

# CLARIN



## Newsletter

Number 9-10, 2010, March-June

### *Language Technologies within A Digital Agenda for Europe and other perspectives*

*DG Information Society and Media of the European Commission supports multilingual technologies through research and innovation programmes. The current level of EU support for language technology projects exceeds 70 million and will reach 130 million in 2012*



**Kimmo Rossi**  
European Commission  
Directorate General Information Society and Media

#### **The challenge**

Future online services, whether addressing business, administrative or social needs, have to bridge language barriers. To achieve this, language technologies still need to find their way into constantly evolving web services and technologies. New ways of communicating and collaborating online are no longer the realm of a few corporate sites or online communities, but an integral feature of most portals and services. Millions of users produce and consume online content which changes shape every second. Language technologies need to adapt to the challenge of delivering transparently, “on-the-fly”, while coping with different forms and uses of written and spoken language. At the

same time, the European map is still fragmented and scattered with too many blank spots.

With the help of our portfolio of projects, we have made a good start – bringing different communities and specialist groups together, stimulating cooperation between users and providers, making supply and demand meet in specific application domains. But we still have a long way to go, and the process will take time, substantial resources, and a common vision.

#### **What happens next?**

The *ICT Work Programme for 2011-2012* operations will be published in the coming weeks. A dozen calls for proposals are planned over the next two years. The 7<sup>th</sup> call launched in late September will provide 50 million for research activities under Objective 4.2 – Language Technologies. In February 2011, another call will be published providing 35 million within Objective 4.1 – SME action on Digital Content and Languages.

The small and medium size enterprises (SME) action is a new concept in many ways. Firstly, it provides an opportunity for fast-moving SMEs, including start-ups and university spin-offs, to test and validate innovative ideas. Secondly, it provides new ways of linking the information management community with the language technology and resources community. Thirdly, since participation is not limited to SMEs, the action provides result-oriented opportunities for new partnerships: public with private, SMEs with well established companies, data providers with developers and leading users. The overall goal is to generate added value from trading, pooling and exploiting data in order to produce novel technologies, products and services.

#### **Looking further into the future**

The recently published *A Digital Agenda for Europe* (DAE) throws a big challenge to the language technology community. The DAE sets out to boost economic growth and exploit

the full benefits of the digital economy by creating a borderless and interoperable digital single market. More concretely, the DAE invites the Commission to “work with stakeholders to develop a new generation of web-based applications and services, including for multilingual content and services, by supporting standards and open platforms through EU-funded programmes.”

In practice, this means broader and closer collaboration between EU projects, businesses, institutions and administrations. We will work on multilingual services and platforms together with other EC departments, EU institutions and international organisations exposed to multilingualism. We will liaise with the web industry to establish standards and best practices to make the web a more language-friendly environment, and to pave the way for widespread deployment of language technologies.

In 2011 we will set up a *Business Forum*, bringing together current and emerging stakeholders from the supply and demand side. The forum is expected to help define the requirements of both providers and users of language technology products and services, with emphasis on innovation and demand-driven technology development. At the same time, we expect ongoing and upcoming EU-funded actions to deliver a compelling and widely supported *Strategic Research Agenda* for the field at large. This agenda will be released in 2012 at a point in time when important decisions are going to be taken on the next R&D framework programme, due to start towards the end of 2013.

Data sharing and portability have become more important than ever for researchers, developers and users. Fifteen years after ELRA was established, new initiatives such as CLARIN and META-NET can promote a new culture and help pool, preserve and reuse valuable resources. While these projects serve different communities and purposes, there is ample scope for collaboration, e.g. implementing open repositories, unifying metadata descriptions, and solving legal issues hampering the wide availability of research results. **C**

# Editors' Foreword



**Marko Tadić  
& Dan Cristea**

*CLARIN Newsletter editors*

It seems that, as we approach the end of the project, the agenda of CLARIN-related activities and events becomes more and more heavy. If you go through this double issue, even from the tip of the finger, you will be surprised to see how many important things has happened from the beginning of this year, and at what speed we are moving forward. In the following we will browse for you the content of this CLARIN Newsletter, the double issue 9-10, not necessarily in the order the articles are included.

First, we are happy to bring to the CLARIN readers a word from the very heart of the European Commission, there where the strategic decisions are being taken in Language Technology. Kimmo Rossi is giving the fresh and very welcome news of a hot autumn and winter: very generous calls of the ICT Work Programme are waiting for us behind a door that we are invited to open and 2011-2012 will bring even more amazing initiatives.

Then, we announce new launches of programmes, projects and collaborations. In Netherlands, the CLARIN-NL Project has received an extremely generous financing, which could be taken as an example by many

CLARIN member states, news offered to us by Jan Odijk. António Branco and Marko Tadić relate about the recent launch of the Portuguese CLARIN in Lisbon. The same António, together with Vera Lúcia Strube de Lima, Thiago Pardo and Steven Krawer, describe, from the position of principal actors, another important event involving the Portuguese language, the PROPOR conference, which can be seen as initiating a promising collaboration of CLARIN on the other side of the Atlantic, in Brasil. From Kaunas, we hear the voice of Rūta Marcinkevičienė, bringing to the CLARIN community news from Lithuania, the new member of CLARIN, and its remarkable progresses in HLT.

Then, as in all our previous issues, a series of articles relate about CLARIN activities. The new face and functionalities of the CLARIN website are commented by Corina Dima, its main implementer (WP6). Ville Oksanen, Krister Lindér and Hanna Westerlund, a Finish team (WP8), present the principal LR distribution licenses standardized within the project. And the state of the art in speech technologies and how are they related to language learning CALL systems, as possible part of the CLARIN infrastructure, is contributed by Catia Cucchiariini.

Finally, a number of articles present the main LR&T events in Europe within the last few months. The LREC series of conferences, due to a rich heritage and an almost perfect conceptualization and organization, has become one of the major global events of our community, attended now by 1,246 registered participants at the main conference and workshops. Less known, perhaps even funny, stories from

the recent Malta event are brought forward by the chair of the local organization committee, Mike Rosner, while the conference main chair, Nicoletta Calzolari, together with Claudia Soria and Riccardo Del Gratta, describe one of the most important side effects of the Malta LREC – the LRT Map. A great idea, which, will certainly have important consequences in the making of a global thesaurus of language resources and tools. Then, the already significant

CICLing, held this year for the first time in Europe, in Iași, is presented by its very inventor, Alexander Gelbukh, and the principal local organizer, Corina Forăscu. One of the most important CLARIN meetings since the beginning of the project was the join gathering of its boards, which included the Executive

Board, the Scientific Board, the Strategic Coordination Board and the International Advisory Board. The coordinator of the project, Steven Krauwer, is giving a short presentation, with a stress on the advances towards establishing the CLARIN ERIC, the new framework under which the project will function starting with the construction phase. You will find a word, then, on the recent join meeting of CLARIN and FLReNet in Stockholm, by Rolf Carlson, Kjell Elenius and David House. The discussions focused on speech and multimodality, as well as on the best practices that would enable to integrate tools and to achieve uniformity in annotation. You can find also a very short survey of the LT Days that took place in Luxembourg, reported by Marko Tadić.

Enjoy reading it! We would be happy to know your opinion, your news, your suggestions. Write us! **C**

## Call for contributions

Dear readers of the CLARIN Newsletter,

If you have ideas, thoughts, comments, additions, corrections, arguments, questions etc. which are connected to the CLARIN project, even remotely, please feel free to send them to us as your contribution at [newsletter@clarin.eu](mailto:newsletter@clarin.eu) or directly to the editors at [marko.tadic@ffzg.hr](mailto:marko.tadic@ffzg.hr) and [dcristea@info.uaic.ro](mailto:dcristea@info.uaic.ro).

## List of national correspondents

### Austria

Gerhard Budin

### Belgium – Flanders

Inneke Schuurman

### Bulgaria

Svetla Koeva

### Croatia

Marko Tadić

### Czech Republic

Karel Pala

### Denmark

Bente Maegaard

Hanne Fersøe

### ELRA/ELDA

Stelios Piperidis

Khalid Choukri

### Estonia

Tiit Roosmaa

### Finland

Kimmo Koskenniemi

### France

William Del Mancino

Bertrand Gaiffe

### Germany

Lothar Lemnitzer

### Greece

Maria Gavrilidou

### Hungary

Tamás Váradi

### Italy

Valeria Quochi

### Latvia

Andrejs Vasiljevs

### Malta

Mike Rosner

### Netherlands

Peter Wittenburg

### Norway

Koenraad De Smedt

### Poland

Maciej Piasecki

### Portugal

Antonio Branco

### Romania

Dan Cristea

Dan Tufiş

### Spain

Nuria Bel

### Sweden

Sven Strömqvist

### UK

Martin Wynne

# The CLARIN-NL Project



**Jan Odiijk**  
Utrecht University

CLARIN-NL forms the Dutch national counterpart of the CLARIN enterprise on the European level (CLARIN-EU). CLARIN-NL covers a period of 6 years, partitioned in three phases of two years: preparation, construction, and exploitation. CLARIN-NL effectively started in April 2009 and has a budget of 9.01 million euro. CLARIN-NL is the first CLARIN-related national project that has been awarded money for the implementation and exploitation phases. Its website (<http://www.clarin.nl>) is mostly in English, so additional information on the project can be easily obtained here.

CLARIN-NL has been set up as a mixture between a programme and a project, providing both the flexibility to adapt the contents to new developments, and to new players in the field (e.g. humanities researchers not reached yet). At the same time it offers opportunities for defining a few longer term projects in selected areas so that knowledge and expertise built up will be sustained in the participating institutes.

## Wide coverage of academic institutions

Currently CLARIN-NL has 23 participants from linguistics and the humanities more broadly, and it is open for new Dutch participants. It includes all universities with a humanities faculty or with expertise in language and/or speech technology, and several institutes from the Royal Netherlands Academy of Arts and Sciences. The institutes carry out research in linguistics construed broadly (10), language technology (6), speech technology (2), culture (2), lexicography (2), social history (4), and literature (1). CLARIN-NL thus covers a large part of humanities research and some social sciences research. Libraries are represented as well (2), and 5 institutes are data cen-

tres. Of these, 4 (INL, MPI, Meertens, and DANS) have expressed the intention to become a CLARIN centre (type A or B), and some others are considering this.

## CLARIN-NL sub-projects

CLARIN-NL has initiated a range of sub-projects. Some of these concern the design and implementation of the infrastructure and the CLARIN centres in the Netherlands, among them a *metadata project* to carry out first tests of the Component Metadata Infrastructure (CMDI) as developed in CLARIN-prep (WP2 and WP5) against a representative

set of interactive interviews in which the surveyor gets into a dialogue with the researcher to get an overview of his/her research questions and may suggest possible tools and data that might facilitate the research. The baseline measurement will inventory the current use or non-use of digital data and tools in the humanities making it possible to determine the impact of the CLARIN infrastructure on the humanities research in a later stage.

A call has been launched in 2009 for *Data Curation and Demonstrator Projects* which has led to 11 small projects carried out by LST and humanities



CLARIN-NL kick-off meeting, Utrecht, 2009-05-27

sample of data residing in Dutch research organizations; the *infrastructure implementation project (IIP)* to design and implement the technical infrastructure with CLARIN data centre candidates as participants; the *Search & Develop* project in which centralized metadata search and distributed content search functionality will be implemented.

A project has been started up to carry out a *User Survey and Baseline Measurement*. The survey will be carried out via

researchers, coordinated by members from the target group (humanities researchers) and focusing on needs derived from the humanities researchers' research questions.

The goal of a *data curation project* is to adapt an existing resource so that it is visible, uniquely referable and accessible via the web, and properly documented.

The goal of a *demonstrator project* is to create a documented web application

Continued on the next page 



starting from an existing tool or application that can be used as a demonstrator and function as a showcase of the type of functionality CLARIN will incorporate and support.

Important goals *common* to both types of projects are (1) apply standards and best practices and make use of the suggested CLARIN architecture and agreements to understand their limitations and the requirements for extensions; and (2) establish requirements and desiderata for the CLARIN infrastructure.

In particular, all projects will have to make metadata for their resources in accordance with the CMDI, and contribute to semantic interoperability by mapping the data categories used to data categories in ISOCAT, or by creating new data categories in ISOCAT.

In this way, curated resources and a range of showcases will become available. At the same time evidence-based requirements and desiderata for the CLARIN infrastructure and supported standards and best practices will be obtained making it possible to influence the final selection of standards and best practices promoted by CLARIN. These data curation and demonstrator projects neatly complement the European CLARIN preparatory project, in which the budget for a work package on these issues was cut by the European Commission. An overview of the awarded projects can be found on the CLARIN-NL website.

### **Dutch-Flanders cooperation**

A 3-year project for *cooperation between the Netherlands and Flanders* has started. The cooperation project consists of multiple aspects, but the core is a project in which existing tools for Dutch are adapted to become web services that can be used in a work flow system. This will be done for two modalities: text and speech. The users involved in this project are researchers from literary studies and archaeology on the text side, and social history on the speech side. Offering web services in a work flow will become one of the most important functionalities the CLARIN infrastructure will offer. The added value of the coop-

eration with Flanders can be found in the fact that the tools have originally been developed together and concern the shared Dutch language.

### **Education, training and awareness**

The CLARIN-NL project includes a plan for creating awareness and for education and training. These activities (some already undertaken, more in the pipeline) involve active contributions as well as financial and logistic support for a range of events such as workshops, tutorials, summer and winter schools, project web site and wiki, newsflashes, mailing lists, publications in magazines, etc. CLARIN-NL is also working on introducing courses for working with digital data and tools in the CLARIN infrastructure as part of the regular humanities curriculum.

### **Relation with CLARIN Europe**

CLARIN-NL is complementary to the European CLARIN preparatory project in several respects: First, CLARIN-NL does not only cover the preparatory phase, but also the implementation phase and part of the exploitation phase of the CLARIN infrastructure. Second, CLARIN-NL carries out activities in the preparatory phase, such as the metadata project, data curation and demonstrator projects, for which no funds are available in the European preparatory project.

### **Future**

The CLARIN-NL executive board is currently analyzing the situation in the Netherlands. It attempts to identify gaps in topics and disciplines covered that are as yet underrepresented, as well as essential infrastructure functionality that is not covered yet. It will – based on this analysis, the results of the user survey, and after consultation with its advisory panels – work out a proposal for a new call for sub-projects in 2010 with clearly identified priorities. It will also determine the best organisation of this call, both in terms of character (open call, tender, direct assignments), budget, and

timing. It is expected that the plans for a new call will be available in June 2010.

CLARIN-NL aims to create an infrastructure that is supposed to have a long term existence, although it is just a project, finishing in 2014. It is therefore of utmost importance to already start working on embedding the CLARIN infrastructure in the normal research and centre activities and preparing both a governance structure and structural financing to ensure the long-term existence of the infrastructure. Therefore it is closely tracking the developments in Europe, especially with regard to the CLARIN ERIC, and assessing how to best play a role in this organisation.

### **Conclusions**

The CLARIN-NL project can serve as an excellent example for other national CLARIN projects: its set-up as a mix of programme and project, creates flexibility, and is an excellent way of involving as many relevant national parties. Moreover data curation and demonstrator projects offer opportunities to seriously test the standards and best practices promoted by CLARIN, strengthening these standards and best practices, and provide arguments for modifications or extensions. More generally, these projects will yield evidence-based requirements and desiderata for the CLARIN infrastructure. This is the best way to ensure possibilities for influencing a selection of standards and best practices in CLARIN that is compatible with the existing national data. In addition the projects will yield curated data, and, via the demonstrators, a range of showcases that can be used to explain and demonstrate the advantages of the CLARIN infrastructure and the new possibilities it will offer to researchers.

Finally, the project requires cooperation between intended users (humanities researchers) and the technology providers (infrastructure specialists and HLT providers), with a central role for the users' research questions and thus contributes to bringing these different communities together in collaborative projects. **C**

# The New Face of the CLARIN Project Website



**Corina Dima**  
"Alexandru Ioan Cuza"  
University of Iași

The CLARIN project internal website is a space used by the researchers in the project to work on their tasks and collaborate with the other members of their group. There are currently 194 member institutions in CLARIN, from 33 different countries, and a number of 406 registered users.

In March this year the CLARIN project internal website was updated. It received a fresh look and a lot of new functionality was added. In this article we'd like to give a short overview of these changes.

## Site navigation

The site navigation system now features a blue drop-down menu at the top of the page, which contains handy links to the most important pages in the site. Here you can find an overview of CLARIN, its structure and its member institutions, in the *Clarin* menu, a series of published material under *Publications*, links to the home pages of the project's groups under *Clarin Groups*, a list of events relevant for CLARIN under *Events*, a list of metadata inventories hosted by CLARIN under *Resources* and a list of frequently asked questions about CLARIN and a *Website* section under *Help Desk*.

## News, events and newsletter

On the left side of the page, three information blocks contain updates relevant to the CLARIN community:

1. The *News* section brings under the members attention information about new developments in CLARIN and related infrastructures, about important upcoming events and various other news relevant for the community.
2. The *Upcoming events* section offers information about future events.
3. The *CLARIN Newsletter* section highlights the latest newsletter published by CLARIN. The newsletter presents information and new developments from all over the CLARIN world, and is published 4 times a year.



CLARIN web site redesigned

## Sections for registered users

The employees of CLARIN Member institutions are allowed to create an account on the CLARIN internal website. Registered users have access to content creation facilities, which are found in the grey user menu.

## Content creation

Under the *Create content* tab members can find several types of content that they are able to create on the internal site:

- **Book page:** book pages are the simplest type of content on the site; they can be used to add new pages of information.
- **Deliverable:** the deliverables content type is used to add the project deliverables and milestones to the site.
- **Event:** the event content type allows the members to add new events to the internal website; the events can be either public (like a conference) or organized by CLARIN.
- **News item:** relevant news items can be added to the website using this form; ideally, the added information should be gener-

al and relevant for all the CLARIN members;

- **Presentation:** members can upload their CLARIN-related presentations using the provided form;
- **Resource:** members are encouraged to use this form to add metadata about existing linguistic resources to the CLARIN inventory;
- **Tool:** tools are also harvested in the CLARIN inventory; this form offers the possibility to add metadata about new tools.

## Site outline

All the content on the website was reorganized and added to a hierarchy. It can be accessed via the Site Outline menu link found in the grey user menu. On the root level there are the most important sections of the site, mainly related to the CLARIN groups, but also to the most important types of content. Almost all the created content is added automatically to the hierarchy. For example, all the events are by default placed

Continued on the next page ➔

under the Events node when they are created. The pages are the only type of content for whom the user can choose the place in the hierarchy where it should be added. The site outline can also be edited and the nodes inside rearranged, but this operation is restricted to a group of trusted users.


## Home pages for groups

All the CLARIN groups now have a home page. Here one can find information about the mission of a group and about its members. A *Documents* section presents the pages that contain relevant information for that group. A *News* section is present on the right side of the page. Group administrators can add news items that are only relevant for their group using the *Group News* form and selecting their group as a target for that piece of news. For the larger, work package-level groups, a *Deliverables & Milestones* section is present. *Deliverables* are added automatically to these sections as soon as they are uploaded to the site using the *Deliverable form*. Three special pages are the ones for the CLARIN Members, CLARIN Partners and CLARIN Executive Board groups. The CLARIN Members page contains an extra *Flyers & Posters* section. The Partners page has 4 additional sections:

- Templates, containing various templates with the CLARIN logo (PPT presentation, deliverable, letter);
- Administration & Finances, with administrative and financial information related to CLARIN;
- Consortium Meetings, which presents the past and future CLARIN consortium meetings;
- News from the Executive Board, where small updates, special for partners, coming from the Executive Board can be seen.

The Executive Board page contains sections created especially for the Executive Board *Meetings, Agendas, Minutes and Presentations*. The access to the Partners and Executive Board pages is restricted to the members of these groups only.

This is just a short presentation of the internal site. You can find more detailed information in the *Help Desk* menu under *Website*. Also, if you have a question or need to report a problem regarding the functionality of the website, you can use the *Contact webmaster* form under *Help Desk > Website*.

Enjoy your browsing! 



**Ville Oksanen**  
**Krister Lindér**  
**Hanna Westerlund**  
*University of Helsinki*

One of the most challenging tasks in building language resources is the management of copyright licences. There are several reasons for this. First of all, the current European copyright system is designed to a large extent to satisfy the commercial actors, e.g. publishers, record companies, etc. This means that the scope and duration of the rights are very extensive and there are even certain forms of protection that do not exist elsewhere in the world, e.g. database rights. On the other hand, the exceptions for research and teaching are typically very narrow.

In CLARIN the typical flow of the content is the following: A copyright holder, e.g. a newspaper, licenses its content to a CLARIN Content Provider that distributes the content to the End Users through a CLARIN Service Provider. This means that the license chain has to follow a similar structure. Unfortunately even the first step is often difficult because there is a group of resources, for which there are no written license agreements and individuals familiar with the details are no longer available. Another problem from the CLARIN perspective is the variation in the existing license agreements, which makes it hard to offer a centralized service. To tackle these problems, several sets of agreements have to

be used. For an outline of the resource classification procedure, see Figure 1.

Regarding the license variation, we carried out an extensive survey and found that it is possible to categorize the licenses into three different groups:

- Publicly Available Resources
- Resources for Academic Use
- Resources for Restricted Use

We followed the model used by Creative Commons and created simple icons, i.e. care symbols, making it easier for the end-user to immediately see under which conditions the resource can be used, see Figure 2. In addition, a deed describes the rights in human readable textual form. Finally, there is also the actual license agreement and the metadata, i.e. the machine readable information. However, Creative Commons is not sufficient as such for CLARIN, because Creative Commons does not allow for distribution restricted to academia or even more limited groups of users, which is essential for many of the older resources to be included in CLARIN.

Publicly Available (PUB) is one of the categories endorsed by CLARIN. To belong to this group, the following requirements have to be met:

- the licence should allow distribution of the tools and resources from the CLARIN infrastructure,
- there must be no limitations, e.g. based on status or geographical location etc., on who can access and use the tools and resources and
- there must be no limitations on the purpose for which the tools and the resources are used.

For Academic Use (ACA) the licence agreement includes an additional requirement that the use is somehow related to an academic institution. Here the problem may arise from the definition of academic use. To qualify under this category, the tools and resources:

- should be available at least for anyone doing research or studying in an academic

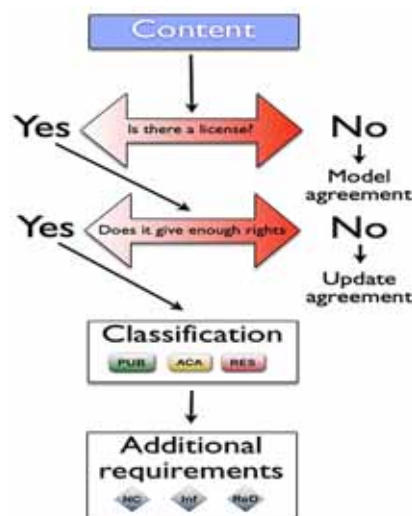


Figure 1: Resource classification task



# Laundry Symbols and Licence Management

## Practical Considerations for the Distribution of LRs based on experiences from CLARIN



Figure 2: Symbols for the main distribution classes

institution recognized by the Identity Provider Federation and

- should be available for studying, research and teaching purposes.

The last category, Restricted Use (RES) includes the resources that do not fulfill the previous requirements but still could be offered to the users if certain additional requirements are met. The most typical reasons for a resource to fall under the scope of RES are:

- a requirement to submit detailed information, e.g. an abstract, on the planned usage or
- specific ethical or data protection-related additional requirements.

In conjunction with the main license categories PUB, ACA and RES, there can also be all or any of three additional requirements:

- A requirement for strictly non-commercial use (NC)
- A requirement to inform the copyright holder regarding the usage of the tools

and/or the resources in published articles (Inf)

- A requirement to re-deposit modified versions of the tools and resources with the Service Provider (ReD)

Figure 3 displays the symbols designed for the additional requirements.

However, this does not solve all the problems. In some cases there either is no licence agreement at all, because such an agreement has never been made. It is also quite common that the existing agreement is somehow problematic, e.g. very low in details, making the categorization impossible. For those situations we created the CLARIN Update Model Agreements with the purpose to procure the required rights. The best option is to re-license the content with the CC0-licence. See the Berlin Declaration (2003) for best scientific licensing practices. It is well-understood and offers enough rights for all parties in different digital and non-digital environments. It is also compatible with most of the other open content licences. Unfortunately it is not always possible to use CC0 due to the demands of the copyright holders. Thus Update Model Agreements for Academic and Non-Commercial Use are also available.



Figure 3: Symbols for additional distribution restrictions

CLARIN LRT inventory. We received an answer for 40 of the resources, i.e. 34.5%. In Figure 4, we see that more than half of the resources were classified into the (RES) restricted category. Approximately one third were classified as (PUB) publicly or (ACA) academically available. Finally, one sixth was found to be exceptional.

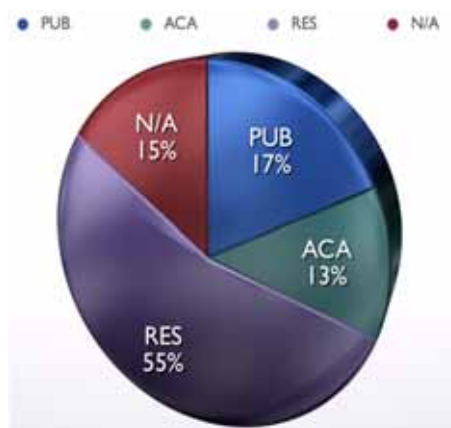


Figure 4: Main distribution categories.

One of the publicly available resources (PUB) was also classified as non-commercial, whereas all of the academically available resources (ACA) were non-commercial. The restricted resources (RES) were roughly equally divided among no additional restrictions (27.3 %), a non-commercial restriction (31.8 %) and a requirement that the license be personally granted by the content owner (40.9 %).

Only one resource was such that the content provider found no applicable distribution type because there was no formal agreement between the content owner and the content provider. In addition, there were 5 resources for which the research project was still ongoing and the question of distribution would be discussed only after the project had finished. All in all, some kind of classification was received for a total of 40 resources in this initial survey. For additional information, see the Oksanen & al (2010). **C**

### References

- Berlin Declaration. (2003). Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Berlin. <http://www.zim.mpg.de/open-access-berlin/berlindeclaration.html>.
- Hietanen, H., Oksanen, V., and Välimäki, M. (2007). Community created content. Law, business and policy.
- Oksanen, Lindén, Westerlund (2010). Laundry Symbols and License Management – Practical Considerations for the Distribution of LRs based on experiences from CLARIN. In Proceedings of LREC 2010. Malta.

# Language Resources, Speech Technology and Language Learning: How to Establish a Virtuous Circles

*State of the art in speech technologies as a part of CLARIN infrastructure*



**Catia Cucchiarini**

*Nederlandse Taalunie, The Hague*

The interest in applying speech technology and in particular Automatic Speech Recognition (ASR) technology to second language (L2) learning has been growing considerably. The addition of ASR technology to Computer Assisted Language Learning (CALL) systems makes it possible to assess oral skills in a second language and to provide corrective feedback automatically. The latter feature appears particularly appealing, because providing practice and feedback on speaking proficiency is extremely time-consuming, with the consequence that the necessary amount of practice is almost never achieved in traditional teacher-fronted lessons. Against this background, ASR-based CALL systems would seem to make for an interesting supplement to traditional L2 classes.

An additional advantage of ASR-based CALL systems that often goes unmentioned is that such systems can also be employed for language resource acquisition and for conducting innovative research on language learning. In this contribution we explain how.

Developing ASR-based CALL systems that can provide accurate and useful feedback on oral proficiency is not trivial, because the speech of non-natives poses special difficulties to ASR technology. In addition, existing systems usually fail to provide corrective feedback that is detailed enough and accurate, especially on L2 pronunciation, which is considered a particularly challenging skill, both for L2 learners and CALL systems. The key factor in developing such systems consists in circumventing the limitations of the technology while taking account of its advantages and of pedagogical criteria (Neri et al. 2009). Recent research shows that this is possible (Cucchiarini, et al. 2009).

Another problem that has hampered the realization of ASR-based CALL systems, especially for the smaller languages, is that

although publishers are willing to use the technology once it is developed, in general they do not have the means to finance the development of such technology and the underlying language resources required. In this connection it may be interesting to refer to an initiative in the Netherlands and Flanders aimed at supporting research and development in the field of language resources and language and speech technology, which, among other things, is likely to have positive consequences for the field of CALL and language learning in general.

In 2004 the Dutch-Flemish programme STEVIN (a Dutch acronym that stands for Essential Language Resources in Dutch) was launched with the aim of stimulating the development of basic language and speech resources and technology for the Dutch language (see: <http://taalunieversum.org/taal/technologie/tevin/english/>). The programme is funded by the Flemish and Dutch governments and is coordinated by the Dutch-Flemish organization for language policy, the Dutch Language Union (<http://taalunieversum.org/taalunie>).

Within the framework of the STEVIN programme various resources have been or are being realized which will be beneficial for the development of ASR-based CALL applications, and, more generally, for research on language learning, such as an ASR toolkit, a corpus of non-native speech and a prototype of an ASR-based CALL system. These resources are briefly discussed below.

## **The SPRAAK speech recognizer**

The STEVIN project SPRAAK extended the work on the ASR package that had been developed for over 15 years by ESAT at the University of Leuven to a) develop a highly modular toolkit for research into speech recognition algorithms and b) provide a state-of-the-art recogniser for Dutch with a simple interface that could be used by non-specialists. (Demuyne et al., 2008). SPRAAK is distributed as open source for academic usage and at moderate cost for

commercial exploitation (for further details, see <http://www.spraak.org/>).

## **The JASMIN corpus of Dutch non-native speech**

The STEVIN project JASMIN (Cucchiarini et al. 2008) produced a corpus of speech by children of different age groups, elderly people and non-natives with different mother tongues (<http://www.esat.kuleuven.be/psi/spraak/projects/JASMIN>), which was collected in the Netherlands and Flanders and is specifically aimed at facilitating the development of speech-based applications for children, non-natives and elderly people. In the case of non-native speakers the applications envisaged were especially language learning applications because there is considerable demand for CALL products that can help making Dutch L2 teaching more efficient.

## **The DISCO prototype of an ASR-based CALL application**

The STEVIN project DISCO (Development and Integration of Speech technology into COurseware for language learning) aims at developing a prototype of an ASR-based CALL system for practicing oral skills in Dutch (Strik et al, 2009; for further details, see: <http://lands.let.ru.nl/~striik/research/DISCO/>). This project makes use of the SPRAAK toolkit and the JASMIN non-native speech corpus. DISCO addresses different aspects of speaking proficiency, detects errors in speaking performance, points them out to the learners and give them the opportunity to try again until they manage to produce the correct form.

One of the interesting features of the STEVIN projects is that the resources and the technology developed are made publicly available to interested users (researchers, HLT companies and publishers) through the Dutch-Flemish HLT Agency (<http://www.inl.nl/tst-centrale>), a central repository for Dutch digital language resources set up and financed by the Dutch





Figure 1: Example of a DISCO pronunciation exercise with feedback

Language Union and hosted by the Institute for Dutch Lexicology in Leiden and Antwerp. These resources can be employed for developing applications and for conducting research.

### ASR-based CALL and language learning research

Once an ASR-based CALL system has been developed, language learners can use it to practice oral skills. The system can thus be used for acquiring additional non-native speech data, for extending already existing corpora like JASMIN, or for creating new ones.

In addition, the CALL system can be designed and developed in such a way that it is possible to log details regarding the interactions with the users. The logbook can contain information on what appeared on the screen, how long the user waited, how (s)he responded, the feedback provided by the system, and how the user reacted to this feedback.

The additional corpus of non-native speech and the log-files can be useful for research on language learning and for developing new, improved CALL systems.

In particular, ASR-based CALL systems allow to create research conditions that were hitherto impossible, thus opening up oppor-

tunities for innovative research on language learning.

For instance, at the moment a project is being carried out at the Radboud University of Nijmegen, which is aimed at studying the impact of corrective feedback on the acquisition of syntax in oral proficiency (<http://lands.let.kun.nl/~cstriik/research/FASOP>).

Within this project an adapted version of the DISCO ASR-based CALL system is used to study how corrective feedback on oral skills is processed on-line, whether it leads to uptake in the short term and to actual acquisition in the long term. This has several advantages compared to previous studies: the learner's oral production can be assessed on-line, near-optimal corrective feedback (clear, intensive, systematic, adapted to the learner's developmental stage and learning style) can be provided immediately, all interactions between learner and system can be logged so that data on input, output and feedback are readily available to be studied from different angles.

The approach chosen in this project indicates how CALL and language and speech technology can offer new opportunities for language learning research (Ellis and Bogart, 2007). In turn, new insights in language learning can inform the development of bet-

ter CALL systems and language and speech technology. In this way a virtuous circle can be established in which these fields profit from each other and lead to increased knowledge and better language learning systems and this systems could be an integral part of CLARIN infrastructure. **C**

### References

- Cucchiari, C., Driesen, J., Van Hamme, H. and Sanders, E., Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus, in Proceedings of LREC, 2008.
- Cucchiari, C., Neri, A. and Strik, H., (2009) Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback, *Speech Communication*, 51, 853-863.
- Demuyne, K., Roelens, J., Compernelle, D. van, and Wambacq, P. (2008). SPRAAK: an open source Speech Recognition and Automatic Annotation Kit, In Proceedings of Interspeech 2008.
- Ellis, N.C., Bogart, P.S.H., (2007), *Speech and Language Technology in Education: the perspective from SLA research and practice*, Proceedings ISCA ITRW SLaTE, Farmington PA.
- Neri, A., Cucchiari, C., Strik, H. and Boves, L.W.J. (2009). The pedagogy-technology interface in computer-assisted pronunciation training. In P. Hubbard (Ed.), *Computer Assisted Language Learning: Critical Concepts in Linguistics, Volumes I-IV* (pp. 140-164). London/New York: Routledge.
- Strik, H., Cornillie, F., Colpaert, J., Doremalen, J.J.H.C. van and Cucchiari, C. (2009). Developing a CALL System for Practicing Oral Proficiency: How to Design for Speech Technology, Pedagogy and Learners. In Proceedings of SLATE (pp. CD-rom). Warwickshire, U.K.. Available from: <http://www.eee.bham.ac.uk/SLaTE2009/papers/SLaTE2009-32.pdf> [05-09-2009].

# The LREC 2010 Map of Language Resources and Tools

LREC2010, Valetta, Malta, May 17-23, 2010



**Nicoletta Calzolari**  
**Claudia Soria**  
**Riccardo Del Gratta**  
*Istituto di Linguistica  
Computazionale "Antonio  
Zampolli", CNRS, Pisa*

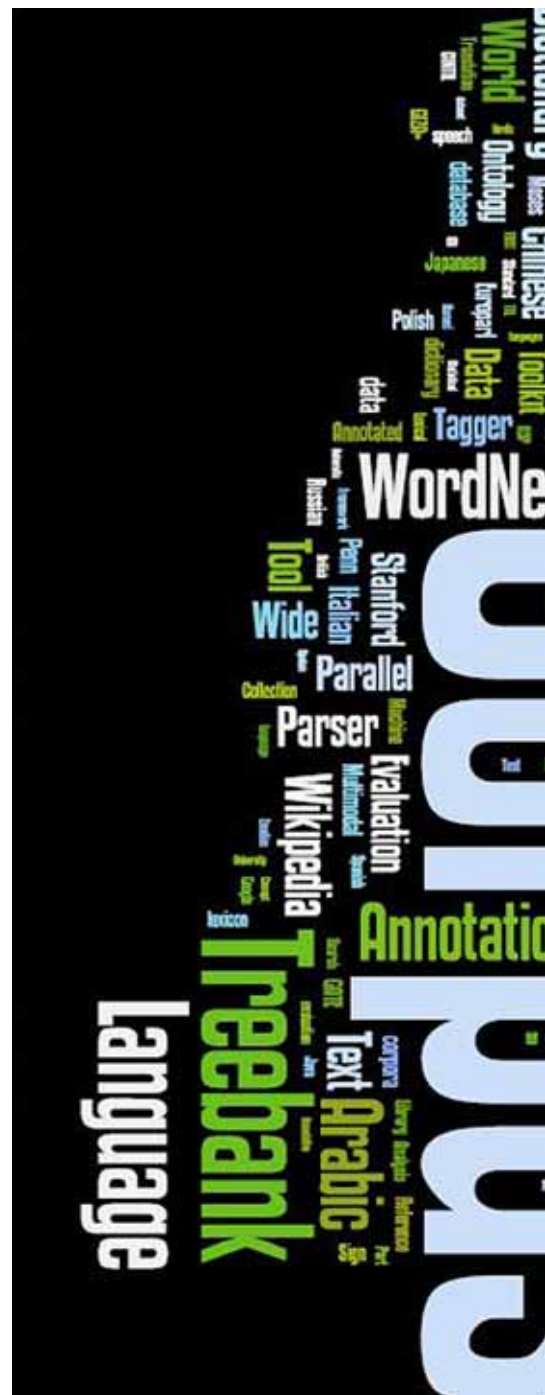
This year, the LREC Conference introduced a very special feature, the LREC Map of Language Resources and Tools, jointly sponsored by FLaReNet and ELRA. The purpose of the Map is to shed light on the vast amount of resources that represent the background of the research presented at LREC, in the attempt to fill in gaps in community knowledge about the resources that are used or created worldwide.

## Basic information about resources in one place

In a nutshell, the Map can be described as a companion to LREC providing basic information about all the resources (in a broad sense, i.e. also tools, standards, evaluation packages, etc.) – either used or created by LREC authors and described in their papers.

Knowledge about existing resources is essential to the overall advancement of research in the field of Language Resources and Technology: it is important to be able to locate and retrieve the right resources for the right applications, and to exploit existing ones before building new ones from scratch. Having a clear picture of which resources are available for which languages and for which use is important in order to identify existing gaps for certain languages at a given time and estimate the amount of investment needed to fill them in. Knowledge about the current use of resources is equally important. Knowing which resources are most used for the various applications will help to better understand the reason behind their success (their intrinsic quality, their wide availability, their licensing model, etc.). Knowing which standards are used in resource representation would help improve the development of standards themselves, by getting them more tuned to actual needs and requirements.

Clear and easy-to-reach information of this type about resources and related technologies is lacking. At the same time, it is very important to stress that most resources are very poorly documented, or not documented at all, thus hindering their accessibility and in the end, their full deployment.



## LREC2010: The Inside Story



**Mike Rosner**  
*University of Malta*

So LREC 2010 is over. And yes, it seems to have run smoothly, to judge from the reactions of some of the 1200 participants. Great! We managed to fool the lot of them. Appearances can be deceptive. Read below for a blow-by-blow account of some of the potential disasters.

The story begins in October 2009 when a delegation of veteran LRECians from Paris and Pisa visited to reconvince themselves that Malta really was an appropriate venue for LREC 2010. We almost blew it on the first

evening. To create a good impression, I had chosen a well-known restaurant for dinner on the Tower Road promenade. Everything was fine until dessert time when one of the Pisa team turned white very fast and had to be escorted from the table supported by two colleagues. Next day, a second member of the Pisa team was looking shaky but stoically maintained a brave but quiet face during our first meeting with the enthusiastic Rector of the University. Was this food poisoning? Was this the delayed effect of some northern Italian bacteria? We shall never know for sure. What we can safely report is that the meal made an impression on all - but perhaps not quite the expected one; and Malta remained the choice.

LREC is dominated by two events: a stand-up welcome reception, usually held on the first



evening, and a sit-down banquet, held on the last. Both present considerable problems for the local organizer, since venues for 1200 people do not exactly grow on trees (particularly in treeless Malta), and also, in May, contrary to popular opinion, the weather is very unpredictable so purely outdoor locations are out of the question.

We decided to adopt an indoor solution for the banquet. But the reception venue remained undecided until the second visit of the Paris/Pisa team in March. On this occasion, accompanied by the Rector, we went to see the President of the Republic in his magnificent Palace in Valletta. The mission had three goals to achieve: to ask for LREC to be held under his Excellency's Patronage, to ask for his Excellency to open the conference, and also to solve the problem of the reception







# CICLing Comes to Europe

CICLing2010, Iași,  
March 21-27, 2010



**Alexander Gelbukh**  
*Centro de Investigación en  
Computación, Instituto  
Politécnico Nacional, Mexico*



**Corina Forăscu**  
*University Alexandru Ioan  
Cuza, Iași*

From 21<sup>st</sup> to 27<sup>th</sup> March 2010, the 11<sup>th</sup> CICLing, International Conference on Intelligent Text Processing and Computational Linguistics (<http://profs.info.uaic.ro/~cicling2010> and <http://www.cicling.org/2010>) took place at the Alexandru Ioan Cuza University of Iași. The conference was jointly organized by the Faculty of Computer Science of the Alexandru Ioan Cuza University of Iași (UAIC), Romania, and the Natural Language Processing Laboratory of the CIC (Center for Computing Research) of the IPN (National Polytechnic Institute), Mexico. Romania was chosen for many reasons, from its touristic attractions to the host institution's experience in organizing international events. Most importantly, however, with this event CICLing paid tribute to a nation that gave the world many of the greatest wonderful computational linguists, of which some have been Program Committee members

or keynote speakers at past CICLing events. Yet another reason for holding CICLing in Iași was the sesquicentennial jubilee of the Alexandru Ioan Cuza University of Iași, the oldest higher education institution in Romania.

## European edition

Held for the first time in Europe, this year's CICLing event received the record high number of submissions in its 11-year history: 271 submitted papers from 47 countries, the previous record being 232 papers received for 2008 event in Haifa, Israel. The 61 accepted oral papers and the three invited papers were published in *Computational Linguistics and Intelligent Text Processing*, a special issue of Springer's Lecture Notes in Computer Science. In addition, two other journals dedicated special issues to this CICLing event: 27 papers were published in *Natural Language Processing and its Applications*, volume 46 of *Research in Computing Science*, IPN, México, and 18 papers in the first inaugural issue of *International Journal of Computational Linguistics and Applications*, Bahri Publications, India.

Almost 150 participants, including registered participants, organizers, students, and the great student volunteers, enjoyed the invited lectures delivered by Shuly Wintner of the University of Haifa, Israel, who presented the keynote talk *Computational Models of Language Acquisition*, and James Pustejovsky of Brandeis University, USA,

who called his keynote talk *The Recognition and Interpretation of Motion in Language*. In addition, each keynote speaker organized very interesting "special events" – this is a distinctive feature of CICLing conferences. In this year both special events took the form of a discussion on *Computational Linguistics vs. Natural Language Engineering*. Gheorghe Grigoraș, the Dean of the Faculty of Computer Science, and Dan Cristea of the Al. I. Cuza University of Iași, who kindly accepted to be the Honorary Local Chair of CICLing 2010 and delivered a very appreciated presentation about dictionaries, addressed the participants warm welcome speeches at the opening ceremony. A very dear friend of CICLing, Nancy Ide of the Vassar College, USA, who was with CICLing from its early days and is a past CICLing keynote speaker, kindly accepted to address some words to the CICLingers right after the ceremony organized by the Al. I. Cuza University where she was awarded the title of Honorary Professor of the Al. I. Cuza University of Iași.

## Best paper awards and live internet coverage

The scientific program included, apart from the keynote lectures and special events, oral presentations of the CICLing authors. The CICLing sessions were organized under the following sections: Lexical Resources, Syntax and Parsing, Word Sense Disambiguation and Named Entity Recognition, Semantics and Dialog, Humor and Emotions, Machine Translation and Multilingualism, Information Extraction, Information Retrieval, Plagiarism Detection, Text Summarization, Speech Generation. All scientific activities took place in the Mihai Eminescu Aula Magna of the Al. I. Cuza University of Iași (see the picture to the left). The poster session, which started with a brief oral presentation of all posters, was combined with the welcome reception and was held in the Hall of the Lost Footsteps of the University, see the other picture.

Based on the Program Committee votes, three best papers awards were offered:

- First place: George Tsatsaronis, Iraklis Varlamis, and Kjetil Nørvåg, *An Experimental Study on Unsupervised Graph-Based Word Sense Disambiguation*; according to a ballot among all attendees, these



authors also won the best presentation award.

- Second place: Paolo Annesi and Roberto Basili, *Cross-lingual Alignment of FrameNet Annotations through Hidden Markov Models*;
- Third place: Lieve Macken and Walter Daelemans, *A Chunk-driven Bootstrapping Approach to Extracting Translation Patterns*.

Another special award was offered for the best student paper: *Integer Linear Programming for Dutch Sentence Compression*, co-authored by Jan De Belder and Marie-Francine Moens.

For the first time in the history of CICLEing and also in the history of the University Aula, all the CICLEing 2010 scientific activities were transmitted live over the internet;

er, with only small rains, and only during the days of scientific program. This was good also for the cultural program of CICLEing 2010—from the very first CICLEing event its famous touristic program is one of its most attractive features. This year CICLEingers participated in three tours to the most attractive touristic areas of Romania. The first tour, on Sunday, March 21, was dedicated especially to nature and history of the Neamț county. Wednesday, March 24, was dedicated to the myths: we visited the old city centre of Brașov, with the Black Church, and after a tour at the Peleş Castle in Sinaia we had an excellent lunch with Romanian cuisine; those CICLEingers who were very keen at the idea and still had energy, went also to the Bran Castle, famous for its connection with the legend of Count

Association of Students in Informatics in Iași (ASII).

### Abundant local support

We also greatly appreciate the support of the Romanian Academy. In particular, Dan Tufiş was closely involved in the organization of PROMISE conference (*Processing Romanian in Multilingual, Interoperational and Scalable Environments*), a satellite event of CICLEing 2010 co-organised under the aegis of the Romanian Academy together with the Al. I. Cuza University of Iași. Very special thanks go to Rada Mihalcea, one of the greatest and oldest friends of CICLEing, who was, like many times before, guiding and helping the organizing team in many



Group photo of CICLEing participants

the recordings of all the sessions are available for download at <http://profs.info.uaic.ro/~cicling2010/historyCic.html>. Other pictures and videos are available at <http://picasaweb.google.com/cicling2010.iasi>; all participants willing to share their memories can contribute their files to this page.

Since it was the first CICLEing held in Europe, the usual dates of CICLEing, which used to be in mid-April, had to be changed taking into account other important scientific events as well as the local Romanian program and, of course, the Romanian Spring weather. Luckily, even though the usual weather in the Spring is a rainy one, during CICLEing we had a very nice weath-

Dracula. The last trip, on Saturday, March 27, was dedicated to the Romanian painted monasteries in the Bucovina county.

From all the feedback received from the CICLEingers during and after the conference, we can consider that the conference was a success, especially taking into account the world economic crisis. This success is in a great degree due to active participation and constant support of many people. First of all we want to thank all people involved from the Alexandru Ioan Cuza University, starting from the leaders of the University, passing through all the University departments and ending with the wonderful student volunteers – most of them from the

activities, behind the scene but always present.

Finally, after many months of close collaboration from the opposite sides of the Atlantic – and now separated by it again – we would say a lot of very warm words about each other if we were not both the authors of these lines – it's a pity we have to omit here this part of our emotions and gratitude...

Some months after CICLEing 2010 we are happy we survived all the arduous work, and we are looking forward for the next CICLEing in 2011, and for the courageous local organizers willing to contribute to yet another successful CICLEing event. **C**



# The First Joint Meeting of the CLARIN Boards

Utrecht, March 3-4, 2010



**Steven Krauwer**  
CLARIN coordinator

As many of you may know, CLARIN has, in addition to the Executive Board, three other important Boards. The *Scientific Board* consists of high-level scientists, one from each participating country and appointed by the national funding agency. It will monitor the execution of the programme and ensure the overall sci-

collaboration, coordination and harmonization with initiatives at the international level. On <http://www.clarin.eu/external/index.php?page=about-clarin&sub=4> one can find a table that presents the composition of the three boards.

## All-boards meeting

On March 3rd this year the three Boards met for the first time with the Executive Board in order to discuss the project results and deliverables of the first two years of the project. On March 4th there was a meeting between the Strategic Coordination Board and the Executive Board, with representatives from the Dutch Ministry of Science, where the focus was on the future structure

outside the field of linguistics. Even if most of the activities related to this specific issue had to be deleted from our original work plan at the request of the EC it was generally felt that this was one of the most pressing problems that should be urgently addressed in the following phases of CLARIN, both at the European and at the national level. Collaboration with existing and emerging infrastructure activities within and outside Europe could be extremely beneficial, and the creation of CHAIN (Coalition of Humanities and Arts Infrastructures and Networks, <http://www.chaincoalition.org/>) was seen as an important step towards this.

## Heading for ERIC framework

On March 4th the focus was on the future governance and financing of CLARIN. The current vision to create an ERIC for CLARIN was presented for the Strategic Coordination Board. An important feature of the ERIC is that it is a consortium of governments (as opposed to universities, digital archives or research organizations). The role of the present project can only be to make the proper preparations for this, so that the governments can take over to conduct the final negotiations. At the meeting the current plans for the organisation and the implementation of the ERIC were presented and discussed. The presence of representatives from ministries or funding agencies from most CLARIN countries made it an excellent opportunity for an open discussion of possible obstacles. Representatives of the Dutch ministry of research confirmed their intention to take the lead in this process and to act as the host country for CLARIN. The participants expressed their agreement with the general direction taken and their willingness to continue the discussions. One of the important conclusions from this meeting was that the original time schedule whereby the ERIC would become operational on Jan 1st 2011 was recognized as unrealistic, as most countries needed more time to finalize their national roadmaps and to start the funding allocation process. It was agreed that the project would apply for an unfunded extension till July 1st 2011, to avoid a gap between the end of the preparatory phase and the construction phase of CLARIN.

The meeting on March 3rd and 4th was a very important one for CLARIN, because this was the first opportunity for a direct exchange of views and ideas between CLARIN and its three Boards. The fact that there was broad support for the directions the project is taking (along its various dimensions) was very encouraging and makes us feel confident that we will succeed in making CLARIN happen! **C**



A picturesque building of the Utrecht University where boards meeting took place

entific soundness, coherence, completeness, consistency and feasibility. It determines the overall scientific strategy and reviews the project deliverables. The *Strategic Coordination Board* consists of representatives appointed by the funding agencies, one per country. It will monitor the execution of the programme of work with a view to compliance with national governments' and funding agencies' policies. It determines the overall governance and financial strategy. The Strategic Coordination Board will review all project deliverables.

The *International Advisory Board* consists of high-ranking international experts, and will give advice to the Executive, Scientific and Strategic Coordination Boards on issues of common interest, such as opportunities for

of CLARIN after completion of the Preparatory Phase project.

The various components of the project were presented by the Work Package leaders, and discussed with the participants. On the whole the members of the Boards expressed their satisfaction with the results presented and were happy to endorse our deliverables. It was encouraging to hear that some of the members of our international advisory board told us that they were impressed by our achievements, especially in comparison with what was happening outside Europe.

The discussion on the first day brought up many different topics, both at the global and at the detail level, and we collected a large number of comments and suggestions. A recurring topic was, as could be expected, how to reach the users, especially users from



# Establishing Portuguese CLARIN

Lisbon, March 19, 2010



**António Branco**

University of Lisbon



**Marko Tadić**

University of Zagreb

The CLARIN Lisbon Meeting on *Common Language Resources and Technology Infrastructure for the Humanities and Social Sciences* took place in March 19, 2010. It was requested by the Portuguese funding agency FCT-Fundação para a Ciência e Tecnologia, from the MCTES-Ministério da Ciência, Tecnologia e Ensino Superior (Portuguese Ministry for Science, Technology and Higher Education) in order to establish the Portuguese CLARIN community. Also, the ultimate goal of this meeting was to support the national decision process on the joining of the CLARIN infrastructure.

The meeting was divided in four sessions where each was conducted by a chair person and a rapporteur.

## Session A

This introductory session started with CLARIN coordinator Steven Krauwer and his overall presentation of the CLARIN project, its goals, achievements, challenges and present state of CLARIN community that is shaping up at the European as well as at the different national levels. Krauwer's presentation was followed by three talks from three different partners presenting their national situations.

The first one was given by Adam Przepiórkowski (Polish Academy of Sciences, Warsaw) and it was a report from a national network with a status of language resources and technology development and maturity quite similar to the one of Portugal. The second talk was given by Núria Bel (University Pompeu Fabra, Barcelona) where she presented a report from a national network with a major language from the neighbouring country, whose global projection matters. The third one was given by Erhard Hinrichs (University of Tübingen) and he reported on a national network that is a major driving force of the infrastructure, with a large companion national funding programme of support. All three talks presented different approaches and sources of inspiration to organize a CLARIN community at the national level thus giving examples to the Portuguese CLARIN initiative.

## Session B

The main purpose of Session B was to plot the situation concerning the Portuguese language and its position in the CLARIN project. The speakers focused on the Portuguese participation in the CLARIN project and the formation of the national CLARIN network, the Portuguese and Brazilian NLP and Linguistics community working on Portuguese language, and the objective of fostering collaboration between Portuguese and Brazilian research centres under CLARIN (see also contribution in this issue of CLARIN Newsletter at page 18). This session was closed with the presentation of the view of the Portuguese Funding Agency on the ESFRI projects and on the involvement of Portugal in the forthcoming development of the CLARIN framework.

The introductory talk in session B was given by António Branco (University of Lisbon), coordinator of the CLARIN-Portugal network titled *The CLARIN project in Portugal*. He was followed by Vera Lúcia Strube de Lima (FACIN-PUCRS, Porto Alegre, Brazil) with the talk *Research in NLP in Brazil: an overview*. The closing contribution in this section was given by Isabel Figueiredo (on behalf of Lígia Amâncio, FCT Vice-President and Portuguese delegate at ESFRI) under the title *The Portuguese participation in ESFRI projects*.

## Session C

This session had two main goals: description of examples of use cases and presentation and discussion of technical problems and possible solutions.

The description of use cases was provided in the following presentations:

1. *Language databases in the psychologist's toolbox*, by São Luís Castro (University of Porto);
2. *Research in History meets LT: the case of Portuguese Royal Inquiries of 1258*, by Luís Gomes (University of Azores);
3. *Multilinguality: global access to information, entrepreneurship and research: a global Portuguese perspective*, by José Gabriel Lopes (New University of Lisbon.)

The discussion on technical problems was initiated by Daan Broeder (Max-Planck Institute for Psycholinguistics) who made a detailed analysis of the main issues involved:

- An ecology of "e-infrastructures", from network services to community services
- Technical, organizational and juridical issues
- What is needed (in terms of ICT infrastructure) to allow researchers to find, process, and store language resources?

## Session D

The last session started with *Overview of Contributions from 16 CLARIN-PT members on their own RUs*, by Maria Francisca Xavier (Centro de Linguística da Universidade Nova de Lisboa) and was followed by a round table. The discussion in the round table was organized around the following questions the Research Units (RUs) had been previously given, and their position papers triggered by such questions:

- the kind of research that should be supported by CLARIN;
- the expected role of CLARIN in preserving knowledge and scientific results;
- the resources and services that CLARIN should make available;
- the expected impact of CLARIN on research and on the support of Portuguese language;
- the contributions to CLARIN that the RU will be ready to make available.

The CLARIN Lisbon meeting finished with a fruitful discussion around these key issues thus undoubtedly leading to an expected Portuguese participation in ERIC. This meeting yielded three deliverables, available for distribution: *Memorandum with the institutional profiles of the Portuguese CLARIN members* (24 pp.); *Memorandum with the position papers of the Portuguese CLARIN members for the round table* (23 pp.); *The Proceedings of the meeting* (128 pp.). **C**



Participants of Portuguese CLARIN kick-off meeting

## Following the Previous Success LT Days, Luxembourg March 22-23, 2010



**Marko Tadić**  
Editor

Following the event successfully organised in January 2009 (see CLARIN Newsletter No 4 for full report and introductory details on pages 1 and 8) where Language Resources and Technologies were strongly supported by the programmes that EC grants through FP7 ICT and CIP ICT-PSP calls, this year the same formula was applied.

### New calls

LT Days 2010 aimed at informing potential proposers about announced ICT-PSP 4th Call, Theme 6: *Multilingual Web* and the ICT-FP7 Work Programme 2010-2011. The details of each programme and call were presented by EC officials in Sessions



3 and 4 preceded by contextualisation of this calls within existing projects/initiatives in Session 1: "Different initiatives, different target communities, different business models... one common effort" and Session 2: "Enterprise Language Processing: mega-trends and developments" where high-level speakers gave presentations covering varied related topics, including newly started projects, the commercial viewpoint and future prospect.

In this introductory session Steven Krauwer successfully presented CLARIN and its achievements so far, particularly when it comes to sharing LRT. Common metadata schemes that would allow different projects and initiatives to coordinate and cooperate in order to avoid the doubling of efforts, were listed as one of the top priorities. **C**

## CLARIN and FLaReNet workshop, Stockholm November 25-26, 2009

**Rolf Carlson**  
**Kjell Elenius**  
**David House**  
KTH, Stockholm

In November 2009, CLARIN and FLaReNet organized a two day workshop at KTH in Stockholm. The theme was "Best practices for speech and multimodal databases". The participants were specially invited to cover different aspects of infrastructure for spoken and multimodal databases. The workshop web page can be found on <http://www.speech.kth.se/clarin/> with links to all presentations. The following is a short report from this meeting including a few snapshots from the subjects that were presented and discussed during these intensive two days.

The background for the workshop was to discuss research needs and share ideas and views among spoken language researchers not yet heavily involved in the CLARIN mission. It is obvious that current advances and rapid development of speech-based interfaces particularly have to a large degree benefited from successful collection and collaborative use of large-scale speech and multimodal databases.

The spoken and multimodal research community is large and varied, and there is a great need to define the requirements for infrastructures which can be fruitfully and easily shared. The annotation of interesting corpora in this field is, however, divergent and would benefit from finding ways for harmonization and interoperability. Thus, it is of importance to look into standards and best practices to encode various relevant features of these corpora to facilitate their use for a wide diversity of researchers. The features may be intonation, facial expressions, gestures, turn-taking and emotions.

When we look at speech and multimodal research topics in the humanities and social sciences such as child-directed speech, accent shifts, dialects and sociolects and dialogues in different environments, we do not have a strong tradition

for collection and distribution of relevant databases. Currently it is very much the researchers' own task to collect the data or try to do research on a corpus collected for some other purpose. To be able to easily locate, access and share data oriented to the humanities and social sciences would be a big step for many researchers. Thus an important goal towards increasing our understanding of human communication and making our applications more intelligent is to make language resources and technology available and readily usable for all kinds of scholars. Common standards are essential for achieving this.

### CLARIN framework

The first session, chaired by Erhard Hinrichs, was opened by Steven Krauwer explaining the CLARIN framework and setting up the challenges we are facing to include or engage the spoken language and multimodal community. While it is so far the case that most CLARIN players originate from text processing, it is acknowledged that speech and multimodal resources are also important, for e.g. phoneticians, linguists, historians of the future and social scientists. Despite this, speech and multimodal actors are underrepresented in the CLARIN activities. Steven pointed out that sharing speech and multimodal resources pays off because of the high cost (in time & money) of annotation, and that consultation with the community, especially about standards is urgently needed. This is in the common interest of both CLARIN and FLaReNet.

Several presentations were focused on current work on data collection, annotation and analyses. The aim of Peter Wittenburg's contribution was to review some of the multimodal annotations in use and point to the heterogeneity of many of the encoding schemes even when based on a harmonized format for annotation structures. Anton Batliner's presentation was entitled "Annotations and beyond – what (not) to standardize, and how come that sometimes, it just hap-



# Best practices for speech and multimodal databases

pens” and also detailed some other types of (missing) standards, wrong standards, and lacunae (such as agreed upon performance measures) where we badly need some conventions which, however, possibly should not be pinned down as “new standards”. Concluding the first session was Volker Steinbiss with a talk entitled “Best practices that could help avoiding the mess”. Among several aspects he presented a few observations and best practices that have been helpful to keep the stress of working with large amounts of data at an acceptable level.

The tools for spoken data collection developed by Christoph Draxler have the

development of freeware acoustic-phonetic analysis tools that will allow student or occasional professional users to quickly learn how to do specific measurements, possibly manually, yet also appeal to expert researchers working with corpora, including those who are competent programmers. Such a common platform should aim to build on and integrate, rather than replace, existing software, and to facilitate cross-disciplinary collaborative research. Such a development is also of importance for sharing speech data from small languages discussed in the presentation by Ingunn Amdal. Emotions in spontaneous conversation and in music were

that using these tools can be tedious and restrictive. Rather than prescribe a standard at this time, he suggested that we might benefit more from creating a support group whereby people who annotate data regularly can communicate and share samples, tools, and formats.

## Questions of ethics and privacy

During the course of the workshop, ample time was devoted to general discussion which covered many of the topics reviewed above. Especially lively discussion points concerned ethics and privacy as well as tools and standards. The final discussion was led by Lou Boves and focused on concrete objectives for CLARIN in meeting the challenges of engaging the spoken language community to a much larger extent. Lou summarized the important medium-term topics discussed during the workshop as the following:

- Taxonomy of research objectives and the required degree of granularity in annotations;
- Different resources and different tools for different research objectives;
- Re-purposing of resources;
- How dangerous is easy-to-use software?;
- Better understanding of the interplay of research and infrastructure;

while the long term issues concerned are:

- Legal Issues;
- Data preservation;
- Additional Centres;
- Sustainable business model;
- Requirements of journals and funding agencies to make data available;
- How can we create a research environment in which research on small languages makes similar contributions to science as the NIST competition.

It is our conviction that the workshop resulted in a deeper understanding of research issues and needs within the two research communities dealing with either text or spoken language and also helped to more clearly articulate and define the basic challenges we need to face to make further progress in the future. **C**



KTH campus

potential to become important web based services for efficient and economic corpus work. Data accessibility using web based services were also presented by Anders Eriksson, together with a description of the extensive dialect project SWEDIA. Important aspects of all corpus work are the ethical or privacy issues which was the final subject in the session discussed by Martti Vainio.

Sarah Hawkins' presentation “Tools for work that uses acoustic-phonetic analysis” gave a closer view on the detailed work in phonetics and especially within the S2S network. Of special interest is the devel-

opment of freeware acoustic-phonetic analysis tools that will allow student or occasional professional users to quickly learn how to do specific measurements, possibly manually, yet also appeal to expert researchers working with corpora, including those who are competent programmers. Such a common platform should aim to build on and integrate, rather than replace, existing software, and to facilitate cross-disciplinary collaborative research. Such a development is also of importance for sharing speech data from small languages discussed in the presentation by Ingunn Amdal. Emotions in spontaneous conversation and in music were

## Hot topic: Multimodality

Multimodality is a current hot topic in human interaction research and the need for annotation on multiple levels was a common denominator for the sequence of presentations by Jean Carletta, Kirsten Bergmann, Jonas Beskow, Jens Edlund and David House.

Based on his experience of working with several different tools for manual annotation and also talking to people who use them regularly, Nick Campbell concluded



# Bridging the Portuguese Speaking Margins of the Atlantic with CLARIN

PROPOR, Porto Alegre,  
April 27-30, 2010

**António Branco**

*University of Lisbon*

**Vera Lucia Strube de Lima**

*PUCRS Porto Alegre*

**Thiago Pardo**

*University of São Paulo*

**Steven Krauwer**

*Utrecht University*



Steven Krauwer and António Branco presenting CLARIN

**P**ROPOR is the Portuguese verb 'to propose'. It is also the acronym of the international conference on the computational processing of Portuguese, the sixth language with the largest number of speakers, spread over the American, African and Asian continents.

This conference is held every other year, alternating between Portugal and Brazil and gathering over one hundred participants. It features the full range of initiatives of a mature scientific event of its size, with oral and poster presentations, plenary keynote presentations, demo sessions, tutorials, best dissertation awards, etc. Its selected papers are published in Springer's Lecture Notes in Artificial Intelligence, with a very competitive acceptance rate, scoring 27% in the last edition. Of course, it is also a festive occasion where old acquaintances meet again and new friends are made.



## A relevance to CLARIN

In terms of research community mobilization, the PROPOR conference plays the role of what in other communities is the attribution of some formal associative structure. It is the reference focal point of this research community from both margins of the Atlantic, and from all over the world, whose work is on the language resources and technology for Portuguese.

As its acronym suggests in the Portuguese reading, the PROPOR conference is the place to propose new ideas and lines of action. That was what happened once more in the last edition, which took place in last April 27-30 in Porto Alegre, Brazil (<http://www.inf.pucrs.br/~propor2010>). And this time with special relevance for our CLARIN project.

At the end of the morning of the second day, the conference accommodated a one hour special plenary session devoted to CLARIN. This provided a unique opportunity to present the project and its progress to an audience including both future researchers – starting now as PhD students – and today's major players in the language resources and technologies from Portugal and Brazil.

This special session included three parts. In a first slot, Steven Krauwer presented a broad overview of the project and described its results and current status. Next, António Branco reported on the achievements of the Portuguese CLARIN network and underlined the interest

of enlarging the participation of the Portuguese language in CLARIN with the contribution of the Brazilian colleagues. Finally, there was time to answer questions and clarify issues raised by the audience.

## Initiative for cooperation of CLARIN with Brazilian centres

In the following day, we were invited by the conference organizers to a side meeting with Brazilian colleagues, including the major players and group leaders in Brazil. The goal was to get a deeper understanding of the project and a better grasp of the possibility for Brazil to join CLARIN and the measures needed, both in Brazil and in the project, to eventually achieve that goal.

At the end of the meeting the enthusiasm and confidence among participants was very high about these new chances of cooperation across continents. It is our belief that very soon we will hear about more developments coming out of these seeds of collaboration sown in Porto Alegre.

As Brazil is a South-American country, this would clearly represent a first opportunity and test bed towards making CLARIN a truly international endeavour across continents. Naturally, it brings about also the challenge for CLARIN to find ways to expand its initial parameters and our accrued responsibility of making it an open and a reliable organization for colleagues from all over the world! **C**

# The State of Lithuanian HLT



**Rūta Marcinkevičienė**  
Vytautas Magnus University,  
Kaunas

## Introduction

The last two decades have been important for Lithuania as a number of activities relating to Human Language Technologies (HLT) have been carried out, including:

- localization of general tools,
- digitization (including adaptation of digitized resources),
- compilation of tools, language resources and knowledge bases,
- training and research,
- documentation and dissemination.

The first two types of activities, i.e. localization of the user interface and digitization of cultural heritage, cannot be classified under HLT proper. However, some types of digitized products can be used as linguistic resources, e.g. Database of Old Lithuanian Writings, Dictionary of Lithuanian Language, Dictionary of Contemporary Lithuanian Language, Dictionary of Toponyms, Database of Lithuanian Dialects etc.

However, digitized resources are of limited use as resources, therefore a greater prominence is given to the third type of activity, the compilation of general and special corpora and language processing tools.

## An overview of Lithuanian HLT

The most obvious outcome of the programme for Lithuanian HLT was compilation of the corpus of 100 million running words and some tools (e.g. corpus query system and collocation extraction tool, a system of morphological annotation and disambiguation) available for public use. Later developments in the field, financed mostly by national foundations, resulted in the production of the following tools and resources:

- a morphologically annotated corpus (115 million running words),
- an annotated manually checked corpus of one million words,
- a set of parallel corpora:
  - a bi-directional Czech-Lithuanian and Lithuanian-Czech corpus of five million words
  - English-Lithuanian corpus of 18 million words in size,
- a database of Lithuanian nominal collocations, extracted from the corpus of 100 million words,

- a number of tools such as:
  - a tool for the automatic identification of text functions for the Lithuanian language,
  - the tool for the extraction of collocations,
  - a Lithuanian tagger,
  - the Aligner2067,
  - an automatic accentuation tool for the Lithuanian language,
- a corpus of Spoken Lithuanian language,
- a universal annotated database of speech recordings.

The above list is confined to the tools for language resources made at Vytautas Magnus University and sponsored mainly by two national funding agencies, the Lithuanian State Language Commission and the Lithuanian State Science and Studies Foundation.

Other institutions have developed a set of tools and databases for public use or purchase. The State Commission of the Lithuanian Language is monitoring an open terminological database. The Institute of Mathematics and Informatics has digitized term dictionaries from 27 disciplines into one database. A private company

Before the automatic machine translation system there was an automatized translation tool *Vertimo Vedlys* incorporated in text editor *Tildės biuras*. together with a spellchecker and multi-lingual support software. It translates noun phrases and simple sentences.

## National policies related to research infrastructures

On a national level various research and development programs continue to promote HLT related activities. However, the most important development and support of resources is foreseen in the framework of the National Research Infrastructure (NRI) compatible with ESFRI requirements for national states. The strategy of NRI includes documentation and unification of existing national resources as well as support for trans-national initiatives such as CLARIN, CESSDA and other similar joint infrastructures for the Social Sciences and Humanities (SSH). National support for research infrastructures in general and HLT in particular is timely since “SSH researchers rely on new technologies, and real overhead costs



The building of Vytautas Magnus University in Kaunas

*Fotonija* is known for its electronic dictionaries *Interleksis*,  $\mathbb{T}\square$ ; English-Lithuanian dictionaries *Alkonas* and *Anglonas*, French-Lithuanian dictionary *Frankonas* and a spellchecker *Juodos avys*. A corpus of academic discourse has been started at Vilnius University, Faculty of Philology.

The most recent jointly developed tool was a rule-based machine translation system for the translation of English internet texts into Lithuanian. It was developed by a group of companies including Promt (St. Petersburg), *Fotonija* (Vilnius), Alna Software (Kaunas). They co-operated within the framework of a project financed by EU Structural Funds.

for SSH research have increased dramatically over the past 20 years, without government subsidies necessarily reflecting these changes. Consequently, more and more SSH research depends on capital injections to develop cutting edge data sets and develop retrieval systems” (METRIS report 2009).

It can be concluded that most of Lithuanian HLT related activities, mentioned in the Introduction, are taken care of on national level. Training and research, however, remain the least attended activities. Therefore, it is expected that our partnership in CLARIN consortium will boost the development of both research infrastructures and research itself. **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

**Austria (NCP: Gerhard Budin)** University of Graz (Graz): Austrian German Research Centre (C.Rudolf Muhr); Institut für Romanistik (C.Stefan Schneider)  
Austrian Academy of Sciences (Vienna): Austrian Academy Corpus (C.Christoph Benda); Department of Linguistics and Communication Research (C.Sabine Laaha); Institute of Lexicography of Austrian Dialects and Names (C.Eveline Wandl-Vogt)  
Secure Business Austria (Vienna): (C.Edgar R. Weipp)  
University of Vienna (Vienna): Center for Translation Studies (C.Gerhard Budin)  
**Belgium (NCP: Ineke Schuurman)** University of Antwerp (Antwerp): Center for Dutch Language and Speech (C.Walter Daelemans)  
Vrije Universiteit Brussel (Brussels): Laboratory for Digital Speech and Audio Processing, Department of Electronics and Information Processing (C.Werner Verhelst)  
Gent University (Gent): Digital Speech and Signal Processing research group at the Electronics and Information Systems department (C.Clean-Pierre Martens)  
University College Ghent (Gent): Faculty of Translation Studies, Language and Translation Technology Team (C.Véronique Hoste)  
Katholieke Universiteit Leuven (Leuven): Center for Computational Linguistics (C.Frank Van Eynde); ESAT-PSI/Speech (C.Patrik Wambacq); Language Intelligence & Information Retrieval (C.Marie-Françine Moens)  
Katholieke Universiteit Leuven (Leuven - Kortrijk): iTeC (Interdisciplinary research on Technology, Education & Communication) (C.Hans Paulussen)  
**Bulgaria (NCP: Kiri Simov)** University of Plovdiv (Plovdiv): Faculty of Mathematics and Informatics (C.Veska Noncheva)  
Bulgarian Academy of Sciences (Sofia): Department of Computational Linguistics, Institute for Bulgarian Language (C.Svetla Koeva); Institute for Parallel Processing of Bulgarian Academy of Sciences (Sofia); Linguistic Modelling Department (C.Kiril Simov); Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (Sofia); Mathematical Linguistics Department (C.Ludmila Dimitrova)  
St. Cyril and St. Methodius University (Veliko Turnovo): (C.Boryana Bratanova)  
**Croatia (NCP: Marko Tadić)** Institute of Croatian Language and Linguistics (Zagreb): (C.Damir Cavar)  
University of Zagreb (Zagreb): Department of Linguistics, Faculty of Humanities and Social Sciences (C.Marko Tadić); Zagreb University Computing Center (C.Zoran Bekić)  
**Cyprus (NCP: -)** Cyprus College (Nicosia): Research Center (C.Antonis Theodorou)  
**Czech Republic (NCP: Eva Hajičová)** Masaryk University (Brno): Faculty of Informatics (C.Aleš Horák)  
Charles University (Prague): Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics (C.Eva Hajičová)  
The Institute of the Czech Language, Czech Academy of Sciences (Prague): The Institute of the Czech Language (C.Karel Oliva)  
**Denmark (NCP: Bente Møgaard)** Copenhagen Business School (Copenhagen): Department of International Language Studies and Computational Linguistics (C.Peter Juel Henriksen)  
Dansk Sprogævn - Danish Language Council (Copenhagen): (C.Sabine Kirchmeier-Andersen)  
Society for Danish Language and Literature (Copenhagen): (C.Jørg Asmussen)  
The National Museum of Denmark (Copenhagen): (C.Birgit Rønne)  
The Royal Library (Copenhagen): (C.Anders Conrad)  
University of Copenhagen (Copenhagen): Centre for Language Technology, Faculty of Humanities (C.Bente Møgaard)  
University of Southern Denmark (Kolding): Faculty of Humanities (C.Johannes Wegner)  
**Estonia (NCP: Tiit Roonmaa)** University of Tartu (Tartu): Institute of Computer Science (C.Tiit Roonmaa)  
**Finland (NCP: Kimmo Koskenniemi)** CSC - the Finnish IT Center for Science (Espoo): (C.Piyo-Leena Forsström)  
Lingsoft Inc. (Helsinki): (C.Luhani Reiman)  
The Research Institute for the Languages of Finland (Helsinki): (C.Toni Suutari)  
University of Helsinki (Helsinki): Department of General Linguistics (C.Kimmo Koskenniemi)  
University of Joensuu (Joensuu): Department of Foreign Languages and Translation Studies (C.Jussi Niemi)

University of Oulu (Oulu): Faculty of Humanities, Finnish Language (C.Marketta Harju-Autti)  
University of Tampere (Tampere): Faculty of Information Sciences, Department of Information Studies and Interactive Media (C.Eero Sarmento)  
**France (NCP: Jean-Marie Pierrat)** Centre de ressources pour la documentation de l'oral (Aix-en-Provence) (C.Bernard Bel)  
National Center for Scientific Research (NRS) (Marseille): Laboratoire d'Informatique Fondamentale de Marseille (LIF-CNRS) (C.Michael Zock)  
Centre National de Ressources Textuelles et Lexicales (CNTRL) (Nancy): (C.Bertrand Gaiffe)  
National Center for Scientific Research (CNRS) (Nancy): Analyse et Traitement Informatique de la Langue Française (ALTI) (C.Jean-Marie Pierrat)  
National Center for Scientific Research (CNRS) (Orsay): Institute for Multilingual and Multimedia Information (IMMI-CNRS) (C.Joseph Mariani)  
Evaluations and Language Resources Distribution Agency (ELDA) (Paris): (C.Khalid Chouki)  
National Center for Scientific Research (CNRS) (Paris): Traitement Electronique des Manuscrits et des Archives (TELMADIS) (C.Florence Clavaud)  
Université Paris 4 Sorbonne (Paris): Centre de linguistique théorique et appliquée (CELTA) (C.Andre Włodarczyk)  
Université de Strasbourg (Strasbourg): Equipe de recherche LiLPa (Linguistique, Langues, Parole) (C.Amalia Todinauc)  
National Center for Scientific Research (CNRS) (Vandœuvre les Nancy): L'Institut de l'Information Scientifique et Technique (INIST-CNRS) (C.Fabrice Lecoca)  
University Paris Est/Paris 12 (Vilry Sur Seine): LISSI Laboratory (C.Yacine Amirat)  
**Germany (NCP: Erhard Hinrichs)** University of Augsburg (Augsburg): Philologisch-Historische Fakultät (C.Ulrike Gut)  
Berlin-Brandenburg Academy of Sciences (Berlin): (C.Alexander Geyken)  
Humboldt-University Berlin (Berlin): Institut für deutsche Sprache und Linguistik (C.Anke Lüdeling)  
Technische Universität Darmstadt (Darmstadt): Ubiquitous Knowledge Processing (UKP) Lab (C.Lynda Gurevich)  
TU Darmstadt University (Darmstadt): Institute for German Language and Literature (C.Michael Beilwenger)  
Universität Duisburg-Essen (Essen): Fakultät Geisteswissenschaften / Germanistik / Linguistik (C.Bernhard Schröder)  
University of Frankfurt/Main (Frankfurt/Main): Comparative Linguistics Department (C.Jost Gippert)  
University of Giessen (Giessen): Institut für Germanistik (C.Henning Labin)  
DFG-Project "Language Variation in Northern Germany" (Sprachvariation in Norddeutschland - SiN) (Hamburg): (C.Ingrid Schröder)  
University of Hamburg (Hamburg): Faculty for Language Literature and Media, Arbeitsstelle "Computerphilologie" (C.Cristina Vertan); Fakultät für Geisteswissenschaften, Fachbereich Sprache, Literatur, Medien (C.Angelika Redder); Institute of German Sign Language and Communication of the Deaf (C.Thomas Hanke); SFB 538 Multilingualism (C.Thomas Schmidt)  
University of Heidelberg (Heidelberg): Computational Linguistics Department (C.Anette Frank)  
University of Cologne (Köln): Institut für Linguistik - Phonetik (C.Dagmar Jung)  
Max Planck Institute for Evolutionary Anthropology (Leipzig): Department of Linguistics (C.Hans-Jörg Bibiko)  
University of Leipzig (Leipzig): Institut für Informatik, Abteilung Automatische Sprachverarbeitung (C.Codrina Lauth)  
Institut für Deutsche Sprache (Mannheim): (C.Marc Kupietz)  
Westfälische Wilhelms-Universität Münster (Münster): Institut für Allgemeine Sprachwissenschaft (C.Gabriele Müller)  
University of Potsdam (Potsdam): Department of Linguistics (C.Manfred Stede)  
German Research Center for Artificial Intelligence (Saarbrücken): Language Technology Lab (C.Thierry Declerck)  
University of Stuttgart (Stuttgart): Institut für Maschinelle Sprachverarbeitung (C.Ulrich Heid)  
Universität Trier (Trier): Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften (C.Andrea Rapp)  
Universität Tübingen (Tübingen): Asien-Orient-Institut (C.Ulrich Apel); Seminar für Sprachwissenschaft (C.Erhard Hinrichs)  
**Greece (NCP: Stelios Piperidis)** Institute for Language and Speech Processing (Athens): Department of Language Technology Applications (C.Stelios Piperidis)  
**Hungary (NCP: Tamás Váradi)** Hungarian Academy of Sciences (Budapest): Research Institute for Linguistics (C.Tamás Váradi); Institute for Psychological Research of the Hungarian Academy of Sciences (Budapest): (C.Bea Ehmann)  
Budapest University of Technology and Economics (Budapest): Department of Sociology and Communications, Media Research Center (C.Peter Halacsy)  
Department of Telecommunication and Media Informatics, Laboratory of Speech Acoustics (C.Klára Vicsi)  
MorphoLogic Ltd. (Budapest): MorphoLogic Ltd. (C.László Tibanyi)  
University of Szeged (Szeged): Department of Informatics, Human Language Technology Group (C.Dóra Csendes)  
**Iceland (NCP: Eiríkur Rögnvaldsson)** Icelandic Centre for Language Technology (Reykjavík): (C.Eiríkur Rögnvaldsson)  
University of Iceland (Reykjavík): Institute of Linguistics (C.Eiríkur Rögnvaldsson)  
**Ireland (NCP: -)** National University of Ireland (Galway): Department of English (C.Seán Ryder)  
**Israel (NCP: -)** Technion-Israel Institute of Technology (Haifa): Computer Science Department (C.Alon Itai)

**Italy (NCP: Nicoletta Calzolari)** European Academy Bozen/Bolzano (Bolzano): Institute for Specialised Communication and Multilingualism (C.Andrea Abel)  
Università di Pavia (Pavia): Dipartimento di Linguistica Teorica e Applicata (C.Andrea Sansò)  
National Research Council (Pisa): Istituto di Linguistica Computazionale (C.Nicoletta Calzolari)  
University of Rome "Tor Vergata" (Rome): Department of Computer Science (C.Fabio Massimo Zanzotto)  
**Latvia (NCP: Inguna Skadina)** Tilde Language Technologies (Riga): Tilde Language Technologies (C.Andrejs Vasilevs)  
University of Latvia (Riga): Institute of Mathematics and Computer Science (C.Inguna Skadina)  
**Lithuania (NCP: Ruta Marcinkevičienė)** Vytautas Magnus University (Kaunas): Center of Computational Linguistics (C.Ruta Marcinkevičienė)  
Institute of the Lithuanian Language (Vilnius): (C.Daiva Vaisnienė)  
**Luxembourg (NCP: -)** European Language Resources Association (ELRA) (Luxembourg): (C.S.Piperidis/K.Chouki)  
**Malta (NCP: Mike Rosner)** University of Malta (Malta): Department of Computer Science (C.Michael Rosner)  
**Netherlands (NCP: Jan Odijk)** Meertens Institute (Amsterdam): Meertens Institute (C.H.J. Beentjes)  
University of Amsterdam (Amsterdam): Intelligent Systems Lab Amsterdam (ISLA) (C.Maarten de Rijke)  
Vrije Universiteit Amsterdam (Amsterdam): Computational Lexicology, Faculteit der Letteren (C.Piek Vossen)  
Data Archiving and Networked Services (Den Haag): (C.Henk Hammen)  
Huygens Instituut KNAW (Den Haag): (C.K.van Dalen-Oskam)  
University of Twente (Enschede): Human Media Interaction Group, Department of Electrical Engineering, Mathematics and Computer Science (C.Roeland Oudelman)  
University of Groningen (Groningen): Faculty of Arts, Center for Language and Cognition (C.Wyke van der Meer)  
Digital Library for Dutch Literature (Leiden): (C.C.A. Klapwijk)  
Institute for Dutch Lexicology (Leiden): Instituut voor Nederlandse Lexicologie (C.Remco van Veenendaal)  
Universiteit Leiden (Leiden): Leiden University Centre for Linguistics, Faculty of Humanities (C.Jeroen van de Weijer)  
Max Planck Institute for Psycholinguistics (Nijmegen): (C.Peter Wittenberg)  
Radboud University (Nijmegen): Centre for Language and Speech Technology (C.L. Boves / N. Oostdijk); Centre for Language Studies (C.Pieter Mulsen)  
Tilburg University (Tilburg): ILK Research Group, Department of Communication and Information Sciences, Faculty of Humanities (C.Antal van den Bosch)  
University of Utrecht (Utrecht): Utrecht Institute of Linguistics OTS, Faculty of Humanities (C.Steven Krauter)  
**Norway (NCP: Koenraad De Smedt)** Norwegian School of Economics and Business Administration (NHH) (Bergen): (C.Gisle Andersen)  
Unifob AS (Bergen): (C.Eli Hagen)  
University of Bergen (Bergen): Language Models and Resources group (C.Koenraad De Smedt)  
SINTEF (Oslo): (C.Diana Santos)  
The Language Council of Norway (Oslo): (C.Torbjørn Brevik)  
The National Library of Norway (Oslo): (C.Kristin Bakken)  
University of Oslo (Oslo): Department of Linguistics and Nordic Studies, Faculty of Humanities (C.Lanne Bondi Johannessen)  
University of Tromsø (Tromsø): Det humanistiske fakultet (C.Trond Tosterud)  
Norwegian University of Science and Technology (Trondheim): Department of Electronics and Telecommunications (C.Torbjørn Svendsen)  
**Poland (NCP: Maciej Ptaszek)** University of Lodz (Lodz): Institute of English Language (C.Piotr Pezik)  
Polish Academy of Sciences (Warsaw): Institute of Computer Science, Department of Artificial Intelligence (C.Adam Przepiórkowski); Institute of Slavic Studies (C.Violetta Koszka-Joszewa)  
Polish-Japanese Institute of Information Technology (Warsaw): (C.Krzysztof Marosek)  
University of Wrocław (Wrocław): Instytut Informatyki Stosowanej (C.Maciej Ptaszek)  
Wrocław University of Technology (Wrocław): Institute of Applied Informatics (C.Maciej Ptaszek)  
**Portugal (NCP: Antonio Branco)** Universidade Católica Portuguesa (Braga): Centro de Estudos Filosóficos e Humanísticos (C.Augusto Soares da Silva)  
University of Minho (Braga): Centro de Estudos Humanísticos (C.Pilar Barbosa)  
New University of Lisbon (Caparica): Faculdade de Ciências e Tecnologia (C.José Gabriel Pereira Lopes)  
Instituto de Telecomunicações (Coimbra): Polo de Coimbra (C.Fernando Pardalão)  
University of Coimbra (Coimbra): Centro de Estudos de Linguística Geral e Aplicada (CELGA) (C.Cristina Martins); Centro de Investigação do Núcleo de Estudos (C.José Augusto Simões Gonçalves Leitão)  
Universidade de Évora (Évora): School of Sciences and Technology (C.Paulo Quaresma)  
INESC-ID, Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa (Lisboa): (C.Nuno Mamede)  
Instituto de Linguística Teórica e Computacional (Lisbon): (C.Margarita Correia)  
New University of Lisbon (Lisbon): Centro de Linguística (C.Maria Francisca Xavier)  
University of Lisbon (Lisbon): Centro de Linguística da Universidade de Lisboa (LUL) (C.Amália Mendes); Natural Language and Speech Group (NLX-Group), Department of Informatics (C.António Branco)  
University of Azores (Ponta Delgada (Azores)): (C.Luis Mendes Gomes)

University of Porto (Porto): Centro de Linguística (C.Fátima Oliveira); Laboratory of Artificial Intelligence and Computer Science (C.Miguel Figueiras)  
**Romania (NCP: Dan Tufiş)** Romanian Academy of Sciences (Bucharest): Research Institute for Artificial Intelligence (C.Dan Tufiş)  
University Babeş-Bolyai (Cluj-Napoca): Faculty of Mathematics and Computer Science (C.Istari Doina)  
"Al. I. Cuza" University of Iaşi (Iaşi): Faculty of Computer Science (C.Dan Cristea)  
Romanian Academy of Sciences (Iaşi): Institute of Computer Science (C.Horia-Nicolai Todeorescu)  
University of Piteşti (Piteşti): Faculty of Letters (C.Mihaela Mitu)  
West University of Timişoara (Timişoara): Faculty of Mathematics and Informatics (C.Viorel Negru)  
**Serbia (NCP: -)** University of Belgrade (Belgrade): Faculty of Mathematics (C.Duško Vitas)  
**Slovakia (NCP: -)** Slovak Academy of Sciences (Bratislava): 'L. Štúr' Institute of Linguistics (C.Rodovan Garabik)  
**Slovenia (NCP: Tomaž Erjavec)** Alpinson d.o.o. (Ljubljana): (C.Cerjeja Zganeč Gros)  
Josef Stefan Institute (Ljubljana): Dept. of Knowledge Technologies (C.Tomaž Erjavec)  
**Spain (NCP: Núria Bel)** University of Alicante (Alicante): Departamento de Lenguajes y Sistemas Informáticos (C.Patrick Martínez-Barca)  
Institut d'Estudis Catalans (Barcelona) (C.Joan Soler i Bou)  
Technical University of Catalonia (UPC) (Barcelona): Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla (TALP) (C.Asunción Moreno)  
Universitat Autònoma de Barcelona (Barcelona): Facultat de Filosofia i Lletres, Dpt. de Filologia Anglesa i de Germanística (C.Ana Fernández Monraveda)  
Universitat de Barcelona (Barcelona): Departament de Linguística General (C.Irene Castellón)  
Universitat Oberta de Catalunya (Barcelona): Department of Languages and Cultures (C.Salvador Climent)  
Universitat Pompeu Fabra (Barcelona): Institut Universitari de Lingüística Aplicada (C.Núria Bel)  
University of Barcelona (Barcelona): Facultat de Filologia - Ramon Llull Documentation Centre (C.Joana Alvarez)  
Autonomous University of Barcelona (Bellaterra): Facultat de Letres, Dept. Filologia Espanyola (C.Carlos Subirats)  
Girona City Council (Girona): Records Management, Archives and Publications Service (C.Joan Boadós i Raset)  
University of Jaén (Jaén): Escuela Politécnica Superior, Departamento de Informática, SINAI group (C.Maria Teresa Martín Valdivia)  
University of the Basque Country (Leioa): Computer Science Faculty, Natural Language Processing Group (C.Arantza Diaz de Ilarza)  
University of Lleida (Lleida): Departament d'Anglès i Lingüística (C.Gloria Vázquez)  
Autonomous University of Madrid (Madrid): Laboratorio de Lingüística Informática (C.Manuel Alcántara Pita)  
University of Málaga (Málaga): Facultad de Filosofía y Letras, Dept. of English, French, and German Philology (C.Antonio Moreno Ortiz)  
Universidad Politécnica de Valencia - ITAC (Valencia): Grid and High Performance Computing Group (C.Vicente Hernández García)  
University of Vigo (Vigo): Facultade de Filoloxía e Tradución, Department of English, Research group LVTC (C.Javier Perez-Guerra); Seminario de Lingüística Informática, Departamento de Traducción e Lingüística, TALG Research Group (C.Xavier Gómez Guinovart)  
University of Zaragoza (Zaragoza): Facultad de Filosofía y Letras (C.Carmen Pérez-Llantada)  
**Sweden (NCP: Lars Borin)** University of Gothenburg (Gothenburg): Department of Linguistics, Faculty of Arts (C.Anders Eriksson); Språkbanken, Dept. of Swedish Language (C.Lars Borin)  
Linköping University (Linköping): Department of Computer and Information Sciences (C.Lars Ahrenberg)  
Lund University (Lund): Humanities Laboratory (C.Sven Strömquist)  
KTH Royal Institute of Technology (Stockholm): Department of Speech, Music and Hearing, CSC (C.Rolf Carlson)  
Language Council of Sweden (Stockholm): (C.Richard Domeij)  
Swedish Institute of Computer Science AB (Stockholm): (C.Björn Gambäck)  
Umeå University (Umeå): HUMLAB (C.Patrik Svensson)  
Uppsala University (Uppsala): Department of Linguistics and Philosophy (C.Joakim Nivre)  
**Turkey (NCP: Gülşen Eryiğit)** Istanbul Technical University (Istanbul): Elektrik-Electronics Faculty, Computer Science Department, Natural Language Processing Group (C.Gülşen Eryiğit)  
Sabanci University (Istanbul): Human Language and Speech Laboratory, Faculty of Engineering and Natural Sciences (C.Kemal Oflazer)  
**United Kingdom (NCP: Martin Wynne)** Bangor University (Bangor): Language Technologies Unit (C.Brony Williams)  
University of Birmingham (Birmingham): Department of English (C.Oliver Mason)  
University of Surrey (Guildford): Department of Computing, Faculty of Engineering and Physical Science (C.Lee Gillam)  
Lancaster University (Lancaster): Department of Linguistics and English Language (C.Paul Rayson)  
National Centre for Text Mining (Manchester): National Centre for Text Mining (C.Bill Black)  
Oxford Text Archive (Oxford): Oxford University Computing Services (C.Martin Wynne)  
University of Sheffield (Sheffield): Natural Language Processing group, Department of Computer Science (C.Wim Peters)  
University of Wolverhampton (Wolverhampton): Research Institute of Information and Language Processing (C.Constantin Orasan)