CLARIN

A European Research Infrastructure

Newsletter

Number 4, 2008, December

And the Winner Is...?

Dan Tufiş
Romanian Academy of Sciences, Bucharest

The last decade has seen significant advances in machine translation, cross-lingual information retrieval and multilingual information distillation areas. This has provided research policy makers with the evidence that the time is ripe for large-scale concerted actions towards eliminating the language barrier in the Information Society. DARPA's GALE project (2005-2010) was a major sign of this awareness. The project is focussed on two languages and the ultimate performance targets are to translate Arabic and Chinese speech and text with 95% accuracy and to extract and deliver key information matching or exceeding humans' proficiency.

European view

In 2005, the European Commission, facing a much larger language diversity, launched a new policy framework, called "i2010 – A European information society for growth and employment", embracing all aspects of the information, communication and audiovisual sector. The first objective of i2010 is to establish a Single European Information space offering affordable and secure high-bandwidth communications, rich and diverse content and digital services. The EC Competitiveness & Innovation Framework Program, established in October 2006, has been given the responsibility for the development and the implementation of the Single European Information Space where the language barriers should exist no more. With the establishment of the Language Technologies and Machine Translation unit (INFSO.E1) in July 2008, the research and development related to language and multilinguality got a new momentum, and brought machine translation and the multilingual web onto the priority list of the EC recent and future calls for project proposals.

The role of CLARINers

The CLARIN project, bringing together specialists from 148 institutions in 32 countries, probably represents the largest coordinated world community in the area of language technologies and language resources. As such, one would expect a very pro-active reaction of our community to the new initiatives of the European Commission.

CLARIN members presenting at the Language Technology Days, Luxembourg

The Call 4 of the ICT2008, under the Challenge 2 – Objective 2.2 "Language-based Interaction", published in November last year, establishes very ambitious targets for the European language research and development community. Serious discussions have already started for mounting highly competitive consortia able to submit winning proposals.

Many CLARINers attended the Language Technology Days in Luxembourg 14-15 January, en event which provided very useful hints from EC officials and opportunities for the last moment contacts for consortia forming or adjustment.

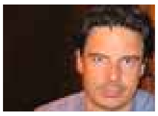
Another highly relevant language centered event will be the Call 3 of the ICT Policy Support Programme within the Competitiveness & Innovation Framework Program (CIP ICT-PSP Call 3) expected for February this year (the ICT-PSP Info Day will take place on 26 January in Brussels). The Draft ICT-PSP Work Programme 2009, under Theme 5: Multilingual Web, pragmatically and very clearly defines the major requirements and expected impact of the work to be funded within this call. The ICT PSP addresses technology and non technology innovation that has moved beyond final research demonstration phase. The ICT-PSP does not support research activities, although it may cover, when needed, technical adaptation and integration work in order to achieve the target objectives. Many of the advanced application-oriented objectives in CLARIN can be found in the ICT-PSP Work Programme.

CLARIN and Single European Information space

Although it has no overt plans for MT, several usage and workflows scenarios of the CLARIN WG5 are built on MT services. My strong conviction is that project proposals generated by consortia formed within the CLARIN community could be not only highly professional and thus eligible for funding, but would also create a synergic long term basis for the construction phase of the CLARIN infrastructure. The multilingual web of the Single European Information Space cannot be dissociated from the common language infrastructure aimed at by the CLARIN project. And vice-versa, the CLARIN-envisaged multilingual services cannot ignore the vision and priorities of the Single European Information Space. C

[1] http://ec.europa.eu/information_society/activities/ict_psp/documents/ICT%20PSP%20WP2009%20-%20v21nov08.pdf

Editors' Foreword



Marko Tadić
& Dan Cristea
CLARIN Newsletter editors

Dear readers,
At the end of the first year of the project, it seems to us that CLARIN has passed the maturity exam and is already on the lips of every builder of LT resources or tools in Europe. It is perhaps less known among the "consumers of LT", which gives us a good hint about an area in which progress is clearly wanted.

The central topic in this issue is the new calls of interest for CLARINers. Indeed, these very days, new calls for research, have been launched in Europe, and we wanted to offer to our readers a short guide, seen through CLARIN eyes, to help you orientate through the calls, understand their philosophy, and prepare the making of consortia for applications.

With this context in mind, we have invited the Scientific Board member Dan Tufiş to open the issue by presenting the EC research horizon, as it is reflected in the new i2010 policy framework. Then, as CLARIN itself launches a call for collaboration with colleagues from the humanities and social sciences, we have asked Koenraad de Smedt and Tamás Váradi to comment on the context of this call. The text of the call is also included.

Closely connected to this call is how to show HSS researchers the advantages that they can gain from CLARIN in their daily research activities. A great challenge of the CLARIN preparatory phase is the definition of representative usage scenarios. This CLARIN activity, which is under development right now, is presented by Valeria Quochi, Lothar Lemnitzer and Marc Kemp-Snijders.

The space in the middle of the issue is, as usual, dedicated to new events which are considered important from the CLARIN perspective. Two very successful workshops organised by the end of 2008 and a European-level meeting, which happened in January 2009, occupy this space. The workshop organised by CLARIN as a WP2/WP5 joint event in October is commented on by Erhard Hinrichs and Peter Wittenburg, the D-Spin-CLARIN joint Workshop in Munich is illustrated by Núria Bel and Marc Kemp-Snijders, while Marko Tadić describe the Language Technology Days in Luxembourg.

Finally, the last several pages are reserved for reports by collaborators from CLARIN partners that bring the news from their countries. In this issue these are Hanne Fersøe, who is commenting the Danish CLARIN project, Eva Hajičová, giving a report on the Czech NLP tradition and horizon, and Dan Cristea, presenting the state-of-the-art in LRT in Romania.

Enjoy the reading. C

Call for contributions

Dear readers of the CLARIN Newsletter,
If you have ideas, thoughts, comments, additions, corrections, arguments, questions etc. which are connected to the CLARIN project, even remotely, please feel free to send them to us as your contribution at newsletter@clarin.eu.

List of national correspondents

Austria
Gerhard Budin

Belgium – Flanders
Inneke Schuurman

Bulgaria
Svetla Koeva

Croatia
Marko Tadić

Czech Republic
Karel Pala

Denmark
Bente Maegaard
Hanne Fersøe

ELRA/ELDA
Stelios Piperidis
Khalid Choukri

Estonia
Tiit Roosmaa

Finland
Kimmo Koskenniemi

France
William Del Mancino
Bertrand Gaiffe

Germany
Lothar Lemnitzer

Greece
Maria Gavrilidou

Hungary
Tamás Váradi

Italy
Valeria Quochi

Latvia
Andrejs Vasiljevs

Malta
Mike Rosner

Netherlands
Peter Wittenburg

Norway
Koenraad De Smedt

Poland
Maciej Piasecki

Portugal
Antonio Branco

Romania
Dan Cristea
Dan Tufiş

Spain
Nuria Bel

Sweden
Sven Strömqvist

UK
Martin Wynne

Koenraad De Smedt
University of Bergen
Tamás Váradi
CLARIN EB member

The CLARIN project has reached an important milestone at the beginning of 2009. It has launched a call for collaboration with humanities and social science projects. It intends to invite applications from projects (actual or in a mature state of development) that may benefit from applying language technology. In the collaboration CLARIN will offer language resources and tools and provide advice on how these may be used to enhance the work of the humanities or social sciences project.

As it is an important document intended for the widest circulation to reach its target audience, we are publishing the text of the proposal below. Hopefully, the call will lead to fruitful cooperation that will benefit the research projects and, through the experiences gained from the collaboration, the CLARIN project as well. The call is also available at the following URL: http://www.clarin.eu/wp3/wp3-doments/call_final-version.

Context

The CLARIN project (http://www.clarin.eu) is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable. CLARIN will ultimately offer scholars the tools to access and process language data, addressing one or more of the multiple roles language plays (i.e. as carrier of cultural content and knowledge, instrument of communication, component of identity and object of study) in the Humanities and Social Sciences (HSS).

One of the objectives in the CLARIN project's current preparatory phase is an in-depth assessment of the technological, methodological, and organizational requirements involved in serving research communities. While language resources and technology (LRT) have great potential in facilitating research in many areas, the Humanities in particular are still an area where few communities have large-scale experience in exploiting the benefits. In the light of this situation, it is essential for the CLARIN project to establish an active interaction with HSS research communities at large through crucial direct experience about user needs, objectives, data and methods used in their research.

The chosen means for gaining this experience is through actual collaboration with HSS col-

Call for Proposals for Collaborating with Humanities and Social Science Projects

leagues in some well chosen projects, the initiative for which must primarily come from HSS researchers themselves. CLARIN is committed to the idea of collaboration with HSS projects on a suitable scale as the best means of identifying needs and removing any potential obstacles from the way of future synergies between the two fields.

Targeted applicants

The current call is targeted at research institutions or consortia with a high research capability, but who need to complement their own expertise and resources with specific LRT components in order to achieve an advance beyond the established state-of-the-art. CLARIN participation is intended to strengthen an already good research plan by providing LRT support. Projects may have theoretical or applied relevance. Projects should not involve core LRT development, but should use LRT to enhance HSS research questions.

Example 1: A literature project wants to study censorship in translation. It has access to uncensored and censored translations of novels. To support the analysis, the project may benefit from producing a searchable parallel corpus where different versions of each sentence are aligned. CLARIN participation could involve access to a corpus alignment tool and transfer of skills in using the tool.

Example 2: A history project wanting to study cultural attitudes in Medieval Northern Europe wants to search through runic inscriptions. CLARIN participation might assist in locating existing digitized corpora of runes in different countries and providing assistance for converting the different materials to a common encoding.

CLARIN consultancy and technical support will be provided at various stages of the execution of the accepted projects. Especially the following target groups are addressed:

1. Groups of individual researchers with basic institutional funding
2. Early stage researchers in funded PhD positions, with their supervisors
3. Research groups or consortia in an advanced pre-proposal stage with prospects of external funding
4. Research groups or consortia that have secured external funding.

CLARIN intends to select between 5 and 10 projects in a range of different Humanities and Social Sciences fields.

Support provided

Although CLARIN does not have the budget to directly finance Humanities projects, it will provide consultancy and technical support to selected projects that are otherwise financed but lack the necessary resources and expertise to enhance their activities with LRT. The contribution of CLARIN to selected projects will therefore consist of providing guidance and access to LRT. This will involve advice on standards and the technologies to adopt for the particular objectives of the selected projects. CLARIN technical support and consultancy will be given through various schemes, including, but not limited to, the following:

Pre-Proposal Closing Date: February 15, 2009

Pre-proposals are invited for Humanities research projects that would benefit from access to language resources and technology (LRT). Research institutions or consortia with funding but little or no access to LRT or related expertise are targeted in this call.

- Access to digital language resources such as archives, corpora, wordnets, lexicons, termbases, etc.
- Access to tools for managing and exploring corpora and other language resources; tools for making word lists and extracting multi-word units, terms, names; tools that convert spoken data to written text and vice versa, etc.
- Assistance to convert legacy formats into formats that can be handled seamlessly.
- Assistance in creating a methodologically sound workflow with data, tools and modeling approaches for innovative research and development.
- Site visits by experts acting as consultants on methodology and the use of resources and tools.
- Hands-on training in the use of specific tools and methods.
- Redefining or extending research plans so as to include LRT aspects where appropriate.
- Dissemination of the results and outcomes from the CLARIN cooperation in the selected projects, using various CLARIN dissemination chanels.

The extent of CLARIN collaboration and technical support will be contractually defined.

Example 1: A project unable to acquire and utilize a text alignment tool may be given access to relevant software and may receive expert advice and training regarding its effective use from a CLARIN partner institution.

Example 2: For a project in need of a database of runic inscriptions, a CLARIN partner institution might negotiate access to existing databases and, if necessary, assist in converting them to a common encoding standard to make them searchable.

Eligibility

Applicants must be organizations or consortia including at least one organization established in a Member State (MS) or Associated Country (AC) with respect to the European Commission Seventh Framework Programme. Applicants may be from all sectors, including universities and colleges, research institutes, industry, international European interest organisations, civil society organisations, and any other legal entities. They must demonstrate their capacity towards performing the proposed research.

The project objectives and theme should be within one or more HSS subject areas, including, but not limited to, language, literature, history, philosophy, history of art, archeology, culture, religion, anthropology, psychology, pedagogy and sociology. The emphasis will not be on developing LRT in itself, but to explore its usefulness for central HSS research.

The period for direct CLARIN cooperation should generally not extend after December 31, 2010, although projects themselves may extend beyond that time.

Research activities should primarily consist of research or technological development within HSS, addressing new and pertinent research questions, and may also include 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 LRT. It is important to justify how the proposed LRT support fits in the overall objectives and methodology of the project.

Projects must be financially viable, either enjoying financial support or clearly demonstrate potential to raise funding for the project.

CLARIN will normally not pay for software and data licences owned by third parties, but may contribute with LRT that is owned by CLARIN partners.

Example 1: A PhD student with a stipend and her supervisor, both at the University of Malta, need LRT support to enlarge its capabilities of handling enough data to get statistically significant results. The project is relevant to CLARIN. Furthermore, CLARIN partners can provide the necessary tools and data. This project is likely to be eligible.

Example 2: A research consortium is led by a Romanian university, with a Dutch research institute and a Turkish company as project partners. The project's bid for support from the National University Research Council in Romania has been successful. The project is relevant and needs additional LRT support which CLARIN can provide. This project is likely to be eligible.

Selection procedure

Proposals will be selected in a fast two-step procedure. In the optional first step, pre-proposals are invited that contain project sketches. During this step, CLARIN will perform an elegibility check and will provide feedback to proposers, taking into account the gist of the project, but will not reject proposals.

In the second step, full proposals will be invited. Having submitted a pre-proposal is no prerequisite for submitting a full proposal. Full proposals will be reviewed non-anonymously by three experts including the national representative for the proposal coordinator's country. Proposals will be judged according to the following criteria:

1. Exemplary LRT needs and use of LRT towards research goals.
2. Capacity of CLARIN to provide the needed LRT and expertise.
3. Relevance of the proposed projects for testing the CLARIN infrastructure.
4. Adherence to CLARIN standards and best practice.
5. Capability to demonstrate the potential of the CLARIN infrastructure to HSS projects.
6. Multilinguality or cross-boundary dimension (projects crossing language and/or national boundaries are particularly welcome but monolingual projects and projects with national relevance will also be encouraged).
7. National and European needs and priorities (to the extent these are formulated).

Only a limited number of projects will be able to receive support, and a good spread of the selected projects across subject areas will be a goal. The national representatives for CLARIN will provide advice on how well the proposals fit in national priorities. The final decision on selection will be taken by the CLARIN Executive Board on the basis of the expert reviews. Decisions about each individual proposal will be documented and a detailed feedback will be sent to its proposers.

Contract, responsibilities and funding scheme

A contract will be drafted that defines the support provided. CLARIN will not assume any other responsibility for projects other than for the tasks agreed on in the contract. CLARIN will have a consultancy role, only for the parts and issues within its scope and competency. CLARIN will not play any part in steering the project, but will assume an advisory role. CLARIN will not be represented on a project's steering committee, but may assume an advisory role only for parts and issues within its competency. The project management shall not make any decisions against the interests of the CLARIN project or of those CLARIN partners involved in providing support. Projects normally retain the full right of use and dissemination of any project results (“foreground”). CLARIN will fund only its own activities in the project and will not provide any other financial support to the project. Proposed projects should aim at disseminating LRT results through CLARIN.

Pre-proposal submission procedure

The pre-proposal should be submitted in two separate email copies addressed to Tamás Váradi [varadi@nytud.hu], coordinator of WP3 and Koenraad De Smedt [clarin@uib.no], chair of Humanities Project Selection Committee, as a single PDF file with the following headings:

1. Proposal acronym
2. Proposal title
3. Duration in months and proposed start date
4. Coordinator: legal organization, organization type, legal address, department or unit, contact person, telephone and email.
5. List of partners: each with legal organization, organization type, legal address, department or unit, contact person, telephone and email.
6. Financing plan (budget and funding sources)
7. List of languages covered
8. Brief description of work, including main objectives, methods and data, plan of activities, and need for CLARIN LRT support (max. 4 pages)
9. Brief description of how the project addresses national and/or European priorities (max 1/2 page)
10. References (bibliographical, websites)

Applications from PhD students should include their supervisors as partners. The preproposal should preferably not have appendices, but references may include web links to background information which, however, should not be essential to the evaluation of the proposal.

Important dates

– February 15, 2009 (noon UT): deadline for pre-proposal submission
– March 7, 2009: feedback is provided to proposers
– April 1, 2009 (noon UT): deadline for full proposal submission
– April 21, 2009: final decision

Further information and assistance on the application

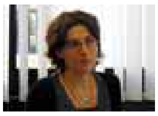
Interested parties are advised to contact one of the following persons:

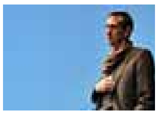
– Koenraad De Smedt [clarin@uib.no]
– Eva Hajičova [hajicova@ufal.mff.cuni.cz]
– Carla Parra [carla.parra@upf.edu]
– Jean-Marie Pierrel [jean-marie.pierrel@atilf.fr]
– Valeria Quochi [valeria.quochi@ilc.cnr.it]
– Paul Rayson [rayson@exchange.lancs.ac.uk]
– Marko Tadić [marko.tadic@ffzg.hr]
– Tamás Váradi [varadi@nytud.hu]
– Martin Wynne [martin.wynne@oucs.ox.ac.uk]

Alternatively, or in addition, the national representatives for CLARIN (http://www.clarin.eu/national_contact_points) may be contacted for further information. C

The Challenge of Defining Usage Scenarios for CLARIN

Reconciling User Demands and Technical Requirements





Valeria Quochi
ILC, Pisa
Lothar Lemnitzer
University of Tübingen
Marc Kemp-Snijders
MPI, Nijmegen

One of the main challenges of the CLARIN preparatory phase is the construction of a prototype demonstrating the great advantages offered by the infrastructure to the targeted research community. A key activity for tackling this challenge is the definition of representative usage scenarios that will 1) provide guidelines and use cases for the implementation of a prototype of the infrastructure, and 2) demonstrate the actual advantages that users can gain from CLARIN in their daily research activities. Usage scenarios will also help CLARIN to describe in more detail the requirements relative to language data, tools and interoperability and to assess the added value of standard-conformant and interoperable resources as well as workflows. Ideal scenarios for CLARIN describe and provide feasible solutions to real-world tasks that Humanities and Social Science researchers face while performing their studies and which they want to automate without much effort or the need to acquire themselves the necessary technical competences. The scenarios primarily address smaller groups or individual researchers and students with low to medium computational competences and without in-house support. For these users the infrastructure is expected to help reduce in the time needed for their research and improve the quality of the results of their investigations.

A call for contribution of scenarios was launched on the CLARIN website. Thanks to zealous members, several interesting scenarios were submitted and the best of these will be promoted in the implementation phase. The winning scenario(s) will be presented soon. We received 29 interesting scenarios from various groups, reflecting different types of users, needs and tasks. Most of these scenarios address a specific research question, while four of them are broadly applicable. Linguistics is the preferred discipline, mostly sociolinguistics and historical linguistics; a few address sociology, history, or school teaching. The proposals cover a quite broad range of languages (Spanish, Catalan, French, English, Swedish, Polish, German, Alsatian, Danish, Norwegian, Italian) and multilinguality is a key feature of many.

Some of the scenarios are illustrated through the following examples. A historian studying the habits of a population through its recipes could profit from a functionality that automatically extracts relevant information from a specific collection of recipes on a conceptual/semantic basis. A sociologist or a sociolinguist who studies the different linguistic strategies of politicians, might benefit from access to political speech transcriptions which would allow him/her to speed up the investigation by automating part of the analysis thanks to using natural language processing modules to perform discourse analysis.

A key issue for the selection of relevant scenarios for the infrastructure is twofold: on the one hand they have to represent real needs of typical users, i.e. researchers in humanities disciplines. On the other hand, the required workflows need to be practically feasible at acceptable costs for the prototype construction.

When looking at technical issues, critical for the implementation of the prototype, one fundamental requirement for the selection of scenarios is first the availability of the resources and tools needed to process them. These may already be available through web services or easily transformable into such services and described in the CLARIN web service inventory. Each step of the workflow should then be further fleshed out, and finally interactions with a GUI should be possible.

Focusing on taking the user-end, one of the limits of most scenarios proposed is the excessive auto-referentiality of the research questions, which reflects the Language Resource and Technology orientation of the consortium. Many scenarios address research questions within various branches of linguistics and language technologies. Viewed from a technical perspective, some of the scenarios proposed are appealing, but appear to be futuristic in the sense that no adequate tool support is yet available and further research and development is needed for their implementation.

Finally, a serious challenge for the realization of the workflows will be input/output data formats, for which we certainly cannot expect to have full convergence for different tasks and modules.

An immediate step to following the selection of the scenarios to be implemented will be the analysis of formats and the devise of conversion procedures, taking into account existing standards. Preference should be given to compatibility between tools and resources and/or adoption of widespread standards and best practices.

Report from the joint W2/W5 Focus Workshop

Berlin
October 6-9, 2008

Erhard Hinrichs
CLARIN EB member
Peter Wittenburg
CLARIN EB member

Work package 2 (Technical Infrastructure) and work package 5 (LRT overview) of CLARIN held an important joint workshop on a number of key issues for the current preparatory phase of CLARIN:

1. the establishment of data centers and center types for the emerging CLARIN infrastructure,
2. the construction of a Language Resources and Technology Federation and its underlying technologies,
3. the development of a flexible metadata infrastructure for resources and tools,
4. intellectual property rights and copyright issues concerning language resources and tools.

It was the first workshop where these aspects were discussed with a broad group of experts from the CLARIN members and can certainly be seen as a milestone in the work of CLARIN.

The workshop was jointly sponsored by CLARIN and by the German national counterpart project D-SPIN. The Max-Planck Society generously provided the venue at the Max-Planck Institutes for the History of Science and for Molecular Genetics.

Preparing deliverables

In a mixture of plenary sessions and working group meetings over 60 participants from CLARIN and D-SPIN took part in this important and utterly productive event. The discussion was based on a number of draft documents that had been distributed prior to the meeting by WP2 and WP5. The participants offered intensive feedback that was extremely helpful in working out more comprehensive versions that will lay out the CLARIN approach to the above-mentioned issues. As a result of this event five documents will be ready as first external versions:

1. CLARIN centres;
2. Centre Types;
3. Language Resource and Technology Federation;
4. Persistent and Unique Identifiers;
5. Metadata Infrastructure.

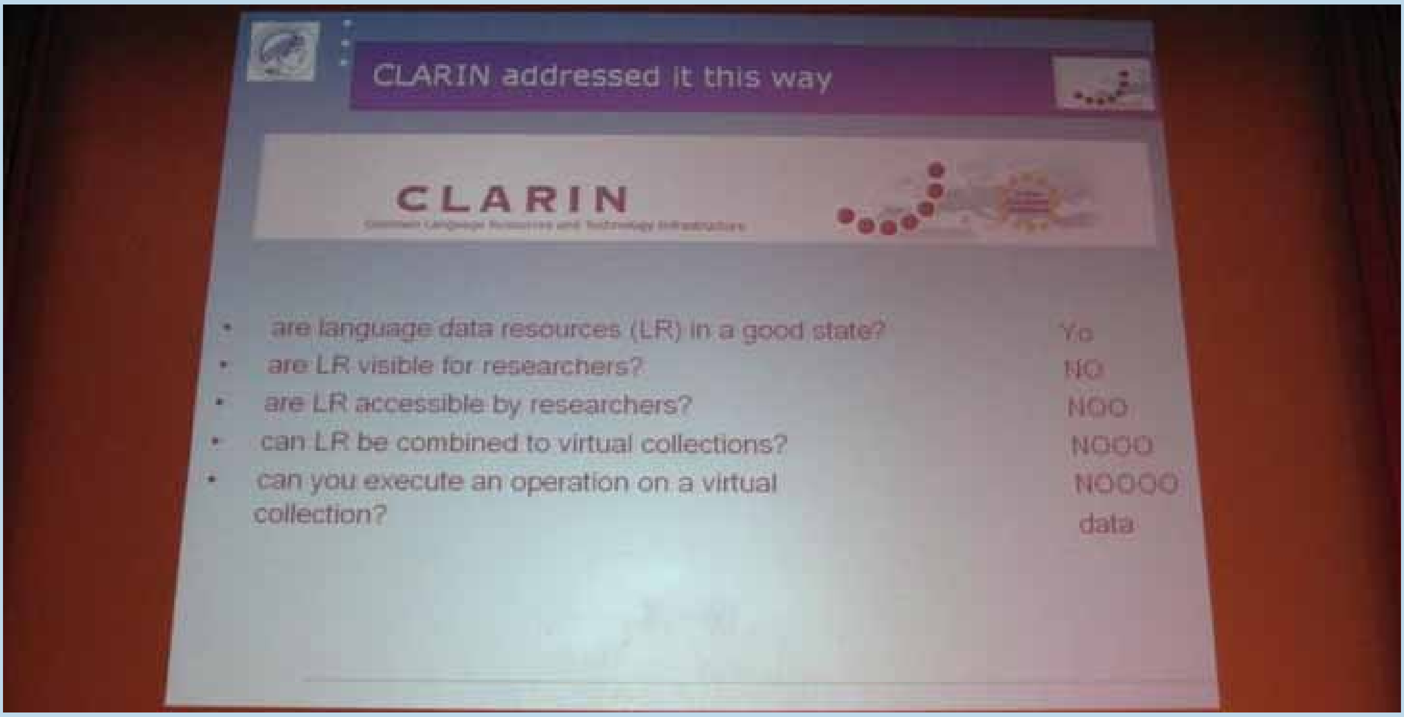
These document versions will be made available as CLARIN deliverables on the CLARIN website.

Follow-up workshops

The Berlin workshop led to a number of follow-up workshops focusing on particular themes highlighted during the Berlin workshop event: A D-SPIN expert workshop on “Web Service Architecture in CLARIN” held in Munich on November 10-11 (see separate report on this event by Nuria Bel in this issue of the CLARIN newsletter); a follow-up WP2 workshop to be held at Oxford in February 2009 to settle all above items, to distribute the work and to exactly describe the goals for 2009; and a WP5 workshop to be held in April that will focus on aspects of the structure of language resources and of the linguistic encodings used within them with the goal to specify how interoperability can be achieved in workflow systems. In November 2009 a WP2/5 workshop will be organized to check the progress of work and to derive plans for 2010.

MICROSOFT eSCIENCE WORKSHOP 2008 (INDIANAPOLIS, DECEMBER 7-9, 2008) PICTORIAL REPORT

Peter Wittenburg
CLARIN EB member



CLARIN addressed it this way

CLARIN

Common Language Resources and Technology Infrastructure

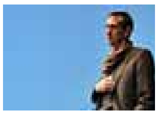
- are language data resources (LR) in a good state? Yo
- are LR visible for researchers? NO
- are LR accessible by researchers? NOO
- can LR be combined to virtual collections? NOOO
- can you execute an operation on a virtual collection? NOOOO data

Peter Wittenburg’s slide giving some answers to questions about usage of language resources in humanities and social sciences

Web Service's architecture: D-Spin Expert's workshop

Munich November 10-11, 2008

The participants of the D-Spin Expert's workshop in Munich



Núria Bel
University Pompeu Fabra, Barcelona
Marc Kemp-Snijders
MPI, Nijmegen

CLARIN is devoted to the creation of a persistent and stable network serving the needs of the European humanities research community. Humanities and Social Sciences researchers will be able to efficiently access distributed resources and apply analysis and exploitation tools relevant for their research questions. For very few languages however, there can be found an abundance of language resources and it is certainly true for all languages that whatever resources do exist, are dispersed and not interoperable. The CLARIN infrastructure strongly relies on the SOA (Service Oriented Architecture) approach to bring together the range of resources and tools available in the research community and make them available to others.

The D-Spin project (the German CLARIN branch) organized an experts' workshop to study the list of requirements derived from web service objectives. The aim of the workshop was to sum up lessons that can be learned from related experiences, especially from different frameworks for process chains in NLP applications, but also from other related initiatives in scientific research applications. The different sessions of the workshop consisted of short presentations by invited experts about their experiences and recommendations for CLARIN. The outcome can be summarized as follows.

Focus on web services

Development of the CLARIN infrastructure should clearly focus on services that are able to deliver the highest added value to our research community. Services should be registered in web service registries that require some level of semantic descriptions to allow both human and automated lookup and extraction. Semantic descriptions should provide at least a basic level of interoperability by mapping information in service input and output specifications to well defined concepts, such as those listed in the Data Category Registry for example. Web service registries such as those constructed for the eGov-Share and the X-road projects may serve as examples for setting up CLARIN web service registries. These were presented by Christoph Ludwig and Peep Küngas respectively. Operation of webs service registries will require organizational support and should take into account the various actors involved in the interaction process.

Web service composition aspects can be illustrated by considering the example of cumulative annotations of text. The LAF framework was introduced by Nancy Ide and Laurent Romary as a framework for linguistic annotation of resources that can serve as a reference or pivot model for different annotation schemes. LAF is one of the working draft standards from ISO/TC 37/SC4. GrAF is the XML serialization of the generic graph structure for linguistic annotations described by LAF. Current practices from the NLP domain for work flow solutions were represented by Ian Roberts and Thilo Goetz, who presented GATE and UIMA respectively. These are the two main systems used for describing processing pipelines in the NLP domain. There is a certain level of interoperability between these two through a module developed by the GATE team that allows GATE and UIMA processing pipelines to be interchangeable. Dan Cristea and Ionut Pistol (University of Iasi) presented ALPE, a new processing platform which offers, among other features, a solution to the integration of different format problems.

When considering work flow composition it is important to get a feeling for the kinds of web services that may be encountered and the volume of requests that these web services will handle. Leipzig Linguistic Services offers a range of web services that have been in operation for the last four years. With over 36 million requests over the past year these services represent the high end of web service usage within our domain. Marco Büchler provided an insight into the challenges that are encountered when dealing with these high request volumes. Interesting to note here is that the top three web services accounted for over 88% of the request volume.

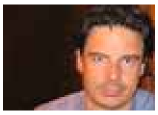
Describing workflows

Experiences from the world of bibliographic services were presented by Malte Dreyer from the eSciDoc project. He strongly adviced to take a bottom-up approach and carefully consider the technological implications when considering work-flow systems. Implementing the WS* specification stack may prove to be a daunting task which requires expert support and extensive man power. This last point was argued by Dimka Karastoyanova in her presentation on BPEL. Only those specifications from the stack that are of direct relevance to the application domain should be dealt with. She further provided a guided tour through different elements of the BPEL specification and related standards. Milan Agatonovic described the experiences made with jBPM, a competing work-flow language. He argued that from his experience jBPM was easier to use and better suited for the NLP domain. The GATE Teamware project is currently using this to describe work flows which include GATE exposed as a web service

The workshop concluded with, Andreas Witt and Peter Wittenburg presenting their views with respect to the CLARIN project. CLARIN can already count on the availability of a registry, metadata for linguistic resources and tools, and lots of resources and data. At present, resources and metadata are rather static in nature and need to be adapted to meet the dynamic requirements for this web service and workflows based view. Eventually, it was proposed that the next CLARIN workshop in February addresses the typical processing chains for combining resources and tools, with special emphasis on the formats used, to start the design of processing chains. C

New wind in sails?

LT Days, Luxembourg, January 14-15, 2009



Marko Tadić
Faculty of Humanities and Social Sciences, University of Zagreb

After several previous years that have witnessed the shift of focus in ICT away from the development of LRT, it looks as if the focus of research is turning back to language and multilinguality issues. In July 2008 a new unit INFSO.E1 Language Technologies & Machine Translation was established within DG Information Society and Media. It was furthermore supported by Council resolution (2008-11-21) with statements such as "…encourage the *development of language technologies*, in particular in the field of translation and interpretation…". Its initial focus is to overcome the language barriers in EU by dealing with multilingual (i.e. cross-lingual) technologies, services and applications.

Language Technology Days, Luxembourg, January 14-15, 2009

Two calls in 2009

This newly established unit is responsible for two calls for project proposals that have been issued in 2008-11 and 2009-01 respectively:

1. FP7-ICT Call 4, Challenge 2: "Cognitive Systems, Interaction, Robotics"; Objective 2.2 "Language-based Interaction";
2. ICT-PSP Call, Theme 5: "Multilingual Web" (Policy Support Programme within Competetiveness & Innovation Framework Programme adopted in 2006-10).

A meeting in Luxembourg was organised with several goals in mind – advertising the calls, giving suggestions by EC officials pinpointing the details of calls, opportunity for researchers to ask for additional information and establishing preliminary consortia by meeting other researchers, were just few of them.

And the community responded in a grand manner – more than 200 researchers from 29 countries and several international institutions attended the event.

The FP7-ICT call is oriented towards developing systems that deal with omnipresent European multilinguality. The idea is to develop MT beyond existing mostly static M(A)T technologies to fully dynamic and web-oriented MT services that could be embedded in products/services also filling the gap of missing EU languages. The instruments that are expected within this call are: one IP (up to 4 years, 5-8 Meuro), one or two NoE (up to 3 years, 3-6 Meuro) and several STREPs (up to 3 years, 2-3 Meuro).

The ICT-PSP call is oriented towards innovative usage of existing methodologies and their recombination to achieve the new values. It is highly geared to further development of single European Information Space under the i2010 policy framework. Its aim is to provide seamless access to ICT-based services taking into account multilingual & cultural diversity in the EU and accessing countries. Its aim is also to support and enhance interpersonal and business communication and information access and publishing in a multilingual Web environment. This common theme – Mulitlingual Web – has three distinct "objectives": MT for multilingual Web (pilot projects); multilingual Web content management (pilot projects); best practices & standards for multilingual Web (thematic network).

The total budget for the two calls is around 40 million euro.

CLARIN's role

In the last decade or so we have seen large and important integrative projects and initiatives that have been oriented to strenthen the field of language and speech processing such as ELSNET, HLT Central and, finally, agencies like ELRA or ELDA which were offprings of these initiatives.

Can CLARIN play similar role in this fresh start for LRT? The LRT field is mature enough to start serving as an infrastructure for text-sciences i.e. humanities and social sciences which have text as whose object of research or their object of research is partially or completely mediated by text. The CLARIN project gives a very good example of this, and with its overall European dimension it is certainly in the position to play the integrating role not just in the LRT community, but also beyond that, in communities of researchers in humanities and social sciences moving towards the eScience paradigm.

But is the LRT field mature enough to support not just researchers as highly educated and motivated users, but also laypersons who lack necessary expertise and/or motivation? It certainly seems that these calls are oriented towards bringing about a positive answer to this question.

Hanne Fersøe
Centre for Language Technology, University of Copenhagen

The Danish CLARIN project plans to deliver both a technical platform (an infrastructure with search and retrieval facilities) and platform content. The content will take the form of examples of many different and interesting types of resources: monolingual written corpora, aligned multilingual written corpora, monolingual spoken corpora, dictionaries, word nets, images, videos, etc.

One such resource for the use of the humanities research community is a monolingual written Danish corpus of approximately 250,000 words composed of extracts from non-literary texts for every-man's use from the period 1500 to 1750. The texts will be extracted from rare books only obtainable from The Royal Library in Copenhagen, and they will cover subjects such as ethics and moral issues, geography and topography, history, housekeeping and cooking, medical science, mathematics and astrology, natural sciences, pedagogics, etc.

Through the CLARIN platform the texts will be made available electronically, marked up, and with a dictionary as a lexical key to the corpus.

How to look up a word with orthographic variation?

The texts are written by different authors and over a period of time when orthographical rules for written Danish had not been stabilized, so there will be a high level of orthographic variation in the corpus. Examples of such variation can be seen in the Danish word currently spelled *sygdom* (eng: illness), which can have the following spellings in older texts: *sigdom, siugdom, siugedom, sygdom, sygdomme, sygdommer, siuge, syge, syuge*. It is necessary to neutralize such spelling variations in order for the researcher to be able to search for and find instances of certain words or phrases in the texts.

The DUDS research group at the Department of Scandinavian Studies and Linguistics at the University of Copenhagen, which is in charge of producing the corpus, has developed a neutralization method and a mark-up system suitable for texts with orthographical variation. They have developed the method and successfully

Knowledge for Everyman from the Renaissance to Modern Times

implemented it on the Danish ballad manuscripts of the 16th century thus making the complete textual tradition before 1591 available electronically with a dictionary.[1]

The method is called multilevel text, MLT, and it consists of providing three linked levels of markup on each word in the corpus: source level, neutral level, and lemma level. The source level is the original word form as written in the text or manuscript; the neutral level represents a neutral word form close to modern Danish spelling, and the lemma level gives the lemma form of the source word with the associated part-of-speech.

Just to illustrate the richness and complexity of orthographic variation that have to be dealt with, we present an extract of a search result for the neutral form *hjertet* (eng: the heart) from the MLT marked-up ballad corpus (only source level forms shown) in a form of concordance within the coloured box below.

An opening of the ballad manuscript called Hjertebogen (The heart book) (with permission from The Royal Library, Denmark)

```
       Then første hand var y hiertiidtt gladtt
            dj iegh er ham aff hierttett huldtt,
 och saa den frue, du haffuer y hiertet kier.
            the hellede wore y hierttit trøst,
            och elsk hanom aff herttiitt
                   det oden i hierted vende:
           da maan ieg ham aff hiertted gaa,
           y maa vell sige aff hiertet frÿ:
     hun hagde stoer soriig y  hierthet sin,
              i haffuer meg i  hierthed saa kier.
     de haffde huer-andenn udi hiertid saa kier,
         tro myg, yeg dyg aff hyerthett for nogin mand well vntt..
```

Research Examples

The Knowledge for Everyman corpus is still being built and no research has yet been based on find-ings in the corpus. The possible research themes, however, are many, for example the following ones suggested by the DUDS researchers: *Perceptions of and attitudes to health and illness in the period 1500-1750* (based on medical books and cook books, using search terms for e.g. fresh, health, ill, illness, medical.), *The use of medical herbs* (based on medical books and cook books, using the names of the herbs as search terms), *Religion in everyday life* (based on catechism, prayer books, ethics, and with search terms for e.g. christian, pray, prayer, enemy), and *Knowledge about the world* (based on descriptions of exotic countries, pamphlets about sensational incidents and the like, and using the geographical names as search terms).

The corpus of ballads from before 1591 with the ballad dictionary is already a rich resource of in-formation for scholars from different disciplines in the humanities. Many scholars have made searches in the corpus and used the results in their research covering themes such as *Formulae*, e.g. *The poetic formulae of the ballads* which studied the flower terms involving roses and lilies used to refer to young maidens. Other themes were *Weapon* studied by a historian (what is said about weapons in the ballads, how does this correspond to weapons, killing and war in the society), *Pragmatics*, for instance *The social variation in pronouns of address* (studied through key words in greetings combined with forms of personal pronouns), *Orthography, Stylistics, Genre definition*. Precise references to the research mentioned here can be found in a dictionary[1].

Conclusion

Today the corpus of ballads and its dictionary is available on CD-ROM together with volume 3[2]. The 18 most popular renaissance ballads and their textual variants are available with neutral word forms at http://duds.nordisk.ku.dk/tekstresurser/aeldste_danske_viseoverlevering/visernes_top-18/, and the remainder of the ballad corpus is being prepared for publication on the Internet. The corpus-to-be of knowledge for everyman is not yet available electronically.

The CLARIN infrastructure aims to become the common vehicle for taking traditional texts for humanities research into the electronic future where they can be made available to other researchers, not merely in their original source form, but also with mark-up and other enhancements produced by research colleagues and as well as tools for further enhancement.

[1] Ruus, Hanne: A Corpus-based Electronic Dictionary for (Re)search, in EURALEX 2002 Proceedings, pages 175-185.

[2] Lundgreen-Nielsen & Ruus, Hanne (Ed.): Svøbt i mår, bind 1-4, København 1999-2001.

Thanks to Hanne Ruus and Dorte Duncker of the DUDS group for inspiring conversations and useful comments.

CLARIN in the context of NLP research in the Czech Republic

Eva Hajičová
Charles University, Prague

1 NLP research in Czech Republic (former Czechoslovakia) has a rather long tradition. Theoretical foundations for formal description of Czech were laid as early as at the beginning of the sixties (based on the well-known Praguian linguistic tradition and formulated as an original dependency- and functionally oriented alternative to Chomskyan generative transformational grammar) with applications (such as machine translation) always in mind. The activities were first anchored at Charles University in Prague, but we have always been aware of the necessity to combine forces. In the middle of the eighties we therefore initiated the creation of an informal group of Computational "Fund" of Czech, gathering researchers from different working places interested in NLP. It was only after 1989 that these efforts could be fully "institutionalized" under the roofs of Czech universities and the Academy of Sciences.

2 Czech belongs to "small" languages and the Czech Republic – as well as the Czech language community – is a very small playground; therefore cooperation and/or division of labour are a key issue. The work on NLP and computerized language resources is basically the main concern of the following four main centres.

Institute of Formal and Applied Linguistics, Faculty of Mathematics, Charles University in Prague – Head: Prof. Jan Hajič

NLP for most different applications incl. machine translation; corpus annotation (The Prague Dependency Treebank of Czech, English and other languages) with a complex and integrated scenario of annotation on three language layers incl. underlying semantico-syntactic relations; spoken language reconstruction and understanding and corresponding corpus annotation; build-up of parallel annotated treebanks; focus on statistical and 'hybrid' methods; formal description of language focussed on dependency syntax, information structure of the sentence, dialog systems and discourse.

Institute of the Czech National Corpus, Faculty of Philosophy, Charles University in Prague – Head: Prof. František Čermak

Build-up and continuous release of the Czech National Corpus, one of the largest in Europe, made up of synchronic, diachronic and spoken parts; build-up of parallel and comparable corpora; corpus methodology and seat of a PhD study branch of corpus linguistics, in addition to articles and number of book publications (series Studies in Corpus Linguistics).

Charles University, Mala Strana Campus, site of the Institute of Formal and Applied Linguistics

Natural Language Processing Centre, Faculty of Informatics, Masaryk University in Brno – Head: Dr. Karel Pala

Theoretical and applied research in automatic analysis of written text on all language levels; corpus management and lexical databases; semantic representation of natural language expressions; semantic web, ontologies, knowledge representation and reasoning; synthesis and recognition of speech (Czech); dialogue representation and management; applications of machine learning methods to disambiguation of corpus data.

Department of Cybernetics, Speech technology section, Faculty of Applied Sciences, University of West Bohemia in Pilsen – Head: Prof. Josef Psutka

Theoretical and applied research into speech analysis, recognition (LVCSR, real-time applications) and TTS systems; spoken dialogue systems; speech understanding, keyword spotting; speaker identification; information retrieval from spoken documents and large spoken archives etc.

3 All the four institutions mentioned above take part in the CLARIN project in one way or another. Charles University is one of the partners and Prof. Eva Hajičová is the national contact person (hajicova@ufal.mff.cuni.cz). Our direct formal involvement is within workpackages 5.3 and 5.8, but we are very much interested also in the work being covered by workpackage 5.5. During the first months of the project, our particular contribution was an overview of funding possibilities currently existing in Czech Republic for NLP and related fields and a constant "revitalisation" of research contacts, both formal and informal, within the NLP oriented Czech research network.

We also organized some events with a respectable international participation. As for the work in progress, we are aiming to collect information on administrative possibilities of international cooperation within the Central and Eastern Europe ("where to look for contacts") and to disseminate it among CLARIN partners. In addition, we are collecting information on legal issues connected with dissemination of language resources.

We are preparing the first seminar in the Czech Republic, that should take place in March 2009 in Prague, demonstrating the possibilities NLP and corpus-oriented research results can offer to the humanities in the broad sense of the domain. C

Romanian language resources and tools

Dan Cristea
University of Iaşi

Romanian CLARIN will soon be on the stage. Applying for funding is our concern for this year. I see two major aspects around which CLARIN-RO should be depicted: continue the acquisition of resources of types that are lacking now and improve/diversify/reorganise the pool of existing NLP tools and make them work as web-services.

Number of centres

However, CLARIN-RO will not be a seed spread on a barren field. A lot is there already and only awaits modernisation, reorganisation, and opportunities to create links with other fields and the industry. Development of language resources and tools for Romanian is being pursued in a number of centres in Romania and elsewhere. The main ones in Bucharest are at the Research Institute for Artificial Intelligence (RACAI[1]) and the I. Iordan – A. Rosetti Institute of Linguistics, both belonging to the Romanian Academy, and the University of Bucharest. In Iaşi, important centres are found at the Department of Computer Science of the Alexandru Ioan Cuza University (UAIC-FII), at the Institute of Computer Science[2] and the Institute of Romanian Philology Alexandru Philippide, both belonging to the Iaşi branch of the Romanian Academy. Another centre can be found in Cluj-Napoca, at the “Babeş-Bolyai” University). Outside Romania, work on Resources is also done in Chişinău, Republic of Moldova, at the Institute of Computer Science of the Moldavian Academy, but also in other centres outside this area (as, for instance, at the University of North Texas, the University of Sheffield and the University of Hamburg). In 2001, research groups from Iaşi, Bucharest and Chişinău founded the Consortium for the Romanian Language Resources & Tools (ConsILR)[3] – an initiative aiming to promote software tools for linguistic processing developed by computer scientists to linguists and to allow the computer scientists to use the resources created by linguists. The Consortium has put up a portal where many of the resources and tools are hosted or mirrored and each year it organises conferences as forums for dissemination of new achievements.

EC and national projects

Development of many Romanian language resources and tools originated from previous projects, founded by the EC or the Romanian Government. One example is the Romanian wordnet, whose development begun inside the FP6 project BalkaNet (whose aim was to build a collection of WordNet resources for 5 Balkan languages and to aligned them with the English WordNet). Since the BalkaNet finised, the development of the Ro-WordNet steadily continued at RACAI and currently it has more than 53.000 synsets[4].

Some tools

It is a known fact that most tools that process language at different levels today are language independent pieces of software. Their specific behaviour is intrinsically dependent on the quality of resources they are fuelled with. One example is the Romanian POS-tagger, based on the Tiered Tagging model, a system which uses an optimal hidden corpus tagset automatically generated from a lexical of morpho-lexical descriptors (Multext-East compliant), whose accuracy reaches almost 99% and which has at its base a very accurate “golden standard” corpus, developed as part of another European Project.

Participation in competitions

Another example is the Romanian version of the Acquis Communautaire, sentence aligned with all the other European languages by JRC. It has also been automatically aligned at word level with its English version, using the language independent lexical aligner COWAL, a system which has been ranked the first in the EN-RO word alignment competitions organised by ACL in 2003 and 2005.

One way to keep up with the world level in language technology is by participating in international competitions. The two main PhD schools in NLP in Romania are RACAI and UAIC-FII. The research interests range from development of general models related to discourse, time, semantic roles to mixed methods in machine translation, question-answering, textual entailment, as well as to automatic generation of processing architectures, human-computer multimedia interfaces, and computational lexicography. Enrolment in international competitions has been a part of PhD student experience, almost every year since 2003. In addition to the already mentioned alignment results obtained by RACAI, also the participation in the Word Sense Disambiguation contest (ranked first among the unsupervised systems at AVL 2007) can be added to their success. Teams of PhD and master students from UAIC-FII also done very well in competitions like Answer Validation Exercise for English (ranked the first out of 7 participants at CLEF 2007 and 2008), the Anaphora Resolution (ranked the first out of 4 participants in ARE-2007) and Textual Entailment (ranked the second out of 26 participants in the 2-way task, and the first out of 13 participants in the 3-way task in the 2008 competition for English). C

[1] http://www.racai.ro

[2] http://www.iit.tuiasi.ro/

[3] http://consilr.info.uaic.ro/

[4] http://nlp.racai.ro/wnbrowser/

CLARIN calendar of events

Here is a list of CLARIN events and events from the fields of language resources and language tools that may be of interest to CLARIN members.

December 2008

2008-12-08 to 2008-12-10: European Conference on Research Infrastructures, Versailles, France

2008-12-13: NaLELA, Natural Language Engineering of Legal Argumentation, Florence, Italy

2008-12-17 to 2008-12-12: IEEE e-Humanities Workshop, Indianapolis, USA

January 2009

2009-01-06 to 2009-01-07: Building a Virtual Humanities Collaboratory, Cambridge, UK

2009-01-07 to 2009-01-09: 8th International Conference on Computational Semantics, Tilburg, Netherlands

2009-01-14 to 2009-01-15: Language Technology Days, Luxembourg

2009-01-23 to 2009-01-24: The Seventh International Workshop on Treebanks and Linguistic Theories, Groningen, Netherlands

2009-01-30: Digital Preservation Workshop, The Hague, Netherlands

February 2009

2009-02-25: AAI hands-on workshop, Oxford, UK

2009-02-26 to 2009-02-27: WP2 follow-up workshop, Oxford, UK

March 2009

2009-03-01 to 2009-03-07: CICLing 2009, Mexico City, Mexico

Join CLARIN

The CLARIN project is a combination of Collaborative Projects and Coordination and Support Actions, registered at the EU under the number FRA-2007-2.2.1.2. It started with the preparatory phase in 2008 that will make the grounds for the next phases and it will cover the generic, language independent activities. In order to do our work properly we have to rely on a much wider circle than just the formal consortium partners in the project. For this reason we have opened up all our project working groups for participation by organizations that are not part of the consortium.

Members

Country; Institution; Location; Contact person

Austria: University of Vienna; Vienna; Gerhard Budin **(NCP)**

Belgium: ALT (Acquiring Language through technology); Leuven – Kortrijk; Hans Paulussen

Center for Computational Linguistics ; Leuven; Ineke Schuurman **(NCP)**

Center for Dutch Language and Speech, University of Antwerp; Antwerp; Walter Daelemans

ELIS-DSSP; Gent; Jean-Pierre Martens

Legal Informatics and Information Retrieval, Katholieke Universiteit Leuven; Leuven; Marie-Francine Moens

Laboratory for Digital Speech and Audio Processing – VUB – ETRO/DSSP ; Brussels; Werner Verhelst

ESAT-PSI/Speech; Leuven; Patrick Wambacq

Bulgaria: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences; Sofia; Svetla Koeva

Institute for Parallel Processing; Sofia; Kiril Simov **(NCP)**

Mathematical Linguistics Departement, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences; Sofia; Ludmila Dimitrova

Croatia: University of Zagreb, Faculty of Humanities and Social Sciences; Zagreb; Marko Tadić **(NCP)**

Institute of Croatian Language and Linguistics; Zagreb; Damir Ćavar

Cyprus: Cyprus College / Research Center; Nicosia; Antonis Theocharous

Czech Republic: Charles University; Prague; Eva Hajičová **(NCP)**

Faculty of Informatics, Masaryk University ; Brno; Aleš Horák

The Institute of the Czech Language, Czech Academy of Sciences; Prague; Karel Oliva

Denmark: Center for Sprogteknologi, University of Copenhagen; Copenhagen; Bente Maegaard **(NCP)**

Dansk Sprognævn – Danish Language Council; Copenhagen; Sabine Kirchmeier-Andersen

Society for Danish Language and Literature; Copenhagen; Jørg Asmussen

Estonia: University of Tartu; Tartu; Tiit Roosmaa **(NCP)**

Finland: CSC – the Finnish IT Center for Science ; Espoo; Tero Aalto

University of Helsinki; Helsinki; Kimmo Koskenniemi **(NCP)**

Department of Foreign Languages and Translation Studies, University of Joensuu; Joensuu; Jussi Niemi

University of Tampere; Tampere; Eero Sormunen

The Research Institute for the Languages of Finland; Helsinki; Toni Suutari

France: ALTIF; Nancy; Jean-Marie Pierrel **(NCP)**

TELMA/DIS CNRS; Paris; Florence Clavaud

CNTRL; Nancy; Bertrand Gaiffe

Evaluations and Language resources Distribution Agency (ELDA); Paris; Khalid Choukri

Université Paris 4 Sorbonne / CELTA ; Paris; Andre Wlodarczyk

LIF-CNRS ; Marseille; Michael Zock

Germany: Berlin-Brandenburg Academy of Sciences; Berlin; Alexander Geyken

Deutsches Forschungszentrum für Künstliche Intelligentz; Saarbrücken; Thierry Declerck

Institut für Deutsche Sprache; Mannheim; Marc Kupietz

Max Planck Institute for Evolutionary Anthropology; Leipzig; Hans-Joerg Bibiko

University of Frankfurt/Main Comparative Linguistics; Frankfurt/Main; Jost Gippert

University of Leipzig; Leipzig; Codrina Lauth

University of Stuttgart; Stuttgart; Ulrich Heid

Universität Tübingen; Tübingen; Erhard Hinrichs **(NCP)**

University of Giessen; Giessen; Henning Lobin

Computational Linguistics Department, University of Heidelberg; Heidelberg; Anette Frank

University of Augsburg ; Augsburg; Ulrike Gut

Greece: Institute for Language and Speech Processing; Athens; Stelios Piperidis **(NCP)**

Hungary: Academy of Sciences; Budapest; Tamás Váradi **(NCP)**

Budapest University of Technology and Economics Media Research (BME MOKK); Budapest; Peter Halacsy

University of Szeged, Department of Informatics, Human Language Technology Group; Szeged; Dóra Csendes

Iceland: Institute of Linguistics, University of Iceland; Reykjavík; Eiríkur Rögnvaldsson

Icelandic Centre for Language Technology; Reykjavík; Eiríkur Rögnvaldsson

Ireland: National University of Ireland; Galway; Sean Ryder

Israel: Technion-Israel Institute of Technology; Haifa; Alon Itai

Italy: Dipartimento di Linguistica Teorica e Applicata, Università di Pavia; Pavia; Andrea Sansò

Istituto di Linguistica Computazionale; Pisa; Nicoletta Calzolari **(NCP)**

Department of Computer Science, University of Rome "Tor Vergata" ; Rome; Fabio Massimo Zanzotto

European Academy Bozen/Bolzano; Bolzano; Andrea Abel

Latvia: Institute of Mathematics and Computer Science, University of Latvia; Riga; Inguna Skadina **(NCP)**

Tilde; Riga; Inguna Skadina

Lithuania: Institute of the Lithuanian Language; Vilnius; Daiva Vaisniene

Center of Computational Linguistics, Vytautas Magnus University ; Kaunas; Ruta Marcinkeviciene

Luxembourg: European Language Resources Association (ELRA); Luxembourg; Bente Maegaard

Malta: University of Malta, Dept. of computer science; Malta; Michael Rosner **(NCP)**

Netherlands: Meertens Institute; Amsterdam; H.J. Bennis

Data Archiving and Networked Services; Den Haag; Henk Harmsen

University of Twente, Human Media Interaction Group; Enschede; Roeland Ordelman

Center for Language and Cognition; Groningen; Wyke van der Meer

Digital Library for Dutch Literature; Leiden; C.A. Klapwijk

Instituut voor Nederlandse Lexicologie; Leiden; Remco van Veenendaal

Leiden University Centre for Linguistics; Leiden; Jeroen van de Weijer

Centre for Language Studies, Radboud University; Nijmegen; Pieter Muysken

Centre for Language and Speech Technology, Radboud University; Nijmegen; L. Boves / N. Oostdijk

Max-Planck-Institute for Psycholinguistics; Nijmegen; Peter Wittenburg

University of Utrecht/Netherlands Graduate School of Linguistics; Utrecht; Jan Odijk **(NCP)**

ILK Research Group ; Tilburg; Antal van den Bosch

Huygens Instituut KNAW ; Den Haag; Karina van Dalen-Oskam

Norway: Dept. of Culture, Language and Information Technology; Bergen; Koenraad de Smedt **(NCP)**

Department of Linguistics and Nordic Studies, University of Oslo; Oslo; Janne Bondi Johannessen

Det humanistiske fakultet, Universitetet i Tromsø; Tromsø; Trond Trosterud

Norwegian University of Science and Technology; Trondheim; Torbjørn Svendsen

The Language Council of Norway, Oslo, Torbjoerg Breivik

Norwegian School of Economics and Business Administration (NHH), Bergen, Gisle Andersen

Poland: University of Wroclaw ; Wroclaw; Adam Pawlowski

Institute of Applied Informatics, Wroclaw University of Technology; Wroclaw; Maciej Piasecki **(NCP)**

Institute of Computer Science, Polish Academy of Sciences ; Warsaw; Adam Przepiórkowski

Institute of English Language, Univeristy of Lodz; Lodz; Lukasz Drozdz

Institute of Slavic Studies, Polish Academy of Sciences ; Warsaw; Violetta Koseska-Toszewa

Portugal: University of Lisbon, NLX-Natural Language and Speech Group; Lisbon; António Branco **(NCP)**

Romania: Al.I.Cuza; Iasi; Dan Cristea

Institute for Computer Science, Romanian Academy of Sciences; Iasi; Horia-Nicolai Teodorescu

Research Institute for Artificial Intelligence, Romanian Academy of Sciences; Bucharest; Dan Tufiş **(NCP)**

University Babes-Bolyai; Cluj-Napoca; Doina Tatar

Serbia: Faculty of Mathematics, University of Belgrade; Belgrade; Duško Vitas

Slovenia: Josef Stefan Institute; Ljubljana; Tomaž Erjavec

Alpineon d.o.o. ; Ljubljana; Jerneja Žganec Gros

Spain: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra; Barcelona; Núria Bel **(NCP)**

Universitat de Lleida ; Lleida; Gloria Vázquez

TALG Research Group (University of Vigo) ; Vigo; Xavier Gómez Guinovart

Sweden: Lund University; Lund; Sven Strömqvist

Språkbanken, Dept. of Swedish Language, Göteborg University; Gothenburg; Lars Borin **(NCP)**

Dept. Speech, Music and Hearing, CSC, KTH ; Stockholm; Kjell Elenius

Uppsala University, Department of Linguistics and Philosophy; Uppsala; Joakim Nivre

Department of Linguistics; Göteborg; Anders Eriksson

Department of Computer and Information Sciences, Linköping University; Linköping; Lars Ahrenberg

Swedish Institute of Computer Science AB ; Stockholm; Björn Gambäck

Language council of Sweden ; Stockholm; Rickard Domeij

HUMlab, Umeå University ; Umeå; Patrik Svensson

Turkey: Sabanci University – Human Language and Speech Laboratory; Istanbul; Kemal Oflazer

UK: Department of Linguistics and English Langauge, Lancaster University; Lancaster; Anna Siewierska

Oxford Text Archive; Oxford; Martin Wynne **(NCP)**

University of Sheffield; Sheffield; Wim Peters

University of Surrey; Guildford; Lee Gillam

Research Institute of Information and Language Processing at the University of Wolverhampton ; Wolverhampton; Gina Sutherland

Language Technologies Unit, Bangor University; Bangor; Briony Williams

Department of English, The University of Birmingham; Birmingham; Oliver Mason

CLARIN newsletter • *publisher* CLARIN • *editor-in-chief* Marko Tadić • *editorial board* Gerhard Budin, Dan Cristea, Koenraad De Smedt, Laurent Romary, Martin Wynne • http://www.clarin.eu