CLARIN Newsletter

Number 5, 2009, March

Looking back, looking forward

Steven Krauwer
CLARIN coordinator

After a slow, or rather, uneven start the project is now moving at full speed. With over 150 participating institutions in 32 countries, and a project consortium of (now) 33 partners in 23 countries (welcome Latvia!) we have managed to mobilize a huge and enthusiastic army of many different types of players from all over Europe, including both research institutions and owners of digital repositories. We see that participating groups are getting organized at the national level, which we feel is crucial for the success of CLARIN as CLARIN is rooted in and based on existing national structures and infrastructures. When looking at the flourishing CLARIN community I have one main concern: the large majority consists of providers rather than users, and even if we had to drop our plans to create a user community from our original work plan we should work very hard to establish firm and structural liaisons with our potential users in order to ensure that all the wonderful things we have to offer will actually be used.

National funding

In parallel with the start of the project the funding agencies in the participating countries have started (and in some countries even completed) their national research infrastructure roadmap process. The good news is that as an intermediate result of these ongoing processes we have now obtained financial support for the CLARIN preparatory phase in 20 out of 23 countries represented in the consortium. The amount of support varies per country, but at this moment the 23 countries together have committed ca 8.0 M€ to the CLARIN preparatory phase, and in addition 7.3 M€ to parallel projects that do not feed directly into the present CLARIN project but that will help preparing the grounds for the CLARIN construction phase. The bad news is that these national roadmap processes, which take place at the government level, and which are completely beyond the control of simple scientists, takes a lot more time than we had originally anticipated and that we will have to ensure that we take proper measures to ensure a smooth and seamless transition from preparatory phase to construction phase, even if not all countries may be ready by that time.

Work done so far

If we look at the work we have done in the first year we can say that at the general project level much of the exploratory groundwork has been done, resulting in a wealth of ideas and insights that have to be consolidated, put together and tested. In terms of the execution of our contract with the EC we can say that most of the first year contract deliverables were completed, though in most cases with some delay caused by the late signature of the contract. Some of the delays may still be propagated into year two, but on the whole there is no reason to assume that the technical work will not be finished on time. In March the CLARIN Executive Board had a meeting with the Scientific Board and the Strategic Coordination Board, where the achievements and deliverables of the first year were discussed, and the result of this meeting, which took place in a very constructive and productive atmosphere, was that both Boards fully endorsed our deliverables.

Interest outside Europe

We have attracted a lot of attention from the outside world, and both in Asia, in the US, in South America and in European countries outside the EU there is a lot of interest to exchange expertise and to collaborate with CLARIN. Also within the EU it has become clear that many of the existing and emerging research infrastructures in other fields are continuously looking for ways to share problems and solutions and for ways to use these infrastructures to explore new research avenues across discipline borders. CLARIN is playing an active role here.

We have now started on the second year of the project and this will necessarily be a period of convergence: strands developed in the various work packages will have to come together. This will not be an easy task in a project where so many parties collaborate and where so many talented and experienced researchers are eager to contribute to the shaping of the common infrastructure we want to build. In order to make everything fit together we will have to agree on the standards (for representation and interoperability) that will form the basis for the design and the specifications on which the construction phase will build, and this will have to happen in the coming months.

The first sketches of infrastructure

The first outlines of our design principles are already becoming visible and we anticipate that towards the end of 2009 we will have (or rather must have) a solid sketch of the future infrastructure in all its aspects, detailed enough to be able to make realistic cost estimations for the construction and exploitation phase, and to be able to take firm decisions about the best possible structure for CLARIN at the governance and operational level.

In brief: we are doing very well in terms of the creation of a thriving community, but we need better contacts with users, our national support is promising but slow, and in our technical work we will now see a rapid shift from exploration and investigation to convergence and consolidation. Let us hope that the second year will be as successful as the first year!

Editors' Foreword

Marko Tadić
& Dan Cristea
CLARIN Newsletter editors

Dear readers,

more than the whole year has passed since our project has started. It may seem that it passed too quickly, but if we turn around and count all consortium meetings, workshops, dedicated thematic and technical meetings at the level or the whole CLARIN project only, I am confident that we can be satisfied by their number. In that respect, we have already achieved a remarkable set of goals, but we are not even in the middle of our first, preparatory and planning phase of the whole CLARIN infrastructure.

The retrospect of our first year and prospect for the following two years is the topic of this number's title page, written by the author who is certainly the most appropriate for that — Steven Krauwer, our coordinator. His ability to put in concise yet perfectly understandable way topics that usually demand several pages to be explained is always striking. In this contribution he also gave us a warm push in the back that is always needed in order to finish our job properly.

Our initial editorial intention for this issue was to organise it around the report on selected humanities project that would exemplify the potential and role of LRT in humanities research (one of primary tasks of WP3). Due to the change in schedule in that working package, this original idea had to be abandoned. Instead, we are covering a broad range of different events organised as thematic or national meetings.

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.tadicr@ffzg.hr and dcristea@info.uaic.ro.

The first one is an expert meeting on metadata (presented by Peter Wittenburg, Maria Gavrilidou and Erhard Hinrich) where this topic of utmost importance for building the federation of different e-archives was tackled.

The establishing of CLARIN-CAT, the Catalan offspring of the CLARIN project in the form of round tables is covered by the contribution by Eva Revilla from University Pompeu Fabra, Barcelona.

The report from the FLaReNet Launching Event that took place in Vienna is presented by Marko Tadić. The FLaReNet project is closely related to CLARIN so we should keep an eye on it constantly, particularly having in mind that both of project share a great deal of participants.

The central part of this issue, pages six and seven, are covered with the thorough report from the WP2/WP5 three day workshop held in Oxford from the keyboard of Dieter van Uytvanck. In this workshop some relevant problems were presented and important decisions were made and it is certainly worthwile to get acquainted with them.

The following pages are covered by our national correspondents. They are either presenting a particular case-studies, like the project INQ1258 from Portugal (page eight), or they are presenting a situation and development in the field of LRT in different countries. Thus on page nine we have a brief description of many activities in Sweden, covering main LR projects/institutions that will play a crucial role once the CLARIN infrastructure will be established.

On page ten you can read about the situation in France and which institutions are involved in LRT production.

The development of Latvian CLARIN, its prospects for national funding and first national seminar is given on page eleven.

We hope that you will, as always, enjoy the reading of this newsletter. C

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

CLARIN Metadata Expert Meeting

Athens Januray 31, 2009

Peter Wittenburg
CLARIN EB member

Maria Gavrilidou
ILSP, Athens
Erhard Hinrichs
CLARIN EB member

Recently the first metadata document (www.clarin.eu/specification-documents) was made available that mainly explains the new component based framework as a consequence of almost a decade of experience in describing resources with the help of keywords. Metadata profiles (schemas) for the different resource types such as annotations, texts, media, lexica etc. will be built up of components and CLARIN will suggest an indicative number of components which may make sense for most of the researchers. To accommodate the wishes of researchers to tailor their schemas to their requirements, however, every researcher is free to create his/her own components. Finally all components are aggregations of elements that must be taken from accepted registries such as ISOcat or Dublin Core.

There are now two eminent questions which we need to address in the coming weeks: (1) which descriptive elements (concepts) do we need to describe our resources and (2) which are useful components. With respect to the elements we need to come up with proper selections and definitions, since the goal is to populate the ISOcat data category registry which is based on the largely stabilized ISO 12620 standard. This also implies that all our suggestions need to pass the decision structure for ISOcat, i.e. they need to be accepted by the Thematic Domain Group for metadata and by the DCR Board. Once we have defined a proper set of elements we can start suggesting useful components.

The metadata expert meeting at ILSP in Athens had the goal to come up with a first set of elements for the most important resource types (media, annotations, texts, lexica, lists). Therefore we invited a number of "experts" covering deep experience and knowledge about the different types. We followed a canonical procedure, i.e. (1) we discussed which type of information we need to cover (identification, creation, access, content, resource technicalities, participants, others) and (2) we discussed per resource and information type which elements are necessary also covering the need to be compliant with the installed base such as defined by IMDI and OLAC. Based on the resulting lists we produced a unified list of elements since many information types such as identification are identical for all resource types.

Critical issues such as semantic granularity of the elements and context dependency of their interpretation were addressed and discussed with the ISOcat experts. The principles that we will follow is to provide a higher granularity where necessary, i.e. we will include a number of "dates" such as "birth date", "creation date", "publication date" etc. However in some cases we will need to rely on the context while searching. Component creators will use the element "description" at many different places and we can assume that descriptions will be provided in various languages. It does not make sense to foresee all contexts of the description field and come up with more fine-grained elements. Users will need to use the context in searches if they would want to distinguish between descriptions.

The next step for the work are: (1) have another period for group internal discussions and clarifications, (2) make the selections and definitions open for all CLARIN members for discussion, modification and extension, (3) start building components to demonstrate compliance with the installed base and appropriate coverage (4) start making suggestions for the resource types which we did not discuss (tools, services, schemas, ontologies etc), and (5) entering the selected elements into the ISOcat user space and suggest it to the ISOcat boards to be accepted.

Lively discussion on the crucial role of metadata in building the federation or resources

The members of the expert group are: Maria Gavrilidou, Elina Desipri, Victoria Arranz, Iris Vogel (secretary), Nelleke Oostdijk, Erhard Hinrichs, Thierry Declerck, Florian Schiel, Wim Peters, Bertrand Gaiffe, Peter Wittenburg, Jean Claude Martin and Martin Wynne.

In particular due to the excellent and productive atmosphere created by the ILSP team, every member of the group stated to be willing to continue as a member.

Catalan CLARIN meeting

Barcelona

February 6, 2009

Eva Revilla
UPF, Barcelona

The *Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya* (Ministry of Innovation, Universities and Enterprise of the Catalan autonomous government) and the *Universitat Pompeu Fabra* (henceforth UPF), as a member of the CLARIN project, signed an agreement on May, 21st 2008, for the integration of Catalan language resources and tools into the European infrastructure. Thus, the Catalan government has granted funds to ensure the presence of Catalan language resources and tools from the initial phase of the CLARIN project. Within the framework of this agreement, on February, 6th 2009 UPF organized the CLARIN-CAT Meeting to present the project to Humanities and Social Sciences researchers from universities in the Catalan-speaking area.

Governmental support

In the opening of the Meeting, the UPF chancellor highlighted the importance of such an initiative for the whole Humanities and Social Sciences research community. The chief of the Commission for Universities and Research of the Catalan government expressed the will of the government to support the building of research infrastructures for sectors such as the ones CLARIN is aimed at, in which the availability of resources and tools for the Catalan language must be the same as for the other European languages.

After the opening, the CLARIN coordinator, Steven Krauwer, made a general presentation of the project, which was followed with great interest by the audience. Besides the presentation of the project, the Meeting was organized in two round tables.

The first round table

The first round table, entitled *Humanities, Social Sciences and e-science*, gathered researchers from different disciplines, that have been, in many respects, pioneers and visionaries of the possibilities that technological innovations offer to the research in Humanities and Social Sciences. The participants were the following:

The first round table *Humanities, Social Sciences and e-science*

- Joan Santanach, professor of Catalan Philology at the Universitat de Barcelona, editor of *Editorial Barcino* (Barcino Publishing House) and an active member of the Ramon Llull Documentation Centre (http://cdocllull.narpan.net/en/).
- Josep Pich, professor of Contemporary History in the UPF's Humanities Departament, and a member of the University Institute of History "Jaume Vicens Vives". He is a specialist in the history of the catalanist movement and a pioneer in the use of the web as a teaching tool (http://www.upf.edu/huma/professors/jpich/index.html).
- Jordi Vallverdú, associate professor in the Universitat Autònoma de Barcelona's Philosophy Department, where he teaches Philosophy and History of Science and Computing (http://www.vallverdu.cat/).
- Daniel Cócera, as the person in charge of the web sites management in the PADICAT project (Digital Heritage of Catalonia) of the Biblioteca de Catalunya (Catalonia Library). The objective of PADICAT is to archive Catalan web sites (http://www.padicat.cat/en/index.php).
- Ignacio Blanquer, researcher in the Universitat Politècnica de València's High Performance Networking and Computing Group, and coordinator of the Applications area in the e-Science Spanish network, funded by the Spanish Ministry of Science and Innovation (http://www.e-ciencia.es/).
- Laura Borràs, professor of Humanities and Philology Studies at the Universitat Oberta de Catalunya, Director of the Hermeneia Research Group: Literary Studies and Digital Technologies (http://www.uoc.edu/in3/hermeneia/eng/index.html).

The debate focused on the possibilities offered by having digitally accessible resources and the need to develop more tools to access these resources in order to facilitate data harvesting. Moreover, besides the purely instrumental aspect of technology, it was emphasized that e-science has brought changes in the manner of doing research to experimental disciplines, and, likewise, it will entail changes in the research in less

experimental areas. In order that Humanities and Social Sciences researchers can contribute to these changes, researchers attending the Meeting were asked to put forward the needs of their research so that they can be covered by the CLARIN infrastructure. In this respect, the Spanish CLARIN web site has got an email box where ideas and suggestions can be sent, as well as collaboration proposals, to develop specific aids for research projects that may help to show the infrastructure potentiality. This participation space can be found at http://clarin-es.iula.upf.edu/cat/ (Catalan) and http://clarin-es.iula.upf.edu/es/ (Spanish).

The second round table

The second round table, entitled *Resources, tools and services for the research in Catalan language*, was aimed at showing already accessible resources in Catalan, and tools prepared to analyse Catalan texts. In this session the following presentations were made:

- Textual resources of the Institut d'Estudis Catalans (Catalan Studies Institute): *Corpus Textual Informatitzat de la Llengua Catalana* (Textual Computerised Corpus of the Catalan Language), *Corpus Lexicogràfic* (Lexicographic Corpus), and *Diccionari Descriptiu de la Llengua Catalana* (Descriptive Dictionary of the Catalan Language). All these resources were presented by Dr. Judit Feliu, researcher in the Institute (http://www.iec.cat).
- *Corpus Informatitzat del Català Antic* (Computerised Corpus of Old Catalan), presented by Dr. Joan Torruella, researcher and professor at the Universitat Autònoma de Barcelona (http://seneca.uab.es/sfi/cica/).
- *Corpus of the Use of Catalan on the Web*, presented by Dr. Toni Badia, professor at the UPF and researcher in the Barcelona Media Foundation (http://ramsesii.upf.es/cgi-bin/cucweb/search-form.pl).
- *Corpus Tècnic de l'IULA* (IULA's Technical Corpus), presented by Dr. Jorge Vivaldi, researcher in the UPF's Institut Universitari de Lingüística Aplicada (University Institute of Applied Linguistics) (http://bwananet.iula.upf.edu/indexen.htm).
- *FreeLing* text analysis tools, presented by Dr. Lluís Padró, professor at the Universitat Politècnica de Catalunya (http://garraf.epsevg.upc.es/freeling/).

Finally, Dr. Marta Villegas, IULA-UPF researcher for the CLARIN project, showed a simulation of computer applications based on the concepts of the CLARIN infrastructure in order to demonstrate its potentialities.

FLaReNet started

Vienna

February 12-13, 2009

Marko Tadić
Editor

FLaReNet (Fostering Language Resources Network) is an EC eContent Plus Thematic Network whose aim is to create a shared policy and to foster a European strategy in the field of Language Resources (LRs) and Language Technologies (LTs).

By creating consensus among major players in the field, the mission of FLaReNet is to identify priorities as well as short, medium, and long-term strategic objectives, sustain international cooperation and provide consensual recommendations in the form of a plan of action for EC, national organisations and industry.

FLaReNet is bringing together leading experts of many research institutions, companies, consortia, associations, funding agencies, public and private bodies both at European and international level. Anyone can subscribe to the FLaReNet website, joining any of the working groups and participating in their activities. This will offer the advantage of playing a role in the definition of recommendations for future actions, thus shaping the future with respect to the new challenges.

The *FLaReNet Launching Event* combined the FLaReNet themes with the i2010 objectives to address some of the technological, market and policy challenges to be faced in a multilingual digital Europe. The Forum was composed of a series of working sessions where leading experts were invited to present their vision on hot topics in the field of LRs and LTs. A new formula was experimented, whereby the FLaReNet Steering Committee prepared for each session a background document highlighting a set of relevant issues, and in particular a number of question to be addresses by the speakers. In all the sessions discussants (some invited in advance) and participants actively contributed to the ongoing debate about priorities in the sector.

After the opening session, the whole event was divided in six thematic sessions:

- S1. Broadening the Coverage, Addressing the Gaps;
- S2. Automatic and Innovative Means of Acquisition, Annotation, Indexing;
- S3. Evaluation and Validation;
- S4. Interoperability and Standards;
- S5. Translation, Localisation, Multilingualism;
- S6. Enhancing Market Places/Models for Lrs: New Challenges, New Services

followed by the Closing Session that gave the general overview on the whole event with a series of closing remarks and directions for future steps.

This event was closed by the final session that was dedicated to a round-table on International Cooperation, mainly with non-European participants, where future policy and priorities were discussed in a global context.

The links with CLARIN need not to be explained in detail. It is sufficient to say that FLaReNet and CLARIN are cooperating so close that they already had a joint workshop on usage scenarios in Athens, on April 4th and 5th, 2009.

Technical Infrastructure workshop

Oxford
February 25-28, 2009

Dieter van Uytvanck
MPI, Nijmegen

Between 25th and 28th of February, Oxford became the gathering place of a varied group of CLARIN folks. Their mission: bringing the technical infrastructure of the LRT universe to a level where no man has gone before.

¡AAI caramba!

Starting Wednesday, about 25 persons attended a thematic specialist session about Authentication and Authorization Infrastructure (AAI). The main idea behind this theme is addressing the wide proliferation of login credentials in this internet era, colloquially known as the "help – I forgot my password" problem. Instead of creating multiple username/password combinations at each place where one needs to login, it is better to have one strong "identity". Usually this will be that of the organization one is affiliated with – in the case of CLARIN this will be probably an academic institution. The main challenge thus comes down to connecting a distant service (let's say a corpus X that only can be accessed by researchers) to the user database of several organizations (university Y).

As simple as it might sound, practice shows that it is often pretty complex to connect all the loose ends in such a way that all parties can "talk" to each other. Luckily Sebastian Rieger and Matthias Egger – both actively involved in the AAI project of the Max Planck Society – agreed to share their experience with interested members from the CLARIN community. Giving an overview of how to setup an Shibboleth Identity and Service Provider, the public was guided step-by-step to a concrete installation.

In the afternoon Joost van Dijk (SURFnet) shed a light on a more lightweight approach for the same theme. His hands-on presentation demonstrated the nuts and bolts of SimpleSAMLphp. Luckily the two aforementioned systems can live happily together, so the choice is left to the implementing centers what to choose.

The participants of the Oxfo

Centers

On Thursday morning, about 35 people discussed the building up of a network of CLARIN centers. A draft report – result of the numerous talks with candidate institutes – was presented by Peter Wittenburg, and quite some participants took the opportunity to raise remarks and corrections. All in all the conclusion was positive: there will be a satisfying base of about 30 participants and about 6 of them strive to fulfill and infrastructural role. However it was agreed that the number of services to be offered is of a higher importance than the mere number of centers. In a next step the concrete timing plan for each center will be determined.

Metadata

As there was a lot to talk about, the metadata slots were spread over both Thursday afternoon and Friday forenoon. Daan Broeder explained the component-based metadata architecture, which aims to provide a flexible and generally useable way to describe linguistic resources. An important concepts that was dealt with is the Data Category Registry, so to say the semantic glue between the individual metadata elements and components that can be used. A

rd workshop on British lawn

preliminary list of these data categories, especially fit for metadata, was presented with the explicit request for feedback.

Next to that the plans for a prototype of a "Spartan" XML-toolkit were brought forward, which is directed towards those users that want to start right away with creating metadata components and combine the latter to metadata profiles that match their linguistic resources. Further work on this will be done and soon a start-up kit will be available for those interested.

The participants were asked as well to provide examples of metadata instances, as this greatly enhances the understanding of how LRT metadata is used in daily practice.

Meanwhile, a central metadata repository is in place that can harvest resource descriptions from the CLARIN centres. The resulting metadata catalogue could also be integrated in a portal website. It was agreed that DFKI will construct such a portal based on its LT World infrastructure.

It was made clear that for those centers that are willing to start contributing metadata now, either to deliver IMDI or OLAC metadata is the best option. WP2 can immediately start harvesting the metadata and any investments in creating these formats are secure since future compatibility and migration paths are planned.

Finally some information was given about the actual software development process: a call for volunteers was made, and it was agreed that Java will be the implementation language. A further small-scale meeting among the developers has been planned.

Federations

An overview of the current state of affairs regarding the federated access infrastructure was given on Thursday morning by Dieter Van Uytvanck. In that respect it was explained that rather than setting up yet another federation the final goal is to re-use the existing identification gatekeepers at the academic institutions for the access to resources as hosted by the CLARIN centres. This approach is in line with the eduGAIN project and therefore there will be further intensive talks with the GEANT3 project that is going to develop the next generation eduGAIN – for which CLARIN could offer an excellent real-world use case.

Afterwards Antti Arppe focused on the federation within the access rights context. The different roles depositor, content provider and service provider were introduced. Furthermore the need for a centralized license acceptance store was discussed. In a related discussion it was pointed out however that the access rights themselves will be stored normally at the resource providers themselves.

Web Services

The last part of the workshop, on Friday, was devoted to the web services aspect. Marc Kemps-Snijders and Núria Bel presented the challenges that this subject comes with. Some of the main subjects that were dealt with are user interaction, the creation of workflows and the need for a pivot data model. Several existing web services were demonstrated (by UPF, RACAI, BBAW, MPI and D-SPIN) which gave the participants an impression of today's state of play. The accompanying discussion made it clear that in this context synchronization is not only an important technical issue, but that the same surely goes for the intra-project communication.

Conclusion

During three intensive days, participants to the workshop were given a pretty comprehensive overview of the building up of the technical infrastructure . Blueprints for the future were presented and discussed. All in all, the starship WP2 is on a steady course. C

* Interested readers can find all workshop material at:
- http://www.clarin.eu/node/1306 and
- http://www.clarin.eu/node/1307

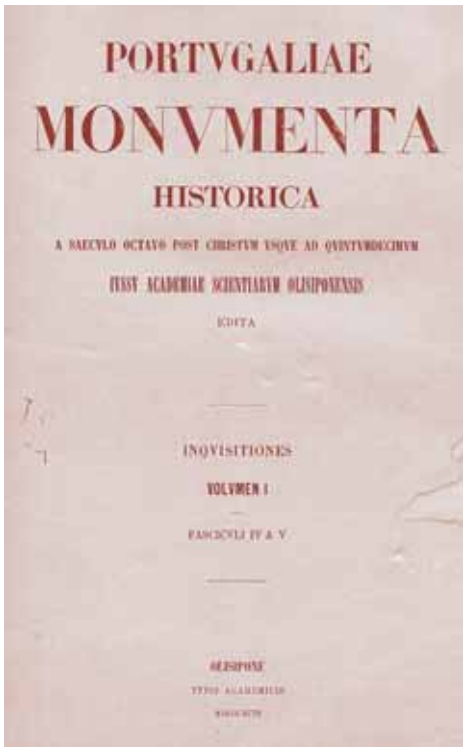
LT meets History in the middle of the Atlantic

A sea of discoveries ahead

Mário Viana
University of Azores
Luís Gomes
University of Azores
António Branco
University of Lisbon

Portuguese is the fifth language with the largest number of native speakers, most of them spread around the margins of the Atlantic. Half way between Europe and America, North Pole and Equator, some live in the middle of this Ocean, in the exceedingly beautiful islands of Azores. This article is about an inspiring project taking place there, at the crossroads between Language Technology and one the most ancient disciplines in the Humanities, History.

The Royal Inquiries (Inquirições Reais) are a set of historical records of paramount importance for the study of the History of Portugal in the Middle Ages. They encode the result of the surveys ordered by different kings to determine who held what (land, livestock, etc.) and what taxes would be liable. By 1888, the result of these inquiries got published in a truly monumental document, which was named, reflecting its oustanding status, as the *Portugalia Monumenta Historica*.



The title page of
Portugaliae Monumenta Historica

The access to this compilation of data, however, is made very difficult or even hampered for a number of reasons, most notably because it is contained in a literally enormous document – the Royal Inquiries of the year 1258 alone, for instance, span over 1100 pages; that document is a mechanical print from the XIX century, in a format close to A3, in 2 columns; and last but not least, it is out of print. A few attempts have been made to rescue this invaluable source of knowledge about the past to the present day productive historical research. It soon became clear that even before getting into the extraction of patterns, generalizations or other findings, the daunting work ahead, with old paper & pencil based technology, just to bring it into a reproducible and manageble format, will hardly permit more than a very fragmented progress.

The example from the manuscript page of *The Royal Inquiries*

By the middle of 2006, a multi-disicplinary team of researchers in the University of Azores, from the History and the Computation Departments, decided to join forces and start a project that aims at improving this state of affairs. After getting a digital version of the document, the ultimate goal is to make it available in the web via an online service that supports historical research, by permiting advanced search of information.

In a first two year phase, this project has focused on the Royal Inquiries gathered in the XIII century, more precisely in the year 1258. This part of the document was scanned into files in PDF format and went through an OCR conversion. The output of this process was filtered out by automatic scripts that handled systematic errors of that conversion, before it entered a final step of validation by human readers.

On a par with this conversion work to a reusable digital format, the project team has been developing a software application that permits to extract important quantitative information from the document. The document is topographically arranged in a such way that many correspondences can be established between the way the pieces of information are displayed (in colums, boxes, etc.) and its semantic value (parish of the land, number of olive trees in it, amount received in taxes, etc.). Natural language technology plays a key role here through many tools – sentence splitters, lemmatizers, taggers, definition extractors, named entity recognizers, etc. – which make it possible to get hold of linguistically significant stretches of text.

This project received the designation INQ1258, and its first phase is funded by the Regional Government of Azores until the end of 2009 under the contract M2.1.2/1/008/2006. After this first pilot phase, the time and expertise gathered will be ripe to address the rest of that document, which turns out to be as much clumsy as monumental. One of the next steps will be to develop an online service, supported by the application mentioned above, which will open the access to the Inquiries and their secrets, in the past only for royal eyes, to the nowadays research community all around the web.

For further details on the INQ1258 project, get in touch with Prof. Mário Viana (mviana@uac.pt) or Prof. Luís Gomes (lmg@uac.pt). The University of Azores is one of the members of the Portuguese network of CLARIN. C

CLARIN activities in Sweden

Lars Borin
*Språkbanken, University of Gothenburg**

CLARIN has two Swedish partners (Centre for Speech Technology, KTH and the Humanities Laboratory, Lund University) and a considerable number of members, including the sites of the authors of this document. However, the Swedish Research Council (VR) has decided not to allocate national funds for Swedish involvement in the preparatory phase of CLARIN. On the other hand, there are a number of ongoing activities which are directly relevant and explicitly linked to CLARIN, funded as national projects mainly by the VR Committee for Research Infrastructures (KFI) and its subcommittee DISC (Database Infrastructure Committee). Some of these projects are decribed below.

An infrastructure for Swedish language technology (National consortium; funding: KFI)

As a result of a planning grant awarded in 2007 to a national Swedish consortium, a proposal was prepared for an integrated basic Swedish language technology research infrastructure, consisting of a *Swedish national corpus* and a Basic LAnguage Resource Kit (BLARK). The proposal is now being reviewed by international experts. The funding needed for realizing the SNK and Swedish BLARK in parallel is estimated at 130 MSEK (about 12 M€) over 7 years. However, we estimate that pursuing the two separately would cost on the order of 50 MSEK more, i.e., there is considerable synergy in the proposal. No doubt in large part as a result of the work in this planning project, VR has listed language technology as one of a small number of national research infrastructure areas of highest national priority in its latest *Roadmap to research infrastructure*. This spring, a one-off call has been issued for proposals by national consortia in exactly those areas. Thus, it seems there is a good chance that the two years of dedicated work laid down in this project might pay off.

Safeguarding the future of Språkbanken (Språkbanken, University of Gothenburg; funding: DISC)

Språkbanken (the Swedish Language Bank; http://spraakbanken.gu.se) provides a service to the research community since 1975, whereby language resources are made freely available online. The aim of this project is to achieve an integration of the resources and tools in Språkbanken in a way that takes into account international standardization work. CLARIN is seen as so important by Språkbanken – whose day-to-day activities will be profoundly influenced by the recommendations, etc., which will emerge from CLARIN preparatory phase work – that part of the funding for this national project is used to participate in CLARIN; at the present time, this is one of the best ways of safeguarding the future of Språkbanken.

The Bank of Swedish and its outstanding web interface for concordancing

Spontal: Multimodal database of spontaneous speech in dialog (Centre for Speech Technology, KTH; funding: DISC)

The ongoing Swedish speech database project, *Spontal: Multimodal database of spontaneous speech in dialog* takes as its point of departure the fact that both vocal signals and gesture involving the face and body are important in everyday, face-to-face communicative interaction. Our understanding of vocal and visual cues and interactions in spontaneous speech is growing, but there is a great need for data with which we can make more precise measurements. The goal of the Spontal project is the creation of a Swedish multimodal spontaneous speech database rich enough to capture important variations among speakers and speaking styles to meet the demands of current talk-in-interaction research. Currently, about 25% of the database has been recorded.

SweDia 2000 – A Swedish dialect database (Phonetics, University of Gothenburg; funding: DISC)

The *SweDia* database consists of recordings of speech from 107 Swedish dialects, made in 1999 by a previous research project, *SweDia 2000*. The database has until now primarily been used by the SweDia group. The goal of the present work is to make the database available to a much wider circle by placing it online. A first version of (nearly) the entire database already exists hosted on an IMDI-server at Lund University. A part of the database that comprises informal interviews and semi spontaneous monologues will be simultaneously hosted on a server at the University of Oslo. This part of the database will be combined with data collected by the Scandinavian Dialect Syntax project.

Litteraturbanken (Språkbanken, University of Gothenburg; funding: the Swedish Academy)

Litteraturbanken (the Swedish Literature Bank; http: //litteraturbanken.se) is a public digital repository of classical Swedish literature – a cultural heritage project with permanent funding by the Swedish Academy. Its relevance to CLARIN is twofold: (1) Technically, Litteraturbanken is included in Språkbanken; (2) Litteraturbanken has been developed with the aim that it can serve as a primary data source for research in a number of disciplines in the humanities and social sciences, using integrated language technology tools.

Summing up

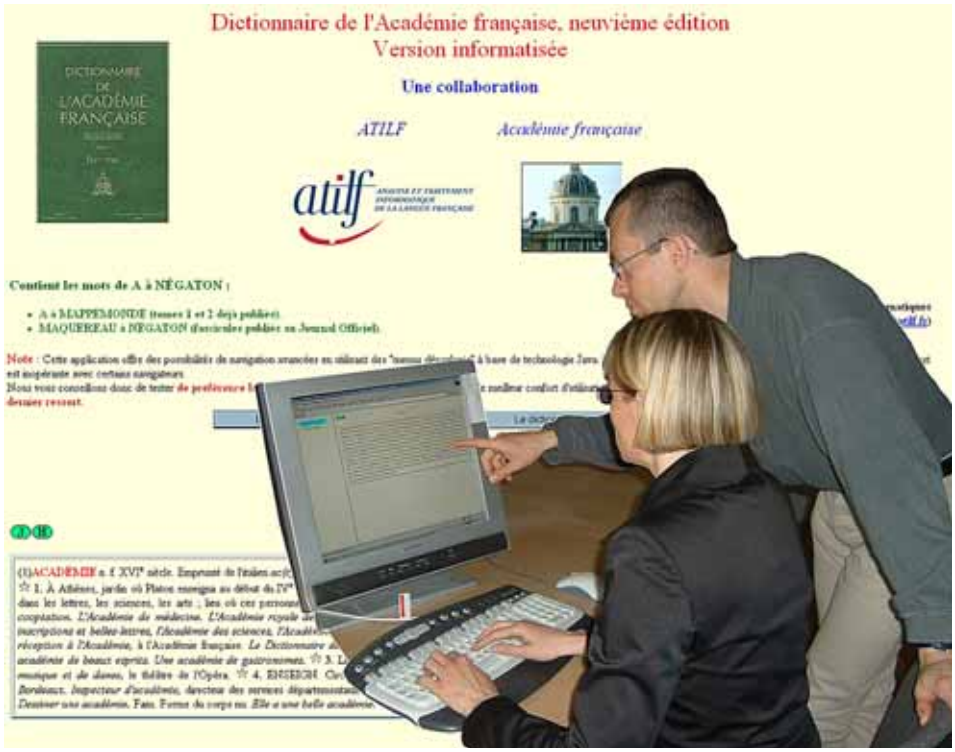
Even though VR has not set aside funds explicitly for CLARIN work, the projects described in the preceding section together represent VR funding of 10.6 MSEK (about 1 M€), plus about 2.5 MSEK annually to Litteraturbanken. The resources being realized with this funding will be extremely valuable when CLARIN enters its permanent phase.

* With contributions from the Swedish CLARIN community

The French activities in LRT and their connection with CLARIN

Jean-Marie Pierrel
ATILF Nancy University and CNRS

The Language Resources and Tools activities in France are both numerous and widely spread over the national territory. The main actors are located in Aix-Marseille (LPL and LIM laboratories), Bordeaux (LABRI and INRIA Bordeaux), Grenoble (GIPSA /ICP anc CLIPS), Lyon (ICAR and Dynamics of the language), Nancy (ATILF, INIST and LORIA), Paris (LIMSI, MODICO, LATTICE, LIPN, LDI and ITEM), Rennes (IRISA), Toulouse (CLEE/ERSS laboratory).



The digital version of the 9th edition of the *Dictionnaire de l'Académie française*

These laboratories are connected to two scientific associations:

- ATALA (French association for computational linguistics). ATALA aims at encouraging the studying of theoretical and practical issues related to computational linguistics and to help along scientific exchanges in the domain. In particular, ATALA supports the French conference TALN each year.
- AFCP (French Association for Spoken Communication, i.e. French chapter of ISCA) dedicated to supporting, distributing and developing research and resources connected to spoken communication from phonetics to interaction studies and automatic treatment.

In 2005 and 2006, CNRS set up "national centers for resources" among which three are directly connected to language:

- CRDO: Resource Center for Oral Data (http://crdo.up.univ-aix.fr) managed by both the laboratory Speech and Language (LPL-UMR 6057) and by Linguistic Laboratory for Languages with Oral Tradition (Lacito-UMR 7107).
- CNRTL: Resource Center for Text and Lexical Resources (http://www.cnrtl.fr), managed by the Analysis and computerized processing of French laboratory (ATILF-UMR 7118) who is a member of the Clarin project.
- TELMA: Electronic Processing of Manuscripts and Archives (http://www.cn-telma.fr/) managed by the Institute for research in the History of Texts (IRHT-UPR 841) and the "Ecole Nationale des Chartes".

More recently, the French Ministry of Research published its roadmap for Big Research Infrastructures. This roadmap makes an official statement for three "TGIR" (French acronym for "big research infrastructure") in the field of the humanities. These TGIR are working in the same field as CLARIN on a national basis.

- the TGIR Adonis is a CNRS big equipment (in social sciences and humanities) dedicated to producing new knowledge through navigation facilities in digital documents. From a technical point of view, this amounts to developing tools based on research from both computer science and social sciences in order to access corpora (and more generally data). From a more sociologic point of view, Adonis promotes interdisciplinarity and interactions between different communities.
- The TGIR BSN (Digital Library Platform for Scientific Literature) aims at promoting the usage of digital scientific literature. This includes what used to be called "Technical and Scientific Information" adding to it facilities for navigating between documents (either digitally born or digitalized paper versions of scientific journals or open archives). (http://cleo.cnrs.fr; http://www. persee.fr)
- The TGIR CORPUS/SHS (Cooperation of Research Operators For Using Digital Resources in the humanities and social sciences). CORPUS is a cooperation platform for accessing the initial data (images, sounds, texts) produced by researchers in linguistics, psychology, history, archeology, geography, literature and arts. CORPUS aims at promoting the sharing of resources between researchers and promoting the use of digital resources in social sciences and humanities. This implies using well defined norms for documents and annotations, proper identification of resources, distribution, long term preservation and help in the polling of resources. Moreover, CORPUS has to support French participation in international projects, and to ensure long term preservation of big corpora of documents (images, sounds, texts) gathered or produced by researchers. CORPUS supports the ESFRI roadmap in what concerns linguistics, that is CLARIN (Common Language Resource and Technology Infrastructure). All disciplines of human and social sciences using digital documents for their research, in particular linguistics, psychology, history, archeology, geography, literature studies and arts are aimed at by this TGIR. (http://www.cnrtl.fr; http://www.msh-reseau.fr/spip.php?article34).

The first year of CLARIN in Latvia

Inguna Skadiņa
Institute of Mathematics and Computer Science, University of Latvia

The first year of the CLARIN project was very important for activities in Latvia. This year, two significant activities have been initiated, i. e., the CLARIN Latvia project and the National Corpus of Latvia. Since these initiatives are closely related and the target audience is very similar, we have joined our forces in dissemination activities and in activities related to assessing the current state of the art in language resources and language processing.

CLARIN presented to the Latvian State Language Commission

On April 2, 2008 the CLARIN initiative was presented at the workshop organized by the Latvian State Language Commission. It was the first time the CLARIN initiative was presented to the researcher community in Latvia and it received positive feedback from the participants.

The current situation in language technologies and resources was presented and discussed in a workshop. This workshop gathered ca. 30 participants – representatives of research institutes, universities, publishing houses, libraries and companies working on Latvian language resources.

The Latvian State Language Commission was established in 2002 by the president of Latvia with the aim to analyze the situation of the official language and to design recommendations for strengthening and developing the status of Latvian as the official language.

Latvian National Corpus initiative

For a number of years development of the Latvian corpus has been among the priorities in the language policy of Latvia. Still practical implementation was hindered by a lack of funding, coordination and a limited awareness in the humanities community. To coordinate the activities of different institutions and raise general awareness the Latvian National Corpus initiative was launched and a working group established.

The Latvian National Corpus (LNC) initiative has linked efforts of the Latvian State Language Commission, the National Library of Latvia, the biggest resource holders and the universities. The basic balanced corpus of 1 million words has already been created by the Institute of Mathematics and Computer Science (IMCS) of the University of Latvia. A huge contribution to LNC will come from the National Library of Latvia which has started a massive digitalization of printed publications.

The Latvian National Corpus has been envisioned as a Latvian building block in CLARIN's common language resource infrastructure. It will be an open online resource providing access to federated resources from different research institutions and content providers. "Collection of every piece of Latvian text that ever has been made public either in printed or online form" – this is the dream of the promoters of the LNC initiative.

CLARIN Latvia receives funding

In September the Institute of Mathematics and Computer Science has signed an agreement with the Ministry of Education and Science for supporting CLARIN activities in Latvia. The funding is granted for the initial period (until April, 2009) of the preparation phase with the aim to support the participation of Latvia in activities of WP2, WP3, WP5 and WP8, as well as for the organization of CLARIN related activities at the national level and for the creation of a national contact point and a national network of expertise.

National seminar

On November 3, the seminar CLARIN project and the National Corpus was organized by IMCS, the Latvian State Language Commission and the National Library of Latvia with the aim of bringing together the potential CLARIN community of Latvia – owners and developers of resources, language technology developers and users of linguistic resources and tools. The seminar was opened by prof. Andrejs Veisbergs, Chairman of the Latvian State Language Commission who explained the necessity for the Latvian National Corpus and the importance of the CLARIN initiative.

The morning session was devoted to the CLARIN project. The CLARIN project coordinator Steven Krauwer presented the mission of the CLARIN initiative and its role for languages, emphasizing that all languages (widely used or small) are equally important in CLARIN. Inguna Skadiņa (IMCS) introduced with CLARIN project activities in Latvia, aims and tasks, asking participants actively participate in the creation of a CLARIN network of expertise in Latvia. The CLARIN National Advisory Board was established during the seminar with the aim of facilitating the creation of a CLARIN Latvia infrastructure.

The afternoon session was devoted to the Latvian National Corpus initiative. Prof. Franti]ek Čermák introduced Latvian scientists to the experience of the Czech National Corpus Institute. Andrejs Vasiļjevs (Latvian State Language Commission, Tilde) presented a work of the National Corpus initiative group. Everita Andronova (IMCS) introduced a corpus concept and sketched the current status of the Latvian corpus. Inga Grīnfelde (National Library of Latvia, NLL) presented the on-going work to develop a Latvian National Digital Library which will be the biggest repository of Latvian culture – comprising newspapers, texts, photos etc.

The good point of this meeting was that representatives of regional universities were acquainted with the aims and possibilities of CLARIN, which will encourage them to use new technologies and to create their own resources. The meeting was closed by a very interesting discussion on issues related to corpus, copyright issues and access to language tools.

Future activities

Now, when the potential contributors and users of CLARIN infrastructure have been introduced to the project, IMCS will continue work on fulfilling the aims of the CLARIN preparation phase. Our researchers are actively contributing to WP2, WP3, WP5 and WP8. During the seminar questionnaires on WP3 and WP5 were distributed to participants. They have been collected and we plan to have an overview of the state of the art in LT and Latvian resources in the spring of 2009. C

* With contribution by Andrejs Vasiļjevs (Latvian State Language Commission, Tilde)

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.

March 2009

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

2009-03-23 to 2009-03-25: Text Mining Services - TMS Conference Leipzig, Germany

2009-03-30 to 2009-04-03: EACL 2009, Athens, Greece

April 2009

2009-04-04 to 2009-04-06: CLARIN – FLaReNet Usage Scenario Workshop, Athens, Greece

2009-04-06 to 2009-04-08: PALC2009, Lodz, Poland

2009-04-22 to 2009-04-24: 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt

May 2009

2009-05-07 to 2009-05-08: Research Connection 2009, Prague, Czech Republic

2009-05-11 to 2009-05-13: CLARIN Consortium Meeting, Barcelona, Spain

2009-05-14 to 2009-05-16: EAMT2009, Barcelnoa, Spain

2009-05-14 to 2009-05-16: NODALIDA 2009; the 17th Nordic Conference of Computational Linguistics, Odense, Denmark

June 2009

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

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