Detailed Research Cooperation Planning in Assisted Humanities and Social Sciences Projects

Editors: Tamás Váradi, Piroska Lendvai, Koenraad De Smedt, Paul Rayson, Lars Borin

www.clarin.eu
The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organisational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) and to secure sustainable support from the funding bodies in the (now 23) participating countries for the subsequent construction and exploitation phases beyond 2010.
Detailed Research Cooperation Planning in Assisted Humanities and Social Sciences Projects

CLARIN-2009-12

EC FP7 project no. 212230


Responsible: Tamás Váradi

Contributing Partners: HASRIL, UIB, UCREL, Göteborg University

© All rights reserved by HASRIL on behalf of CLARIN
**Scope of the Document**

CLARIN provides assistance for nine ongoing projects from the Humanities and Social Sciences discipline. This document describes the procedures that resulted in the cooperation, and provides the detailed research planning of each individual project.
# Common Language Resources and Technology Infrastructure

## Table of Contents

1. Introduction.......................................................................................................................................................... 7

2. Description of Planning in Assisted Projects................................................................................................. 9

   2.1 CKCC: Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic. A Web-based Humanities Collaboratory on Correspondences ....................................................... 9

      Administrative Details ......................................................................................................................................... 9

      Topic ................................................................................................................................................................. 10

      LT components ............................................................................................................................................... 10

      Joint work ..................................................................................................................................................... 10

      Publications .................................................................................................................................................. 10

   2.2 CAP: A Hierarchical Lexical Function related to Proper Nouns ................................................................. 12

      Administrative Details ....................................................................................................................................... 12

      Topic ............................................................................................................................................................... 12

      LT components ............................................................................................................................................... 12

      Joint work ..................................................................................................................................................... 13

      Publications .................................................................................................................................................. 14

   2.3 CONPLISIT: Consumption patterns and life-style in Swedish literature – novels 1830-1860 .................. 15

      Administrative Details ....................................................................................................................................... 15

      Topic ............................................................................................................................................................... 15

      LT components ............................................................................................................................................... 15

      Joint work ..................................................................................................................................................... 16

      Publications .................................................................................................................................................. 16

   2.4 DATIST (Diagnostic And Terminology In Speech Therapy): Studying speech and language therapy diagnosis, using statistical analysis and textual statistics ........................................... 18

      Administrative Details ....................................................................................................................................... 18

      Topic ............................................................................................................................................................... 18

      LT components ............................................................................................................................................... 18

      Joint work ..................................................................................................................................................... 19

   2.5 DID-Cph: Consultancy for the Dictionary of Danish Insular Dialects......................................................... 21

      Administrative Details ....................................................................................................................................... 21

      Topic ............................................................................................................................................................... 21

      LT components ............................................................................................................................................... 22

      Joint work ..................................................................................................................................................... 22

   2.6 HISTOPOL: Narrative Social Psychological Studies of European History ............................................. 24

      Administrative Details ....................................................................................................................................... 24

      Topic ............................................................................................................................................................... 25

      LT components ............................................................................................................................................... 25

      Joint work ..................................................................................................................................................... 26

      Publications .................................................................................................................................................. 27

   2.7 HTM4Emode: Historical text mining for assisting the study of discourses in Early Modern England ............. 28

      Administrative Details ....................................................................................................................................... 28

      Topic ............................................................................................................................................................... 28

      LT components ............................................................................................................................................... 29

      Joint work ..................................................................................................................................................... 29
2.8  LaMeCos: Database and Database Analyzer of Medieval Latin Scientific Terminology

Administrative Details

Topic

LT components

Joint work

2.9  MLT-Cphil: Multilingual language technology for classical philology research

Administrative Details

Topic

LT components

Joint work

Appendix

The Best of Several Worlds: Cross-community knowledge acquisition during research collaboration
1. Introduction

Support for the Humanities and Social Sciences (HSS) research community is central to the CLARIN mission. In the preparatory phase, it is important to build bridges to our target communities and it is vital to gain first hand knowledge about the research interests, methods, and conceptual frameworks based on which HSS scholars conduct their research. This knowledge is used in tailoring the services of the CLARIN infrastructure to suit the needs of our target communities. One way of securing this was to engage in joint collaborative work with a handful of ongoing HSS projects. The actual projects were selected in an open call for collaboration.

Accordingly, at the end of 2008 a Call for Proposals for Collaborating with Humanities and Social Science Projects was issued by WP3 of CLARIN, targeted at research institutions or consortia featuring high research capability, but in need of complementing their expertise and resources with specific language technology (LT) components for advancing beyond the established state-of-the-art. The contribution of CLARIN to this end was meant to consist of guidance about and providing access to language resources and LT, desirable to adopt for the particular objectives of a project.

Evaluation of the received proposals was based on the following criteria. Applicants had to be organizations or consortia from any sector, including at least one organization established in a Member State or Associated Country with respect to the European Commission Seventh Framework Programme. Research activities were to primarily consist of research or technological development within HSS, addressing new and pertinent research questions, and may also have included demonstration activities, designed to prove the viability of new approaches or their exploitation. The research plan should have reached some maturity and should profit from one component involving a specifically described need for language resources technology.

The national representatives for CLARIN provided advice on how well each proposal fit in national priorities. The final decision on selection was taken by the CLARIN Executive Board, on the basis of expert peer reviews. From the total candidate applications CLARIN has received, nine projects were selected. The goals of these pose intriguing challenges to current LT infrastructure, concepts, and standards, and as such are capable of demonstrating the potential of LT infrastructure when applied to data from HSS disciplines. Except for the projects (CONPLISIT, HISTOPOL, HTM4EmodE) that had already some prior collaboration with CLARIN members, each project was assigned to a CLARIN WP3 partner. The assignment of projects to a WP3 supervisor partner is displayed in Table 1.

<table>
<thead>
<tr>
<th>CKCC</th>
<th>UCREL, Lancaster University</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP</td>
<td>HASRIL, Research Institute for Linguistics</td>
</tr>
<tr>
<td>CONPLISIT</td>
<td>Göteborg University</td>
</tr>
<tr>
<td>HISTOPOL</td>
<td>HASRIL, Research Institute for Linguistics</td>
</tr>
<tr>
<td>MLT-Cphil</td>
<td>Charles University, Prague</td>
</tr>
<tr>
<td>DID-Cph</td>
<td>Bergen University</td>
</tr>
<tr>
<td>DATIST</td>
<td>Bergen University</td>
</tr>
<tr>
<td>HTM4EmodE</td>
<td>UCREL, Lancaster University</td>
</tr>
<tr>
<td>LaMeCos</td>
<td>EPF, Barcelona</td>
</tr>
</tbody>
</table>

Table 1. Humanities Projects and their supervising CLARIN partner
In the startup phase of the assistance (May – September 2009), the individual needs of each project were charted, and preliminary negotiations were conducted between CLARIN representatives and the respective HSS researchers. Subsequently, the actual research collaborations between projects and CLARIN partners have started. The cooperation is expected to last till October 2010, following which each project coordinator and CLARIN partner will evaluate and write up the conclusions from their collaboration. These reports will form the basis of the final report on this activity within WP3; it has been agreed on that the projects should aim at disseminating LT results through CLARIN.

Based on the information exchanged during the bilateral preparatory meetings, CLARIN have identified an important issue that originates from the emerging goal-oriented research collaboration between LT experts and HSS researchers. During such interaction distinct disciplines come together, whereby there is a clear need that the representatives of these different communities engage in in-depth dialogue with each other, in order to gain systematic knowledge about each other’s research priorities, methods, domain concepts, and language use. Focusing on this topic, a panel proposal was submitted for the Digital Heritage 2010 Conference, titled "The Best of Several Worlds: Cross-community knowledge acquisition during research collaboration". The detailed text of the proposal, currently pending acceptance, is in the Appendix of the present document.

It holds in general, and is thus not mentioned explicitly at each individual project description, that a major expected impact of the collaboration between HSS projects and CLARIN is that the specific LT expertise to which the HSS participants are exposed shall bring novel insights not only with respect to a project’s data, but also possibly to their data analysis and programming practice, thus to a certain extent LT expertise and knowledge shall be transferred to both HSS and IT specialists. In turn, the LT community will learn about research methods and types of resources utilised in HSS projects.

Specifically, the cooperation shall provide practice for HSS researchers in elaborating on the following issues:

(i) What are common methodologies in LT, what are best practices to locate and to develop basic computational resources for text-based research?
(ii) How could the adaptation of language-specific tools tools (currently typically existing for processing widely-spoken Indo-European languages, e.g. English) be performed to enable processing other languages (e.g., as featured in 17th century scholarly letters from The Netherlands)?
(iii) Along which dimensions and how can one make extensions to existing domain-specific LT approaches for full applicability to data from a project's target domain?
(iv) To what extent can LT techniques (e.g., keyword and concept extraction) be applied to a specific task aimed at by a specific project (e.g., to analyzing knowledge circulation)?

The remainder of this document outlines the context of each of the nine projects, and provides the planned cooperation details.
2. Description of Planning in Assisted Projects

2.1 CKCC: Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic. A Web-based Humanities Collaboratory on Correspondences

Administrative Details

Coordinator/Main applicant:
Name: Prof. dr. W.W. Mijnhardt
Function: Director
Organization: University of Utrecht, Descartes Centre for History and Philosophy of the Sciences and the Humanities
Address: Janskerkhof 3A, 2512 BK Utrecht
E-mail: w.w.mijnhardt@uu.nl
Tel: +31 (0)30-253 8911
Role(s): Executive

Project manager:
Name: Drs. Guido Gerritsen MMO
Function: Project Manager ICT&Text
Organization: Huygens Institute (KNAW)
Address: Prins Willem-Alexanderhof 5, NL - 2595 BE Den Haag
E-mail: guido.gerritsen@huygensinstituut.knaw.nl
Tel: +31(0)70-3315811
Role(s): Project Manager CKCC; Team Manager WP Analysis Tools

Composition of the Project Team:

Name: Dr. E.J Bos, Researcher University of Utrecht, Philosophy Faculty
Role(s): User

Name: Ing. R. Haentjens Dekker, Head Development Huygens Instituut (KNAW)
Role(s): Technology Provider

Name: Drs. Guido Gerritsen MMO, Project Manager ICT&Texts Huygens Institute (KNAW)
Role(s): Team manager; Data Provider

Name: Dr. C.M.J.M. van den Heuvel, Senior Researcher Virtual Knowledge Studio for the Humanities and Social Sciences - VKS (KNAW)
Role(s): Technology Provider

Name: W. Ravenek, Lead Developer Huygens Institute (KNAW)
Role(s): Technology Provider
Name: Dr. D. Roorda, Information specialist Data Archiving and Networked Services - DANS (KNAW)  
Role(s): Technology Provider

Topic

A consortium of Dutch universities and cultural heritage institutions is building a web-based collaboratory around a corpus of 20,000 letters of scholars who lived in the 17th-century Dutch Republic to answer the research question: how did knowledge circulate in the 17th century? CKCC aims to build an analysis environment that visualizes constellation patterns involving people, places and domain concepts occurring in a corpus of letters of selected 17th century Dutch scholars.

LT components

Based on this (extendable) corpus, a content processing workflow will be implemented, which involves several LT approaches, and consists of iterative cycles of (i) conceptual analysis, (ii) enrichment with several layers of annotation, and (iii) their visualization. Details are provided in the below subsections. CLARIN support is solicited for the work package “Analysis Tools”.

Joint work

Completed

Having benefited from counselling by CLARIN in the early phase of the project, a detailed work plan has been jointly prepared that specifies three phases of the cooperation trajectory. Note that an application for extra funding for additional (external) LT and IT specialists has also been jointly filed, and been recently submitted to CLARIN-NL, the Dutch counterpart of the CLARIN project.

Planned

In the first phase of the CKCC project, a demonstrator application will be developed, which implements techniques of keyword extraction (Deadline: 1 October 2010). The result of the extraction procedure is a list of words weighted by their relevance, as well as so-called word clouds. Such output enables statistical characterisation of the corpus by identifying major content-carrying elements in the philosophers' letters, as well as the comparison and linking of individual letters based on the keywords.

The project's goal hereby is to enable and realize visualisation of major moments in knowledge production, such as the emergence of a novel concept in philosophy, an invention, or a scientific discovery. Importantly, LT in this phase of the project needs to address the issue of historical variation in spelling, and dealing with several languages (Dutch, French, Latin, Greek) within individual documents of the corpus.

The second cooperation phase consists of evaluating existing, complex LT tools and techniques that are candidates for enabling the targeted grammatical, content-related, and network complexity
analysis and annotation. This exploratory phase shall identify a set of tools that can be readily utilized in CKCC, as well as tools that need to be adapted or extended to the needs of CKCC.

In the third stage (that starts at the end of phase one and ends in November 2012) the more advanced project toolset is to be developed, focusing on techniques that allow for concept extraction. This is planned to involve (a pipeline of) higher-level text analysis modules, such as named entity recognition, coreference resolution, record linkage, relation extraction, but also possibly the merging of various terminological resources, and modelling of discourse structure. The sophisticated goal to be reached by these means is to track down entities that are used in the texts to express domain concepts (of various levels of importance) and their different epistemic stadia within scientific and philosophical discussions, and to plot their geospatial and temporal dynamics.

The tools tackled within the CKCC project are to be adapted and/or implemented by the IT development team of the Huygens Institute, who will work in close cooperation with the involved humanities researchers (i) towards the establishment of conceptual categories and terms that form the annotation and evaluation schemes and methods, as well as (ii) on assessing the performance of both such schemes and tools interfaces, involving performance checking, and suggestions for adaptation and extension.

One of the main tangible results of the project is a demonstrator application. This piece of software will feature the above techniques in order to visualize various types of named entities (e.g. persons, institutions, geographical locations, dates, scientific or philosophical terms or concepts), and eventually relations among these, in 17th century letters. The demonstrator prototype serves testing purposes, but its enhanced version shall be incorporated in the final analysis and visualization toolset of the project. The developed tools, the demonstrator, and (meta)data will be available via the servers of a KNAW CLARIN Centre (DANS and/or Huygens Instituut).

Publications

2.2 CAP: A Hierarchical Lexical Function related to Proper Nouns

Administrative Details

Coordinator:

Prof. Thierry GRASS ; Université de Strasbourg ; UFR LSHA, LILPA ; 22, rue Descartes ; BP 80010 ; F-67084 Strasbourg Cedex ; Phone : +33 388 417468 ; email : thierry.grass@yahoo.fr

Dr. Amalia TODIRASCU ; Université de Strasbourg ; UFR LSHA, LILPA ; 22, rue Descartes ; BP 80010 ; F-67084 Strasbourg Cedex ; Phone : +33 388 417429 ; email : todiras@umb.u-strasbg.fr

List of partners:

Equipe LiLPa (Linguistique, Langues et Parole) ; Université de Strasbourg ; 14 rue René Descartes ; F-67084 Strasbourg Cedex ; Contact Person : Prof. Catherine SCHNEDECKER; Phone : +33 388 417885 ; email : Catherine.Schnedecker@umb.u-strasbg.fr

Institut de Traducteurs, d'Interprètes et de Relations Internationales ; 22, rue Descartes ; BP 80010 ; F-67084 Strasbourg Cedex ; Contact Person : Eckhart HÖTZEL ; Phone : +33 388 417480 ; email : hoetzl@umb.u-strasbg.fr

Laboratoire d’Informatique de Tours, Université François Rabelais ; 64, Avenue Jean Portalis ; F-37200 Tours ; Contact Person : Prof. Denis MAUREL ; Phone : +33 247 361435; email : denis.maurel@univ-tours.fr

Topic

Several projects study proper nouns to find criteria for an automatic extraction of Named Entities; the CAP project wishes to complete existing classifications and descriptions by proposing a linguistic description that accounts for the ways of expressing the CAP relation in a multilingual context. Two areas of linguistics, lexical semantics and discourse, are targeted, by studying the properties of the hierarchy-founding “head_of” (CAP) semantic relation. The relation CAP refers to nouns that express a function or a position in some hierarchy (e.g., ‘general manager’, ‘chief executive’), but also to verbs expressing a hierarchical relation. This hierarchy-founding relation is considered as a fundamental relation in the structure of the mental lexicon, since hierarchical relations exist in any kind of social organization.

From a syntactic perspective, this relation links common nouns to proper nouns. The phenomenon is so far insufficiently described, and is thus also missing from computational lexical resources, and, consequently from analytic tools. CAP explores corpora in three languages (French, German and English) to this end. The output shall contribute to complete existing linguistic descriptions as well as to extend existing electronic dictionaries, for example to extend the multilingual ontology “Prolexbase”. In addition, the defined lexical patterns are to automatically identify the Cap relation in further HSS studies: sociologists, historians might be interested in analyzing hierarchical structures synchronically or diachronically, and could get linguistic data for their research by using
the lexical patterns the project aims to create. At a European level the use of the patterns that automatically identify the CAP relation could be extended to operate on data that the CLARIN infrastructure intends to provide.

**LT components**

A systematic study of the linguistic expression of the CAP relation, using various monolingual and multilingual corpora, is necessary. The requested advice and technical support pertains to selecting relevant (tagged and lemmatised) corpora in German, English and French. The corpora need to contain documents from the economy or politics domain, featuring named entities (names of persons, organisations, parties, enterprises). Automated named entity recognition is an important component of the research. Expected technical support from CLARIN is guidance with respect to querying textual resources. Since project plans include preparing a scheme for newly created metadata, which are to be published with the resulting tagged multilingual corpus, CLARIN advice about standards and best practices will be applicable.

**Joint work**

**Completed**

A monolingual corpus is prepared by manual selection of articles from the domain of economics and politics, extracted from online newspapers. A new multilingual corpus is compiled from scratch, based on newsletters from multinational enterprises (Swiss, Lufthansa, Air France). Alignment on the sentence level, as well as lemmatisation and enrichment of the parallel corpus is in progress, using the "Unitex" tools.

In order to ensure the use of standardised corpora, additionally a few monolingual and multilingual online corpora were selected from CLARIN’s resource and tool registry (in certain cases also obtaining access to these via CLARIN), which are provided with tagging and lemmatization information, and are available for the three languages, from the economics or politics fields. A context-based bootstrapping method is used to identify pairs of named entities known to be related by a hierarchical relation. Person and organisation names are extracted using CasSys, a free named entity recognition tool.

**Planned**

Exploitation of some of the monolingual and multilingual on-line corpora is planned as follows.

Phase 2, till June 2010:
Analysis of the extracted data in order to describe the CAP relation and to propose examples to illustrate this relation: several linguistic expressions of the CAP lexical function are (being) identified in French, based on which translation equivalents in German and in English are extracted. Classification of new proper noun categories: the linguistic means (affixation, lexicalization, modification, etc.) employed to express the relation CAP.
Development of the web service for the corpus to integrate it into CLARIN resource federation: during corpora analysis, several problems and limitations of the available Web interfaces are noted. These observations are collected, and will be provided as suggestions for defining an improved, uniform query interface to the language resources.

Phase 3, July 2010 – October 2010:
Definition and evaluation of lexical patterns for English, German and French, based on which several proper noun categories are designed, and subsequently added to the Prolexbase ontology.

Publications

### 2.3 CONPLISIT: Consumption patterns and life-style in Swedish literature – novels 1830-1860

#### Administrative Details

Coordinator:  
University of Gothenburg  
HEI  
Göteborgs universitet, Box 100-405 30 Göteborg  
Department of Historical Studies  
Professor Christer Ahlberger  
+46 31 786 4504  
<christer.ahlberger@history.gu.se>

List of partners:

Litteraturbanken (the Swedish Literature Bank)  
Non-profit organization  
Litteraturbanken ideell förening  
c/o Svenska akademien  
Box 2118  
10313 Stockholm  
Professor Mats Malm (Director of Litteraturbanken)  
+46 31 786 5285  
<mats.malm@lit.gu.se>

#### Topic

The project focuses on consumption patterns and lifestyle in Swedish literature, based on 19th century literature. Earlier research targeted the describing and explaining of concrete changes in patterns of consumption during this period, however, there is still little knowledge of the contexts in which the new commodities were used, since the use of consumer items was to a high degree individual. The research program has two specific goals: to deepen the knowledge of the creation of the consumer society, and to develop new methods by the use of literature during the period 1830-1860 as a main source.

In addition to and alongside with the Weberian “protestant ethic”, during the nineteenth century, a corresponding “romantic ethic” was created, the former characterised by cold “profit maximisation” and the latter by fantasies and “satisfaction maximisation”. There is an extremely rich and currently vastly underutilised source material for understanding and analysing the modern consumption society to be found in the modern novel emerging at this time. The concrete question that drives the research program is: what consumption fantasies are expressed in the modern novel?
LT components

In order to accomplish the project goal, it is crucial to analyze a large amount of books. One of the most important means for this is to use full text databases. There is therefore a need for intelligent search tools that go beyond the string/text word search offered currently by the site. The project envisions semantically enhanced queries, and thus asks to develop a semantically organised lexical resource with a morphological analysis component. The project is perhaps best seen as a contribution to methodology development in Humanities e-Science.

There is at present no morphological analysis module for historical language variety in Swedish, but one could be developed fairly quickly on the basis of the modern Swedish morphology module developed for the SALDO modern Swedish lexical resource (the morphology of 19th century and modern Swedish differ in details, but agree in general structure). This would allow text access on the level of lexical entries, rather than text words, already a gain for researchers working with these texts. The 19th century lexical resource is a digitised version of the printed dictionary (i.e., Dalin 1850–1853). In addition, with a limited amount of additional work, most entries in the Dalin dictionary could be linked to the corresponding entries in SALDO, which would provide access to the thesaurus structure of SALDO, enabling a kind of semantic search.

Joint work

Språkbanken has been planning to devise a computational morphology for 19th century Swedish and link Dalin’s dictionary to SALDO, in collaboration with the historians at the University of Gothenburg and Litteraturbanken, and drawing upon expertise and resources available in the CLARIN network. The Dalin dictionary needs further processing in order to make it into an LT resource, including some part-manual, part-automatic cleaning of the structure. The combined resource (Dalin–SALDO with inflectional morphology modules for both language varieties) will come about as a result of this collaboration. It is part of Språkbanken’s plans to harmonize all its language resources internally as well as with regard to emerging standards for language resources and tools. For this reason, the collaboration in CLARIN, especially with the group at MPI Nijmegen on the use of LMF, is very valuable.

Completed

Because of some delay at the start, certain modifications have been made to the order of project activities. Activities concentrated primarily on preparing the basic lexical resource, and deferred the specification of use cases until somewhat later. Thus, the cleaning of the digitised version of Dalin is almost finished, and some experiments have been conducted with automatic methods to integrate Dalin and SALDO). The result is made available on the web under the address <http://spraakbanken.gu.se/swefn/resurser/dalin_saldo.html>.

The literature digitisation effort on which CONPLISISIT is dependent is proceeding apace in the Swedish Literature Bank <http://litteraturbanken.se>. Currently, almost 250 works are available as e-text (i.e., TEI P5 XML), many of which fall into the time period targeted in this project. Thus, there will be a sufficient amount of text material for user testing in the later stages of the project.
Planned

A detailed resource and needs inventory is being carried out, together with planning of activities and specification of scenarios and use cases. Currently, work targets Dalin: cleaning of the lexical resource and definition of the 19th century Swedish morphology is in progress.

Further time plan for the project in 2010 is:

2010-04 – 2010-06:
The integration of Dalin and SALDO, specification of scenarios/use cases, and implementation of a user search interface prototype is planned.

2010-07 – 2010-12:
User testing, as well as the reassessment of scenarios and use cases; preparation of reports and other publications.
2.4 DATIST (Diagnostic And Terminology In Speech Therapy): Studying speech and language therapy diagnosis, using statistical analysis and textual statistics

Administrative Details

Coordinator:
Frédérique Brin-Henry
PhD student in Linguistic Science
Department : Linguistic Science, UFR Sciences du Langage, Université de Nancy2.
ECOLE DOCTORALE «LANGAGE, TEMPS, SOCIETES».
frederique.brin-henry@univ-nancy2.fr

List of partners:
Supervisor (1) : Pr Richard DUDA, Professor in Linguistic Science
  Richard.Duda@univ-nancy2.fr
Supervisor (2) : Dr Hervé ADAMI, Doctor in Linguistic Science
  Herve.Adami@univ-nancy2.fr

Topic

This project aims at validating some hypotheses about inconsistency in the terminology used in speech and language therapy in France when carrying out a diagnostic process. Very naturally, the use of modern means of analyzing language became obvious in dealing with the collected data, a 270 000 word corpus, made out of 433 speech and language therapy assessment reports. This study is part of a PhD research process started in 2006, in which the aim is to demonstrate that making a diagnosis is directly linked to the current conception of pathologies, as it appears through the classifications currently used in Europe.

Close links may be made with philosopher's concerns about the relationship between language, thought and reality. It is investigated what creates the reality of a language disorder, and wish to measure the way terminological evolutions contribute to labelling what then becomes a disorder. Issues on classifications of pathological conditions addressed by speech and language therapy are of interest to researchers of linguistics, psychology and social sciences, as well as of the medical domain. These results would indeed be the base on which a questionnaire aimed at speech and language therapists will be compiled. This questionnaire would support some hypotheses about the link between the terms from the corpus, based on the data and the results from the statistical and textual analyses, exploring speech and language therapists' personal representations of the pathology.

LT components

Help is sought to analyse terminology in the corpus. Metadata in this corpus is little, hindering proper analytic access. The approach aims to extract occurrences, concordances and to evaluate the impact of the adjective in the nuances brought to the main diagnosis.
Joint work

The CLARIN cooperation is executed by UniDigital (formerly Unifob AKSIS), an affiliate partner to the University of Bergen. Øystein Reigem has been contact person and main scientist at UniDigital assigned to this cooperation. Prof. Koenraad De Smedt has been coordinator on behalf of the University of Bergen.

Completed

The cooperation was started in June 2009. After some discussion it was agreed that the text file needed some structure to make computerised processing possible. The file was converted to XML, with a minimum of structural information. Each report got its own element and the content was divided into various sections: "objet", "bilan", "anamnese", "diagnostic", and "plan". The sectionizing was pure manual work; there was neither a common layout to the reports nor any useful explicit markup that could be used for an automatic segmentation into sections.

Then the diagnostic terms were located and tagged, which was again a manual process. The initial manual markup work was performed by Frédérique Brin-Henry, although not in XML, but with a simplified syntax we agreed on, which was fast to use and didn't require her to learn the intricacies of XML; Øystein Reigem wrote a program to convert this simplified markup to valid XML, and made several products that were useful for Frédérique Brin-Henry's study on the reports and their terms. More specifically, XSLT stylesheets were written and used for processing the XML data, and the results put on a web server. After that, there were a few rounds of modifying the reports file and adjusting the web site.

Planned

Help is requested in the automatic transcription of spoken material. A very specific annotation framework needs to be constructed and applied, since the corpus is unique and valuable, but also very highly domain-specific. CLARIN is asked to advice on corpus query tools, as well as to provide training on the statistical program R, and on the use of other statistical tools and methodology to answer the project’s research questions. It is planned to enlighten the links between terms and concepts of language and communication pathology used by French speech and language therapists, leading to more pragmatic research using oral corpora.

The cooperation is now in a phase where Frédérique Brin-Henry works with analyzing the occurring diagnostic terms; she looks at different syntactic and semantic aspects, and classifies the terms into various categories. When she is finished, statistical analyses will be performed of the occurring terms and their categories. Apart from simple statistics, a correspondence analysis will be attempted to see how the terms correspond with other variables. In this statistical work it is likely that Christer Johansson from the University of Bergen will be involved. During the final phase of the cooperation, it will be useful to have a meeting of the participants in order to work together intensely during a couple of days.
Preliminary evaluation of the cooperation

Since the cooperation is interdisciplinary, scientific conversations on the object and methods of the study initially had to be simplified to the core before a plan of activities could be made. This communication process had some positive side-effects. On the one hand, the process helped Frédérique Brin-Henry to be more efficient and down-to-earth regarding the project by forcing the conversion of vague ideas into well-described, operational objectives. These attitudes and thought processes were instrumental in the explicit formulation of concrete needs that initially were desiderata considered possible.

On the other hand, Øystein Reigem's focus on the translation of needs into software use was not passive but led to the discovery of new possibilities that might be useful and therefore new suggestions for looking at the data in new ways. In the end, a common language was found and a fruitful cooperation was established for working with the data, but there is still a lot to do in terms of statistics for finding patterns and correlations.
2.5 DID-Cph: Consultancy for the Dictionary of Danish Insular Dialects

Administrative Details

Coordinator:
Assistant professor Henrik Hovmark
Dir.: +45 35 32 84 97
hovmark@hum.ku.dk
Department of Scandinavian Research
Dialectology Section
University of Copenhagen

List of partners:

DID is participating as observer in DK-CLARIN (WP4 Technological resources – aiming at establishing improved access to, linking between, and usability of dictionaries, wordnets, and corpora through the construction of new technological resources).

Senior researcher Bolette Pedersen
Dir.: +45 35 32 90 78
bspedersen@hum.ku.dk
Centre for Language Technology
University of Copenhagen
Njalsgade 140-142
DK-2300 Copenhagen S
Denmark

Associate professor Viggo Sørensen
Dir.: +45 89 42 19 16
jysvs@hum.au.dk
The Peter Skautrup Centre for Jutlandish Dialectology
Aarhus University

Professor Frans Gregersen
Dir.: +45 35 32 83 54
fg@hum.ku.dk
The LANCHART Centre
University of Copenhagen

Senior researcher Lars Trap-Jensen
Dir.: +45 33 13 84 20 (+9638)
ltj@dsl.dk
Danish Society for Language and Literature
Common Language Resources and Technology Infrastructure

**Topic**

DID is a well-established, major lexicographic project designed to capture cultural heritage, i.e., ethnographic descriptions, by describing the Danish dialects on Zealand, Funen, and surrounding islands in Denmark. It covers the period from 1750 to 1945, but the core period is 1850 to 1920. The language recorded in DID is the spoken vernacular of peasants and fishermen, dialecticisms as well as words in common with standard Danish. Due to the status and nature of the dictionary, emphasizing the close connection between language and culture, and integrating linguistic and ethnological knowledge, DID prepares a rich resource to scholars within a number of disciplines (linguistics, history, cultural anthropology).

**LT components**

The data are extremely diverse, due to the long period of data collection. DID is a dictionary of spoken language, which has to be represented by a standardized phonetic notation. DID uses Dania, the traditional Danish standard phonetic notation. The different kinds of data on the specific words and word forms are extracted manually (paper slips) as well as electronically (transcribed tape recordings). Two important LRT tools are currently in use by the project: an editor, T3, which has allowed for systematic and efficient work with and printing of the phonetic notation Dania, and a corpus managing tool, WordCruncher, which is used to handle the transcribed tape recordings, making it possible to search for words and collocations.

At present, DID is ready to update its use of modern technology in order to be able to continue editing and to make the huge amount of information accessible. Updating of LT platforms and tools has high priority in order to sustain exchange and co-operation with colleagues in Denmark and internationally, especially the Nordic countries. An upgrade of the technological platform would make this knowledge more accessible to for instance museums, local historians or fellow researchers abroad.

**Joint work**

A new editor based on a common standard like XML is needed, alongside with the conversion of the T3 format into modern standard formats. Working with Dania has special typesetting requirements. DID would welcome support and guidance from CLARIN in the process of providing a new and updated corpus query tool allowing for multiple queries and further development of the corpus of transcribed tape recordings. The ethnological descriptions are planned to be linked with an existing image database. Merging of terminology across resources is also necessary for improved access. Advice regarding standards and best practice is requested.

**Completed**

Detailed description of daily editing workflow and demands at DID; detailed descriptions of data set; exchange of data.
Preparatory meeting: definition of work tasks and schedule; consultancy on standards and best practice by CLARIN advisors.
T3 and WordCruncher: product specification of new editor/programme; definition of test phase and procedure.

Planned

December 2009 – January 2010:
Contact with specific CLARIN advisors, exchange of detailed information about possible LT solutions; definition of pilot project, including tests.
T3: definition of test sample, definition of testing (editing, typing, keyboards, typesetting, Dania font; proof-reading, cross-referencing, correction of misprints; conversion into different formats, printing); check-up on existing editors (features, availability).
WordCruncher: definition of test sample and testing (typical searches, text retrievals and statistical uses; printing); check-up on existing text retrieval programmes

February – May 2010, Work phase I:
prototype implementation of possible LT solutions; possible visits by CLARIN advisors.
T3: new editors at DID; advice by CLARIN; testing; attempt at converting T3 data.
WordCruncher: new programmes at DID; advice by CLARIN; testing; importing/exporting data.

May 2010, follow-up meeting:
Check up on problems and progress; general advice by CLARIN partners on LT tools and future LT initiatives at DID.
T3 and WordCruncher: conclusion on first test results; choice of final editor/programme candidates

June – September 2010, Work phase II:
further implementation of LT tools.
T3 and WordCruncher: testing.

September 2010, Final meeting:
results; final advice and input by CLARIN advisors; evaluation.
2.6 HISTOPOL: Narrative Social Psychological Studies of European History

Administrative Details

Coordinator:
Institute for Psychological Research of the Hungarian Academy of Sciences
H-1132 Budapest, Victor Hugo str. 18-22, HUNGARY
Department of Cognitive Social Psychology
Contact Person: János László
Phone: +36 1 239 4740
Fax: +36 1 239 6727
Email: laszlo@mtapi.hu

List of partners:

Université de Toulouse II - Le Mirail
Organization Type: University and Research Institute
Legal Address: 5 allées Antonio Machado 31058 Toulouse, France
Department of Unit: Institute of Psychology
Contact Person: Denis Hilton
Phone: +335 61504981
Fax: +335 61 50 42 09
Email: Hilton@univ-tlse2.fr

Johannes Kepler Universitaet, Linz, Institut für Paedagogik und Psychologie
Organization Type: University and Research Institute
Legal Address: A-4040 Altenbergerstrasse 69, Linz
Department of Unit: Abteilung für Sozial- und Wirtschaftspsychologie
Contact Person: Wolfgang Wagner
Phone: +43 732 2468 8577 (Office); +43 732 2468 8225 (Secretary)
Fax: +43 732 24689315
Email: w.wagner@jku.at

Université Libre de Bruxelles
Organization Type: University and Research Institute
Legal Address: 50 Avenue Franklin Roosevelt; B-1050 Bruxelles
Department of Unit: Unité de Psychologie Sociale
Contact Person: Laurent Licata
Phone: +32 2 650 32 37
Fax: +32 2 650 40 45
Email: licata@ulb.ac.be

Università degli Studi di Roma "La Sapienza"
Organization Type: University and Research Institute
Legal Address: Palazzo Ateneo, Via Crisanzio 1, I-70110 Bari
HISTOPOL investigates how shared historical events are narrated in ways that promote or hinder empathy and understanding across borders. It targets the HSS subject areas of history, social psychology, narrative psychology, and language, and aims at a user group of historians, social psychologists, social scientists, teachers, media experts and politicians. CLARIN support is solicited for the implementation of a work pack called "Representation of history in European textbooks and novels" of a full proposal submitted to the EU 7th Framework Programme. Unfortunately, this proposal was rejected; therefore the project is being carried out at a reduced size and pace.

HISTOPOL focuses on how the past is remembered and narrated by citizens themselves, investigating several factors that participate in this process, thereby statistically testing social psychological theories of identity construction (relation between historical representations and identity, and real life inter-group conflicts).

LT components

The methodology, which draws on verbal behaviour as registered in written sources, is narrative Psychological Content Analysis (NPCA), which is able to map the linguistic and compositional/structural features of texts (self-narratives of individuals and/or societal groups/nations) in order to explore psychological processes of identity formation. HISTOPOL targets the use of the NooJ tool, in which experts can manually define lexico-syntactic patterns to express a domain concept that is judged significantly content-bearing in some context.
**Joint work**

**Completed**

As a computational resource, the researchers were provided with a verb frequency list from HASRIL. The extension of the lexicon and morphology has been undertaken, tailoring the tool more towards the specific domain, and the specific resources used by HISTOPOL (e.g., lemma-based encoding of semantic properties), which now provides improved coverage for the psychology domain than originally foreseen in NooJ. The HISTOPOL team's request about a semantic disambiguation module was realised by the NooJ developers.

**Planned**

In the coming phase, frame-like constructions of verbs are going to be constructed, alongside with the encoding of thematic roles of verbs, to achieve richer semantic modelling and analysis in NooJ. CLARIN will inform the researchers about automatic means to construct pattern-based rules, as well as about automatically extracting relevant domain-specific lexicon of entities and relations from texts, and the implications on trade-off in recall and precision that is to be expected between the manual and automatic methods. Advice will be given to bring more methodological soundness in the LT part of the research by investigating the possibility of automatically setting up stopping criterions in the generation of rules.

Another aspect the HISTOPOL researchers are going to be trained on is the requirements and preparation of a so-called gold-standard reference corpus, which is currently is missing from the project, serving the purpose of properly measuring the impact of changing one or another factor in a model, as well as the proper tracking of changes in lexicons and other utilized resources. The role of CLARIN hereby is going to be to guide investigators about identifying, compiling, and preparing (i) statistically representative text samples as input data in such studies, as well as (ii) parallel monolingual and multilingual texts. The latter can be based on different editions of certain schoolbooks and historical novels, respectively the translations of these.

Yet another of the project’s objective is to provide insight into how social emotions such as guilt, shame and regret operate in the formation of identity and elaboration of the past. CLARIN will point out LT studies conducted in the field of research on emotion, affect, and factuality -- as well as available resources from these, such as SentiWordNet --, and the necessity to compare and measure the overlap between employed features as designed by LT respectively HSS specialists in these studies. Time should also be invested in ensuring that domain concepts, terms, synonyms, and related taxonomic artefacts are properly treated.

For illuminating the role of representations of the past on present inter-group relations and conflicts, the plotting of tendencies along a timeline is an informative method, CLARIN shall thus point out relevant infrastructure with which both the synchronic and diachronic aspects of these concepts can be investigated.
Publications


2.7 HTM4EmodE: Historical text mining for assisting the study of discourses in Early Modern England

Administrative Details

Coordinator:
Dr Michaela Mahlberg, Lecturer, School of English, University of Liverpool, Cypress Building, Chatham Street, Liverpool L69 7ZR
Tel.: +44 (0)151 7942729   Email: m.mahlberg@liverpool.ac.uk

List of partners:

Professor Justin Champion, Department of History, Royal Holloway, University of London, Egham, Surrey TW20 0EX
Tel.: 01784 443749  Email: j.champion@rhul.ac.uk

Dr Mike Scott, Reader, School of English, University of Liverpool, Cypress Building, Chatham Street, Liverpool L69 7ZR
Telephone: +44 151 794-2928 or 2735  Email: Mike.Scott@liv.ac.uk

Alistair Baron, PhD student in UCREL research centre.
Department of Computing, Infolab21, Lancaster University, LA1 4WA, UK.
Tel: +44 (0)1524 510348 Fax: +44 (0)1524 510492
Email: a.baron@comp.lancs.ac.uk

Topic

The work in this proposal is divided into two driver projects. Each one aims at exploring historical discourse from different perspectives with different research questions in mind. The driver projects will explore the possibilities of using historical text mining by applying corpus linguistics techniques to the very large-scale language resources available in Early English Books Online. The EEBO collection represents virtually all books published in the English-speaking world between 1473 and 1700, the period during which Modern English emerged fully from later Middle English and the printing press as a means of mass distribution of discourse emerged. The latest update of the corpus, received in December 2007, contains approximately 540 million words. The main beneficiaries of the projects will be early modern historians and students in the field, who will be introduced to new methodologies that exploit the increasing availability of computer-readable texts.

Driver project 1:
The research in this project aims to study key concepts in early modern religious discourses in a cross-disciplinary approach that draws on current trends in corpus linguistics and historical discourse analysis. The focus of this study will be on the question of how early modern society dealt with and related to the religious "other" with respect to the languages used in early modern England both in relation to Muslims and Jews, but also towards a variety of Christian groups labelled religious
‘dissenters’, ‘non-conformists’, ‘sectarians’, ‘fanatics’ or more specifically ‘puritans’ or ‘papists’, or towards critics of religion variously referred to as ‘anabaptists’, ‘blasphemers’ or even ‘atheists’.

Driver project 2:
The aim is to investigate visualisation issues along the lines of those proposed in driver project 1, facilitated by the extraction of place names within the texts and then geo-referencing each name, in order to explore spatial patterns within the corpus.

LT components

Spelling varied widely in the Early Modern Period and for the application of quantitative methods regularised spellings are essential to improve the quality of the results. Support from CLARIN is solicited on the technical aspects related to the analysis of historical spelling variation. This aspect is a fundamental requirement for the proposed research as listed below and will enable the better application of corpus and computational linguistics techniques such as frequency profiling, key words and annotation (e.g. POS tagging) to historical data. Previous studies have shown that spelling variation has an adverse affect on the accuracy of part-of-speech tagging and semantic analysis when automatically annotating historical corpora and frequency and key words analyses. Spelling variation also appears in place names.

Joint work

Lancaster University provides support as part of their ongoing work on the CLARIN project. Support in kind from other CLARIN partners is expected via the provision of language resources, e.g. lexicons and historical corpora in Early Modern English, to assist directly in the training of tools for the analysis of historical spelling variation. In addition, if such historical resources can be provided in other languages, it may provide a substantial basis on which to explore similar HSS research questions in other national contexts.

Unfortunately, the original ESRC applications were rejected and a revised submission is delayed, primarily due to the Principal Investigator’s maternity leave and move to the University of Nottingham.

Completed

The EEBO data preparation is progressing. A copy of the final EEBO-TCP version was secured. These files have been cleaned and indexed. Current activities include the implementation of an iterative development plan to discover and solve problems with the EEBO encoding and metadata formatting and structuring, i.e. to make necessary changes in VARD to enable successful processing of EEBO data.

Planned

Driver project 1:
Methods that will be applied in this project combine historical scholarship and corpus linguistic techniques. Labels for religious groupings of the time (e.g. dissenters) will be identified by historians, and will form the starting-point for concordance and collocation analyses. The method of computer key words will be employed to analyse specific discourses as identified on the basis of historical scholarship. One of the main corpus tools to be used is WordSmith Tools.

The VARD tool will be used for the pre-processing of the data as follows in order to carry out further analysis. Requirements for CLARIN cooperation include the sourcing and provision of training materials such as Early Modern English corpora possibly marked with spelling variants and their modern equivalents. Lexicons of spelling variants would also be useful. Further, the provision of equivalent lexicons and corpora for non-English historical data would enable investigation of similar techniques for those languages.

Steps to be taken:
Devise a parallel processing framework to enable the corpus to be processed on multiple computers; Incorporation of contextual rules within VARD to enable ‘real-word errors’ to be detected and corrected; evaluation of initial results and carry out further manual training of VARD as required; pre-processing of full EEBO data set and incorporation of encoding so that original variant and standardised forms can be available for other tools such as WordSmith; indexing of the full EEBO data in a corpus query index system such as CQPweb.

*Driver project 2:*
Problems of place-name disambiguation need to be tackled using gazetteers and contextual information. To be explored techniques: semantic tagging, to identify the theme that is being discussed about each place, such as war, famine, trade, and so on. The results however depend crucially on the initial normalisation of the spelling.
2.8 LaMeCos: Database and Database Analyzer of Medieval Latin Scientific Terminology

Administrative Details

Coordinator:
Legal organization: Universitat de Barcelona; organization type: University; legal address: Gran Via de les Corts Catalanes, 585, 08007 Barcelona; Department or Unit: Department of Latin Philology; contact person: Dra. Rosa Maria Comes Maymó; telephone: 34 93 4020491; email: rcomes@ub.edu

List of partners:
Being part of the MedievalCosmos project “Initial Stages of the Transmission of Science to Medieval Europe, the LatinMedievalCosmos project is individual and will be led by a single Principal Investigator, although it will consider a working team formed by pre- and post-doctorate fellowships.

Topic

The project means to design a database and a database analyzer of medieval Latin scientific terminology, based on the scientific astronomical language of Plato of Tivoli’s translations, with the aim of providing scholars working in the field of History of Science, Historical Linguistics, Anthropology, Sociology and Social Sciences with new resources. The database is planned to cover the scientific terminology used in the Middle Ages, as well as to contribute to the revaluation of scientific and technical knowledge in medieval Europe, taking as a starting point the edition of translations, crucial for the transmission of knowledge and ideas.

LT components

The resource shall include two components:
(i) An open source database in a relational database management system, providing online multi-user access. Metadata will include author, date, thematic area, etc., also enabling data input by authorized scholars, so that it could house, in the future, a number of medieval Latin scientific lexicons.
(ii) A database query tool, enabling access to different types of databases, to support surveys on medieval scientific Latin. It needs to implement features of a multidisciplinary approach.

Joint work

Planned

Setting up the database and database analyzer of medieval Latin scientific terminology. Afterwards, the entry of data is carried out.
2.9 MLT-Cphil: *Multilingual language technology for classical philology research*

**Administrative Details**

Coordinator:
University of Hamburg, Faculty for Language- Literature and Media, Institute for Greek and Language Philology
Von-Melle Park 6, 20149 Hamburg, Germany
Contact person: Dr. Cristina Vertan
Tel: +494042838 4767
Email: cristina.vertan@uni-hamburg.de

List of partners:

Prof. Liviu Franga, Vice-Dean
Classical Languages Dept.
Foreign Literature and Languages
Bucharest University
Tel: +40745 76.24.75.
liviu_franga@yahoo.com,
http://www.unibuc.ro/ro/cd_livfranga_ro

Prof. Gregory Crane
Classical Department, Tufts Distinguished Faculty
Email: gregory.crane@Tufts.edu
http://ase.tufts.edu/faculty-guide/fac/gcrane.htm

Prof. Wendy Scase
Department of English, University of Birmingham, United Kingdom

Dr. Kiril Simov
Bulgarian Academy of Sciences

Prof. Dr. Stelios Piperidis
ILSP, Greece

**Topic**

MLT-CPhil is an ambitious and complex project that aims to maximize use of LT in multilingual access to ancient manuscripts. The Teuchos project forms the frame for the proposal, being an infrastructure project to offer the user community a platform for classical philology, to publish ancient and classical manuscripts as well as accompanying texts and research papers, and to offer researchers the opportunity to exchange ideas via a forum. It needs to address at least five modern languages, alongside ancient Greek and Latin, and be able to support documents from other
languages as well. The platform enables users to view, search, and enter comments about different manuscripts, or provide manuscript descriptions, point to research papers, and so on.

**LT components**

The platform hopes to offer multilingual services to HSS researchers, defined through two components: (i) Multilinguality in communication between users: for the modern target languages “flat”-processing is implemented, such as extracting keywords, providing dictionaries, extracting noun phrases (German, French, English). (ii) Intra-document multilinguality at the level of primary data: manuscripts are accompanied by modern descriptions, critical texts, which, although written in modern languages, often contain passages from the manuscript, or Latin citations. On top of all, information retrieval from heterogeneous datasets is a complex challenge in itself.

All documents tackled in the project are annotated with metadata automatically, but only at the level of distinctively marked structural elements of texts (e.g. well-defined chapter-headers in manuscript descriptions). Further elements to be annotated are: names of watermarks, and their relations with already digitised watermarks; names of copyists, persons involved in the manuscript creations; names of other related manuscripts; time references. Access to named entity recognition tools and shallow parsers are requested for German, English, and French. On ancient Greek and Latin, consultancy from the authors of the tools on these languages is needed. Technical help is solicited on stand-off annotation of documents.

The second aspect necessitates expertise in the area of text mining and semantic processing, for managing complex relationships between objects (even in a mono-lingual setting), including record linkage across heterogeneous data. It is envisaged that only a semantic annotation of the objects and relationships based on semantic web methods can help in modelling the data (extracting semantic relations, annotation of objects with concepts, semantic web technologies, keyword extraction, linking to linguistic and domain ontologies). CLARIN is asked to help with mapping and integration of wordnets, and in particular EuroWordNet as a kind of interlingua.

The third layer involves advice on best practices in IR and query interfaces using natural language, also for the management of the multilingual content. This might imply ontological search.

**Joint work**

The project is delayed, all activities are currently paused. The original planning scheme is included in the next section.
<table>
<thead>
<tr>
<th>No.</th>
<th>Activity</th>
<th>Required CLARIN support</th>
<th>Deliverable</th>
<th>Deadline</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>Semantic modelling of the repository</td>
<td>Expertise in Semantic annotation, WordNet and FrameNET, semantic web; best practices in the domain; keyword extractors, for languages L1</td>
<td>Demonstrator showing possible relations between documents and how these are changing when modifying some documents, adding a forum Best-practice report.</td>
<td>5 Months November 2009 - March 2010</td>
</tr>
<tr>
<td>3.</td>
<td>Cross-lingual search engine</td>
<td>Best practice in CLIR domain (especially CLEF) and query interfaces</td>
<td>Demonstrator Documentation</td>
<td>March 2010 - July 2010</td>
</tr>
<tr>
<td>4.</td>
<td>Integration of lexical material for Romanian</td>
<td>Access and integration to Romanian lexical ressources,</td>
<td>Documentation, Integration of lexical material into our system</td>
<td>July 2010 - September 2010</td>
</tr>
</tbody>
</table>
Appendix

The Best of Several Worlds: Cross-community knowledge acquisition during research collaboration

Panel Proposal submitted to the Digital Heritage 2010 Conference

In recent years there has been growing recognition of the importance of research infrastructures as the pillars of modern-day e-science. Language technology has, accordingly, found a new mission in addressing the needs of the wide-ranging disciplines of the Humanities. Central among the objectives of the CLARIN project is the building of a general and interoperable computing infrastructure (e.g., access to digital archives and corpora, language and speech processing tools and standards) especially for the benefit of HU researchers that lack the technological basis to operate existing resources or create new ones. No less important than the technical task of building the infrastructure proper and populating it with tools and resources, is the task of ensuring that CLARIN makes sufficient impact within the Humanities field. This involves a two-way dialogue where CLARIN may advise Humanities scholars about best practice, standards, availability of suitable resources and tools on the one hand and in return it can learn from Humanities about its prevalent research paradigms, methods, conceptual frameworks, etc.

In such cases close, goal-oriented research collaboration between language technology experts and Humanities researchers takes place, where distinct disciplines come together, for example philosophy or sociology, and computational linguistics. It is very important that the knowledge of the different communities involved is hereby optimally merged, to facilitate research and maximise the success of the given project. One crucial aspect in this is that specialists belonging to the different communities need to make sure they understand each other's goals, methods, and, last but not least, terminology. This requirement however necessitates obtaining several insights about a new domain for all parties, which can be realised only by consciously investing time and effort in this process.

There is a clear need at several phases of a research project that representatives of different communities engage in in-depth dialogue with each other in order to gain new knowledge about each other's research foci, priorities, domain concepts, and language use. The establishment and practical implementation of this important layer of collaboration call for the identification of prerequisites that allow for optimally managing and realising transdisciplinary research, involving simultaneous coordination and evaluation of step to be taken during what we see as recursive exposure to knowledge from a new domain for both HU and LT specialists.

It was decided that this cross-community knowledge acquisition is best tried in actual research collaboration with a handful of projects selected from the Humanities fields in an open call. As a result, CLARIN is currently engaged in research collaboration with nine projects at various levels of intensity. By the time the panel takes place this collaboration will be drawing to the end of its first phase, which provides a good opportunity to discuss and share the experiences accumulated and formulate some conclusions that may be of interest to participants of the DH2010 conference.

The panel seeks to explore the following questions:
- What types of dialogue arose between Humanities scholars and language technology specialists during the process of getting acquainted with each other's field (forms of communication: interview, tutorials, pointing to literature);
- What LT methods were used for fast modelling of a Humanities domain (knowledge acquisition and representation) for getting a quick overview, an alternative for reading up on a domain from textbooks and publications?
- Which Humanities research methodologies should language technology expert be aware of (e.g. available tools for automatically processing data on a certain linguistic level implies a must for applying the tool, even if HU researchers initially think this information is irrelevant for their project);
- To what extent can one expect a non-specialist colleague to understand domain phenomena: pitfalls to avoid and best practices to implement during a project's lifetime (negotiating about task sharing and responsibility between LT and HU specialists; personnel assessment and recruiting);
- How to ensure the same language is spoken: consistency checking and agreement on terminological usage;
- What is the workflow of cooperation between specialists of different fields (investing in a longer initial learning process will pay off in the long run? recursive implementation based on predefined development phases?)
- Do new research paradigms emerge from cross-fertilization among communities (success stories).

The panel will be moderated by Tamás Váradi (Hungarian Academy of Sciences), member of the CLARIN Executive Board, who is coordinating the research collaboration with Humanities project within CLARIN.

Members of the panel:
Paul Rayson, UCREL, Lancaster University, CLARIN advisor to Humanities projects
Koenraad De Smedt, University of Bergen, Head of Humanities Projects Selection Committee, CLARIN
Guido Gerritsen, Huygens Institute, The Hague (CKCC project)
Amalia Todirascu, University of Strasburg (CAP project)
Frédérique Brin-Henry, University of Nancy (DATIST project)