

CLARIN



Newsletter

Number 3, 2008, October

Language Infrastructures: what happens outside EU?



Nicoletta Calzolari
ILC-CNR, Pisa, Italy

The setup of CLARIN in Europe was the result of a long series of initiatives and attempts from many of us, starting already at the beginning on the 6th Framework Programme. That time is finally ripe for such an infrastructure is shown also by other initiatives outside Europe that share objectives and ideas with CLARIN. I mention here just a few.

Probably the most similar is a 5-year program just finished in Japan: the 21st Century COE (Center of Excellence) Program "Framework for Systematization and Application of Large-scale Knowledge Resources", led by Sadaoki Furui at Tokyo Institute of Technology¹. It aimed at systematising and relating a variety of multimedia information to make use of them as 'knowledge': this is one of the key issues in the 21st century. I was member of the International Board and could observe the breadth of the areas covered to construct, integrate and use large-scale knowledge resources (from spontaneous speech, written language, to materials for e-learning and multimedia teaching, classical literature and historical documents) in many domains of research for Human and Social Sciences. I also witnessed one of the most difficult problems in such interdisciplinary research combining humanities and technology, i.e. communication among researchers belonging to different communities and with very diverse backgrounds. I think this is a problem that CLARIN should expect and to which it must dedicate attention now already.

Another ongoing Japanese project with similarities with CLARIN is Language Grid, led by Toru Ishida at the National Institute of

Information and Communications Technology (NICT)² and Kyoto University³, with partners also in Europe (DFKI, ELDA, ILC-CNR). The Language Grid is an infrastructure built on top of the Internet to allow not only professionals but also end users to conquer the language barriers. The project includes language resource (LR) and computation resource providers, as well as language service users, and is based on Web service technologies that enable users to freely combine software distributed via the Internet.

Semantic Web technologies enable the collaboration among LRs and language processing functions for intercultural activities, to improve the accessibility and usability of existing language services and to easily develop new language services by combining existing ones. It also offers an infrastructure where stakeholders can provide and/or use LRs by mutual consent, with understanding and resolution of the intellectual property issues. The main goal is to allow a better understanding of Internet content written in different languages and by people from different countries.

I add to this picture an US-funded effort, the NSF CISE-CRI (Computer and Information Science and Engineering-Computing Research Infrastructure) "Towards a Unified Linguistic Annotation", led by James Pustejovsky at Brandeis, with Martha Palmer, Adam Meyers, Mitch Marcus, Aravind Joshi and Jan Wiebe. This project, developing a Unified Linguistic Annotation (ULA) that integrates in one framework different layers of annotation (e.g., semantics, discourse, temporal, opinions) and several

existing resources, including PropBank, NomBank, TimeBank, Penn Discourse Treebank, and coreference and opinion annotations, aims at providing a large corpus with balanced and annotated data. The project also aims at achieving an international consensus on a meta-specification framework allowing individual annotations to cohabit with each other, as well as a language-independent methodology and widely accessible tools and guidelines. The activity enhances infrastructure for research and

education by providing a resource that could lead to major advances in robust, broad coverage semantic processing.

The set of these initiatives, sharing partially similar perspectives and high-

lighting the value and the need of new language infrastructures is, on one side, a sign of the timeliness of the CLARIN effort, and, on the other, invites all of us to a more global collaboration. As a last remark, I add that efforts toward the cooperation of these and similar initiatives all over the world will also be one of the aims and tasks of two new projects, the European e-Content-plus FLaReNet (Fostering Language Resources) Thematic Network, led by me, and the American NSF INTEROP project, just started, led by Nancy Ide and James Pustejovsky. Worldwide collaboration on these infrastructural issues is an essential step towards better exploitation of all the resources and technologies we develop and therefore towards higher impact of our field in the society. **C**



Participants of Unified Linguistics Annotation Workshop
(<http://verbs.colorado.edu/ula2008/>)

¹ <http://www.coe21-lkr.titech.ac.jp/>

² <http://langrid.nict.go.jp/en/>

³ <http://www.langrid.org/association/indexe.html>

Editors' Foreword



Marko Tadić & Dan Cristea

CLARIN Newsletter editors

Dear readers, As we all entered into a second half of CLARIN's first year, we felt that several key issues had to be presented at this point of our project. The first one is comparison, correspondence and relations with infrastructure building initiatives and projects similar to CLARIN. Nicoletta Calzolari is covering this topic on the front page because we believe that we can learn from others as well as others are learning from us.

The second important topic – covered by Steven Krauwer and Bente Maegaard in an opening article to this issue – is the role of EC and national funding in the phase where CLARIN currently is i.e. preparatory phase. The list of steps for building up the national CLARIN teams is given thus providing a recipe how to establish firm CLARIN communities at the national level that would easily connect to the European level. This article is clearly demonstrating the priorities and best practices to finish this task.

In this issue our regular two-fold contribution, where users and developers share their needs and solutions, covers the topic of endangered languages and the way these language data are recorded, transferred, compressed, archived and used. The contributors on the users' side are Jost Gippert, Sebastian Drude and Peter Wittenburg, while the contribution from the developers' side is by Florian Wittenburg.

The two centerfold pages are devoted to a report (by Tamás Váradi, Marko Tadić, Peter Wittenburg and Peter Tindemans) from an ESF supported workshop of the Alliance for Permanent Access – *Keeping the Records of Science Accessible: Can We Afford It?* The Alliance for Permanent Access was established to ensure that research data (and not just the publications with the results of research) is freely accessible to other researchers. In this workshop different business models for long-term preservation of research data were presented and discussed. Peter Wittenburg presented CLARIN's idea of federated language archives, and three

other members of CLARIN project also attended the workshop.

This issue of our Newsletter finishes with a short note by Jan Šnajder about this year ACL that took place in Columbus, Ohio and four important national correspondents' reports: Mike Rosner from

Malta, Koenraad De Smedt from Norway, Montserrat Marimon from Spain and Maciej Piasecki from Poland. Each of them is giving a survey of LRT situation and CLARIN activities in their countries thus shaping up the European landscape that we have started to observe in previous issues of CLARIN Newsletter.

We wish you a pleasant reading. **C**



Steven Krauwer
CLARIN coordinator
Bente Maegaard
WP8 coordinator

In this article we would like to explain very briefly the relationship between EC and nationally funded activities in CLARIN. We will present (non-exhaustively) a number of ways in which national funding is essential in order to complement the CLARIN activities at the national level.

Background

When the European Commission prepared the call for proposals for research infrastructures, there was an expectation that the EC funding would be complemented by national funding, but it was not explicitly a requirement. National funding is highly desirable as there are many activities that will only be performed if there is national funding to support them. These activities are important for the national teams and for CLARIN as such.

For the CLARIN Preparatory phase we have asked all participating countries for a letter of support from their relevant funding agency, without specification of details about the level of funding and the nature of the contribution. The main reason for this was that many countries were (and still are) in the process of establishing their policies and strategies with respect to the creation of or participation in European research infrastructures.

Furthermore funding of research infrastructures at the national level can take many shapes (e.g. as a stand-alone activity, as part of a national programme, as part of trans-national activities, etc), each of which may require different funding models.

CLARIN activities at the European level

In our work plan for the CLARIN preparatory phase the EC contribution (4.1 million euro for 36 months) will in essence be spent on the generic, language independent tasks, such as the technical specification of the infrastructure, the construction of a prototype, collecting requirements from our users, providing overviews of existing resources and technologies, agreeing on standards, addressing IPR and related issues, coordination of national activities, dissemination, creation of awareness, and the formulation of a draft agreement between the participating countries on the construction and exploitation of the CLARIN infrastructure.

National CLARIN activities

1. Participation in Working Groups
Throughout this phase it has to be ensured that whatever is proposed (standards, tools, technologies, services, licences, etc) will be

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

Relation between EC and National funding in the CLARIN Preparatory Phase

adequate to serve all the language communities (large and small) and all potential user communities.

To this end we have set up a number of working groups to gather information, discuss standards, adapt existing tools and resources to the CLARIN specifications, conduct experiments with the prototype, etc. There are a number of reasons why these tasks cannot be carried out by the consortium partners alone: the working capacity of the consortium is limited, discussions about e.g. standards require broad participation and support, and not all languages and potentially relevant areas of expertise are represented in the consortium. Our solution to this problem has been to open up our Working Groups to participants from all CLARIN member sites (at this moment 109 institutions in 32 countries).

Due to our limited budget we cannot offer any financial support to these participants to pay for their labour or travel expenses to meetings. We would therefore like to urge the national funding agencies to provide *financial support to participants in CLARIN working groups*, both from consortium partner sites (some of whom have a very limited budget from the EC funds) and from member sites. The main justification for this is that this would serve to protect the interests of the national language(s) and the national humanities and social sciences research communities.

2. Demonstrators

Demonstrator services and applications are an excellent instrument to show the potential of the infrastructure for future users. These demonstrators can play a crucial role in both the promotion of the infrastructure and the discovery of user needs.

There is no room in our budget for the implementation of such projects but in our humanities and linguistics oriented work packages we have a modest budget to coordinate the execution of some demonstrator projects and we hope that we will be able to launch a coordinated call for (small) *project proposals for demonstrators in all participating countries*, based on funding from national sources (no cross-border funding is foreseen, but we are hoping for joint projects).

3. Prepare for future role

Keeping in mind that the main purpose of the preparatory phase is to prepare for the construction and exploitation phase, it is also important to look to the future and to use this phase to address issues like the following and (if the answer is affirmative) start making *preparations for the specific role your country wants to play* in the European CLARIN infrastructure:

CLARIN Newsletter (2)

1. Convenio para el desarrollo del demostrador CLARIN-CAT

El Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya y la Universitat Pompeu Fabra han firmado un convenio para la financiación del desarrollo de un demostrador catalán para CLARIN.

El objetivo principal del convenio con la Generalitat es construir el demostrador catalán CLARIN-CAT, que integrará recursos lingüísticos en y para la lengua catalana en la e-infraestructura desde la fase inicial del proyecto CLARIN. Con este objetivo está previsto la colaboración con organismos que ya disponen de estos recursos y que dan acceso, como por ejemplo el Institut d'Estudis Catalans. Por otra parte, CLARIN-CAT también tiene que ser una acción que impulse la investigación de excelencia en el ámbito de las humanidades y ciencias sociales creado las condiciones necesarias para dar a conocer y asesorar en el uso de esta nueva infraestructura.

2. Nuevos colaboradores en CLARIN

La Universitat Pompeu Fabra ha firmado el convenio de colaboración del proyecto CLARIN con las siguientes universidades, que se han convertido en miembros colaboradores de CLARIN:

An example of a strong national engagement: Spanish CLARIN community newsletter

- Would your country want to host one of the main hubs in the future federation of archives?
- Would you like to connect your existing archives to the infrastructure?
- Would you want to host (or participate in) one of the centres of expertise that will be created?
- Would you want to create a national or regional network of expertise?

4. Essential resources

The CLARIN preparatory phase does not aim at the creation of new resources or technologies. Yet it might be worth while investigating the possibility of launching projects or programmes at the national level that would run in parallel with CLARIN and that would aim at the creation (and maybe even exploitation) of *essential resources* that do not exist yet or

essential resources that would support future national research programmes or participation in international programmes.

We see an enormous potential for cross-fertilization between such activities and CLARIN.

5. Events

The organisation of CLARIN related activities at the national level is an important instrument for bringing national players together.

This would include organising meetings and workshops, bringing together providers (language and speech technologists) and users (humanities and social sciences scholars), awareness events, attending conferences, supporting mobility for researchers and students, etc.

We hope that national funds will be available for the (co-)organisation of national, regional or international events related to CLARIN. **C**

CLARIN and Endangered Languages



Jost Gippert

University of Frankfurt



Sebastian Drude

Museo Goeldie Belem



Peter Wittenburg

MPI Nijmegen

Currently there are about 5500-6500 languages spoken all over the world, 96 % of which are spoken by only 3 % of all humans, i.e. most of these languages are spoken by only a few persons – and worse, these are often elders who will pass away together with their language before long. Most of these languages are therefore highly endangered and with them all knowledge about culture and environment that is encoded in them. Language change is by no means a new phenomenon; however, now it is the globalization which puts many of the languages under an enormous pressure unseen before.

Maintaining cultural and language diversity

When UNESCO stressed the necessity of maintaining biodiversity, it reminded us of the close relation to cultural and language diversity. Therefore we see two big challenges for linguists here: (1) Documenting the highly endangered languages of which we know that many will become extinct rather soon. (2) Maintaining language diversity wherever possible. It is obvious that documentation work often is a prerequisite for language maintenance or even revitalization, since it is documentation work that can help small communities to come to a written lexicon, to grammar descriptions and much more. While language maintenance and revitalization efforts are directed towards immediate help, language documentation also has the task to preserve part of our cultural heritage for future generations.

During the last decade a number of initiatives (AILLA, DOBES, ELF, HRELP, PAR-

ADISEC, etc.) were taken to document languages or to gather material about languages which are highly endangered and hardly accessible. The primary goal is the creation of a proper and balanced record of a language spoken in a specific environment and culture. Lexica, sketch grammars and other document types are created to describe the language system, based on audio and also video recordings and their annotations.

The role of digital archives

Most initiatives, however, have defined the establishment of a digital archive of their material as another strong pillar. These digital archives have two tasks: (1) Preserving the digital resources for future generations and (2) allowing researchers, community members, students and other interested parties to access the material. The DOBES archive established in Nijmegen is a good example of this approach. However, one question has not yet been fully addressed: how can we ensure that the material and the access software will survive given the high technological innovation rate and the yet unclear question of long-term funding for digital archives?

In this and a few other aspects we can formulate high expectations with respect to a research infrastructure such as CLARIN. In what follows we mention a few major expectations:

– Older materials concerning (endangered) languages are often in a rather bad and fur-

Editors' note

On this page(s) we publish opinions, discussions, views and arguments that usually come from two angles. One illustrates the standpoint of CLARIN users or "consumers", or presents a problem to be solved, while the other focus on ideas that are coming from the direction of LRT developers.

ther degrading condition. An infrastructure such as CLARIN with clearly defined and trustful centres will motivate researchers to deposit their data and thereby bring it into a stable, visible and accessible state. This will increase its re-usage by different types of users.

– Still materials concerning a specific language are often stored in a number of centres distributed around the world. Infrastructures such as CLARIN will finally allow researchers to create virtual collections and by virtually combining the various contributions.

– Transcribing the spoken material and creating, e.g., translations and morphosyntac-

tic descriptions is a highly time-consuming task, since this usually has to be done manually, word by word. CLARIN will increase the chance to find tools or combine existing tools with new ones to be able to work semi-automatically at least when creating the various annotations. Currently it is hardly possible to find suitable tools even for major languages.

– Currently the 'long-term' funding scheme for digital research archives only covers a few years. CLARIN will increase the awareness of policy makers that the material collected by linguists and others needs to be preserved for many years, if not forever. A persistent research infrastructure as CLARIN will need to be equipped with centres that have the task of preserving data for a long time. Since these centers can be shared by several countries, the costs of long-term preservation can be reduced.

– Currently, linguistic theory is biased to the western languages. It is the BABEL project of the ESF that stated correctly that the documentation of the many small languages will help to reformulate our assumptions in particular about how our mind is processing language. The material that is now available in accessible archives can help a broader group of linguists study a larger variety of languages. CLARIN will help fostering this type of work, since the material will become visible and since access software can be maintained.

– Making recordings and processing them (producing annotation, metadata and analysis) is a task that currently needs not only field work competence but also advanced software skills. More and more, however, the smaller communities are involved in cultural and language documentation. CLARIN should generally help to make the new technology more generally accessible and easier to use, contributing to overcoming the digital gap.

CLARIN's role

Summarizing we can say that we see CLARIN as a logical follow up of a process that was started about a decade ago when linguists began to realize that beyond carrying out the linguistic work we need to take efforts to archive our data and take measures to keep it in an accessible state. It is very positive and clearly of mutual benefit that an initiative like CLARIN, which is at the most advanced edge of digital technology, is also so closely linked to traditional communities and people that are usually not acquainted with such innovations. **C**

Techniques of Preservation for Digital Language Recordings



Florian Wittenburg
MPI, Nijmegen

Today, more and more resources are “born compressed” and this is probably a trend that is going to increase in the future. Moreover, the applied degree of compression seems to increase as technology advances.

We all know MP3, but new videotape formats such as HDV (High Definition Video), the successor of DV (Digital Video), and HDCAM SR also feature compression codecs, such as MPEG2 and MPEG4/H.264.

For archivists “born compressed” means that they do not have any choice but to archive the data in that state. With this compression techniques the data is compressed in a lossy manner, meaning that the original information cannot be re-calculated again. Actually, in the case of lossy compression, the well-known notion of data compression is misleading, because the principle is based on reduction of the data-flow. So-called perceptually irrelevant data is actually “thrown away” and cannot be retrieved during the decoding process.

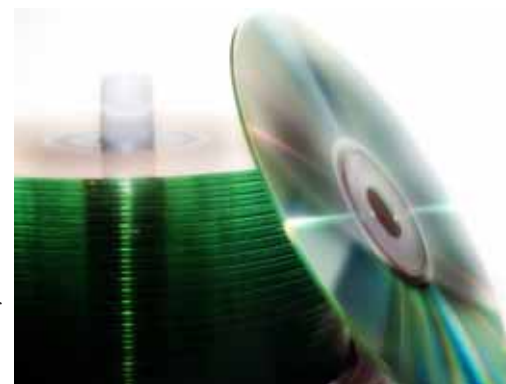
Usage of compressed data

Is this entire situation satisfactory, or does it have to change, improve? Viewed from an archivist perspective this is definitely not an ideal situation. Data-reduced signals should

only be archived when they are “born compressed/reduced”, that is, when the source format is already compressed. All other signals should go into the archive in linear, unreduced form. Why?

For simplicity let me confine this short analysis to sound: on first glance the MPEG and also Dolby codec families seem to be fairly attractive. Elaborate, thorough hearing tests showed that in most cases the sound quality of linear, unreduced audio-signals was almost equalled.

But in critical cases differences remained clearly audible. Although many codecs apparently can be applied pretty satisfactorily for hearing purposes, small errors remain, which under the following circumstances give birth to unpleasant appearances:



Cascading: multiple successive coding and decoding of signals, which produces audible artefacts after a number of cycles. This depends on the complexity of the signals, the applied degree of data reduction and the type of algorithms applied.

Code-switching: the switch from one to another codec also results in audible artefacts.

Post-production: editing of data-reduced signals, for example filtering or mixing several signals together, again results in an audible degradation of the signal quality. This depends again on the complexity of the signals, the applied degree of

data-reduction and the sequence of algorithms.

Thus, data compression may lead to restrictions and limitations for future use.

Limitations of our perception

Furthermore it should be stated clearly that the data-reduced signal is modelled accord-

ing to human perception. It does not represent the original acoustical signal – though we may not hear better, we know better. Using such data we may not be able to (re)submit it to a real physical, acoustic measurements which are often needed when starting from the basic phonetic level. We may lose or throw away important information that could reveal some physical facts hidden so far. In this way we are losing the very possibility of gaining that knowledge and possibly reinterpret it in future.

Would you not want to have a digital copy that comes as close as possible to the original, and that is flexible, independent, and without restrictions for future use? It is important not to forget that the transition to digital already means data reduction – reducing an analogue change of air pressure to a sequence of discrete numbers.

Also, a better recognition/perception of the imperfections, while comparing linear with reduced signals, can be expected. With future generations, where we don't know what kind of operations they will apply to the data, this could lead to more critical judgements of data reduced signals.

The purpose of compression

It should not be forgotten that data compression has other roots and purposes: it was invented to make it possible to transfer data via limited bandwidths. Digital broadcasting and networks like the internet are good examples of such bandwidths. But is this an issue for archiving? It is true that you have to transfer data to an archive, or from an archive, so in this sense bandwidth is a time/cost-factor, but considering the total costs this is marginal.

The same applies to storage. According to D. Schueller from the Phonogramm Archive Austria the storage costs only amount to 5-10% of the total costs. So the cost-saving aspect of data-reduction is not crucial and does not compensate for the disadvantages I outlined earlier.

For data acquisition this means, in case of future archiving, use and evaluation, that audio data should be recorded in linear, unreduced form. In other words: a “compression birth” should be prevented while for storage purposes lossless compression techniques should be used. **C**



Budapest Meeting of the Alliance for Permanent Access



Tamás Váradi
CLARIN EB member



Marko Tadić
CLARIN Newsletter editor



Peter Wittenburg
CLARIN EB member

In Budapest, 4 November 2008, the Alliance for Permanent Access (APA) organized its third annual conference under the title *Keeping the Records of Science Accessible: Can We Afford It?*. The conference was jointly supported by the European Science Foundation (ESF) and the Hungarian Research Council OTKA as one of the ESF workshops. The key topic of the conference was 'Business models for permanent access' i.e. how to ensure the funding strategies that will allow permanent access to research data repositories and archives and how these strategies can be combined with a number of already existing research infrastructure initiatives. The Alliance's overall mission could be phrased as follows: 'The Alliance aims to create a sustainable organisational infrastructure for permanent access to scientific information.' This involves calculating operational costs, developing real business models and developing a funding strategy for permanent access. The CLARIN view on maintaining linguistic resources in persistent repositories was presented by Peter Wittenburg in one of the two parallel ses-

sions that highlighted examples from humanities and social sciences on the one hand, and the natural sciences on the other in order to explore similarities and differences between disciplines.

"Digital documents last forever – or for five years, whichever comes first." (Rothenberg?)

A large number of research infrastructure initiatives are receiving funds from the European Commission and from national funding councils mainly as a result of the ESFRI process. To a large extent these initiatives are discipline driven, i.e. all these infrastructure initiatives need to address a number of topics that they all have in common. Currently effort is duplicated in a number of areas ranging from how to integrate services with national identity federations, how to setup repositories and archives, how to tackle the gigantic IPR issues etc. CLARIN is a good example of such an approach seeking solutions for a wide range of vertical topics of which only some are specific to the linguistic community.

Therefore it makes sense that there should be a few initiatives that bring people together focussing on issues that we all share. The Alliance for Permanent Access is such an initiative. While the Brussels APA meeting a year ago did not yet have such a clear focus, the central theme for the recent meeting in Budapest was the question whether we can afford to keep the records of science accessible, i.e. what are the business models and the costs for maintaining repositories and archives, for curating data, for maintaining software to manage and access research data etc. Some talks presented general views and some speakers argued from the discipline perspective.

Almost everyone agreed that a sustainable infrastructure of proper and dedicated repositories lies at the heart of data preservation and is the core of any research infrastructure in a time where volume and complexity of data is dramatically increasing in nearly all disciplines. Yet no clear and widely agreed



Peter Wittenburg presenting CLARIN project

funding scheme can be seen. Although vague cost estimates can be made, like they have in CLARIN, we can only speak about an "imperfect market", since no one can estimate the exact value of preserving research data. It may well be seen as priceless in terms of collective memory. Due to this situation it was also widely agreed that only public funding will work and that out of the whole budget for research a small proportion needs to be allocated to preserving the data. Compared to the whole amount spent on research this sum will be a very small one.

"Any data which is not assigned to an Archive for long-term preservation will be lost as a useful scientific asset. This is not just a risk, it is a certainty: only the exact date of disappearance is unknown." (Huc?)

According to a UK overview 42% of the overall costs for preserving our research data needs to be reserved for the acquisition and

Alliance for Permanent Access is there for everyone, including CLARIN



Peter Tindemans
Acting director Alliance for Permanent Access

The Alliance for Permanent Access (APA) has been established to help ensure the creation of a European Digital Information Infrastructure or in US terms a cyberinfrastructure. Basically this consists of a series of repositories or archives where the digital record of science (both documents and data) is stored, curated

and kept accessible. For universities, research organizations, operational agencies, funding agencies and society at large this is rapidly becoming an issue of crucial strategic importance. The Alliance is therefore gathering a number of key players in European science and science information to bring their commitment and expertise to the creation of such an infrastructure: MPG, STFC; key libraries such as the BL, the KB or the DFG; the funding agencies are represented through ESF; the Association of Scientific, Technical and Medical Publishers is a member; as are national digital coalitions. It will work with everyone who is involved in developing these ideas.

Though many of the repositories will be organized in particular communities (CLARIN is one example) they must be interoperable across communities. The infrastructure must also be operational and provide practical services to the community of scientists and other users. The European Bioinformatics Institute in Cambridge is a beautiful example even though it underscores that project funding

does not provide a safe basis for a digital information infrastructure. So far this has, nevertheless, been how most projects developing tools, exchanging information or setting up small testbeds have been funded, often through the IST and digital libraries directorates of the EC. The idea that we are actually speaking of an infrastructure has taken root. Several projects on the ESFRI Road Map demonstrate this, such as CLARIN and DARIAH in the social sciences and humanities and Life Watch in the area of biodiversity.

At the recent APA conference in Budapest, which focused on experience with and ideas for funding models for data infrastructures, a wide variety of ongoing initiatives were presented. At the same time several areas were identified where participants felt the Alliance could be of considerable help in turning their individual efforts into a concerted effort across Europe. A number of them are listed below; they all reflect that the Alliance's major value lies in its ability to bring all the stakeholders together and act as an umbrella organisation.



at the Alliance for Permanent Access workshop

ingestion of new data, 23 % are needed for the real archiving work and 35 % to support access. At the Max Planck Institute the costs of maintaining a complex 45 Terabyte archive amount to 450 K€/year, out of which 85 K€ needs to be reserved for storage hardware (6 copies at three different locations). Regarding personnel costs, the necessary system and archive management costs about 2 fte/year (full-time employee), maintaining the repository software costs about 1 fte/year and for the comprehensive utilization software 2 fte/year are necessary. The latter post is associated with accessing the data. The pure storage costs are decreasing over time due to new technologies – after 10 years the costs are only 10% for about the same capacity. An economy of scale effect can be observed (costs only grow by about 50 % with additional collections), i.e. it seems to be cost efficient to centralize preser-

vation. However, there are other criteria that need to be considered when comparing centralization and decentralization. National political strategies and the availability of specific expertise are important criteria for decentralized services. It was interesting to note that the costs for creating proper metadata descriptions after 10 years are about 30 times higher compared to creating them immediately in the resource creation process. It is strongly advisable not to offer personalized services by repositories since they would increase the costs extremely.

Running a repository/archive with the required quality of service requires expertise, which is expensive. The technological pillars of repositories are stable repository software systems, support for persistent identifiers (PID), efficient mechanisms for curation, the availability of standards for formats and proper metadata descriptions. It has been concluded that those disciplines that lack standards are in general falling behind in data processing capabilities. It is expected that metadata descriptions need to be increasingly more precise to cope with the extreme rate of increase in data. The use of common and consistent concepts and terminology is increasingly important for all disciplines to ensure interoperability. While it is evident that it will not be possible to preserve all data at this stage of available capacities, format coherence and usability of data will form important selection criteria.

“The volume of data generated by linguistics pales in comparison with the volumes generated by natural sciences”, said Peter Wittenburg, “but terabytes are not everything”. The complexity of the data must be taken into account, the all-important semantics. The unique sound recordings of endangered languages are very fragile types of data and it is an inevitable fact that these must also be regarded as important cultural treasures. As most countries prefer to archive their own cultural heritage, a single facility has not

been an option. Instead, the Max Planck Institute set up the technical infrastructure for an international network of data archives, depending on multiple copies and migration as preservation strategies. This work has been combined with several other initiatives and has become adopted as one of the cornerstones of the CLARIN project.

Open Access principle at the research data level

It was also widely agreed that the Open Access principle is very important in science, since it should be the basic rule that the scientific record is free and open for use by all researchers. At least a “fair use” principle needs to be established in Europe as well to create the open scenario that will allow researchers to easily access the data that is needed to answer research questions or generate new ones.

In relation to this the question what APA can do for the existing research infrastructures was raised. Most important is the bridging function with the help of focused workshops. At the Budapest conference a number of ideas were presented. On behalf of CLARIN we asked for help in reducing the IPR complexity, in pushing the federation harmonization and in lobbying for the kind of network of stable centres we need to establish. The question of how many centres will be needed could not be answered but it was obvious that infrastructure services can often be shared across discipline boundaries. Therefore CLARIN offered for example a PID and a terminology service for the others, since it seems that we are ahead of many other in these areas. **C**

¹ See the conference web-site: <http://www.alliancepermanentaccess.eu/index.php?id=3>

² <http://www.alliancepermanentaccess.eu/documenten/dimper.ppt>

³ <http://www.alliancepermanentaccess.eu/documenten/%5Cchuc.pdf>

- The Alliance has a major task in the area of advocacy, promoting the cause of working together for the common good rather than individually, and raising awareness of digital preservation issues with governments, the EU and funding agencies. In this way the Alliance will also contribute to increased alignment and coordination between governments, the different branches of the EU and the funding agencies.
- One role for the Alliance is to generate consensus on an optimal infrastructure of repositories and archives. A suggestion worth investigating is that connecting archives in a network could provide an upstream incentive for international collaboration to cut costs, as not all repositories need to include long-term preservation facilities.
- The advocacy role of the Alliance must involve promoting the value of preservation for the users, whether scientists or societal bodies or companies. Identifying and maybe quantifying the value is as much needed as further investigation of the costs. Funding models must reflect both, and the Alliance

should continue working with the funding agencies to develop such models.

- The Alliance has a very useful role to play in ensuring that common provisions and facilities become available. To facilitate data sharing and seamless interoperability work must be done on registries, terminology and standards. The Alliance is poised to take a coordinating role here. Other examples are a European solution for accreditation of repositories and archives, as well as agreement on persistent identifiers.
- Clearly the Alliance is well placed to work with the various stakeholders on matters of policy. These range from policy within the scientific community to the policies of governments. Sharing data and fair use is an example of the first category. This must somehow become part and parcel of a researcher's workflow, stimulated not only by funding agencies but also by other, less tangible rewards such as citations. As an example of the second category the Alliance must

make a case in Brussels for less restrictive Intellectual Property Rights regimes.

- Cross-community exchange of experience is a very valuable field of action for the Alliance, as the Budapest conference has shown. For example, it was demonstrated that when it comes to permanent access, the humanities and social sciences on the one hand and the natural sciences on the other share much the same problems, only the scale differs.
- European solutions must be part of, and therefore be designed with a perspective to, global solutions. Here the Alliance is in a good position to work with key stakeholders elsewhere in the world, such as the National Science Foundation (NSF) in the United States and organisations in Asia and Australasia.

The Alliance is eager to follow them up. As a concrete step we are happy to work with CLARIN and the other data-focused ESFRI Road Map projects to organize a workshop to establish a common work programme. **C**

Maltese Language Resources Infrastructure



Mike Rosner

*Dept. of Artificial Intelligence,
University of Malta*

The Maltese Language

Maltese is the national language of Malta, spoken by about 400,000 inhabitants, and by a further 100-200,000 speakers outside Malta. Within Malta, the language is used for all types of interaction and communication. Since 2004 it has been an official language of the EU.

Maltese is a so-called 'mixed' language, with a substrate of Arabic, a considerable superstrate of Romance origin (especially Sicilian) and, to a more limited extent, English. Its script, codified in the 1920s, is unusual in that it utilises a mainly Latin alphabet which also includes the following non-standard characters: ċ, ġ, ħ, ġħ, ż. Unlike Arabic and Hebrew, vowels are written. Some treat Maltese as a dialect of Arabic, but many scholars reason that it has changed to such an extent that it deserves the status of an independent language.

Language Resources and Tools

Computational approaches to Maltese began to materialise during the second half of the 1990s, mainly through small-scale undergraduate and masters projects focusing on areas such as verb morphology, spelling correction, domain-specific translation, automated word clustering, legal document classification and speech processing. One notable PhD thesis concerned text-to-speech synthesis.

MLRS

More recently efforts have been concentrated within a programme known as the Maltese Language Resource Server (MLRS)¹, initially funded under a government research initiative, whose main aim is the creation of infrastructure and content for a Maltese National Corpus (MNC), a computational lexicon, and certain language-specific services. The corpus itself currently comprises about 50 million words of different genres including articles from newspapers, legal texts and tracts, European documents, and works of fiction. The infrastructure supports multiple levels of annotation, namely (i) source (ii) standard utf8 text, (iii) text structure, (iv) syntax, and (v) semantics. Currently the first three levels are supported. Migration from (i) to (ii) is mostly handled manually; subsequent mappings are to be automated.

POS Tagging

The mapping from (iii) to (iv) requires a POS tagger in the first instance, so an HMM-based tagger is currently under development, trained initially on a corpus of approximately 10,000 tokens that has been manually tagged using a set of POS categories also developed within the MLRS project. The manual aspects of this process, including the tagset design itself, have been a bottleneck, though we are now at the end of a phase of iterative training involving manual correction and retraining of tagger output.

Informal error analysis of the tagger's performance has revealed that the main source of error for unseen words is the fine-grained morphological distinctions made in the



Valetta

tagset. In a new version of the tagger a distinction is made between three levels of annotation, yielding three possible taggers, giving the user the option of trading tagging accuracy with different levels of tagging granularity. We are also considering the use of symbolic (rule-based) methods in addition to the statistical model to filter the output of the HMM-based tagger using pattern-matching rules.

To date there is no fully defined computational grammar of Maltese, although some promising first steps using HPSG have been taken by Stefan Mueller at the Free University of Berlin.

Multilex Lexicon and Editor

Besides the corpus, MLRS includes Maltilex, a full-form lexicon, initially populated

with words extracted from level 1 of the corpus using Maltitok, a custom-made tokenizer. The current wordlist comprises about 25,000 distinct words, some of which contain orthographic errors. The system therefore includes some rudimentary facilities for wordlist management before words are migrated to the lexicon. Initially entries are empty, the idea being that linguistic information is supplied by linguists. Elexi, an editor client, allows linguists to add or modify lexical information over the internet. The behaviour of the editor is partially controlled by a configuration file which specifies dependencies between the attributes and values of lexical information (e.g. verbs have person whilst nouns don't; plural nouns and adjectives are without gender) and this is used to dynamically generate the form interface that the linguist uses to enter information.

Morphological Analysis


Maltese morphology presents a number of issues. The mixed nature of the language gives rise to two distinct morphological sub-systems: the Romance substrate resembles that of Italian, whilst the Semitic one shares many characteristics with languages like Arabic and Hebrew. Being based on roots and templates, it is non-concatenative, and presents the usual computational challenges that cannot be addressed using standard finite-state solutions. Another challenge is to know which system is operating when faced with a particular word.

We are currently in the process of designing a morphological analyser inspired by the work of Yona & Wintner² on Hebrew. Though the languages clearly share certain morphological phenomena, there are a number of differences so the main challenge will be to find out the extent to which techniques for Hebrew carry over.

Maltese and CLARIN

Much work remains to be done – both at the level of contents and infrastructure. Progress on all aspects of MLRS has been steady, but slow, the greatest problem being lack in continuity of both funding and personnel. Although CLARIN will not solve this problem, we are confident that it will assure a place for Maltese within the emerging infrastructure for language resources in Europe.

Acknowledgements

I wish to acknowledge the contribution of Ray Fabri, Albert Gatt, Mike Spagnol and others to the work reported above. 

¹ Rosner, M., Fabri, R. Attard, D. and Gatt, A. MLRS, a Resource Server for the Maltese Language, Proc. CSAW06, Dept. CSAI, University of Malta, November 2006.

² Yona, S. & Wintner, S. A finite-state morphological grammar of Hebrew, Natural Language Engineering 14.2 pp 173-190, 2008.

Some Background on Language Resources in Norway



Koenraad De Smedt
University of Bergen

For the past 40 years or so, researchers in Norway have built language resources that are of interest far beyond the national boundaries. Several of the early and important contributions by Norwegian researchers to language resources were in fact not concerned with Norwegian. A ground-breaking corpus project which was started in Lancaster moved to Norway in 1977 where it was completed as the Lancaster-Oslo-Bergen corpus in 1978. Later on, it was integrated in the International Computer Archive of Modern and Medieval English (ICAME), distributed from Bergen. ICAME is perhaps one of the earliest examples of a common language resource infrastructure, since it allows simultaneous on-line searches in 9 different corpora with a single sign-on and a common interface.

Another important milestone concerned with a different language was the Norwegian Wittgenstein Project, which started in 1980 and later was to become the Wittgenstein

Archives at the University of Bergen (WAB). In 2000, the project finished the digitization of Wittgenstein's Nachlass, converting more than 20,000 hand-written pages into a complete electronic edition with multiple codings and multiple views. WAB was selected under the Transnational Access to European Research Infrastructures programme, and from 2002 to 2005 it hosted more than 30 international user projects. This shows that large infrastructures, once they are available to a wide audience, tend to quickly generate large amounts of new research.

Although these examples have been valuable in their own right, the experiences gained through them have also been applied to other projects involving both spoken and written Norwegian as well as other languages. The past decade, has seen a plethora of research projects in the area of language resources and their applications, including for instance tagging, named entity recognition, lexicography, automatic proofreading, ontologies and word nets, terminology, language learning machine translation and tree-banking. Adequate funding has been indispensable to these efforts. Two research programmes in language technology that were recently concluded, one Norwegian and one Nordic, have been of special importance.

New national projects

By no means have Norwegian efforts been limited to corpus linguistics in the narrow sense, but they cover the whole range of Humanities disciplines. In the Documentation Project, a cooperative project coordinated at the University of Oslo and concluded in 1997, a very wide range of materials was digitized, including not only lexico-

cated the importance of facial expressions in human communication, and how these can be used to improve human-machine interfaces. Susan Dumais, a researcher with extensive experience in the field of IR, stressed the need to finally move away from traditional search boxes towards interactive, personalized, and context-sensitive search facilities that would better meet users' information needs. This year's ACL Lifetime Achievement Award went to Yorick Wilks, probably best known for his pioneering work on natural language understanding and preference semantics.

Admittedly, with 800 participants this year's ACL did not top the largest ever ACL conference held last year in Prague. Nevertheless ACL 2008 was without a doubt an overwhelming success. ACL 2009, this time in conjunction with the conference of the Asian Federation of NLP, will be held next August in Singapore, the culturally and technologically vibrant Garden City of Asia. **C**

graphical filing card cabinets and old printed dictionaries, but also museum catalogs, archeological catalogs, collections of letters, ethnographical maps, song lyrics, topographical and historical bibliographies, census records, medieval administrative texts, rune archives, photographic archives, a coin catalog, etc. About 50 main collections were digitized and hundreds of person years were invested in this project. The project was followed in 1998 by the National Database Project of Norwegian University Museums, linked to the European ARENA network.

Early adopters of a technology tend to pay a price, and that is also the case for Norwegian contributions to language resources. Since many materials were produced before TEI-XML and Unicode became standards, a multitude of different codings were used, and most materials come with their own access mechanisms. Furthermore, materials are spread among several institutions where access, maintenance and documentation are highly dependent on individual staff members, so the danger is that resources and tools become difficult to access and reuse as time goes by. However, the need for standards was quickly realized. During 2001-2004, Bergen was the head office of the TEI consortium.

Currently, the main actors are the Text Laboratory and the Unit for Digital Documentation, both in Oslo, and the Centre for Culture, Language and Information Technology in Bergen, while some databases are kept by the National Library and other actors. There is a clear need for coordination at the national level. It is also remarkable that despite the production of many corpora in Norway for several languages, there is still no Norwegian national corpus comparable to the British or American national corpora. The Norwegian government has, however, recently announced that a large national corpus project can start in 2009.

Research infrastructures

Moreover, early in 2008, the Research Council of Norway launched a national strategy for research infrastructures spanning the period to 2017. Stimulated by the European infrastructure actions, the Norwegian plan recommends long-term and large-scale investments, financed by a capital of 20 billion NOK which is expected to yield 800 million NOK yearly. Whether and how the Norwegian part of CLARIN will fit into this framework, is as yet unclear. **C**

ACL2008 REPORT

Jan Šnajder

*Faculty of Electrical Engineering and Computing,
University of Zagreb*

This year's Annual Meeting of the Association for Computational Linguistics (ACL) was held in June in Columbus, Ohio. The conference – arguably considered the most important in the field – covered virtually the whole range of NLP topics from morphology and phonology to pragmatics and discourse. ACL 2008 was hosted by Ohio State University, the largest US university, and organized in conjunction with the Human Language Technology Conference of the North American Chapter of the ACL. This year's invited talks were given by Marc Swerts from of Tilburg University and Susan Dumais of Microsoft Research. Marc Swerts, the distinguished lecturer of the International Speech Communication Association, expli-

The Spanish CLARIN development



Montserrat Marimon
University Institute for Applied Linguistics, University Pompeu Fabra, Barcelona

In this contribution we will briefly report on the activities we have completed so far at the Institut Universitari de Lingüística Aplicada of the Universitat Pompeu Fabra (henceforth, IULA-UPF) in Barcelona, Spain, within the framework of the CLARIN project.

Goals

For the preparatory phase, the Spanish Ministerio de Educación y Ciencia (Ministry of Education and Science), within the Program Estudios de Diseño y Viabilidad – Acciones Complementarias, has already funded the first year and we have just submitted a follow-up for the remaining years (2009-2010). In addition to this, the Departament d'Innovació, Universitat i Empresa de la Generalitat de Catalunya (the Catalan autonomous government) has granted the CLARIN project funds for integrating Catalan language resources and tools into the

European infrastructure. These funds will cover the period from September 2008 till December 2010.

One of our goals during this first year of the preparatory phase has been to disseminate information about CLARIN in Spain in order to show the Spanish Ministry that there is a critical mass of Language Resource and Language Technology (LR<), Spanish providers and users within the Humanities and Social Sciences and to demonstrate both the feasibility and the interest of the CLARIN project for Spain.

Covering four official languages

With this goal, IULA-UPF, as a consortium member of the European CLARIN project, is identifying the LR< developed in Spain for the four official languages – Spanish, Catalan, Basque and Galician – and establishing contacts with leading LR< researchers and developers to present the project's goals and calendar. Institutions hosting groups that are interested in joining the initiative and that are willing to integrate their resources and tools, are invited to sign an agreement with the UPF.

In this agreement LR< developers commit to providing IULA-UPF with a list and the technical details (as well as the necessary assistance) of the linguistic resources and tools developed by them, with the aim that IULA-UPF includes the cost of the integration and regular maintenance on with the

costs of the infrastructure development for Spain. Furthermore, both organizations agree to favour and promote the CLARIN project within their own organization and with third parties. Thus, agreements are signed between institutions in order to guarantee persistence of the agreed resources as well as support to the highest level.

Survey of existing HSS projects

The agreement has already been signed by the following six Universities: Universidade de Vigo, Universidad de Málaga, Universitat de Lleida, Universitat de Barcelona, Universitat Autònoma de Barcelona and Universitat Oberta de Catalunya. Note that, since the agreement is signed by the president of the University, more than one department or research group may be included in the agreement. Once the agreement was signed, we asked the University to fill in the member request form.

On the other hand, to demonstrate the interest in the project among researchers working in Humanities and Social Sciences (HSS), we are carrying out an in-depth survey of existing HSS projects and we have established contacts with leading researchers, with the aim of gaining a thorough understanding of the research requirements and needs to set the best ways of collaboration. We have already presented CLARIN at the University of Murcia and the University of Navarra.

The Spanish CLARIN Newsletter

To contribute to the dissemination of the project, we have created a web page – <http://clarin-es.iula.upf.edu/> – where we report (in Spanish and Catalan, for the moment) on events and news related to both CLARIN.eu and CLARIN-ES; e.g., meetings, new collaborators, etc. The FAQ section collects questions raised during the several national meetings both with LR< developers and CLARIN users. There is a download section where people can download both documents and software. We also have a section for users where we collect information about user needs and possible ideas for user scenarios. Latest news are also collected and published in a Newsletter to which people may subscribe at <http://clarin-es.iula.upf.edu/es/newsletter>. And for those that want to know the members of the CLARIN-ES team, they can find us at <http://clarin-es.iula.upf.edu/es/equipo>. **C**



Spanish CLARIN community web site

Polish Clarin: sometimes flying, sometimes walking, but still moving forward



Maciej Piasecki
*Institute of Informatics,
Wrocław University of
Technology*

The second number of the CLARIN Newsletter presented several great national CLARIN projects. This is very good for the whole CLARIN. Congratulations! Unfortunately, we cannot report on any similar Polish project, so far. The discussion of the national program of research infrastructure is under way, the rules for ESFRI projects will be announced, but all requested CLARIN related information is on the ministerial committee desk and we are ready to deliver a formal application. Paradoxically, we do not treat the still unclear status of our attempts as a big danger to CLARIN Poland, as we have got used to difficult, but stimulating conditions of research work in the areas of Natural Language Processing, Computational Linguistics and General Linguistics in Poland. Permanent underfunding and some marginalisation of these fields have always forced researchers in Poland to be very creative. Lacking resources, but with a detailed plan, we will build CLARIN in Poland¹, as it is a perfect and demanding goal for the development of usable language technology. Requirements for the intelligent information extraction tool working on corpora seem to be higher when it is considered as a trustworthy research tool than when it is applied as a user facility for intelligent web mining on the response of the final user. Results generated by the research tool must be valid in any sense.

Existing Polish resources and tools

The implementation of BLARK² for Polish is still quite sparse – many basic LRT components are missing, insufficiently developed or existing but unavailable. Such LRT components are, however, necessary in order to make CLARIN a unique research tool. We need to offer efficient, flexible and high accuracy tools for language content analysis directly to the final users, i.e. researchers from the area of Humanities and Social Sciences. Ideally, the final users should be able, to e.g., perform analysis of the multi-

lingual language corpora tracing occurrences of events of some specified types and get report linked to the relevant parts of the text. In order to make such scenarios realistic we need to implement interoperable LRT components for CLARIN languages, i.e. components matured and well designed enough to be easily combined with a few mouse clicks according to the type of outcome a given user wants to get. This is an ideal situation, but CLARIN is to create future standards in research. That is why, we put the construction of the missing LRT components in the focal point of the CLARIN Poland.

During the last few years, some Polish LRT components have been built or their development has been initiated by the first versions created. The general idea, shared by all Polish CLARIN members is to make the LRT components publicly available, at least for research applications, as it is only then that they can have an impact. It is not possible to list all the components, so only a few are named below to indicate that our web pages are already worth further exploration: The IPI PAN Corpus – a large morphosyntactically annotated corpus of Polish³, Morfeusz – morphological analyser⁴, TaKIPI – a morphosyntactic tagger⁵, Spejd – shallow parsing and disambiguation engine⁶, Świgr – a deep parse implementing a large DCG-based grammar of Polish⁷, plWordNet – a Polish wordnet⁸, SuperMatrix – a universal system for extracting lexical semantic relations and distributional semantics (see the plWordNet web page) or a set of selected tools for speech recognition and synthesis. The construction of the huge National Corpus of Polish and a medium size tree bank has been initiated⁹. A bilingual Polish-Ukrainian corpus¹⁰ and a first version of the Chronological Corpus of Polish Press Texts¹¹ have been also built.

Missing components

Morphosyntactic processing has been completed to some extent, but in other areas numerous blank spaces can still be found. Our main idea for the preparatory phase is to fill as many of the blank spaces as possible and required in relation to other projects which are ongoing or very likely to happen. Better fleshed technical infrastructure skeleton will enable us to do accurate planning of high level user facilities in the future CLARIN system, especially with respect to Polish. The set of LRT components that we have considered it necessary to construct includes: shallow syntactic parser, extended tree bank, monolingual and bilingual sub-categorisation dictionaries, dynamic dictionary of collocations, extended plWordNet and next aligned with BalkaNet, corpus annotated by senses, Polish-Bulgarian parallel corpus, extended chronological corpus, speech corpus and a minimal set of tools for

speech recognition, to name most of them. The amount of work which we will be able to perform depends on the funding we will get, nevertheless, we will follow it independently of funding, as the majority of those LRT components are basically indispensable.

The Polish CLARIN community

The Polish part of the CLARIN network has originated from the informal network of scientific cooperation and is the result of a bottom-up process of self-organisation. However, the group is quite diverse, but well balanced in relation the tasks we are facing. It presently consists of six members, but is still open for new partners. There are three typical LRT teams, namely: II WUT – Polish coordinator, ICS PAS and PJWSTK (for the explanation of the abbreviations see the frame *Join CLARIN* at page 12 or in the CLARIN webpage). However the LRT partners are supported by two teams working simultaneously in the areas of corpus linguistics and general linguistics: ISS PAS and IEL UL. Finally, (at the present moment) there is also one team working in the areas of quantitative linguistics and its applications in Humanities, namely: WU. This last partner is a natural link to the community of the future users of CLARIN in Poland.

In addition to developing a detailed plan of the preparatory phase and continuing work on filling blank spaces in the implementation of the BLARK, the CLARIN-related activities have encompassed a regional CLARIN *Workshop on Interoperable European Language Resources and Technology* which was held in Zakopane, Poland, in June 2008¹². We were happy to welcome CLARIN members from Bulgaria, Czech Republic, Germany, as well as all Polish members at the workshop. **C**

¹ But, to be perfectly honest, we would be very grateful to anyone offering us some substantial money.

² A minimal set of language resources and tools that should be available for all language postulated by Mapelli and Choukri in 2003. BLARK is in CLARIN considered as a starting point for a similar guideline.

³ korpus.pl

⁴ <http://nlp.ipipan.waw.pl/~wolinski/morfeusz/>

⁵ <http://plwordnet.pwr.wroc.pl/g419/tagget/> or <http://nlp.ipipan.waw.pl/TaKIPI/>

⁶ <http://nlp.ipipan.waw.pl/Spejd/>

⁷ <http://nlp.ipipan.waw.pl/~wolinski/swigra/>

⁸ <http://www.plwordnet.pwr.wroc.pl>

⁹ <http://nkjp.pl/>

¹⁰ <http://corpus.domeczek.pl>

¹¹ <http://www.lingwistyka.uni.wroc.pl/ql/>

¹² The proceedings of the whole conference can be found at <http://iis.ipipan.waw.pl/2008/proceedings.html>

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.

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

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

Center for Computational Linguistics ; Leuven; Ineke Schuurman

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

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

Institute of Croatian Language and Linguistics; Zagreb; Damir Čavar

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

Czech Republic: Charles University; Prague; Eva Hajičová

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

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

Society for Danish Language and Literature; Copenhagen; Jørg Asmussen

Estonia: University of Tartu; Tartu; Tiit Roosmaa

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

University of Helsinki; Helsinki; Kimmo Koskenniemi

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: ALTI; Nancy; Bertrand Gaiffe

TELMA/DIS CNRS; Paris; Florence Clavaud

CNTRL; Nancy; Bertrand Gaiffe

November 2008

2008-11-01 to 2008-11-03: Chicago Digital Humanities/Computer Science Colloquium, Chicago, USA

2008-11-04: ESF supported workshop of the Alliance for Permanent Access – Keeping the Records of Science Accessible: Can We Afford It?, Budapest, Hungary

2008-11-10 to 2008-11-11: Web services architecture in CLARIN, Munich, Germany

2008-11-25 to 2008-11-27: ICT 2008, Lyon, France

December 2008

2008-12-09 to 2008-12-10: European Conference on Research Infrastructures, Versailles, France

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 Intelligenz; 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

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

Hungary: Academy of Sciences; Budapest; Tamás Váradi

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; Reykjavik; Eiríkur Rögnvaldsson

Icelandic Centre for Language Technology; Reykjavik; 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

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

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

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

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

2009-01-07 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 **C**

University of Utrecht/Netherlands Graduate School of Linguistics; Utrecht; Jan Odijk

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

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 Nordgård

Poland: University of Wrocław ; Wrocław; Adam Pawłowski

Institute of Applied Informatics, Wrocław University of Technology; Wrocław; Maciej Piasecki

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

Institute of English Language, University of Lodz; Lodz; Lukasz Drozd

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

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 Tufis

University Babeş-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

Universitat de Lleida ; Lleida; Gloria Vázquez

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

Sweden: Lund University; Lund; Sven Strömquist

Språkbanken, Dept. of Swedish Language, Göteborg University; Gothenburg; Lars Borin

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 Language, Lancaster University; Lancaster; Anna Siewierska

Oxford Text Archive; Oxford; Martin Wynne

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