CLARIN
A European Research Infrastructure
Newsletter

Number 6, 2009, June

CLARIN: Communities addressed

Peter Wittenburg
CLARIN EB member

After about a year of hard work after the CLARIN kick-off meeting, this is an excellent moment to reflect on intermediate achievements, especially with respect to the different communities we need to address. The central questions for all ESFRI-related research infrastructure initiatives that received funding for the preparatory phase are the same: can these initiatives show that they are able to contribute to establishing a persistent infrastructure layer that facilitates competitive research in their discipline in Europe and can they help tackle the small and big challenges in the future? Of course, we were aware from the start that these difficult tasks can only be accomplished if we can convince a number of "communities" of our results.

1 First, we have to convince ourselves and our own community. This has proven to be the most challenging part. We started full of dreams and enthusiasm, but underestimated the patience required in an undertaking trying to synchronise 150 member institutes comprising about 500 experts. Until now we have aimed at the involvement of as many experts as possible in our discussions. However, we realized that a big part of the community cannot cope with the speed of concurrent activities and document creation. Nevertheless, a broad group of experts agreed upon the present requirement specification documents so that all interested researchers can now gain a deep understanding of the intentions of CLARIN. Of course, only a few actual services addressing the needs of the language resources and technology (LRT) experts have been set up so far.

2 Second, we want to convince the humanities and social sciences (HSS) communities in particular because these communities use LRT as the basis of their increasingly data-driven work. We should, however, be able to offer concrete services before approaching these communities. We clearly indicated the need and possibility to offer first services. Especially, the Virtual Language World services will be an attractive first step by providing users – among other things – with a browsable and searchable observatory comprising many LRT components in different views (catalogue, facetted browsing, GIS, etc.), a virtual laboratory allowing researchers to participate in testing and contributing to new eScience type frameworks, a consultancy portal referring to experts who can help with various issues. These services will be available to users by September 2009.

3 Third, we need to interact with comparable initiatives such as DARIAH and CESSDA to ensure compatibility of our goals and solutions as well as mutual benefits from each other's work. Interaction has begun at various levels, but due to different foci we cannot speak of a joint and coordinated approach towards a common research infrastructure yet.

4 We need to interact with other existing initiatives and investigate to which extent we could make use of existing solutions, e.g. by eduGain/TERENA and the grid community, and whether our solutions may be of interest to others. The key issue will be finding a good balance between ICT-driven top-down concepts and the bottom-up-driven community needs which per definition require some conceptualization and interaction time. This is why CLARIN actively pursued discussions with various horizontal initiatives such as Alliance for Permanent Access, eduGain/TERENA, e-IRG, DEISA, EGEE and DRIVER. Initial reactions were extremely positive with two exceptions for which there was not enough time for cross-fertilization yet.

5 We have to convince the research organizations, funding agencies and ministries since they are the ones that eventually need to reserve funding for infrastructures. Again, we need to be able to demonstrate concretely how a solid research infrastructure can improve research. It has become obvious that it will primarily be the research organizations that have to decide about reserving funding. The European Commission and national ministries are only willing to provide seed money. The required decision process is complicated and largely depends on the amount of funding required to run infrastructures. Unfortunately, there are too many open ends to establish a realistic budget. Cost calculation and division fails on seemingly simple questions such as: what is community-specific and what is generic? For instance, CLARIN faced the situation that one of its most essential ingredients, namely assigning PIDs to all of its resources, had not been tackled by horizontal activities at all, although it is a very generic bit of technology. Eventually, CLARIN initiated a financially feasible service open to the research community. This is why more horizontal and vertical discussions are required amongst the various key initiatives to come to a balanced proposal that will convince funding agencies.

In summary, we have shown that CLARIN is aware of the necessity to address the various communities mentioned above and that we have achieved a lot in approximately one year of hard work, in particular if we can launch the *Virtual Language World* portal at the NEERI 09 event in Helsinki in September/October 2009. This event will also give us the opportunity to increase interaction with other infrastructure initiatives.

Editors' Foreword

Marko Tadić
& Dan Cristea
CLARIN Newsletter editors

We offer this issue to our readers at the half way of our project. It's time to count some eggs, sure not yet the ducklings… This was indeed what Peter Wittenburg had in mind when writing the opening article. Because CLARIN has become well known, in Europe at least, many people started to become nervous. Those inside the community, as they want to show to the world already some results (which would diminish somehow a certain amount of grumpy gossips about an expensive community doing nothing but travelling), and those outside the community, eager to consume what we have promised. At the half time break (before summer) Peter addresses 5 thoughts to 5 distinct communities.

We continue with two articles that present evaluation results of the two calls issued this year: the one addressing collaboration with HSS projects, article contributed by Tamás Váradi and Koenraad de Smedt, and the usage scenarios call – an article written by Valeria Quochi. As you will see, 3 and, respectively, 4 winners have been preselected in these two contests. And the game has only begun, because CLARIN will have now to prove that it can provide a kind of CLARIN-grounded workflow as solution in all cases.

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 or directly to the editors at marko.tadic@ffzg.hr and dcristea@info.uaic.ro.

The middle pages, as well as the next one, are dedicated to the very important Consortium meeting that took place in Barcelona in the second decade of May. We have invited Carla Parra and Eva Revilla, members of the host organisation – Universitat Pompeu Fabra, to make a general presentation of this extremely well organised event. Then, Frank Binder and Dan Cristea describe the training and dissemination session, in which both of us (Dan and Marko) together with 5 other colleagues have been actively involved. Although not financed in this preparatory phase, training will become a big issue in the construction and exploitation phase and therefore merits already a vivid attention.

Dieter Van Uytvanck signs two articles on page 9, the first one alone – announcing the launching of the attractively interesting Virtual Language Observatory (VLO), an MPI achievement, and the second one, together with his colleague Florian Wittenburg – presenting briefly the Research Connection event in Prague, at the beginning of May.

The last two articles bring into focus activities in LRT performed by an European agency and at a national level. In what way ELRA – the already 14 years old well-known association – connects to CLARIN and why such a marriage is fortunate is commented by Victoria Arranz and Khalid Choukri. Finally, Svetla Koeva presents the Bulgarian LRT sector, including research and teaching performed in universities, state institutes and in the private sector.

Enjoy the reading!

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

The Results of the Call for Collaboration with Humanities and Social Sciences projects

Tamás Váradi
CLARIN EB member
Koenraad de Smedt
University of Bergen

CLARIN aims to bring language technology to the benefit of humanities and social sciences (HSS). In order to provide this service efficiently it is important that we gain first hand experience in collaborating with actual projects in our target community. For this reason we issued a Call for Collaboration, as was reported in Newsletter No 4 (pp. 2-4). This article reports on the results of the Call and discusses current work in connection with it.

Ten applications

We received as many as ten applications to the Call. The table below shows an alphabetical list of the applications. Each application was reviewed by at least two members of Working Group 3.3, the results were presented at the meeting of WP3 partners in Barcelona and submitted to the EB at its meeting in Iaşi.

The applications for cooperation were very instructive in their own right, as many pointed out clear needs of humanities research using digital language materials and computer tools. The need for strategic guidance on standards and best practices for annotation clearly emerged as an important issue, while other questions related to the suitability of specific approaches and tools for the purposes of a particular study.

Demonstrative proposals

We decided to set up three groups among the applications. The first group comprised those that best demonstrate the use of LRT and would show the potential of a research infrastructure in the humanities. It is therefore in the interest of CLARIN to closely monitor and advise these projects. This group includes CAP, CKCC and MLT-CPhil.

LRT dependent proposals

The second group, DATIST, DID-Cph and LaMeCos, have been selected for limited cooperation and support, as it was felt that while the projects are of interest to CLARIN, the scope and nature of the required language technology work is quite specific, allowing relatively little room for testing the capabilites of the CLARIN infrastructure.

Cooperation with CLARIN partners: 3 out of 10

The third group consists of CONPLISIT, HISTOPOL and HTM4-EmodE, three projects which already have a planned or ongoing cooperation with some CLARIN partners, which is hereby endorsed with a formal arrangement. Such cooperation typically does not require central CLARIN funding but may well rely on local resources.

We should emphasize that our selection criteria reflect the very specific purpose for which this call for cooperation was

CAP	A Hierarchical Lexical Function related to Proper Nouns	Université de Strasbourg
CKCC	Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic.	Utrecht University
CONPLISIT	Consumption patterns and life-style in Swedish literature – novels 1830-1860	Göteborgs universitet
DATIST	Studying speech and language therapy diagnosis, using statistical analysis and textual statistics	ATILF-CNRS
DID-Cph	Consultancy for the Dictionary of Danish Insular Dialects	University of Copenhagen
HISTOPOL	Narrative Social Psychological Studies of European History	Hungarian Academy of Sciences
HTM4EmodE	Historical text mining for assisting the study of discourses in Early Modern England	University of Liverpool
LaMeCos	Database and Database Analyzer of Medieval Latin Scientific Terminology	Universitat de Barcelona
MLT-CPhil	Multilingual language technology for classical philology research	University of Hamburg
Pro_Trans	The professionalisation of translation in the North of Portugal	University of Minho

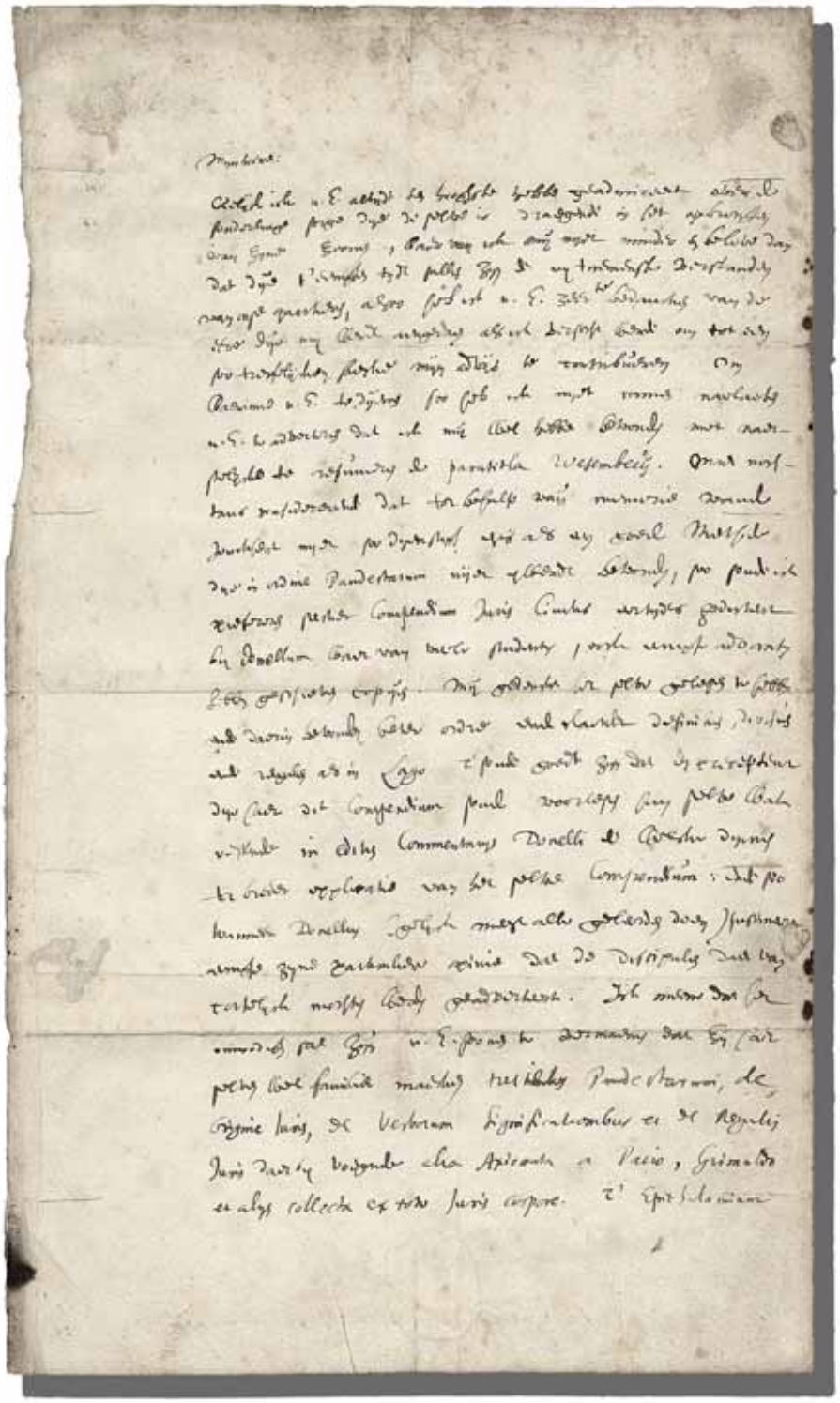
The list of project proposals submitted to CLARIN call

issued and are by no way intended to imply anything about the intrinsic value of the projects concerned. With this in mind, let us briefly review the three projects CAP, CKCC and MLT-CPhil.

CAP focuses on the analysis of a lexical relationship "head of" as defined by Igor Mel'cuk in English, German and French. Extracting data from the analysis of parallel corpora, the project aims to set up a lexical database reflecting hierarchical relationships in these languages, providing a resource that may be important for sociologists and historians as well as linguists.

CKCC is a very ambitious project aiming to investigate the complex issue of how knowledge was shared and disseminated in the Netherlands and indeed throughout Europe in the seventeenth century. It intends to pursue this research goal through compiling and analyzing a corpus consisting of thousands of letters that scientists and eminent figures of the day wrote to each other. This project plans to use a whole array of fairly advanced language technologies ranging from named entity recognition to automatic topic detection etc. Thus it has the potential of demonstrating the widespread deployment of the CLARIN infrastructure in a convincing manner. The figure aside illustrates one of the sources the project will focus on.

MLT-Phil is also a large-scale project that intends to provide multilingual query and retrieval services to a database archive of ancient documents. This project, too, is very ambitious in its intended use of language technologies and intends to make a great impact in the humanities and the cultural heritage domain. MLT-Phil, just as CKCC can expect to have further support from the local CLARIN network.



CKCC project: a letter from Christiaen Huygens

Negotiation process

Currently, Work Package 3 partners are engaged in negotiations with the coordinators of the above mentioned projects with a view to specifying the scope and the particular areas of collaboration with CLARIN. We decided at our meeting in Barcelona to explore common aspects of the workflow required by the projects and pool our resources together within WP3 workgroups and outside, namely with the effort going on relating to the User Scenarios.

The reader may wonder how is this line of work related to the user scenarios reported on by Valeria Quochi in this issue of the Newsletter. Although the user scenarios and the humanities projects have started from different angles, it is hoped that they supplement each other and eventually converge, in particular if some of the workflows addressed in the user scenarios may find ready deployment in some of the humanities project plans.

Let's hope that this will indeed happen with many scenarios. C

Usage scenarios and basic workflows

Valeria Quochi
CNR-ILC, Pisa

Language is the primary means for expressing and communicating ideas, results and findings. In CLARIN, we all know that language technologies are the key to accessing content in a quick and efficient way and, if deployed massively, they could really make the difference especially in SSH research. Our task and goal in CLARIN is to show this to the community. For these reasons, the first scenarios we are aiming at have to propose feasible solutions to real-world tasks in SSH research.

At the last consortium meeting, Barcelona 11-13 May 2009, we closed the Usage Scenario activity by awarding the 4 best detailed scenarios participating in the "contest": UPF's (University Pompeu Fabra, Barcelona) "Language Indicators" scenario, IMS' (Institut für machinelle Sprachverabeiterung, Stuttgart) "Multiword and term extractor" scenario, ILSP's (Institute for Language and Speech Processing, Athens) "Translation repository" scenario, and University of Rome "Tor Vergata"'s "What we are is what we eat".

These four scenarios have been selected as representatives of groups of scenarios proposed, and therefore, we can say that, by collaborating at this task, everybody won!

Four selected representative scenarios

The "Language Indicator" scenario targets a variety of users and comprises several functionalities common to many of the scenarios proposed. Its main purpose is to provide tools for monitoring and analyzing language use, for example extracting keywords, frequencies, etc. It provides, therefore, assistance in quantitative and qualitative analysis of various kinds of texts, and thus is useful in various sub-fields of linguistics and of social sciences, where they need to perform qualitative analysis of speeches, interviews etc. For example, a historian or a politologist studying political speeches of a given time or a given statesman. In terms of workflows, the scenario provides a detailed description of the various steps and tools required (for Spanish and English at least), some of which are already realized as webservices.

The "Multiword and term extractor" scenario, presents a showcase for multilingual extraction of terms and multi-word expressions. Beyond theoretical linguists, it addresses the needs of terminologists and lexicographers. The scenario includes several (multilingual) core components, on top of which then different and more complex functionalities could be added, and which are already working as webservices.

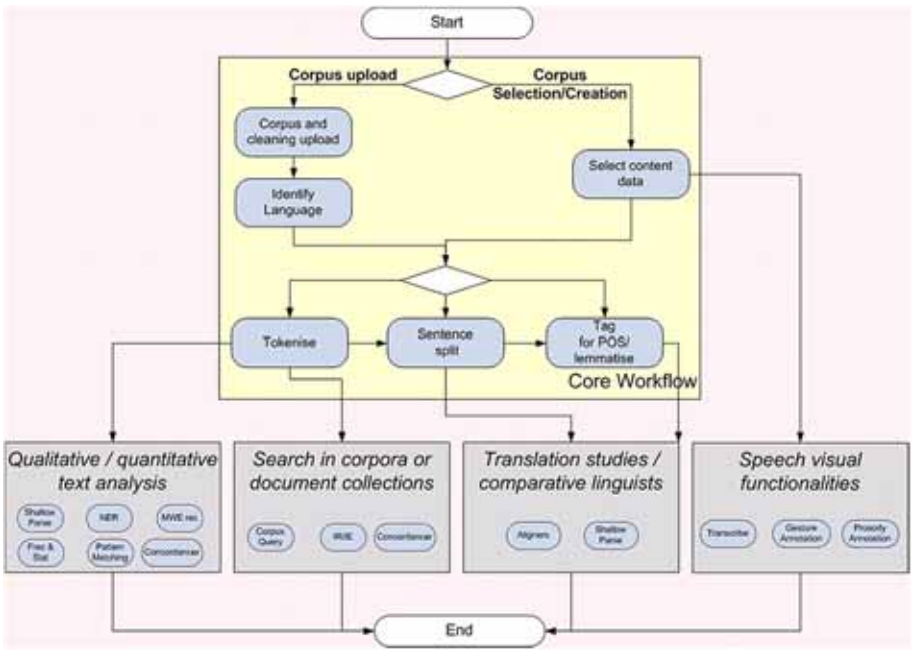


Diagram of the basic workflow

The "Translation Repository" scenario actually addresses various multilinguality issues and targets not only translation studies, but also other linguistics or literary research where cross-language comparative or contrastive studies are required, by including alignment functionalities.

The "What you are is what you eat" scenario, targets humanists, like historians, and describes how basic text processing functionalities can form a useful pipeline that performs tasks that were previously hand-made, without requiring users to deeply understand the technology behind: relevant information found in text is stored into a database and searched for by the user. The intended target users are historians, but other user groups could benefit from the scenario as well.

Basic Workflows

All contributed scenarios, in fact, were fundamental to reach a better understanding of potential user needs and of the big challenge CLARIN is facing. In particular, they helped identifying basic core functionalities and pipelines (simplistically described in the figure) that constitute a basis for performing the more complex tasks. A core pipeline (yellow box in the picture) that are shared by most of the scenarios require functionalities for corpus upload or selection, cleaning and eventually indexing, as well as basic NLP tools such as tokenizers, language identifiers, sentence splitters, POS taggers and/or lemmatizers. On top of this core box, other tools can then be plugged in to perform more complex task of different nature. Some tasks require further language processing tools, others seem to involve (multilingual) information access technologies, others are more oriented to speech and multimodal data. In all cases, only relatively robust technology has been considered at this stage, leaving it for the near future to integrate more sophisticated technologies.

Challenges for the prototype

Defining basic pipelines gives us a good starting point for the construction of the prototype. One of the first steps in this direction, together with the availability of tools or pipelines as webservices, is to tackle interoperability obstacles of various nature (data formats, interfaces, compatibility, I/O formats etc.). Interoperability is in fact a huge challenge for CLARIN. As it appears clearer and clearer as we proceed, interoperability among different components is already hard to achieve with very "simple" workflows, it is therefore of primary importance to concentrate on these "simple" functionalities first.

One of the main take-home messages from the scenario experience therefore is:

Start small to grow stronger!

CLARIN Consortium Meeting

Barcelona May 11-13, 2009

Carla Parra
Eva Revilla
University Pompeu Fabra, Barcelona

From 11 to 13 May the Institut Universitari de Lingüística Aplicada at the Universitat Pompeu Fabra in Barcelona (IULA-UPF) hosted the annual CLARIN Consortium Meeting. For the IULA this event was really important, as CLARIN is one of its main current research projects, and therefore everyone involved did their best so that everything would be arranged up to the smallest detail. During the intensive two and a half days of the meeting, the CLARIN members who attended it discussed the current status of the project and the next moves towards our final aim.

The opening ceremony with Steven Krauwer, Nuría Bel and representatives of Generalitat de Catalunya and Spanish Ministry

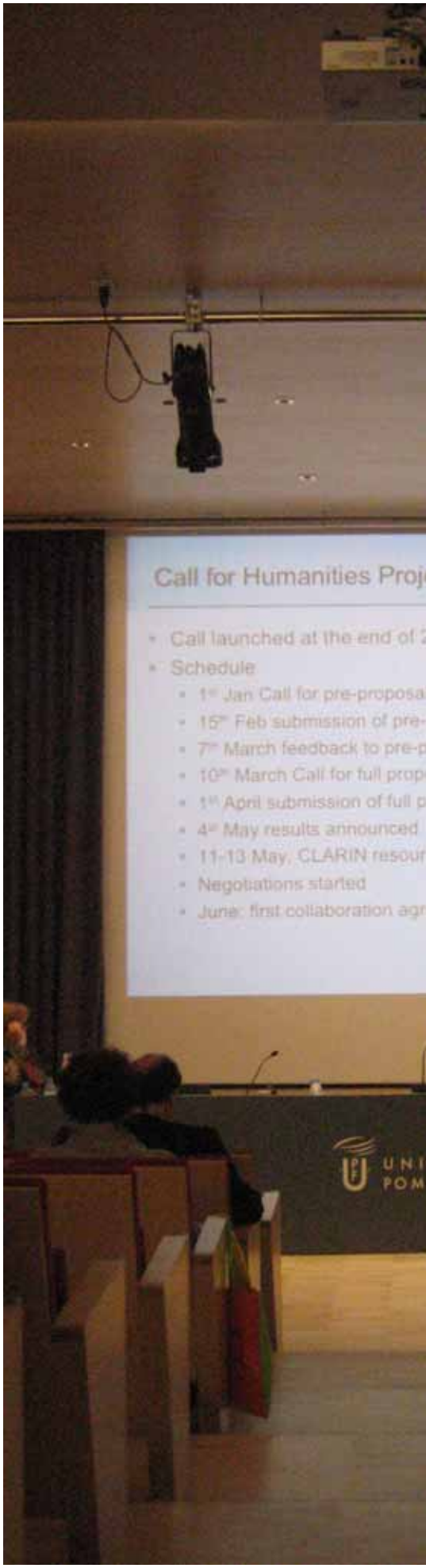
Key points of the project

On Monday morning the different site managers chaired by Steven Krauwer (UU), CLARIN Coordinator, gathered together and discussed the current situation of the project. Besides a general overview of the present situation and planning offered by Steven Krauwer, the morning session concentrated on some key points for the project. During her presentation, Hetty Winkel (UU), CLARIN's project manager, pointed out the requirements for the first intermediate report, to be submitted in the near future. Tamás Váradi (HASRIL), manager of Work Package 3 (Humanities Overview), explained the use cases selection which has been going on in this work package and Bente Maegaard (University of Copenhagen), manager of Work Package 8 (Construction and Exploitation Agreement), stated the state of affairs with regard to governance.

CLARIN call for project proposals

During lunch, the CLARIN members who are not site managers also joined the meeting and in the afternoon the thematic sessions started. Monday's session was focused on the future CLARIN ssers. Tamás Váradi (HASRIL) and Koenraad De Smedt (UIB) presented to the audience the projects received during the Call for European Projects carried out by WP3, while Erhard Hinrichs (UTU) and Valeria Quochi (CNR-ILC) gave an overview of the main user scenarios gathered in the survey carried out by Work Package 5 (LRT Overview). Finally, the authors of the "prototypical user scenarios", Marta Villegas (UPF), Fabienne Fritzinger (University of Stuttgart), Fabio Zanzotto (U.Roma), Prokopis Prokopidis (ILSP-ATHENA RC) and David House (KTH) gave a presentation on their use cases and how they are being implemented at their different institutions as show cases of the CLARIN infrastructure potential.

On Monday evening all the participants of the meeting were invited by the IULA-UPF and the Language Resources Technologies



Támas Váradi presenting the results of

f the CLARIN call for project proposals

Research Group to a dinner at the restaurant "La Miranda del Museu", located on the fourth floor of the Museum of History of Catalonia, offering a wonderful view of the port of Barcelona and Montjuic.

Legal issues

Tuesday morning was devoted to one of the key topics to be addressed by CLARIN: legal issues. As we all know, when building an infrastructure such as CLARIN, crucial aspects like licensing, authorisation and authentication will be problematic if we haven't foreseen and studied all possible scenarios. Furthermore, we shall also bear in mind the different legislations existing in the partner members' countries. While Kimmo Koskenniemi and Antti Arppe (UHEL) concentrated on authorisation and authentication issues, their colleague Marjut Salokannel shared with the audience a thorough study on the differences in legislation amongst the different partner countries and stated the things we shall bear in mind and work on so that the infrastructure won't be facing legal problems in the implementation phase.

Education and dissemination

In the afternoon, the topic changed towards education and dissemination, a session chaired by Dan Cristea (UAIC) and Frank Binder (University of Giessen), and in which Martin Wynne (OTA), Arjan van Hessen (University of Twente), Koenraad De Smedt, Thierry de Clerck (DFKI) and Marko Tadić (FFZG) actively participated as panellists and presented their own experiences and views.

After this short session, the participants in the meeting split up in four different groups to work on the different tasks to be done in the near future. The members of Work Package 3 met to discuss the call for European Projects and work out the schedule for the tasks in this WP during the next months. On the other side, the 3 newly established Working Groups in Work Package 5, Taxonomy, BLARK and Integration, had their "kick-off meetings", in which the WGs' aims were established, as well as their working schedules to meet the overall time schedule within the CLARIN project. WP3 meeting was chaired by Tamás Váradi; WG 5.4. BLARK was chaired by Kathrin Beck (UTU) and Erhard Hinrichs, and had Mike Rosner (UMT) as rapporteur; WG 5.5. Taxonomy was chaired by Iris Vogel (UTU) and Erhard Hinrichs; and finally WG 5.6. Integration was chaired by Adam Przepiórkowski (IPIPAN).

Metadata and standards

After these intense and productive sessions, the second day of the meeting was finished. The last session, held on Wednesday morning, focused on metadata and resource formats in CLARIN. Before focusing on the hot topic of metadata, the chairs of the internal WPs and WGs meetings from Tuesday afternoon briefed their discussions so that everyone was updated on the work being planned and those members who could not join all the meetings, as they were carried out in parallel, could do valuable contributions. During the session on metadata, chaired by Erhard Hinrichs and Peter Wittenburg (MPI), the metadata to be used in CLARIN were explained and some discussions on the topic arose.


A lot of fruitful discussions were going on during breaks

With this session, the CLARIN Consortium Meeting was closed. Once again, on behalf of the Barcelona team, we would like to thank you for letting us be your host during the Consortium Meeting. It was a pleasure.

Training and dissemination in CLARIN – the potential to bridge gaps

Frank Binder
University of Giessen
Dan Cristea
University of Iaşi

The Education and dissemination session at CLARIN consortium meeting

Through its network of partners and members, CLARIN connects a large number of institutions and personalities who have substantial experience in training and supporting users and providers of language resources and technology. A thematic session of the CLARIN consortium meeting in Barcelona, in May 2009, was dedicated to sharing and developing perspectives on training experience and plans within and around CLARIN.

Sketches from a session in Barcelona

On the second day of the CLARIN consortium meeting in Barcelona, a collaborative thematic session on "training and dissemination in CLARIN", chaired by Dan Cristea and Frank Binder, featured seven experts from six countries, who presented their perspectives on LRT-related training.

Martin Wynne and Dan Cristea talked about their long time involvement in supporting and training researchers at the Oxford Text Archive and through the EUROLAN series of summer schools, respectively. Koenraad De Smedt drew the attention to CLARA, a Marie Curie Research Training Network dedicated to LRT-enabled linguistic research and undergoing final negotiation at that time (now approved). Arjan van Hessen (CLARIN-NL) and Frank Binder (D-SPIN) presented training plans and activities of two national CLARIN support initiatives, from the Netherlands and Germany. Thierry Declerck suggested registries of expertise, such as lt.world.org, as an entry point and information hub for interested LRT-users. Marko Tadić stimulated the subsequent discussions by warning against a serious gap between top-down and bottom-up approaches in CLARIN's relation to its prospective users from the Humanities and Social Sciences. Close interaction and feedback from the user communities is vital for CLARIN to succeed in the long run. Since training activities involve two-way face-to-face communication between CLARIN and its prospective users, community training is a key opportunity to bridge this gap.

Considering Steven Krauwer's guiding questions, the session offered different perspectives and an exchange of ideas. Many concrete issues, such as how to enter humanities curricula at various levels across Europe, remain to be solved. For subsequent phases, CLARIN will need to have a strategy on training that supports and connects ongoing national activities.

Background: Training in CLARIN

As a Pan-European research infrastructure CLARIN involves various aspects of communication. A major communicative goal is to inform CLARIN users about our technology and services, which is a key factor to success as CLARIN progresses. Hence, community training is a fundamental aspect.

Training and user support, however, is currently not covered at CLARIN level. Through funding decisions taken by the EC, issues of community contacts and awareness have largely been delegated into the national realms of CLARIN. Accordingly, for training activities, CLARIN currently has to rely to a large degree on national support initiatives. As discussed above, the need and potential to communicate, support and coordinate training activities at CLARIN level has recently become obvious at the latest CLARIN consortium meeting.

For CLARIN there are at least three relevant perspectives on training:

1. Training activities should aim at offering immediate interaction between CLARIN and its users. A bottom-up channel of communication should also be exploited, one which is too valuable to be ignored. User experience and suggestions need to be considered for setting goals, acquiring practical experience and offering it back to the public. Identifying usage scenarios is just one important step in that direction.
2. CLARIN has certain responsibilities regarding its own human resources. New staff and researchers will be needed to work for CLARIN throughout its long running perspective. CLARIN has an interest in training and knowing its future staff and collaborators.
3. User training and support is vital for the sake of substantial interest in CLARIN services. LRT-providers will have to be informed and trained at various levels about the recommended standards that their resources should comply with and how to make use of the CLARIN integrating technology in order make available their products to the widest audience. On the other side, the LRT-users will have to know how to access and work with the CLARIN platforms and its resources and tools.

From a user-centric perspective, the closely linked objectives of support, training and community awareness may well be addressed through national or local events. These events then should involve CLARIN members as well as local developers and users of language resources, possibly within the respective research communities. Besides CLARA, such events and activities are now solely taking place locally at national levels.

It is clear therefore that, if we want a unifying vision in training and education (components of common curricula at master or bachelor level, inclusion of LRT-related topics in humanity curricula, mobility schemes for students and teachers, common criteria of evaluation, maybe also issuing certificates about CLARIN skills), in the future phases, CLARIN should provide support for the training-related activities and should also find ways to connect the distributed local training efforts to its pan-European vision. C

The Virtual Language Observatory

Dieter Van Uytvanck
MPI, Nijmegen

Libraries are often considered to be the natural habitat of the prototypical language researchers. But does the same go for their 21st century counterparts: the digital repositories? Probably not. Although access to them is not bound anymore to a specific location, their user interface often cannot compete with the charm of glancing through a book.

To address this lack of attractiveness the concept of the Virtual Language Observatory was born. Just like astronomers can make a voyage of discovery these days from behind their computer, the intent is to provide humanities researchers with a digital atlas where they can find their way to location-related language resources and tools.

A corpus for French? Just zoom in on France and click around. What members does CLARIN have in the Baltic area? You can identify them immediately by the icon on the map. Or just play around and learn more about the linguistic software created in Sweden. All points can in turn contain references to websites, repositories, online annotation viewers, etc.

CLARIN members as viewed from the Virtual Language Observatory

Needless to say that this is of course only one of the many access methods to the CLARIN universe of digital resources. But it's clearly one without high barriers.

At the time of writing the Virtual Language World just passed the big bang phase and is still getting into its final form but you can expect it to be ready soon. It has come here to stay.

To be continued... C

Research Connection 2009

Prague

May 6-7, 2009

Florian Wittenburg
MPI, Nijmegen
Dieter Van Uytvanck
MPI, Nijmegen

Networking our way to a research future – that was the slogan of the Research Connection 2009 conference, organized by the European Commission. On May 6 and 7, the Prague Congress Hall attracted a variated mix of researchers, policy makers, journalists and students. Next to a broad choice of thematic sessions and press conferences the visitors were also welcomed at an exhibition of research infrastructures. That's were CLARIN came into play: as the sole representative of the humanities, an interactive booth based on the theme "language and mind" managed to catch the attention of quite some public.

Visitors could learn about grammar, phonology, pragmatics and much more in a multi-media presentation that has been created by the French Quai-Branly museum. Many were surprised to discover that only by listening to a weather forecast one can make its first steps in learning Chinese. An image-naming reaction time experiment ensured the playfulness of the whole stand. The interactive website about the documentation of endangered languages within the DOBES programme illustrated clearly what a language resource actually is. Finally all CLARIN members and their contributions to the LRT inventory were brought into the spotlight by a first version of the Virtual Language Observatory (see elsewhere in this newsletter for details).

The practical "discover-it-yourself" setup was clearly appreciated by the conference participants: we received lots of enthusiastic reactions. Presenting the humanities and social sciences next to impressing models of fusion reactors certainly was a challenge. Thanks to the interactive approach the positive CLARIN energy was nevertheless very noticeable in Prague. C

ELRA & CLARIN: Sharing Language Resources for Different Communities

Victoria Arranz
Khalid Choukri
ELRA / ELDA

For more than 14 years, since its creation in February 1995, ELRA (the European Language Resources Association) has focused its activities on a central point of interest: Language Resources (LRs).

ELRA's Mission and Services

One of the main rationale that lays behind that orientation is to bring into focus the need for a mutual exchange and use of the LRs that are required for research and development works in the Human Language Technology (HLT) world. Thanks to the funding of the European Commission during its first three years of activity and with the support of very active experts of the field, ELRA succeeded in providing the HLT community with a now internationally known platform of services for the identification, collection, validation and distribution of LRs. In order to answer to the ever increasing needs of this growing community, ELRA worked out at its early stage of development (October 1995) on the creation of an operational body to help with its everyday life activities: ELDA (the Evaluations and Language resources Distribution Agency).

ELRA Board

For those not familiar with the way the association is run, we will just add a few lines on this. As a non-profit organisation, the management of ELRA is entrusted to a Board. The Board consists of 9 elected members who reflect the various dimensions of the Language Resources and Evaluation field. They are elected by an open vote of all the ELRA members.

Based on the long expertise and knowledge of its members, the Board is responsible for defining the strategy and activities of the association. Its decisions are then implemented by its operational body, ELDA.

At present, the ELRA Board is led by its President Stelios Piperidis, from the Institute for Language and Speech Processing (ILSP), Greece.

ELRA's Services

ELRA, through ELDA, carries out a wide variety of activities related to LRs. These activities, further developed over the past few years, can be distributed over the following groups of services, where each of them consists of a long series of sub-activities:

- Identification and Distribution/Sharing of LRs, and related sub-activities: these can be represented by a two-direction relationship with two different types of entities, namely the providers and the users.
- Legal Support & IPR Clearing: where ELRA may undergo activities of IPR clearing either in the framework of its own negotiation and distribution activities or simply as support to LR users/owners to handle their own needs.
- Production of LRs: further to its interaction with both providers and users, ELRA also invests considerable efforts on the production of LRs.
- Technology Evaluation: ELRA plays an important role in the area of Technology evaluation, combining its skills acquired with LRs and adapting them for the improvement of Language Engineering products.
- Information Dissemination: Both for its members and general users, ELRA carries out tasks of Dissemination, in order to inform on the resources available and any relevant activity (with the maintenance of catalogues, edition of the ELRA Newsletter, the organisation of the LREC Conference, and the maintenance of the HLT Portal, among others).
- Involvement in "community building": through a number of initiatives like FlareNet and CLARIN, ELRA remains very active in its work towards Language Resources and Human Language Technologies.

A Joint Ground for Evolution and Collaboration both for ELRA and CLARIN

As we all know, CLARIN is in its second year of work, advancing well and learning a lot from its partners and the community. As we have seen in the numerous meetings, workshops and Newsletter articles, work is moving ahead in a variety of directions, in terms of metadata, integration, dissemination, etc. As a member of the consortium, ELRA is also participating in these activities, aiming to support a rather neglected community in terms of language resources: the Humanities and Social Sciences (H&SS) community. For ELRA, CLARIN's mission to help this community brings back the memories of opening this path for the Human Language Technology community, a path that the Association initiated more than 14 years ago.

Throughout these years, ELRA has gained expertise in a number of areas, and this expertise has been put to good use in the services described above. Looking at these services and the needs the CLARIN community may encounter, we quickly come up with a few ideas and proposals for collaboration with CLARIN:

- Identification of LRs: CLARIN is doing hard work on the identification of language resources and ELRA would be happy to take part in this. As it has been mentioned during earlier meetings, ELRA maintains 2 language resource catalogues: the ELRA Catalogue, with language resources whose rights have been discussed and negotiated and which are available to the public; and the Universal Catalogue, which aims to be a repository for all identified LRs and plays a very important role towards the locating and negotiating of LR distribution rights for its users.

The Universal Catalogue is also part of a larger international initiative, run jointly with institutions in the USA and Asia, to achieve a unique reference catalogue of language resources for HLT. This is a unique opportunity for CLARIN, who can spare the efforts of redeveloping another LR repository for HLT and could thus focus on one such tool for H&SS. As an initial step to benefiting from this already existing service, an extraction of all the resources that may be used by the CLARIN community will be done by ELRA from both the ELRA Catalogue and the Universal Catalogue, and provided to CLARIN so as to be inserted in the Virtual Language Observatory, together with its metadata descriptions.

- Handling of Legal Issues: Another important issue in the availability of LRs is the handling of legal issues. With its long expertise in the area, ELRA is ready to take responsibility for this, if asked.

At ELRA, the basic principles of language resource licensing have been worked out with the support of lawyers. One of ELRA's priority tasks was to simplify the relationship between producers/providers and users of LRs. In order to encourage producers and/or providers of LRs to make such data available to others, ELRA drafted generic contracts defining the responsibilities and obligations of both parties. Such contracts establish, among other things, what usage is allowed for the LRs, whether for both research and technology/product development or only for research purposes. In any case, these contracts protect the providers and their LRs. These licenses have already been shared with CLARIN and a more detailed discussion could take place with WP7 so as to help them adapt the documents for CLARIN's particular audience, if asked.

- Standards and metadata: As we all know, a meeting by the CLARIN metadata expert group took place in Athens last January, 31st. We are pleased to say that the group is composed of an important number of language resource description experts, who already carry the expertise of struggling through earlier initiatives in the past. ELRA also joined the group, reflecting upon critical issues with the rest of the team, and looking into its experience from its own metadata as developed in agreement with that of INTERA. C

Bulgarian Language Resources and Applications

Svetla Koeva
Institute of Bulgarian Language, Bulgarian Academy, Sofia

Bulgarian is spoken by approximately 12 million native speakers, mainly in Bulgaria, but also in Greece, Macedonia, Romania, Serbia, Turkey, Ukraine, Australia, Canada, USA, Germany and Spain. The official alphabet is Cyrillic, one of the three official alphabets in the EU. Bulgarian is the first Slavic language with its own writing system dating from the 9th century.

Bulgarian belongs to the family of South Slavic languages and it is a part of the Balkan linguistic union (Balkan Sprachbund) – consequently Bulgarian displays similarities with both language groups. As a Slavic language Bulgarian possesses a rich inflectional and derivational morphology, verb aspect pairs, etc. Due to the mutual influence of Balkan languages, Bulgarian has lost noun cases (except of rear vocative), completely has lost the infinitive, etc.

Natural Language Processing in Bulgaria is a relatively new scientific branch. Until recently there have been no complex large scale applications for Bulgarian, comparable to those of the more studied European languages, and aimed at serving the electronic society, government, education, business and communications. It is only for the past years that some significant electronic language resources were created (wordnet, dictionaries, corpora, lexical databases), as well as tools for their processing (Part Of Speech tagging, spelling correction, advanced text searches, etc.). There are several major institutions in Bulgaria where language resources are developed: in Sofia University (spoken language corpora, dialect archives, etc.), in the Institute for Mathematics and Informatics (parallel corpora, multimedia digital archives, etc.), in the Institute for Parallel Processing of Information (a Treebank, parallel corpora, electronic dictionaries, etc.), in Plovdiv university (electronic dictionaries, archives with literary texts, etc.), in the Institute for Bulgarian language (the Bulgarian national corpus, Bulgarian wordnet, Bulgarian FrameNet, etc.). These investigations however were not synchronized until now (as in many other European countries) – resources were often duplicated at different research centres (there are at least three Bulgarian electronic morphological dictionaries) – which leaded to serious losses of labour time.

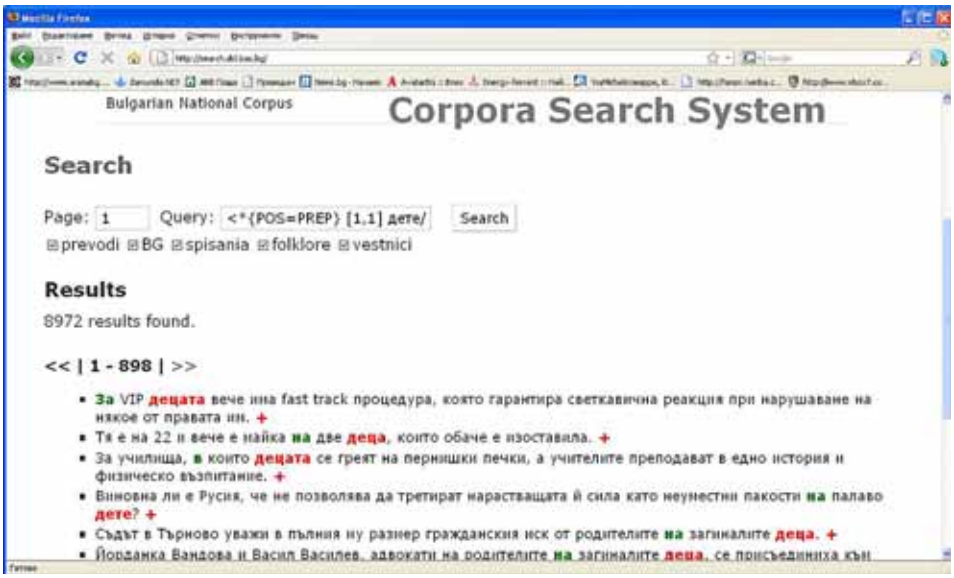
Similar conclusion (fast increase and relative dispersion) can be made for the NLP applications for Bulgarian as well, which can be generally divided into several types – for research optimization (for example, computer aided dictionaries development) and for society in general (for spelling and grammar checking, for information extraction, etc.). There are a number of programs in the first group. However, these are mainly for linguistic purposes: for example, there are several applications assisting the development of morphological dictionaries, corpora annotation, etc. Within the software products for non-specialized use, a central place is given to the so-called Proofing Tools – for check-up of the text correctness (there are several commercial applications with a good quality: Slovnik by Sirma Group, Kirila by AKT Soft, FlexWord by Datecs), while high quality programs for text processing on higher levels (for example grammar correction or automatic translation) are subject to further development. The Bulgarian Association for Computational Linguistics is an attempt at synchronization of the work of researchers from various scientific institutions, which so far has resulted in the creation of the Bulgarian speech synthesizer SpeechLab 2.0.

Education in Computational Linguistics is offered by the University of Plovdiv, at the undergraduate level (for instance, courses in Computational linguistics) and by the University of Sofia at the master level (master programs in Computational Linguistics. Internet technologies in Humanities and Artificial Intelligence).

Bulgaria is well known with the organization of some valuable scientific events: RANLP (Recent Advances in Natural Language Processing) – a biennial conference with an internationally recognized very high level, a meeting point of scientists from all over the world), and FASSBL (Formal Approaches to South Slavic and Balkan languages – a biennial international conference).

Due to space limits, no detailed specification and comparison of the characteristics of the existing language resources, technologies, and software products for Bulgarian can be presented here. Generally, there is a critical mass of already developed electronic language resource and tools. However, the following general conclusions can be drawn, concerning needs of: better coordination between research centres, common standardization and/or convertibility of the resources, availability of a documentation and clarity on the copyright issues, information on the possible applications of the language resources in scientific research, better connection between the commercial products and the scientific developments, unified policy for education and further qualification of human resources in the field. The scientific community in Bulgaria however realizes the fact that language resources and technologies can be successfully used to increase significantly the effectiveness and the quality of the research (especially in the humanities and social sciences).

As a conclusion, CLARIN has a significant role for the establishment of a functioning research infrastructure for creation, compilation, maintenance, standardization, and transfer of language resources and tools for Bulgarian (as well as for evaluation and comparison of their characteristics).



The search system of the Bulgarian National Corpus

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.

June 2009

2009-06-22 to 2009-06-26: Digital Humanities 2009 – research infrastructures panel, University of Maryland, Baltimore, USA

2009-06-25 to 2009-06-26: eScience Seminar Repository Systems, Munich, Germany

July 2009

2009-07-06 to 2009-07-07: Sustaining Digital Resources in the Humanities, Cambridge, UK

2009-07-20 to 2009-07-24: Text Encoding Initiative Summer School, Oxford, UK

2009-07-26 to 2009-07-27: Sign Linguistics Corpora Network: Data Collection Workshop, London, UK

August 2009

2009-08-02 to 2009-08-02: 47th ACL and 4th IJCNLP 2009, Singapore

September 2009

2009-09-08 to 2009-09-11: International Conference for Digital Libraries and the Semantic Web, Trento, Italy

2009-09-13 to 2009-09-18: 12th International Conference on Text, Speech and Dialog, Plzen, Czech Republic

2009-09-30 to 2009-10-03: NEERI09, Networking Event for European Research Infrastructures, University of Helsinki, Helsinki, Finland

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