1. Project Title & Acronym and Abstract

Title: Linking lexical databases and annotated corpora of signed languages
Acronym: SignLinC

Abstract: This project aims to link two independently evolved data sets for a signed language: the Corpus NGT and the lexical database of the Dutch Sign Centre. The first is a corpus of video texts that is already fully compliant with CLARIN standards, while the latter is an independently evolved Microsoft SQL database. Both are prototypical for the situation of signed language resources in the world: corpora of running signing have only recently been under development and typically use ELAN as the annotation tool, while lexical databases have typically evolved as standalone applications to produce dictionaries in books or on CD/DVD-ROMs. In order to establish the link in both directions a conversion of the lexical database to the CLARIN LMF standard for lexica is planned in the project. Further, enhancements to the existing CLARIN tools ELAN and LEXUS are created to start the exchange between the two domains.

Target Start Date: January 1st, 2010
Target End Date: January 1st, 2011
Type: Resource Curation Project

2. Coordinator

Name: Dr. O.A. Crasborn
Function: Senior Researcher
Organization: Centre for Language Studies, Radboud University Nijmegen (RU)
Address: Department of Linguistics, PO Box 9103, 6500 HD Nijmegen
E-mail: o.crasborn@let.ru.nl
Tel: +31 24 3611377
Fax: +31 24 3611070
Role(s): User, Data Provider

3. Composition of the Project Team

Name: Dr. G.M. Schermer
Function: Director
Organization: Dutch Sign Centre (Nederlands Gebarencentrum, NGc)
Address: JF Kennedylaan 99
E-mail: t.schermer@gebarencentrum.nl
Tel: +31 30 6565407
Fax: +31 30 6565408
Role(s): User, Data Provider

Name: Drs. J. Ringersma
Function: Language Archive Manager
Organization: MPI for Psycholinguistics (MPI)
Address: P.O. Box 310, 6500 AH Nijmegen, The Netherlands
E-mail: jacquelijn.ringersma@mpi.nl
Tel: +31 24 3521102

Fax: +31-24-3521213
Role(s): Technology Provider

Name: Mr. H. Sloetjes
Function: Software developer
Organization: MPI for Psycholinguistics (MPI)
Address: P.O. Box 310, 6500 AH Nijmegen, The Netherlands
E-mail: han.sloetjes@mpi.nl
Tel: +31-24-3521467
Fax: +31-24-3521213
Role(s): Technology Provider

4. CLARIN centre

Max Planck Institute for Psycholinguistics

5.Requested Budget

EUR 60,000

6.Description of the Proposed Project

1.1 Research Question(s)

The proposed project will be crucial for all domains of sign language linguistics. For lexicography, corpus-based lexicon development will become possible on the basis of the Corpus NGT. For phonological and other grammatical research, the NGc lexicon will allow for the clear distinction of variants of lemmata, and make accessible a rich lexical specification of all items in the lexicon, including semantic, morpho-syntactic and phonological properties. Finally, a more consistent lexical annotation of corpus data is urgently needed for research on automatic sign recognition and translation.

Ongoing linguistic projects that will benefit from the project's outcomes include the ongoing lexicographic and related projects at the NGc, research under prof. A. Baker at the University of Amsterdam, and the projects at Radboud University (Crasborn: VIDI/ERC/CLS). The SignSpeak project (EU 2009-2012; Dutch partner: RU) that started in 2009 is one of the first to attempt automatic sign recognition in 'running signing'; consistent annotation of the Corpus NGT data is crucial for this project and can be greatly improved using the outcomes of the present proposal.

1.2 Research Data

The two central sets of research data, a lexical database and a discourse corpus, have been developed by two of the applicants, who own all property rights. The one ethical issue to watch is that some 1% of the sessions in the Corpus NGT is not publicly accessible: the deaf people participating in the recording did not give permission for public use. The materials are securely protected in the Browsable Corpus, but many copies for use by researchers are locally available.

The open access *Corpus Nederlandse Gebarentaal* (Corpus NGT; http://www.corpusngt.nl) is a sign language corpus of video data, totalling more than 70 hours of rich video material of 92 Dutch deaf people. It is a unique resource, being the second linguistic corpus to be developed in the world (after Australia), and it is the first to be published online as open content. The videos are in MPEG format, and have been partly annotated with Dutch glosses and Dutch translations in ELAN

(EAF/XML). Metadata descriptions have been made for the whole corpus according to the IMDI standard including the Sign Language Profile.

The NGT lexicon of the Dutch Sign Centre has been the outcome of a long series of dictionary projects for more than 25 years now. The current database contains over 15,000 lemmata including variants; there are example movies and context examples for about 20% of the lemmata. Part of the database is accessible online (http://www.gebarencentrum.nl). The database is created in Microsoft SQL Server 2005; the data stored in SQL-format. Movies are stored in MPEG format.

1.3 Technology

- **ELAN** is a multimedia annotation tool that allows for the creation of multi-layered annotation documents. Annotations are contained in tiers; tiers can be grouped hierarchically. The annotation values are Unicode text and annotation documents are stored in an XML format, EAF. ELAN's multi-video support makes it a useful tool for sign-language research (amongst others). ELAN's online sibling ANNEX enables web access to EAF documents with streaming media.
- **Lexical Markup Framework** (**LMF**) is an abstract metamodel that provides a standardised framework for the construction of computational lexicons. LMF provides a default lexicon structure and lexicon concept naming based on the proposed ISO data category registry for linguistic resources. It ensures the encoding of linguistic information in a way that enables interoperability across all aspects of electronic lexical resources.
- **LEXUS** is the online lexicon tool of the MPI LAT suite, and is based on LMF. LEXUS allows one to enrich the lexical entries with multi-media resources either stored in the LEXUS database or on an external URL. In LEXUS each lexical entry, and within the lexical entry each element is assigned a unique identifier which makes it relatively easy to connect to lexicon items from other applications. LEXUS provides an XML import and XML-LMF export of the lexicon, and an IMDI metadata file for lexicon and resources can be created during the export process.
- The technology which is used for the **NGT lexicon application** is split up in two parts, a management part and a viewer part. The first one is a web application that is developed in ASP (Active Server Pages). The viewer is a VB (Visual Basic) application which can used as a stand-alone application without any installation required. For its metadata a Microsoft SQL 2005 database is used for the web application, the viewer application works with a set of XML extractions of this SQL database. The database structure and functionality of the applications are fully described in the NGT lexicon application footprint document that is up to date.

1.4 Description

See under 1.1: the linking of the discourse corpus of annotated video and the lexical database will be beneficial for almost all areas of sign linguistic studies and language technology relating to sign language.

The central problem that underlies all sign language corpus annotation is that there is no commonly used writing system for any signed language. This has made consistent annotations (one-form-one-gloss) difficult. The establishment of direct links between the existing lexicon and corpus will thus create a more reliable basis for *any* research on signed languages.

1.5 Plan

Type: *Resource Curation project*

The project has two main aims. First, the current state of the lexical database will be converted to LEXUS/LMF (D3, D4, D8, M2). IMDI compatible metadata will be produced for this resource and included in the MPI archive (D5, M1). As the database itself is not in a specific definite stage, the lexicon itself will not yet be included in the archive, but simply be made accessible through LEXUS. The project builds on experiences with the conversion of the SignPhon phonetic database of NGT to LMF/LEXUS (Crasborn + Ringersma, 2008).

Secondly, existing CLARIN tools ELAN and LEXUS will be enriched so as to create links between the annotation tool and the lexicon tool. It is clear that given the absence of such a link at present, the proposed developments will be first steps in the direction of further communication between the two tools. While it is possible to use the web-based ANNEX for browsing annotations in EAF files, it is not in its current form an annotation tool. The present enhancements will therefore involve the development of the stand-alone application ELAN (D6, D9, M3). Two functionalities will be added. First, the lookup of a given existing annotation value in a LEXUS database (D6). This requires a connection between the ELAN annotation and the LEXUS search engine. Secondly, the possibility will be added for a LEXUS field to act as a constrained vocabulary for a tier in ELAN (D9); thus, entering a new string in a gloss annotation in ELAN would lead to a display of possible candidates, thus enforcing a link between gloss annotations and fields in LEXUS. In this first phase it is foreseen that the LEXUS XML export provides the semi-controlled vocabulary during ELAN glossing as well as the option of opening the lexical entry in the web version, to view its full content. A direct connection between ELAN and the online database during the glossing process is beyond the constraints of this project proposal.

The software development component is crucial in the curation of the lexical database, as there is currently no systematic way of ensuring consistent and persistent reference from the annotation files to lexical entries. Adding the functionality proposed here will facilitate consistent glossing in the annotation phase, aiming at the one-lemma-one-gloss rule. This consistent link forms the basis for any productive use of both the combination lexicon-corpus.

Where possible ISOcat data categories will be used for existing fields in the database. The main focus of the project is on the creation of a LMF-template for signed language lexicons, however. An effort will be made from the beginning of the project to explore where the sign language lexicon touches on existing ISOcat categories. Together with the study of an existing tool that integrates lexicon and corpus (iLex), these will be documented for the wider sign language community and for the other tasks in the project (D1).

Finally, it should be clear that no matter what the enhancements and conversions to LEXUS and ELAN will be, they will only contribute to a well-curated lexical resource, but cannot approach the functionality that is present in the current database used by the Dutch Sign Centre. It will take many years of development time to approach the rich and versatile interfaces that are currently in use at the NGc. A short workshop will therefore be devoted to explaining the way in which persistent identifiers (PIDs) for annotation segments in the Corpus NGT can be created, and added to the lexical database in its current form (D2). This will enable links from the present database to the CLARIN compatible ANNEX, which enables immediate use of the corpus at the NGc.

7. Deliverables and Milestones

Deliverables & Milestones	**Description**	**Responsible**	**Target Date**
D1	Document describing the challenges in the European context, including the use of ISOcat concepts	Crasborn	Month 3
D2	Workshop for NGc on persistent links to EAF annotations in ANNEX	Ringersma	Month 3
D3	LMF specification for the NGc lexicon	Ringersma & Schermer	Month 5
D4	Generation of a clean XML export of the NGc database	Schermer	Month 5
D5	IMDI metadata description of lexicon	Crasborn	Month 5
D6	ELAN: lookup of existing annotation in LEXUS	Sloetjes	Month 6
D7	Document describing linking to ANNEX for a wider audience	Crasborn	Month 6
M1	Publication of IMDI metadata description of NGc lexicon	Crasborn	Month 6
D8	Import of SQL/XML data in LEXUS	Ringersma	Month 9
M2	Publication of lexical resource in LEXUS	Ringersma	Month 10
D9	ELAN: use of LEXUS database as controlled vocabulary to enhance glossing consistency	Sloetjes	Month 10
M3	Release of ELAN incorporating D5 and D6	Sloetjes	Month 12

8. IPR and Ethical Issues: Risks

- No IPR issues expected.
- Ethics: special attention needed to ensure that corpus sessions without public access are indeed well-protected.

9. Expertise of the applicant(s)

- **Crasborn**: expert in sign linguistics (PhD 2001); coordinator of sign language research at the Centre for Language Studies (RU). Recipient of NWO VIDI and ERC Starting grants. Initiator and project manager of the *Corpus NGT* (NWO Investment Grant 2006-2008). Chair of the Sign Linguistics Corpora Network (NWO Internationalisation Project 2008-2011). Involved in development of ELAN since 2003.
- **Schermer**: director Dutch Sign Language Centre since 2002, expert in sign linguistic lexicography (MA Experimental Psychology, Northeastern University, Boston 1981, PhD University of Amsterdam 1990). Head of the research department

of the Dutch Foundation for the Deaf and Hard of Hearing Child (NSDSK, Amsterdam) from 1982 until 2000. Involved as project coordinator in numerous national and international research projects such as the development of various sign language databases, the development of multimedia educational materials and the production of a number of sign language dictionaries.

- **Ringersma**: expert in LMF and lexicon creation in LEXUS (MSc in Artificial Intelligence - computational linguistics, University of Utrecht). Heading the equipment and information management section in the Technical Group of the Max Planck Institute for Psycholinguistics. Involved in the development of LEXUS and ViCoS since 2006, bridging the gap between the software developers of the tool and the pilot users from language documentation teams (DoBeS).
- **Sloetjes**: software developer. Main responsible person for maintaining and further developing ELAN, as well as for providing support to users. Involved in the development of ELAN since 2003.

10. Project budget details

In the table below, cost estimates are based on the strong commitment of all partners to move forward with this project; any additional costs for the partner MMs will be carried by the respective organisations. In addition, assistant and developer costs will be highly contingent on who we can find to work on such a brief project.

Participant	**Organization**	**Effort (PM)**	**Salary Costs/PM (Euro)**	**Salary Costs (Euro)**	**Travel & subsistence (Euro)**	**Total (Euro)**
Crasborn	RU	1	5,000	5,000	250	**5,250**
Schermer	NGc	2	5,000	10,000	500	**10,500**
Ringersma	MPI	1	5,000	5,000	250	**5,250**
Sloetjes	MPI	1	5,000	5,000	250	**5,250**
Student assistant	MPI	2	3,000	6,000	500	**6,500**
Developer	RU	4	6,500	26,000	1,000	**27,000**
Total		**11**		**57,000**	**2,7500**	**59,750**

11. Literature

Crasborn, O. & I. Zwitserlood (2008) *Annotation of the video data in the Corpus NGT*. (Version: November 2008) http://www.ru.nl/corpusngtuk/methodology/transcription/.

Crasborn, O., I. Zwitserlood & J. Ros. 2008. *Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands*. Centre for Language Studies, Radboud University Nijmegen. http://www.corpusngt.nl.

Kemps-Snijders, M., Zinn, C., Ringersma, J., & Windhouwer, M. (2008). Ensuring semantic interoperability on lexical resources. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008).

Ringersma, J., & Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. In H. van Hamme & R. van Son (Eds.), *Proceedings of Interspeech 2007* (pp. 1529-1532). Adelaide: Causal Productions

Schermer, Trude M. (2003). From variant to standard: An overview of the standardization process of the lexicon of Sign Language of the Netherlands (SLN) over two decades. Sign Language Studies 3 (4), 469-486.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods* 41:841-849; doi:10.3758/BRM.41.3.841.