care have been carried out. A seventh project on landscaping and horticulture is in progress. Six of the seven dictionaries are based on a corpus collected from deaf experts in the respective fields. Elicitation methods, such as interviews and picture prompts, corpus design as well as annotation, transcription, sign analysis and dictionary production have been continually developed and refined over the years. Many procedures rely heavily on the use of a relational database system iLex (see other presentation).

The presentation provides an overview of the projects, procedures and products with special attention given to the issues of corpus-building for and corpus-relatedness of the dictionaries at most stages of analysis and production. We focus on the corpus-based selection process which translations to include in the dictionary and on the analysis of single signs.

From 1998 on, the dictionaries do not only provide translations of technical terms into DGS but also include a special section that lists single signs used in these translations in separate entries. The structure of these entries is similar to what you would expect from a general sign language dictionary. Information including lexical status, meaning, use of space, iconic value and cross references to similar signs is given for each sign. However, due to the limited size of each of these corpora and the elicitation methods used, not all information can be drawn from or validated by the corpus.

Within the scope of the projects, assumptions and practical decisions have been made to deal with lexicological and lexicographical issues. These include the identification of lexemes, the degree of lexicalisation, i.e. the lexical status of signs and their meanings, the role of mouthings, and the relations between signs (polysemes vs. homonyms, modifications and variants). One important criterion for these decisions is the iconic value of signs.

The lexicographic solutions applied to specialised sign language dictionaries also provide a solid basis for general sign lexicography as well as corpus annotation and lexical analysis.

Content-Based Video Analysis and Access for Finnish Sign Language – A Multidisciplinary Research Project

Markus Koskela[1], Jorma Laaksonen[1], Tommi Jantunen[2], Ritva Takkinen[2], Päivi Rainò[3], Antti Raike[4]
[1]Helsinki University of Technology, Finland ; [2]University of Jyväskylä, Finland ; [3]Finnish Association of the Deaf, Helsinki, Finland; [4]University of Art and Design, Finland

In this research project, computer vision techniques for recognition and analysis of gestures and facial expressions from video will be developed and the techniques will be applied for processing of sign language. This is a collaborative project between four partners: Helsinki University of Technology, University of Art and Design, University of Jyväskylä, and the Finnish Association of the Deaf. It has several objectives of which four are presented in more detail in this poster.

The first objective is to develop novel methods for content-based processing and analysis of sign language video recorded using a single camera. The PicSOM retrieval system framework developed by the Helsinki University of Technology regarding content-based analysis of multimedia data will be adapted to continuous signing to facilitate automatic and semi-automatic analysis of sign language videos.

The second objective of the project is to develop a computer system which can both (i) automatically indicate meaningful signs and other gesture-like sequences from a video signal which contains natural sign language data, and (ii) disregard parts of the signal which do not count as such sequences. In other words, the goal is to develop an automatized mechanism which can identify sign and gesture boundaries and indicate, from the video, the sequences that correspond to signs and gestures. The system is not expected to be able to tell the meanings of these sequences.

An automatic segmentation of recorded continuous-signing sign language is an important first step in the automatic processing of sign language videos and online applications. It is our hypothesis that the temporal boundaries of different sign gestures can be detected and signs and non-signs (intersign transitions, other movements) can be classified using a combination of a hand motion detector, still image multimodal analysis, facial expression analysis and and other non-manual signal recognition. The PicSOM system inherently supports such fusion of different features.

The third objective is linked to generating an example-based corpus for FinSL. There exist increasing amounts of recorded video data of the language, but almost no means for utilizing it efficiently due to missing indexing and lack of methods for content-based access. The studied methods could facilitate a leap forward in founding the corpus.

The fourth objective is a feasibility study for the implementation of mobile video access to sign language dictionaries and corpora. Currently an existing dictionary can be searched by giving a rough description of the location, motion and hand

form of the sign. The automatic content-based analysis methods could be applied to online mobile phone videos, thus enabling sign language access to dictionaries and corpora.

The Klagenfurt lexicon database for sign languages as a web application: LedaSila, a free sign language database for international use

Klaudia Krammer, Elisabeth Bergmeister, Silke Bornholdt, Franz Dotter, Christian Hausch, Marlene Hilzensauer, Anita Pirker, Andrea Skant, Natalie Unterberger

University of Klagenfurt, Austria

Klagenfurt University has created a database which is described in Sign Language & Linguistics 4 (2001), 191-201. The objective of turning it into a web application was to offer our database for sign languages and users all over the world. In accordance with the self-conception of the Internet community and the rules of linguistic ethics, this is a basic service for sign language communities: the database can be used for free for non-commercial deaf and scientific issues.

As for the procedure: In order to add a sign language to the database, you need to provide a legitimation from the respective deaf organisation (i.e. of those people who use the sign language in question regionally or nationally), then you will be authorised to enter the data for this sign language into the database. By entering data you open them for communities of deaf people, scientists, and learners.

The data of the sign languages entered in the database will be stored on a Klagenfurt server. There is no limitation on calling up sign language data (searching for a certain sign or parameter value(s) etc.). For downloading videos, the users have to disclose their identity.

Main characteristics of the database

The database is designed in a way that everything which should appear in any monolingual or bilingual dictionary can be entered. All descriptive categories are as closely related to phenomena as possible. The analysis of the categories does not have to follow a strictly linear procedure or any assumed phonological or grammatical hierarchy. Additionally it offers:

- Openness of the sets of categories and their values (the users can add new categories or values to the database at their discretion). They can also translate the English terms of the description language into any other language which uses an alphabet.
- Quick production of entries: In order to enable the users to enter signs as fast as possible, we provide the possibility of a "minimum entry": it is sufficient to enter only one item, e.g. a single meaning, and then to store the sign video. The entries can then be amended later on.

Fields of data types within the database

- Type of sign (e.g. one-handed or two-handed symmetrical/asymmetrical; hand shape, location, orientation, type of contact, type of movement, intensity, etc.)
- Non-manual component: facial gestures, mouth gestures, body orientation, eye gaze, etc.
- Semantics: translation equivalents for a bilingual dictionary or explanation in the respective sign language for a monolingual dictionary); connotations or sign etymologies can also be added.
- Pragmatics: use, collocations, or idioms can be documented with video examples.
- Text/context examples
- Morphosyntax: categories of parts of speech, coding properties (e.g. morphological changes or position in a sentence or phrase) and syntactic functions
- Word field(s).

Digital Deployment of the Signs of Ireland Corpus in Elearning

Lorraine Leeson [1], Brian Nolan [2]

[1] Trinity College Dublin, Ireland, [2] Institute of Technology, Ireland

The Signs of Ireland corpus is part of the School of Linguistic, Speech and Communication Sciences' "Languages of Ireland" project. The first of its kind in Ireland, it comprises 40 male and female signers from across the Republic of Ireland, aged 18-65+, all of whom were educated in a school for the Deaf. The object was to create a snapshot of how ISL is used by 'real' signers across geographic, gendered and generational boundaries, all of which have been indicated as

sociolinguistically relevant for ISL (cf. the work of Le Master; also see Leeson and Grehan 2004, Leonard 2005, Leeson et al. 2006). With the aim of maximising the potential of cross-linguistic comparability, we mirrored aspects of data collection on other corpora collected to date. Thus, we include the Volterra et al. picture elicitation task (1984), "The Frog Story", and also asked informants to tell a self-selected story from their own life. To date, all of the self-selected stories have been annotated using ELAN.

Two institutions (CDS, TCD and ITB) have partnered to create a unique elearning environment based on MOODLE as the learning management system. This delivers third level signed language programmes to a student constituency in a way that resolves problems of time, geography and access, maximizing multi-functional uses of the corpus across programmes. Students can take courseware synchronously and asynchronously. We have now built a considerable digital asset and plan to re-architect our framework to avail of current best practice in digital repositories and digital learning objects vis-à-vis Irish Sign Language.

This paper outlines the establishment and annotation of the corpus, and the success of the corpus to date in supporting curricula and research. This paper focuses on moving the corpus forward as an asset to develop digital teaching objects. This paper outlines the challenges inherent in this process, and outlines our plans and our progress to date in meeting these objectives. Specific issues include:

- Decisions regarding annotation
- Establishing mark-up standards
- Use of the Signs of Ireland corpus in elearning/ blended learning contexts
- Leveraging a corpus within digital learning objects
- Architecture of a digital repository to support sign language learning
- Tagging of learning objects versus language objects
- Issues of assessment in an elearning context

References

Lorraine Leeson, John Saeed, Cormac Leonard, Alison Macduff and Deirdre Byrne-Dunne 2006: Moving Heads and Moving Hands: Developing a Digital Corpus of Irish Sign Language: The 'Signs of Ireland' Corpus Development Project. Paper presented at the IT&T conference, Carlow, 2006.

Leeson, L. and C. Grehan 2004: To the Lexicon and Beyond: The Effect of Gender on Variation in Irish Sign Language. In M. Van Herreweghe and M. Vermeerbergen (eds.): *To The Lexicon and Beyond: The Sociolinguistics of European Sign Languages*. Gallaudet University Press. 39-73.

Le Master, B. 1990: The Maintenance and Loss of Female and Male Signs in the Dublin Deaf Community. PhD Dissertation. Los Angeles: University of California.

Le Master,B. 2002: What Difference Does Difference Make? Negotiating Gender and Generation in Irish Sign Language. In S. Benor, M. Rose, D. Sharma and Q. Shang (eds.): *Gendered Practices in Language*. Centre for the Study of Languages and Information Publication. Stanford.

Leonard, C. 2005: Signs of Diversity: Use and Recognition of Gendered Signs among Young Irish Deaf People. *Deaf Worlds,* Vol. 21 (2) 62-77.

Volterra, V., S. Laudanna, E. Corazza and F. Natale 1984: Italian Sign Language: The Order of Elements in the Declarative Sentence. In F. Lonke (ed.) *Recent Research on European Sign Languages*. Svets and Zeitlinger: Lisse. 19-48.

Toward a Computer-aided Segmentation

François Lefebvre-Albaret, Frederick Gianni, Patrice Dalle

IRIT, Toulouse, France

Processing sentences of a sign language corpus requires a first step of temporal segmentation, which is long and tedious. To realize this segmentation more quickly, we propose an innovating method of computer-aided segmentation. This method processes motions of the dominated and dominating hands during the sign realisation. The video treatments are applied in four steps. The first one consists in tracking the hands in a video sequence using particles filtering. Then, in a second step, an operator watches the video sequence and indicates for each sign a time stamp during the sign realisation. Using this information and the trajectories of each hand, our method is able to find the beginning and the end of each sign in a third step. At the end, the operator can eventually apply some rectifications and validate the segmentation.

The presented article explains the different steps, from the calculation of the head and hands 2D positions to the computer-aided determination of the temporal segmentation of the signs. The segmentation exploits a model of French Sign Language and focuses especially on the characteristics of manual sign movements. Our method detects in the video several dynamic properties as the relative hands movement (symmetries, static hands) and the movement primitives

(simple or double repetition, uniform or accelerated straight movement). We also detect time spaces between two consecutive signs. Those transitions must be economical, considering the necessary energy to realize these: a movement with a complex realization will contain a sign. Those elements are then combined to each other to determine the most plausible temporal segmentation of the signed sentences. The result can be represented as a succession of signs and transitions segments.

Other observations can be taken into account to obtain the temporal segmentation. We can mention the determination of the elbows 2D positions, the characterization of hand configurations and the head orientation measurement. We describe how those elements could be used to improve the segmentation reliability.

The proposed method is based on motion analysis and does not use any knowledge about the words used in the processed sentences. Using the characteristics shared by the majority of French Sign Language's signs, it is possible to detect not only standard signs but also other manual iconic signs.

Our segmentation results are finally compared with a traditional manual segmentation produced with an annotation software named AnColin. This comparison exhibits several possible error sources. We focus on the problem of granularity and precision of the segmentation. We also discuss about other qualitative problems such as the detection criteria of the signs start and end. The evaluation protocol of a temporal segmentation is also adressed. Finally we will raise several problems to overcome, to realize a fully automatic segmentation.

Development of Sign Language Acquisition Corpora

Diane Lillo-Martin[1], Deborah Chen Pichler[2]

[1]University of Connecticut and Haskins Laboratories, USA; [2]Gallaudet University, USA

Longitudinal, spontaneous production data have long been a cornerstone of language acquisition studies, but building corpora of sign language acquisition data poses considerable challenges. Our experience began with the development of a sign language acquisition corpus more than 15 years ago and has recently included a small-scale experiment in corpus sharing between our two research groups. Our combined database includes regular samples of deaf and hearing children between the ages of 1;06 to 3;06 years acquiring ASL as their native language. The process through which we generate and share transcripts has undergone dramatic changes, always with the triple goal of creating transcripts with sufficient information for the reader to locate regions of interest, while keeping the video fully accessible and minimizing the time required to generate transcripts. In this paper we summarize the various incarnations of our transcription system, from simple Word documents with minimal integration of video, to a combination of FileMaker Pro software integrated with Autolog, to a fully integrated transcript+video package in ELAN. Along the way, we discuss the potential of ELAN to surmount several obstacles that have traditionally stood in the way of large-scale corpus sharing in the sign language acquisition community.

Use of sign language materials in teaching

Johanna Mesch and Lars Wallin

University of Stockholm, Sweden

We are in the beginning phase of creating a Swedish Sign Language corpus. Some of the material is now used with students in two separate courses: Swedish Sign Language for beginners, and Swedish Sign Language linguistics (for deaf and hearing signers). In this workshop we will present some teaching methods and technical problems. Some examples are shown of how the students use the sign language corpus through the dictionary database, the corpus database and a learning platform for studying and analyzing sign language texts, like for example the small corpus in Bergman & Mesch 2004 and also some old and new recordings. Students can practice sentences, analyze the entries and annotate the texts or their own recordings. Bergman's earlier transcription system for Swedish Sign Language (Bergman 1982) has been updated continuously, and partly adapted for possible use as a standard annotation system. Problems with storing sign language material are also discussed.

References

Bergman, Brita. & Mesch, Johanna. 2004. *ECHO data set for Swedish Sign Language (SSL).* Department of Linguistics, University of Stockholm.

Bergman, Brita. 1982. *Teckenspråkstranskription. Forskning om teckenspråk X.* Stockholms universitet, Institutionen för lingvistik.

Lexique LSF

Cédric Moreau and Bruno Mascret

[1]Institut National Supérieur de Formation et de Recherche pour l'Education des Jeunes Handicapés et les Enseignements Adaptés, France; [2]Université Paris 8, France

The French Sign Language (LSF) was banned in 1880 from all teaching institutions. From then on, it continued expanding in an uncoordinated way throughout special schools. In 1991, a new French law allowed deaf people to choose a bilingual education (French and sign language), and since February 2005 each school is required to integrate every devoted child who wishes it, no matter his handicap. All public websites must also become accessible.

With this new context, the LSF grows using regional differences, and users invent new signs to translate new concepts. However, the sign language cannot count on traditional media to spread out new expressions or words, since it is nor spoken nor written. Therefore the sign vocabulary differs depending on geographical and social situations, furthermore if the concept is specific and elaborate. The website LexiqueLSF wishes to propose users a contributing and efficient tool, allowing a large diffusion of new signs and concepts. A short analysis of the existing supports will lead us to present the main issues and to describe precisely the technical and linguistic solutions we chose, as well as some of the problems we met. This website must absolutely have a relevant and sharp classifying system, must be accessible to everyone, and offer new entries to satisfy all users. Likewise, all the elements composing the website should be considered as a concept in order to imagine complete accessibility to deaf people, and not only to blind people. We do not wish to make a simple dictionary.

Our aim is to allow exchanges between users, to encourage them to invent and spread neologisms, and to make sure that the represented concepts are clear and understandable. Publishing a new notion requires to create a number of descriptors (in french and in sign language, illustrations, examples...) and to relate this notion to others already existing (opposite or similar concepts...). Each new sign proposed will be completely described, therefore it can easily be appropriated. A reliable, but not compulsory, validation system will guarantee only serious suggestions.

Our production is thus very different from already existing paper or digital dictionaries, containing only everyday life vocabulary and almost no definitions, nor use examples. The best ones sort words according to the space location and configuration of the sign, but do not recognise morphological variations. Let us also observe that these dictionaries are not "bilingual" since they are accessible only to french speakers.

According to C. Cuxac 2000, two discursive enunciation strategies co-exist in LSF: through the canal leading from the vision to the sign, you can either choose to say with or without showing. Meaning you can either "make see" your experience with a visually accurate sequence of signs, or you can use the standard signs having no physical resemblance with the experience you are describing. Referring to this theory, our research supposes to organise into a hierarchy all linguistic parameters used in signs as meaning elements.

Construction of Japanese Sign Language Dialogue Corpus: KOSIGN

Yuji Nagashima[1], Mina Terauchi[2], Kaoru Nakazono[3]

[1] Kogakuin University, Japan; [2] Polytecnic University, Japan; [3] NTT Network Innovation Laboratories, Japan

This report presents a method of building corpuses of dialogue in Japanese Sign Language (JSL) and the results of the analysis in co- occurrences of manual and non-manual signals using the corpus.

We have built the sign dialogue corpus by video recording the dialogues between native JSL speakers. The purpose of building corpus is deriving electronic dictionaries such as morphological dictionary, different meaning word dictionary, allomorph dictionary and example dictionary. Example sentences are recorded for every word (key sign) those were recorded in the sign language word data base KOSIGN Ver.2. Until now, we were able to confirm a correlation of manual and non-manual signals or a characteristic appearance of sign language dialogue.

As a result of the analysis, the pointing occurred to the end of sentence at high frequency. It suggested that pointing be one of the ends of sentence, and clarified the role as the conjunctive pronoun. The co-occurrence relation between the manual and non-manual signals acquired confirmed an important role to make the meaning of the expression sign language limited was achieved. Moreover, "Roll shift" and "Sandwich construction" that was the linguistic feature of sign language were confirmed, too. These information is necessary for the hearing person to study sign language.

Documenting an Endangered Sign Language: Creating a Corpus of Langue des Signes Malienne (CLaSiMa)

Victoria Nyst
Leiden University, The Netherlands

Langue des Signes Malienne (LaSiMa) is the local sign language of Mali. It evolved spontaneously in the streets of urban centers outside the context of Deaf education. The Malian 'grins', meeting places where men gather in the afternoon to chat and drink tea seem to be the cradle of this language.

Since about 15 years now, American Sign Language (ASL) has been introduced in Deaf education in Mali. As a consequence, LaSiMa has become the language of non-educated Deaf people and is by many considered to be inferior to ASL. LaSiMa is marginalized and at present, few Deaf adults in Bamako sign LaSiMa without mixing in some ASL signs. It is likely that in few generations, the use of LaSiMa will have reduced or altered greatly.

Little is known about sign languages from the African continent. In view of the evolution of LaSiMa outside the context of Deaf education in addition to its endangered status, a three-year documentation and description project was set up with the help of the Hans Rausing Endangered Language Program (URL). The aim of the project is to establish a corpus of LaSiMa discourse, a lexical database and descriptions of selected structural aspects of the language.

The Corpus Langue des Signes Malienne (CLaSiMa) aims at collecting a large, filmed sample of LaSiMa discourse. The sample is to be diverse and representative with respect to its signers (age, gender) and with respect to its discourse types. So far, 15 hours of discourse have been filmed. The filmed discourse is to be translated in French using ELAN software. A selection of the discourse in the corpus will be glossed. The material will be deposited in a digital archive, where it will be accessible through internet for the academic as well as the Malian Deaf community.

The initial approach to gather the data was inspired by the work on the NGT corpus (Crasborn & Zwitserlood, 2007). Their approach involved among others inviting pairs of signers to a filming location where they discuss preset issues and do specific language-based tasks. A similar approach in the LaSiMa context was challenged in several ways. Controlling the age and gender balance, the "nativeness" of signers was problematic as well as the cultural appropriateness of the material and the tasks, affecting the spontaneity of the signers. An alternative approach was developed, which involved recording one or two signers in their 'grins'. This required training two native LaSiMa signers in film and interview techniques.

References

Onno Crasborn & Inge Zwitserlood (2007) 'The Corpus NGT project: challenges in publishing sign language data'. Workshop 'The emergence of corpus sign language linguistics', paper presented at BAAL 2007, Edinburgh, Thursday 6 Sept. 2007.

The Representation Issue and its Multifaceted Aspects in Constructing Sign Language Corpora: Questions, Answers, Further Problems

Elena Antinoro Pizzuto[1], Isabella Chiari[2], Paolo Rossini[13]

[1]Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Roma, Italy; [2]Dipartimento di Studi Filologici, Linguistici e Letterari, Università di Roma "La Sapienza", Italy; [3]Istituto Statale Sordi di Roma, Italy

This paper aims to address and clarify one issue we believe is crucial in constructing Sign Languages (SL) corpora: identifying appropriate tools for representing in written form SL productions of any sort, i.e. lexical items, utterances, discourse at large. Towards this end, building on research done within our group on multimedia corpora of both SL and spoken or verbal languages (vl), we first outline some of the major requirements and guidelines followed in current work with vl corpora (e.g. regarding transcription, representation [mark-up], coding [or annotation] Chiari, 2007; Edwards & Lampert; 1993; Leech & al, 1995; Ochs, 1979; Powers, 2005, among others). We highlight that a basic requirement of vl corpora is an easily readable transcription that, aside from specialist linguistic annotations, allows anyone who knows the object language to reconstruct its forms, and its form-meaning correspondences. Second, we show how this basic requirement is not met in most current work on SL, where the 'transcription' of SL productions consists primarily of word-labels taken from vl, inappropriately called 'glosses'. As argued by some authors (e.g. Pizzuto & Pietrandrea, 2001; Russo, 2005; Pizzuto et al., 2006), the use of such word-labels as a primary representation tool grossly misrepresents SL, even when supported by specialist linguistic annotations (e.g. Stokoe-based notations, the Berkeley Transcription System [Slobin et al., 2001]). Drawing on a crosslinguistic overview of relevant work on SL lexicon and discourse (e.g. Brennan, 2001; Cuxac, 2000; Cuxac & Sallandre, 2007; Russo, 2004; Antinoro Pizzuto et al., 2007), we illustrate how the 'transcriptions' most widely used for SL do not allow to anyone who knows the specific SL to reconstruct its forms and form-meaning correspondences, and are especially inadequate for representing complex sign units that are very frequent in SL discourse, and exhibit highly iconic, muldimensional/multilinear features that have no parallel in vl. Third, we

present and discuss ongoing research on Italian Sign Language (LIS) in which experienced deaf signers explore the use of SignWriting (Sutton, 1995) as a tool for both composing texts conceived in written form – thereby creating a corpus of written LIS – and for transcribing corpora of face-to-face LIS discourse (Di Renzo et al., 2006; Di Renzo, in press; Lamano et al., in press). The results show that, in both cases, deaf signers can easily represent the form-meaning patterns of their language with an accuracy never experienced with other representation or notation systems. The analysis of the texts produced has also provided new indications on the structure of LIS, highlighting the need of revising the criteria for constructing lexical corpora on the grounds of regularities (and variance) found in discourse corpora. While all of this suggests that SignWriting can be a valuable tool for addressing the representation issue in constructing SL corpora, the present computerized form of SignWriting poses technical problems that severely constrain its use. We conclude specifying the problems that need to be faced for conducting more extensive experimentations.

DGS Corpus Project – Development of a Corpus Based Electronic Dictionary German Sign Language / German

Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, Arvid Schwarz

University of Hamburg, Germany

The poster introduces a 15-year project accepted for funding by the Hamburg Academy of Sciences. The proposed project aims to combine the collection of a large corpus with the development and production of a comprehensive, corpus based electronic dictionary of German Sign Language (DGS).

To this aim, a corpus of approximately 350–400 hours from 250–300 informants will be collected in a variety of elicitation settings. This is, in size and scope, comparable to large spoken language corpora. The design allows the use of the corpus for various tasks. These are, amongst others: (i) the validation by corpus data of a basic vocabulary compiled from different published sources; (ii) research on DGS grammar based on detailed transcription data; (iii) identification of different meanings and collocations of a sign by appropriate contexts. Furthermore, the design anticipates a comparative sociolinguistic study comparable in kind and quality to Lucas et al. (2001) and Schembri/Johnston (2004). The corpus thus provides a starting point for research deep into the structure and lexicon of German Sign Language as well as into the visual-gestural mode of sign languages in general. Parts of the annotated corpus, i.e. transcription files with English translations, will be made available online to the international linguistic community.

The corpus data will undergo two stages of transcription. First, a basic transcription serves to segment utterances and to identify lexical items and thus provides a first access to the data. Second, approximately 50 % of the transcriptions will be transcribed again in more detail. This serves the purpose of clarifying grammatical questions for the dictionary grammar as well as dealing with lexicological and lexicographic issues. The annotation of the corpus will be closely intertwined with the requirements of lexical analysis. A high quality of transcription will be achieved through continuous verification by native signers. A relational database (iLex, cf. Hanke/Storz) supports this process, especially the consistency of type-token matching.

Lexical analysis and lexicographic decisions concerning for example lexical status, language change, and lemma selection will be continuously validated by a deaf focus group and a general voting web interface which will be open for all interested members of the deaf community.

The dictionary will be entirely based on the corpus with respect to the list of lemmas to be included but decidedly exceed a conglomeration of corpus references. Rather, we will systematically abstract from the references to obtain a generalized description of lexical items. Examples of sign uses will be taken directly from the corpus.

For cross-linguistic research and comparability of results across projects, we consider it essential to push standardisation or at least compatibility of annotation and transcription conventions. To reach this, we have arranged cooperations with some other national corpus projects and look forward to cooperate with more projects currently in preparation.

References

Lucas, Ceil / Bayley, Robert / Valli, Clayton (2001): Sociolinguistic Variation in American Sign Language. Washington, DC: Gallaudet Univ. Press.

Schembri, Adam / Johnston, Trevor (2004): Sociolinguistic variation in Auslan (Australian Sign Language). A research project in progress. In: Deaf Worlds 20 (1), 78-90.

British Sign Language Corpus Project: Open Access Archives and the Observer's Paradox

Adam Schembri
University College London, UK

The British Sign Language Corpus Project is a new three-year project (2008-2010) that aims to create a machine-readable digital corpus of spontaneous and elicited British Sign Language (BSL) collected from deaf native signers and early learners across the United Kingdom. In the field of sign language studies, it represents a unique combination of methodology from variationist sociolinguistics and corpus linguistics. The project aims to conduct a studies of sociolinguistic variation, language change and language contact simultaneously with the creation of a corpus. As such the nature of the dataset to be collected will be guided by the need to create a judgement sample of the deaf community rather than a strictly representative sample. Although the recruitment of participants will be balanced for gender and age, it will focus only on signers exposed to BSL before the age of 7 years, and adult deaf native signers will be disproportionately represented. Signers will also be filmed in 8 key regions across the United Kingdom, with a minimum of 30 participants from each region. Furthermore, participant recruitment will rely on deaf community fieldworkers in each region, using a technique of 'network sampling' in which the local community member begins by recruiting people he or she knows, and asks these individuals to recommend other individuals matching the project criteria. Moreover, the data will be limited in terms of situational varieties, focusing mainly on conversational and interview data, together with narratives and some elicitation tasks. Unlike previous large-scale sociolinguistic projects, however, the dataset will be partly annotated and tagged using ELAN software, given metadata descriptions using IMDI tools, and will be archived and made accessible and searchable on-line. As such, we hope that it will become a standard reference and core data source for all researchers investigating BSL structure and use. This means, however, that, unlike previous sociolinguistic projects on ASL and Auslan, participants must consent to having the video data of their sign language use made public. This seems to put at risk the authenticity of the data collected, as signers may monitor their production more carefully than might otherwise occur. As the aim of variationist sociolinguistics is to study the vernacular variety (i.e., the variety adopted by speakers/signers when they are monitoring their style least closely), open-access archives thus may not always provide the best data source. While recognising that this concept of the vernacular represents an abstraction, we discuss the possibility of overcoming this problem by making some of the conversational data password protected for use by academic researchers only, while making other parts of the corpus publicly available as part of a dual access archive of BSL.

Tactile sign language corpora: capture and annotation issues

Sandrine Schwartz
Paris 8 University, France

Sign language, being a visual-gestural language, can also be used tactually among or with deaf people who become blind. When this language is shared between people who are totally blind, non-manual features of signs are totally neutralised, resulting into a purely kinesthetic-gestural variant of sign language. This tactile modality of reception leads to adjustments impacting sign language pragmatics, as well as sign order and to a lesser extent, the way signs are formed. We aim to explore these phenomena by carrying out a systematic analysis of tactile sign language corpora.

Such a corpus has been filmed in 2006, involving six French deafblind informants, all of them using tactile sign language as their primary means of communication. A total of 14 hours of spontaneous discussions, free conversations or elicited data were captured by up to three digital cameras.

In order thoroughly to analyse our corpus, we need the help of a reliable annotation tool. After trying a couple of them, we decided to select Anvil, for its visual layout and flexibility, as well as its temporal granularity. We need a partition annotation system which allows us to create, rename or reorder tracks freely even while annotating. The first steps of our annotation will take us on the lanes of conversational analysis, using a mix of glosses and pragmatic occurrences, eventually to lead us on the more sinuous paths of a syntactic micro-analysis.

Building 3D French Sign Language lexicon

Jérémie Segouat[1,2], Annelies Braffort[1], Laurence Bolot[1], Annick Choisier[1], Michael Filhol[1], Cyril Verrecchia[1]

[1]LIMSI-CNRS, Orsay, France; [2]WebSourd, Toulouse, France

Sign Language (SL) corpora can be made in vivo (in natural conditions, with or without guidelines) or in vitro (in a laboratory, with guidelines), like speech corpora. While some effort has been made to standardise on corpus metadata with the IMDI project, there is not such norm for SL corpus elaboration: the methodology depends on the research goal.

Our aim is to create a 3D French SL (LSF) corpus to be used in different types of software, thus the signs must be considered without context. To fit our specific goals, we have decided on the following methodology:

First of all we look for a *referent gestuel*: a deaf person, whose first language is LSF and whose signing corresponds to what is needed for the infographic part of the process (see step five). The *referent gestuel* has been chosen by the team's infographist. It is important to keep the same *referent gestuel* for all the lexicon, in order to keep a consistency in the 3D corpus. Then, for each scientific topic we consider, we look for an *expert*: a deaf person, whose first language is FSL and who is aware of the terminology used in the field. Because we want many topics corpus, we work with different *experts*. Thirdly we organise a meeting with the *referent gestuel*, the *expert* and, if necessary, other specialists of the topic we are working on, so they can discuss each meaning of each concept and sign it the as accurately as possible. The next step is to film the *referent gestuel* two views of the *referent gestuel* (frontal view and side view). Lastly we animate the 3D avatar by copying each frame of the video, using 3DSMax software.

This 3D corpus will be used in at least three different informational pieces of software. One that can be used in railway stations to inform deaf users in LSF about the delay of a train or any other problem. Another application will be a dictionary: We are developing a LSF-French dictionary where signs would be given in a the form of a 3D avatar. A third example is to display an Embodied Conversational Agent (ECA) on our laboratory website so that our research topics can be explained in LSF if wanted.

References

Michael Filhol, Annelies Braffort et Laurence Bolot (to appear), Signing Avatar: Say hello to Elsi! In: Gesture in Human-Computer Interaction and Simulation, selected revised papers of the 7th International Gesture Workshop (GW'07), LNCS LNAI, Springer.

Onno Crasborn & Thomas Hanke (2003), Metadata for sign language corpora. Background document for an ECHO workshop, 8 + 9 May 2003, Nijmegen University.

Interface Development for Computer Assisted Sign Language Learning: Compact version of CASLL

Saori Tanaka[1], Yosuke Matsusaka[2], Kaoru Nakazono[3]

[1] MID, Japan; [2] National Institute of Advanced Industrial Science and Technology, Japan, [3] NTT Network Innovation Laboratories, Japan

In this study, we introduce e-learning system called CASLL and demonstrate the small interface in order to implement it into the mobile video reproducers. As one of our series of previous studies for developing human interface by using Japanese Sign Language (JSL) contents [2, 3, 4], we proposed a new learning program and compare it with the existing learning program implemented in the Computer Assisted Sign Language Learning (CASLL) system [1]. In the existing learning program, users learn sign words and then try to select the appropriate Japanese translations in a natural conversation expressed by two native signers. In the proposed program, users try to segment each word from a stream of signing by manipulating a control knob on the bottom of a movie screen, and then do the same tasks in the existing learning model. The end of the segmentation task is to know how continuous signs are articulated in the natural discourse. Ten Japanese learners participated in the experiments. Five subjects learned the existing word learning program and the other five subjects learned the proposed segmentation learning program. The mean accuracy rate of the proposed program was higher than that of the existing program. The result has indicated that focusing on transitional movements has an effect for learning JSL as a second-language.

Although the segmentation learning method has been shown as more effective learning method compared to the word learning program by which learners need to just memorize the meaning of words, there were some technical problems. Some learners answered that they could not see each JSL movies at once by using their own laptops to conduct the learning programs. Therefore, we needed to improve how to show JSL movies by using different sizes of screen. We define the size of movie screen as small as possible, and develop use-friendly interface with which learners can recognize whole serious of JSL movies by switching each movie side by side. We will demonstrate the interface development and see if the interface is applicable to the other sign languages.

References

[1] Saori Tanaka, Yosuke Matsusaka, Kuniaki Uehara: "Segmentation Learning Method as a Proposal for Sign Language e-learning", *Human Interface*, 2006 (in Japanese)

[2] Saori Tanaka, Kaoru Nakazono, Masafumi Nishida, Yasuo Horiuchi, Akira Ichikawa: Analysis of Interpreter's Skill to Recognize Prosody in Japanese Sign Language, *Journal of Japanese Society for Artificial Intelligence* (in press).

[3] Kaoru Nakazono, Saori Tanaka: Study of Spatial Configurations of Equipment for Online Sign Interpretation Service, *IEICE Transaction on Information and System* (in press)

[4] Saori Tanaka: A Study of Non-linguistic Information in Japanese Sign Language and its Application for Assisting Learners and Interpreters, Ph.D thesis, Chiba University, 2008

Annotation of Sign and Gesture Cross-linguistically

Inge Zwitserlood [1,2], Pamela Perniss [1,3], Asli Özyürek [1,2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands ; [2] Radboud University Nijmegen, The Netherlands; [3] University College London, UK

In a 5-year project, we compare expressions in the spatial domain between two sign languages (German Sign Language and Turkish Sign Language), the co-speech gestures accompanying two spoken languages (German and Turkish), and the pantomime-like structures used by hearing people (German and Turkish) asked to convey information without speaking. The aim is to discover the similarities and differences in the use of space in expressing referent location and motion between the sign languages, and between the signing, co-speech gesture and no-speech pantomime modes. To this end, we are building a large video corpus of task-related discourse data (about 90 minutes per 15 participants per condition). The data will be described using the IMDI metadata standard and linguistically annotated using ELAN. Parts of the data will be made accessible for research and educational purposes on the Browsable Corpus based at the MPI for Psycholinguistics.

In this presentation, we report the annotation conventions we have been developing based on collected data. There are two levels of annotation: (i) a descriptive level where we gloss signs and gestures according to the movements/positions of the hands, head, face, and body; and (ii) an analytic/coding level where each sign or gesture is analyzed with respect to the function of establishing and/or maintaining reference in discourse (e.g. through the use of pronouns, classifier predicates, modified verbs, and role shift in signing, and similar forms in gestures). Our conventions combine aspects from other annotation and coding systems developed for sign and gesture (e.g. the ECHO, Corpus NGT, and Auslan Corpus conventions; HamNoSys; gesture coding conventions as developed by Kita, Van Gijn and Van der Hulst), but go beyond them in placing special emphasis on coding both systems with the same parameters.

On the descriptive level, we developed a 3-dimensional scheme to identify for hand orientation, location, and direction of signs and gestures, allowing comparison across languages. On the analytic/coding level, we devised ways of categorizing how the various spatial expressions in sign and gesture map onto different coreference devices in discourse.

Parts of the sign and gesture data have now been annotated. We will present some generalizations and conclusions drawn from using our annotation conventions regarding cross-linguistic and sign-gesture comparison. Furthermore, based on our annotation experiences, we will discuss the advantages as well as the shortcomings of our annotation scheme and suggest specific improvements, which the linguistic community needs to consider in terms of ways they can be implemented in the technology of annotation software (such as ELAN).

Author Index