CLARIN Newsletter

Number 7, 2009, September

Towards an ERIC for CLARIN

Bente Maegaard
CLARIN EB member
Steven Krauwer
CLARIN coordinator

CLARIN is currently in the second year of its three-year preparatory phase. When the three years are over (at the end of 2010), the intention is that CLARIN shall make a transition to a more permanent structure. There are two requirements for this: a 'legal form' is needed, and funding is needed.

Legal form

A legal form can be e.g. a company, an association, or the like. The European Commission has investigated available legal forms and concluded that they are not well suited for research infrastructures, see also the CLARIN report Requirements and Best Practice for Governance, D8S-1.1 at the CLARIN website (documents, deliverables). As the various existing legal forms were not suitable, the Commission elaborated a new legal form.

In June 2009 the European Council of Ministers adopted a council regulation for a European Research Infrastructure Consortium (ERIC) providing a community legal framework for research infrastructures. The Commission is still working on the practical implementation of the regulation in collaboration with the governments.

CLARIN is now investigating what it would mean to implement this suggested ERIC as our legal form, and in this article we want to discuss some of the features of an ERIC.

Members of an ERIC

It follows from the nature of research infrastructures that in order for them to be sustainable the members have to be countries. If a country is a member, then all the relevant institutions in that country are part of the ERIC and can use the services. However, we are pretty sure that not all countries will be in a situation where they can join right now; so we are investigating possibilities for also collaborating with institutions from countries that are not a member.

Members' rights and duties

Member countries will gain access to the CLARIN infrastructure and all its services for all their relevant institutions (universities, research institutions, academies, research libraries and the like; the exact list will be according to the agreement to be made). They will have a vote at the General Assembly so that they can influence the development of the research infrastructure. Members will have to pay a fee (with the fee proportional to country size, according to some measure and a formula to be agreed), and they will have to ensure a centre which can provide CLARIN services (or they may pay another country to provide the centre service), and they must have implemented an authentication and authorisation system for access (to control which users get access to what materials).

The current thinking is that CLARIN should collaborate with institutions from non-member countries, as users and/or contributors, and that such institutions could enjoy the same access rights as members. They would have to pay a fee like everyone else, but they would not be able to participate in the General Assembly or in any of the managing bodies.

Funding

One of the important issues to solve is how the necessary funding can be provided. Funding will come from national contributions, from contributions from institutional members in non-member countries, and from use of the infrastructure by non-researchers. In the next framework programme, FP8, there may be a possibility that the Commission can fund up to 20% of the operational costs for an ERIC.

Host country

An ERIC will have a statutory seat in a host country and right now the Netherlands are offering to take that role. The Netherlands has committed funding until 2014, and actually NL is the only country which has secured funding of CLARIN beyond the preparatory phase.

Next steps

The next step for CLARIN is to approach all countries (current partner countries and those who want to become partners) to see their interest in discussing the CLARIN ERIC in further detail with the aim of setting up first a Memorandum of Understanding, and eventually an agreement between those countries which agree to form the "Initial Coalition" or be the founders of the CLARIN ERIC.

Other countries may join when they are ready, but it is of utmost importance that we make sure that CLARIN will carry on from the preparatory phase into the next phase as smoothly as possible. Ideally the CLARIN ERIC should be operational from January 1, 2011. In order for this to work, we will need to submit an application for the ERIC to the Commission at the latest mid-summer 2010, so this will be a hectic time!

Editors' Foreword

Marko Tadić
& Dan Cristea
CLARIN Newsletter editors

We offer the honour to open the issue this time to Bente Maegaard and Steven Krauwer, because we want to emphasize the significant activity that was recently pursued inside our community towards the creation of a legal form to the CLARIN entity. The newly framework approved by the Commission to support research infrastructures related activities is called ERIC and this is what we are heading towards, with an application that has to be submitted in the middle of 2010.

Research infrastructures in linguistics seem to develop widely, not only in Europe, and a signal from USA is brought by Koenraad de Smedt. He reports on the Cyberlink 2009 workshop, held recently at the start of cyber-infrastructure – the American term for research infrastructures – project at UC Berkeley.

Then, continuing our tradition of presenting a research in domains that need the help of CLARIN, Hanne Fersøe offers a very interesting article on a consortium project studying hearing loss, showing how sound and language technologies could feed advanced techniques intended to develop hearing aid devices.

The central pages of the issue, as always intended to present recent important CLARIN events, are contributed by Peter Wittenburg, Antti Arppe, Pirjoleanna Forsstrom and Nicoletta Calzolari. NEERI, the Networking Event for European Research Infrastructures + Standards was organised by CLARIN and hosted by the University of Helsinki, in September this year. By its attendance and the diversity of themes approached (among which: long-term preservation, persistent identifiers, metadata frameworks, semantic interoperability, grid computing and federation technologies used in CLARIN, ethical and legal issues) this workshop configured as one of the most significant events organised by CLARIN since its inception.

Call for contributions

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

We wanted to host in this Newsletter a presentation of one traditional NLP event, that grew in significance over the years, due to the attentive selection of the hot topics of this so rapidly developing field, the quality of the papers presented and the inspired association with a number of workshops: the International Conference Recent Advances in Natural Language Processing. So we invited two of its main organisers, Galia Angelova and Ruslan Mitkov, to author an article describing this September event, held in Borovets, Bulgaria.

Then, Felix Sasaki signs a short article bringing into attention a major topic of interest for the infrastructure that we want to build: the standards to support the diversity of languages, and the language identification software that will have to bridge the language resource information and the Web information space.

And finally, dear reader, you will find a report, written by Ineke Shuurman, on the activities developed by CLARIN-Vlaanderen on the Dutch language, spoken essentially in the northern part of Belgium, Flanders, and partly in Brussels.

Enjoy your reading! C

List of national correspondents

Austria
Gerhard Budin

Belgium – Flanders
Inneke Schuurman

Bulgaria
Svetla Koeva

Croatia
Marko Tadić

Czech Republic
Karel Pala

Denmark
Bente Maegaard
Hanne Fersøe

ELRA/ELDA
Stelios Piperidis
Khalid Choukri

Estonia
Tiit Roosmaa

Finland
Kimmo Koskenniemi

France
William Del Mancino
Bertrand Gaiffe

Germany
Lothar Lemnitzer

Greece
Maria Gavrilidou

Hungary
Tamás Váradi

Italy
Valeria Quochi

Latvia
Andrejs Vasiljevs

Malta
Mike Rosner

Netherlands
Peter Wittenburg

Norway
Koenraad De Smedt

Poland
Maciej Piasecki

Portugal
Antonio Branco

Romania
Dan Cristea
Dan Tufiş

Spain
Nuria Bel

Sweden
Sven Strömqvist

UK
Martin Wynne

Research Infrastructures: The American Way

Cyberling 2009 Workshop Berkeley, July 17-19, 2009

Koenraad de Smedt
University of Bergen

We have really everything in common with America nowadays, except, of course, language.
Oscar Wilde

Cyberinfrastructure is the American term for what Europeans refer to as *research infrastructure*. Hence the name Cyberling for an initiative taken by Emily M. Bender, Jeff Good, Scott Farrar, Laura Welcher and Dan McCloy to promote a cyberinfrastructure for linguistics. The objectives of this American initiative are obviously also close to CLARIN's heart. Large scale infrastructures presuppose a culture of publishing and sharing data and annotations, a set of common standards and architectures, and new long term funding models; since these problems cannot be solved by isolated research projects, they require widespread communication, participation and buy-in from the whole field, according to the Cyberling organizers.

Building a cyberinfrastructure for linguistics

Therefore, with funding from the National Science Foundation, 43 researchers were invited to a workshop at the UC Berkeley campus in California from July 17 to 19, 2009. The goal was to reflect on the processes that will be necessary to build a cyberinfrastructure for linguistics. Most of the invited participants were from the USA, but some came from other continents and a few of those, including the author, represented CLARIN partners. The workshop coincided with the Linguistic Institute.

The structure of the Cyberling workshop promoted high productivity and intense discussion. Each participant was assigned to a working group which was given specific discussion themes and reporting tasks. Working groups were established for each of the following themes: (1) exploring annotation standards, (2) other standards, (3) "killer apps", (4) data reliability and provenance, (5) models from other fields, (6) funding sources, and (7) collaboration structure.

How and Why Approach

On each day of the workshop, working groups presented preliminary conclusions in plenary sessions open to anyone. The first one, with the theme *Data Sharing: How and Why?* addressed the benefits of data sharing and the obstacles to the widespread adoption of sharing practices, from the perspective of a variety of subfields. The second one, *Computational Methods: How and Why?* explored the state of the art in computational methods for linguistic analysis, and the emerging possibilities presented by new technology. The third plenary session saw presentations and discussion of final working group reports. These reports are currently being rewritten into white papers to be published in the near future.

On the one hand, we felt that there are many common concerns on both sides of the Atlantic, and we can learn much from each other. Cyberling workshop participants agreed that an infrastructure should be attractive to users and that efforts should be made to promote a culture of sharing data. The worskhop converged on the ideas that an infrastructure should be based on software as a service and that data sharing should also count as publication.

Cyberling vs. CLARIN

On the other hand, some differences can be observed, particularly in scope. Cyberling is targeted at linguistics, while CLARIN has a broader orientation towards the use of language materials in all the Humanities and Social Sciences. This has consequences not only for the inclusion or exclusion of different kinds of data sources and applications, but also for the consideration of possible organizational anchorings of a future infrastructure.

Cyberling currently does not have the organizational status of a project like CLARIN.

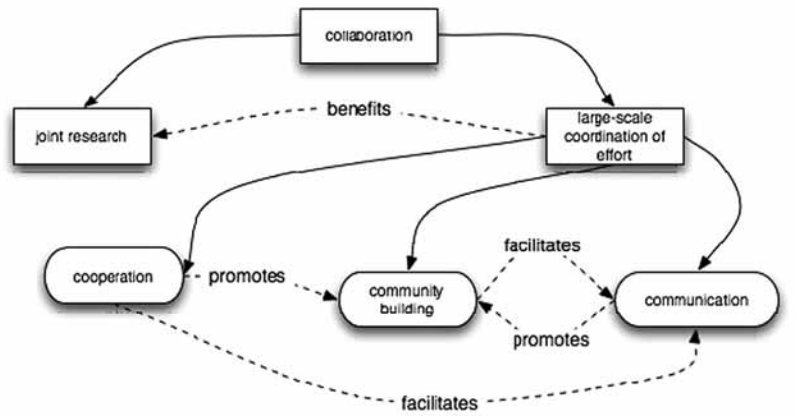


Diagram by Emily Bender, Nancy Ide, Nicoletta Calzolari, and David Robinson.

There are several complementary activities based in the USA, such as Project Bamboo, The Rosetta Project, the LSA's eLanguage Platform and the Open Language Archives Community. Because the Cyberling community is still very open, it seems to be creatively interested in linking up with such initiatives, as well as drawing inspiration from other, more general resources and organizations such as FreeBase, the Internet Archive, the LongNow Foundation, etc.

A wiki has been intensively used throughout the whole process, starting from the preparation of the workshop with the presentation of participants and issues and ultimately resulting in reports. During the workshop itself, the wiki was used to jot down notes and present an almost real-time record of ongoing discussions, while it was also open to non-workshop participants. Now the discussion is taken further on http://blog.cyberling.org/.

I hope for CLARIN to keep a liaison with Cyberling and for continued communication and co-operation between initiatives on different continents to raise language infrastructure issues to a level of global importance.

Hearing loss, perception and annotated corpora

Hanne Fersøe
Centre for Language Technology, University of Copenhagen

Most people have experienced being in a noisy environment where it was difficult to hear and therefore difficult to participate in a conversation. These difficulties may have been caused by other parallel conversations or by background noises. This phenomenon is often referred to as the cocktail party situation. People with a hearing loss experience this not only at cocktail parties, but all the time in group conversations where speakers interrupt one another, and where several people talk at the same time.

Suffering from a hearing loss

A hearing aid does not help much in this type of situation because it both amplifies noise and speech, the irrelevant and the relevant sounds, to the same degree, regardless of what the brain needs for successful perception. Hearing loss is not evenly distributed across a person's entire auditory spectrum (12Hz to 20000 Hz). Typically it affects high frequencies, so since the frequency of speech information is mostly lower than 3000Hz it is possible to conduct person-to-person conversations in quiet surroundings even with a moderate hearing loss. In noisy environments, however, all parts of the signal are needed, i.e. it is necessary for the brain to also have access to the redundancy located in the upper part of the spectrum experienced by a person with normal hearing. So, consequently, it is the lost data from the high frequency spectrum of the speech signal that the brain would need the hearing aid to amplify, and not all the data [1].

Reconstruction strategy of the brain

The hypothesis regarding reconstruction is that the hearing centre of the brain, called the auditory cortex, employs a strategy for reconstructing those parts of the speech signal which are covered or disturbed by noise. With this strategy the brain supplements the acoustic analysis with expectations about the communicative message, knowledge about the language, and general experience and memory in its process of fishing for meaningful units. For the brain, listening, or creating meaning, to a very large extent consists of actively co-creating meaning, and therefore measurements of the sound waves or analysis of the phonetic details of the speech signal are not sufficient to describe what happens in the process of decoding. In the description it is necessary to include more parameters which contribute to enabling the brain to sort through the mess of sounds that it receives and pick out what it needs [2].

One future perspective arising from this research is that it may become possible to build smarter hearing aids. The vision is that with this new understanding of the reconstruction process, the hearing aids may be able to recognize language sounds in a chaotic sound picture and amplify the relevant ones selectively. Moreover, they may also be able to improve their their ability to recognize language sounds over time through a learning mechanism. With a hearing aid like that a person with a hearing loss would be able to listen to and understand a conversation partner even in a noisy cocktail party. People without a hearing loss may find it useful, too.

In Denmark, a research group of phoneticians, computational linguists, psychoacousticians, and cognitive scientists from Copenhagen Business School (CBS), Technical University of Denmark (DTU), University of Copenhagen (UCPH), and University of Southern Denmark (SDU) are currently collaborating in a basic research project called Reconstruction of Speech Events. The project is aimed at understanding how the brain separates the different layers of sound information, and how the reconstruction process of the brain works.

Social occasions are often difficult for a person with hearing loss. Background noise, such as music, or group conversations, can become overwhelming, making it impossible to participate in a conversation (Illustration by Cristian Goila)

Access to large speech corpora contributes to the deeper understanding of some of the issues involved in the fundamental question of how the brain navigates through the various sounds and noises and constructs meaning out of the relevant ones.

The PAROLE spoken corpus

An example of such a large corpus is the spoken Danish PAROLE corpus, which will be made available through the Danish CLARIN platform [3]. The spoken Danish PAROLE corpus is a 100,000 token read-aloud corpus in lab quality with several layers of annotation. *"To speech researchers spontaneous speech corpora are fundamental because the researchers want their sound analysis to function in realistic situations, i.e. on ordinary spoken language with background noise, but in the development phase we also need speech recorded under controlled circumstances (studio microphones and echo-free room) for the calibration of the analysis programs"*, as CBS computational linguist Peter Juel Henrichsen puts it. *"The read-aloud PAROLE corpus is also particularly well suited for the development of methods because certain phenomena are kept constant"* he adds.

After the recording process, the resulting sound files have been annotated with orthographic transcription, two different layers of part of speech (PoS) tagging, phonetic transcription with International Phonetic Alphabet (IPA), time stamps for every phone, and with acoustic parameters such as pitch measurements for every five milliseconds, intensity i.e. volume measured in decibel, for every five milliseconds, and HNR (harmonics-to-noise ratio). More annotation layers currently in the making or being considered for future enhancements of the corpus are English translation, syntax trees for Danish and English, information structure, parallel reading by eight readers, whispered reading, and translations into more languages, including Russian and Tamil. The corpus has been read aloud, phonetically transcribed, and acoustically measured by the researchers of CBS; see an example of annotation layers illustrated in the table below. Interested readers may listen to the spoken PAROLE corpus at http://isvcbs.dk/speechevents.

Annotation in the PAROLE spoken corpus

With the use of this corpus, Peter Juel Henrichsen has developed a tool which, through listening alone, is able to identify how a specific speech event, [s] for instance, sounds. Speech events may or may not be equal to the phones in classical phonetics. Speech events are not defined beforehand; the tool lets the data, in this case the PAROLE corpus data, define the most appropriate events. Examples of other events identified by the tool include hesitations, pauses, emphatic stress, and special personal habits like tongue clicks or sighs. Input to the processing done by the tool is sound waves and a rough approximation of their phonetic transcription; no speaker specific phonetic knowledge about for instance the phone [s] and the surroundings where [s] most frequently may occur is included. The tool is able to recognize all the different events in the corpus based on the sound waves and the crude transcription alone. The output from the processing is a list of events in the corpus with their time stamps. This output is used for re-synthesis, i.e. for creating synthetic sound, which is then subsequently stereo-synchronized with the natural sound so that the researcher can compare the two and thus calibrate the tool and improve its performance. The quality of the synchronization between the natural and the synthetic speech is a measure of how well the tool works, i.e. whether it has identified the right phones.

Like the corpus, this tool will also be made available through the Danish CLARIN platform. It has the potential of becoming useful for other speech technology researchers, since one of the perspectives for its use is that it can be adapted to deliver automatic transcription of speech, including spontaneous speech, as output.

Research themes pursued by other researchers in the Speech Events group include modelling of the auditory cortex where the researchers at DTU use corpora like PAROLE as input to experiments with perception; research into man-machine interfaces and robotics at SDU based on dialogue corpora; phonetic investigations of spontaneous speech at UCPH using multiple spoken corpora, and prosodic investigations of dialogues at CBS.

Speech recording in a sound proof studio at CBS:
Computational linguist Peter Juel Henrichsen (photo by Hanne Fersøe)

Orthography	illusioner	er	Farlige
PoS, general	noun plural/indefinite	verb present tense/active	adjective positive/plural
PoS, PAROLE tags	NCCPU==I	VADR=----A	ANP[CN]PU=[DI]U
Transcription, IPA	[iluS'o:ʔnC]	[a]	[f'A:li:]
Pitch (Hz)	116-108-152-156	112	97-88-110-89

Text: *Illusioner er farlige* (*Illusions are dangerous*)

Industry interest

Today, hearing aid devices base their functionality solely on physical acoustics, i.e. on manipulation of sound waves. The devices neither include nor use any knowledge about phonetics or linguistics. The industry is beginning to understand the limitations of this approach and to recognize that they need to understand how to build phonetic and linguistic knowledge into devices in the future, and to amplify speech events to aid the brain to reconstruct the relevant meaning.

Thanks to Peter Juel Henrichsen from the research project Reconstruction of Speech Events (CBS) for inspiring conversation and useful comments.

References

[1] E. Colin Cherry (1953): Some Experiments on the Recognition of Speech, with One and with Two Ears. The Journal of the Acoustical Society of America, Vol. 25, no. 5, p. 975-979.

[2] Reinier Plomb (2001): The intelligent ear: on the nature of sound and perception. Lawrence Erlbaum Associates Inc.

[3] CLARIN Newsletter 2 and 4. www.clarin.eu/newsletter

NEERI 09 Report

Helsinki 30 September–2 October, 2009

Peter Wittenburg
Antti Arppe
Pirjoleena Forsstrom
Nicoletta Calzolari
CLARIN

NEERI 09 was organized for a number of reasons: (1) to bring together ESFRI research infrastructure and e-Infrastructure initiatives to interact on technological topics to facilitate cross-fertilization; (2) to present and discuss some of the CLARIN approaches to prevent isolated solutions; (3) to launch several activities to signal that we are moving ahead; (4) to disseminate knowledge to achieve a high level of synchronization and (5) to interact with representatives from the EC. Looking back, we can say that the conference met all expectations, and in particular it functioned as an eye opener for many CLARIN colleagues, with the many discussions sharpening our minds.

EC and APA

Hervé Pero (EC) stressed the importance of overcoming the fragmentation and managing the increasing costs and complexity to enable leading-edge research. Research infrastructures can play an important role, if they act pro-actively, solve the data management challenges (a common theme that requires coherent policies), integrate into an eco-system of infrastructures and are managed professionally. Kostas Glinos (EC) looked at the service layers, argued about the need of a proper orchestration between all initiatives and identified a gap for a data services infrastructure and workspaces. Virtual Research Communities are the basis for strengthening the service and user orientation. Keith Jeffery (Alliance for Permanent Access) also stressed the need to take care of our scientific information, the amount of which is increasing continuously and the value of which cannot be specified in market terms. Compared to the costs for acquisition, ingest and access, it is known that those for storage and preservation are the lowest.

Long-term Preservation

Dave Giaretta (STFC) discussed the many threats for the long term preservation and accessibility of data and that this is primarily a social challenge. A variety of organizations are working on key issues of long term preservation, yet overviews indicate that amongst researchers preservation is not a primary concern. What kind of measures need to be taken to go beyond pure bit-stream preservation, but to ensure understanding of formats and semantics? Dany Vandromme (Renater) presented the work of ESFRI and the e-IRG task forces on data management, stressing that proper repositories, interoperability, methods to create metadata and to assess quality are essential. Research infrastructures need to take over responsibility for the data their community is producing. Kimmo Koski (CSC) argued for a data services infrastructure as a cross-disciplinary enterprise based on EU Council decisions and an analysis of the life cycle of data. An infrastructure could foster interdisciplinary research, but certainly would be more efficient in terms of energy efficiency. Providers of e-infrastructure and services succeed only if approaching the technology from user community perspective: trust must be built between the stakeholders.

Persistent Identifiers

Larry Lannom (CNRI) explained the need for persistent references in a world of increasing complexity of the relationships between objects. To make this all work, any registration and resolving services need to meet criteria such as reliability, scalability and flexibility and the organizations offering a service need to make a long term commitment. Software systems such as the Handle System are implementing a set of services, but do not guarantee persistence. Ulrich Schwardmann (GWDG) introduced the EPIC consortium offering redundant persistent identifier services based on the Handle System. A decision and management structure will need to be worked out to establish trust. Rutger Kramer (DANS) showed that PIDs need to be associated with policies such as to determine granularity, fragment addressing and versioning policy. Since various systems are already in use it might be worth to build a resolver of resolvers.

Flexible Metadata Frameworks

Peter Wittenburg and Daan Broeder (MPI) described its essentials of the component based flexible metadata infrastructure currently under development in CLARIN, and put it in the context of the common trends in communities such as Dublin Core. Esssential for interoperability are compulsory registered data categories rather than fixed schemas. The price for this desired flexibility is more complexity of the framework that will have to be built. As a first concrete result the Virtual Language World and Observatory was introduced. Michael Lautenschlager (DKRZ) introduced the comprehensive common information model and its metadata components, developed to serve climate research. This is based on agreements among the comparatively small set of centres involved. Paul Doorenbosch (KB) described the huge interoperability issues that need to be addressed in DRIVER, and in particular in the Europeana projects. While for DRIVER the mapping to Dublin Core needs to be provided by any party, in

Kimmo Koskenniemi and Nicoletta Calzolari

Europeana no such strict agreements could be enforced, requiring a lot of manual curation work. Keith Jeffery (STFC) made clear that metadata of different sorts will become even more important in the future IT world if it is not to collapse. Proper formal syntax and explicitly declared semantics will guarantee manageability as well as a clear separation between data and operations. A number of further IT challenges such as data representativity, quality and permanency, and explicit policy and service descriptions must be tackled to build complex service oriented infrastructures.

Semantic Interoperability

Already in the preparatory phase it has become obvious that many infrastructures need to tackle the issue of semantic interop-

Steven Krauwer

erability, but, in general, they have not proceeded far enough to formulate policies. For many humanities disciplines, this issue is crucial to overcome fragmentation. Sue Ellen Wright (Kent) described the ISO 12620 model and its ISOcat implementation which is mainly driven forward by terminologists and linguists. Nevertheless, the model and its implementation are basically discipline-independent, based on rather generic specifications such as ISO 11179 about metadata categories, and based on the assumption that concept definitions need to be separated from relations between them. This chimes with Jeffery's statement about manageability, and also reflects the fact that relations are often usage driven.

Grid and Federation Technologies

This session was of great importance for the experts from the various research infrastructures, since it not only showed the state of

The full amphitheatre

developments in the cross-community infrastructure initiatives, but also how other communities are dealing with commitments from service centres. Diego Lopez (eduGain) presented, amongst other things, the efforts, problems and solutions so far to come to a harmonized distributed European authentication solution implementing a single identity for all European researchers. Essentially it is trust across countries and cultures that needs to be established and agreements about attributes to make progress. Different attitudes and requirements will require to peel “identity onions”, where concrete community requirements such as those specified by CLARIN may help to focus on practical solutions. From studying CLARIN short guides, Bob Jones (EGEE) explained which kind of services the grid community could give to research infrastructures. In particular for establishing, running and monitoring the network of community centres many tools can be re-used. Although many of the solutions primarily focus on high energy physics other communities have already successfully applied them. Jones suggested continuing the interaction bilaterally with the follow-up organization EGI and to establish a European Infrastructure Forum to bring together experts from all infrastructure initiatives about essential topics. DEISA presented by Johannes Reetz is focussing on bringing together high performance computers and offering a wide variety of software packages that can be used by researchers and that are tuned to the various architectures. The user should not need to rewrite code, and an elaborate support infrastructure is being set up to help the user in easily making use of the virtual supercomputer systems. Monitoring instruments are used to check commitments and to optimize load distribution.

Ethical and Legal Issues

This session addressed issues which are to a substantial extent influenced by strategic decision making and legislation at both the national and EU levels, which can concern the iresaerch infrastructures,e.g. the foundation of the European Research Initiative Consortia (ERICs) as legal entities, or copyright and personal privacy protection. Sami Borg (CESSDA, FSD) discussed first the ethical balancing act of protecting the privacy of individuals in social sciences surveys through anonymization, while at the same time making such survey data available as soon and as widely as possible. Next, Borg discussed how a European network of national social sciences data archives (CESSDA), already long-established on a less formal legal basis, will make use of the recently founded ERIC legal entity. Importantly, the CESSDA ERIC will not be a data archive that will hold or disseminate data by itself – rather, it will operate in a distributed fashion among a number of clusters of expertise, The role of the ERIC will be that of a co-ordinator, facilitator, and broker in developing and implementing common standards and tools. In terms of data discovery and access, the role of the CESSDA ERIC will be that of a gateway service – providing access with common authentication to a distributed set of national data collections. Marjut Salokannel (CLARIN) discussed the short-comings of both EU and national copyright legislation to provide free access to copyrighted works for academic purposes, which the EC has acknowledged in its recently published Green Paper on Copyright in the Knowledge Economy. Salokannel argued that without a

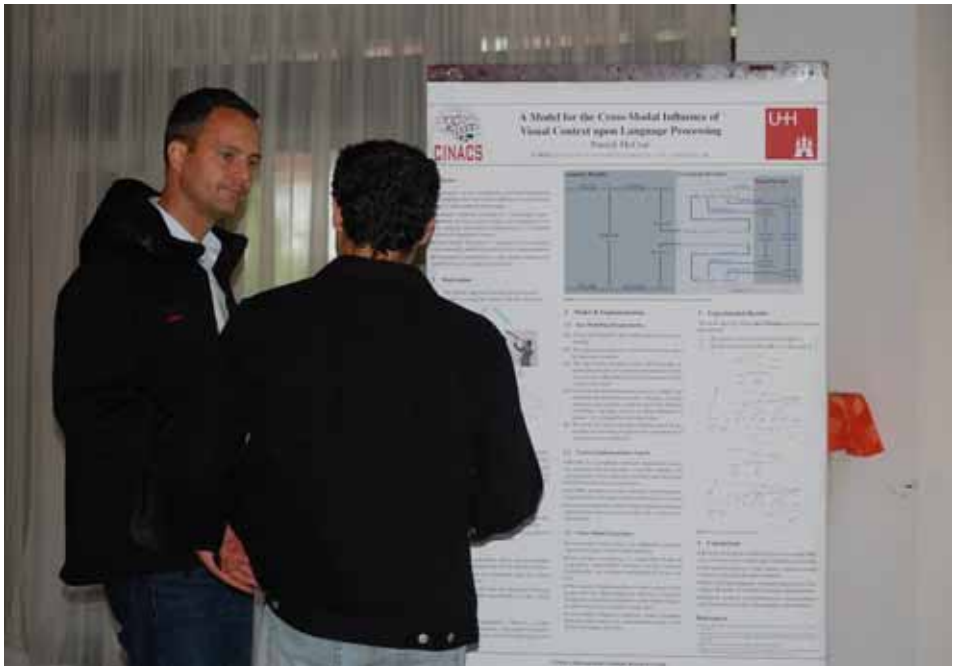
Continued on the next page ☞

harmonized copyright framework it will be virtually impossible to create an efficiently functioning pan-European research infrastructure, and suggested that ESFRI should make a joint proposal to the EC for rectifying this situation. To this end, Ville Oksanen (CLARIN) presented a concrete draft text urging the legislatures at the European level and at the national level to harmonize copyright law in such a way that the free use of copyrighted works for academic purposes is permitted in all member states, with the key qualification that such Academic Use should not unreasonably prejudice the legitimate interest of the right-holders. In the ensuing discussion chaired by Steven Krauwer (CLARIN), the participants unanimously agreed on the urgent necessity of such Academic Use as the NEERI message.

Other Topics

A number of other topics were addressed in short presentations. Kimmo Koski, Peter Wittenburg and Michael Lautenschlager presented the PARADE White Paper that makes concrete suggestions of how to tackle the data services infrastructure gap at European level and that is the result of intensive discussions between a number of structured communities and computer centre experts. Kostas Glinos (EC) briefly commented on this presentation by pointing to the needs and the requirements associated with such an infrastructure, such as its community-driven nature. Diego Lopez (RedIRIS) reflected on thoughts about digital identity and service composition, i.e. topics which are currently addressed in infrastructures working on service-oriented architectures. The transition from a client-rooted to a chained model requires federation services and proper interface specifications. Volker Boehlke (D-SPIN) explained the current work on service chaining in the German CLARIN branch and the use of functional metadata elements registered in ISOcat to do automatic profile matching, i.e. to find matching services for given resource types. Peter Kunszt (SystemsX) discussed the potential benefits of computing clouds for the long term storage of research data and its security aspects, while not neglecting the potential problems of trust and sustainability when making use of service offers from Amazon, Google, etc. Finally, Milena Piccoli (DANS) presented the "Data Seal of Approval", a lightweight method to carry out a quality assessment for repositories, as it is now suggested for DARIAH and CLARIN centres.

Many participants expressed their satisfaction with the meeting and expressed their wish to continue with the cross-fertilization attempts. C



Student Poster Session

RANLP-2009 Borovets, Bulgaria September 12-18, 2009

Galia Angelova
Bulgarian Academy of Sciences
Ruslan Mitkov
University of Wolverhampton

The 7th International Conference *Recent Advances in Natural Language Processing* and its associated events were held in the mountain resort Borovets. RANLP-2009 is the latest event in the RANLP conference series. It continues the tradition of many successful summer schools, the International Conferences RANLP-1995, RANLP-1997, the EuroConference RANLP-2001, RANLP-2003, the Marie Curie Large Conference RANLP-2005 and RANLP-2007. The biennial event RANLP has established itself as one of leading international conferences in the field. In February 2009 RANLP was cited as one of the most successful Computer Science conferences (http://www.cs-conference-ranking.org/conferencerankings/alltopics.html). RANLP is preceded by tutorials and followed up by workshops. The keynote speakers (6 for each conference) represent leading lights in the field.

Some 136 papers were submitted to RANLP-09. The articles were reviewed by a Programme Committee consisting of well-known experts. In terms of acceptance rate of regular papers, RANLP-2009 reported an acceptance rate as low as 13%! For comparison, the selection rate of the regular papers at RANLP-2007 was 11% and at RANLP-2005 – 14%. This year 87 papers were presented at the conference (as regular, short or poster papers), authored by researchers from 28 countries. More than 90 papers were submitted to the RANLP-2009 workshops. For the first time, a Student Research Workshop was held in the form of 3 oral presentations and 13 posters.

Keynote speakers and tutorials

Keynote speakers at RANLP-2009 gave six invited talks:

- Ricardo Baeza-Yates (*Towards Semantic Search*),
- Kevin Bretonnel Cohen (*Paradigms for evaluation in natural language processing*),
- Walter Daelemans (*Robust features for Computational Stylometry*),
- Mirella Lapata (*Vector-based Models of Semantic Composition*),

Still Stable after All This Years

- Shalom Lappin (*Restricting Probability Distributions to Expand the Class of Learnable Languages*),
- Massimo Poesio (*Conceptual Knowledge: Evidence from Corpora and the Brain*),

whereas tutorial speakers and titles were:

- Kevin Bretonnel Cohen (*Biomedical Natural Language Processing: BioNLP*),
- Roberto Navigli (*Graph-Based Word Sense Disambiguation and Discrimination*),
- Constantin Orasan (*Automatic summarisation in the Information Age*) and
- Kiril Simov & Petya Osenova (*In the NLP world of Knowledge Nets*).

RANLP-2009 Opening Ceremony

A number of workshops

The following workshops were accepted and held at RANLP-2009:

- *Multilingual resources, technologies and evaluation for Central and Eastern European languages* (organised by Cristina Vertan, Stelios Piperidis, Elena Paskaleva and Milena Slavcheva),
- *Adaptation of Language Resources and Technology to New Domains* (organised by Nuria Bel, Erhard Hinrichs, Kiril Simov and Petya Osenova),
- *Natural Language Processing methods and corpora in translation, lexicography, and language learning* (organised by Viktor Pekar, Iustina Narcisa Ilisei, and Silvia Bernardini),
- *Events in Emerging Text Types* (organised by Constantin Orasan, Laura Hasler and Corina Forascu),
- *First Workshop on Definition Extraction* (organised by Gerardo Eugenio Sierra Martinez),
- *Biomedical Information Extraction* (organised by Guergana Savova, Vangelis Karkaletsis and Galia Angelova) and
- *Student Research Workshop* (organised by Irina Temnikova, Ivelina Nikolova and Natalia Konstantinova).

Organisers of the Student Research Workshop

The awards

The RANLP-2009 award for Best Student paper was given to Mr. Ravi Sinha from the University of North Texas for his paper "Combining Lexical Resources for Contextual Synonym Expansion".

RANLP-2009 published full proceedings of the conference and all workshops. These proceedings will be publicly available at the ACL Anthology. Traditionally, John Benjamins (Amsterdam) also publishes a volume with RANLP keynote lecturers talks and regular papers. The volume of RANLP-2007 will appear soon in the series "Current Issues in Linguistic Theory".

The informal team behind RANLP-2009 includes:

- Galia Angelova (*Organising Committee Chair*),
- Kalina Bontcheva,
- Ruslan Mitkov (*Programme Committee Chair*),
- Nicolas Nicolov,
- Nikolai Nikolov, and
- Kiril Simov (*Workshop Coordinator*).

Ivelina Nikolova was the Programme Committee Coordinator. Irina Temnikova and Natalia Konstantinova supported the Programme Committee in the review and selection process.

We are sure that we will see you again in 2011. C

The role of language identification in CLARIN infrastructure

Felix Sasaki
University of Applied Sciences, Potsdam

In the Clarin Newsletter Number 4, I recently read the following statement from Dan Tufiş:

"The multilingual web of the Single European Information Space cannot be dissociated from the common language infrastructure aimed at by the CLARIN project. And vice versa, the CLARIN-envisaged multilingual services cannot ignore the vision and priorities of the Single European Information Space."

My initial reaction was "Yes, that is it!", and after calming down a bit I realised why and decided to put the reason in words.

Web infrastructure is evolving constantly, and the factors motivating change are sometimes based on industrial, and not community-specific needs. On the other hand, communities e.g. those represented in CLARIN, develop their own infrastructure. Under these circumstances a mismatch occurs.

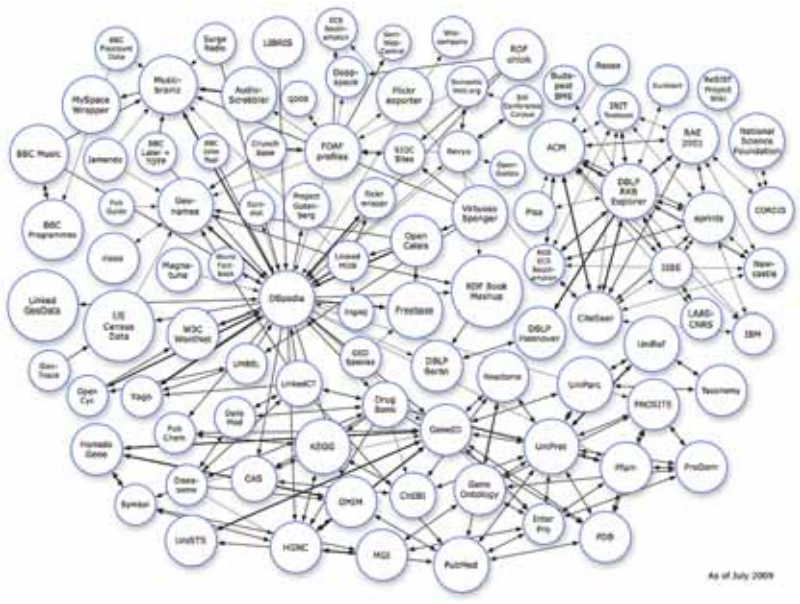
An example of such a mismatch is the area of language identification. 14 years ago the Internet Engineering Task Force (IETF) "Web" (and also "Internet") standard for this purpose was RFC 1766, which encompasses 136 language codes from ISO 639-1. Some metadata schemata for language resources made use of this RFC. Meanwhile other parts of the ISO 639 series emerged, e.g. ISO 639-3. This standard is derived mainly from Ethnologue, with the goal of a comprehensive coverage of languages, supplying thousands of language codes. Of course language resource projects wanted to take advantage of this new standard, so related metadata schemes allowed usages of ISO 639-3 codes.

Coming back to web technology, meanwhile RFC 1766 had been obsoleted by several RFCs (3066 and 4646). Soon there will be a new RFC which will encompass ISO 639-3 codes. With this new RFC the following language identification will be possible (http://www.sasakiatcf.com/felix/lta/language-tags/q?input=ksh) that is, the identification of the Kölsch dialect via an ISO 639-3 code on the Web. "The Web" includes technologies like HTTP, XML or RDF, which all will allow for using the new RFC.

From the language resources perspective, this is not exciting news. After all, these codes have been available before. The interesting task which now lies ahead of us is to work on bridging the language resource information space and the "Web" information space, by finding answers to questions like:

- What is the current state of language identification on the Web and in various communities like language resource users, libraries, terminology management, etc.?
- In what cases is a straightforward, round-tripping comparison of language identifiers for language resources and for web resources possible?
- (How) can we work towards an integration of the various sets of language identifiers, as e.g. envisaged by http://www.lingvoj.org/ ?

Especially the last question should make clear why I am so excited about the statement from Dan Tufiş, because it makes clear that there is a critical mass emerging within the various communities to find answers to these questions, with language identification being just one example of common infrastructure necessary for a truly multilingual, community integrating web. C



Snapshot of linkages between data repositories from Linked Data community

References

Page about language tags http://www.langtag.net/

Current RFC for language identification http://tools.ietf.org/html/rfc4646

Draft of its soon to be approved successor http://tools.ietf.org/html/draft-ietf-ltru-4646bis-23, including ISO 639-3 language codes

Tool for checking language tags with the new new RFC http://www.sasakiatcf.com/felix/lta/language-tags/

About the author

Felix Sasaki has studied Japanese and Linguistics, wrote a PhD about web technologies and their application for language resources, and has worked 4 years at the World Wide Web Consortium, mainly in the area of Internationalization. He currently holds a professorship at the University of Applied Sciences Potsdam, Germany. Felix is also still engaged in W3C and IETF standardisation work, and is the head of the German-Austrian W3C Office located in Potsdam.

Belgium, CLARIN-Vlaanderen and the Dutch language

Ineke Schuurman
Centre for Computational Linguistics (CCL), Katholieke Universiteit Leuven

If you take a look at the last page of this newsletter, it is Belgium, the country, that is involved in the European CLARIN project. But in Belgium, a federal state, academic research is entirely handled at the level of the communities, the Flemish Community and the French-speaking Community. Roughly, the Flemish Community coincides with the region of Flanders, the northern part of the country, while the French-speaking Community roughly coincides with the southern part: Wallonia. In the bilingual town of Brussels, the public institutions (schools, universities, museums and the like) are shared between the two communities.

The Flanders side of CLARIN

This article reports on the CLARIN related activities in the Flanders (Vlaanderen) part of Belgium. The financing body is EWI, the Flemish department of Economy, Science and Innovation.

The groups involved in CLARIN-Vlaanderen are based in Flanders and Brussels. They are all members of CLIF, the Flemish research network for Computational Linguistics, Speech and Language Technology (http://clif.esat.kuleuven.be), funded by the Flemish Science Foundation (FWO). CCL (K.U.Leuven) represents CLIF in the European CLARIN project.

Dutch is the official language in Flanders, and the core of the resources and technologies contributed to the CLARIN community concern this language, or they are language-independent.

Flanders and the Netherlands joint research on Dutch

Of course, Dutch is also an official language in the Netherlands. And for many, many years there has been a close collaboration between Flanders and the Netherlands where the Dutch language is concerned. This has led to two well-known transnational, Dutch-Flemish institutions: Nederlandse Taalunie (NTU) and Instituut voor Nederlandse Lexicologie (INL), taking care of all aspects of the language (including education, literature, official spelling and the use of Dutch in the European Union).

For CLARIN, the joint transnational research programme for language and speech technology STEVIN (a Dutch acronym for 'Essential Speech and Language Technology Resources for Dutch') is of importance. Based on a basic language resources kit (BLARK), priorities for Dutch were identified, and, via a series of calls, projects were selected to create the missing resources and technology.

As a result, at this moment a whole series of tools and resources is already available, or will become available in the years to come, the latest in 2011, when the current programme ends. Two examples: 1) a corpus of Spoken Dutch (CGN: balanced, 1000 hours, ca. 9 million words, enriched with manually corrected PoS-tagging and lemmatization for the whole corpus, plus syntactic annotation for 1 million), and 2) a reference corpus of written Dutch (SoNaR: balanced, 500 million words, with for 1 million words several kinds of manually corrected annotations: PoS tagging, lemmatization, syntactic analysis, named entity and co-reference annotation, semantic role labeling, spatiotemporal annotation) and for 499 million words uncorrected PoS tagging, lemmatization and NE recognition, as well as uncorrected syntactic annotation for 49 million words. SoNaR incorporates the results of two other projecs: D-Coi and LASSY.

Other projects are concerned with an improved wordnet for Dutch; recognition of Dutch names uttered by foreigners; addition to Spoken Dutch Corpus with speech by children, elderly people and non-natives; multiword expressions; semantic overlap detection; a parallel corpus – for a full list see http://taalunieversum.org/taal/technologie/stevin/projecten/.

The tools used in all these projects will also be made available, once licensing issues have been dealt with, for academic use, and in some cases also for commercial exploitation. The Flemish-Dutch TST-Centrale (HLT Agency) has been set up for the management, maintenance and distribution of all STEVIN project results and assistance with IPR.

In the meantime, independently, in both the Netherlands and Flanders other projects are enlarging the size of the existing corpora or annotating new types of texts, showing that the STEVIN way of handling spoken and written language de facto has become a standard for Dutch. In other (planned) projects, new layers of annotation will be added to existing ones.

For example, at this very moment, in Flanders a project proposal, DAMAST, is being submitted for a 500 million word corpus, incorporating SoNaR, while adding to the core 1 million several deep semantic annotations (sentiment, opinion, rhetorical structure), plus translations in English, French and German, plus, for a smaller part, videos and eye-tracking. Keep your fingers crossed!

The next phases of Flanders CLARIN

Recently, we were informed that EWI, our financing body, in 2010 may be willing to finance some costs in an additional Dutch-Flemish CLARIN-pilot project, adapting some joint STEVIN resources to CLARIN standards (including web services, involvement of users from the Humanities and Social Sciences, etc.).

Furthermore, we have just submitted a request for funding of the next phases of CLARIN, covering construction plus the first three years of exploitation. The reports by the non-European reviewers were very positive. But, as we have 7 very strong competitors... Please keep your fingers crossed a second time!

CLARIN calendar of events

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

October 2009

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

2009-10-07 to 2009-10-09: IWPT'09, 11th International Conference on Parsing Technologies, Paris, France

2009-10-22 to 2009-10-23: The Future of the Social Sciences and Humanities, Brussels, Belgium

November 2009

2009-11-04 to 2009-11-06: INFuture2009: Digital Resources and Knowledge Sharing, Zagreb, Croatia

2009-11-06 to 2009-11-08: LTC'09: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań , Poland

2009-11-14 to 2009-11-16: 2009 Chicago Colloquium on Digital Humanities and Computer Science, Chicago, USA

December 2009

2009-12-04 to 2009-12-05: TLT8: The Eighth International Workshop on Treebanks and Linguistic Theories, Milan, Italy

2009-12-07 to 2009-12-09: eScience 2009: 5th IEEE International Conference on e-Science, Oxford, UK

January 2010

2009-01-18 to 2009-01-20: ICGL 2010: The Second International Conference on Global Interoperability for Language Resources, Hong Kong, China C

Join CLARIN

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

Members

Country; Institution; Location; Contact person

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

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

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

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

ELIS-DSSP; Gent; Jean-Pierre Martens

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

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

ESAT-PSI/Speech; Leuven; Patrick Wambacq

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

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

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

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

Institute of Croatian Language and Linguistics; Zagreb; Damir Ćavar

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

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

Faculty of Informatics, Masaryk University ; Brno; Aleš Horák

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

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

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

Society for Danish Language and Literature; Copenhagen; Jørg Asmussen

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

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

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

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

University of Tampere; Tampere; Eero Sormunen

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

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

TELMA/DIS CNRS; Paris; Florence Clavaud

CNTRL; Nancy; Bertrand Gaiffe

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

Université Paris 4 Sorbonne / CELTA ; Paris; Andre Wlodarczyk

LIF-CNRS ; Marseille; Michael Zock

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

Deutsches Forschungszentrum für Künstliche Intelligentz; Saarbrücken; Thierry Declerck

Institut für Deutsche Sprache; Mannheim; Marc Kupietz

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

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

University of Leipzig; Leipzig; Codrina Lauth

University of Stuttgart; Stuttgart; Ulrich Heid

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

University of Giessen; Giessen; Henning Lobin

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

University of Augsburg ; Augsburg; Ulrike Gut

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

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

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

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

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

Icelandic Centre for Language Technology; Reykjavík; Eiríkur Rögnvaldsson

Ireland: National University of Ireland; Galway; Sean Ryder

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

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

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

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

European Academy Bozen/Bolzano; Bolzano; Andrea Abel

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

Tilde; Riga; Inguna Skadina

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

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

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

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

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

Data Archiving and Networked Services; Den Haag; Henk Harmsen

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

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

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

Instituut voor Nederlandse Lexicologie; Leiden; Remco van Veenendaal

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

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

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

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

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

ILK Research Group ; Tilburg; Antal van den Bosch

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

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

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

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

Norwegian University of Science and Technology; Trondheim; Torbjørn Svendsen

The Language Council of Norway, Oslo, Torbjoerg Breivik

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

Poland: University of Wroclaw ; Wroclaw; Adam Pawlowski

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

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

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

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

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

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

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

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

University Babes-Bolyai; Cluj-Napoca; Doina Tatar

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

Slovenia: Josef Stefan Institute; Ljubljana; Tomaž Erjavec

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

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

Universitat de Lleida ; Lleida; Gloria Vázquez

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

Sweden: Lund University; Lund; Sven Strömqvist

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

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

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

Department of Linguistics; Göteborg; Anders Eriksson

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

Swedish Institute of Computer Science AB ; Stockholm; Björn Gambäck

Language council of Sweden ; Stockholm; Rickard Domeij

HUMlab, Umeå University ; Umeå; Patrik Svensson

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

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

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

University of Sheffield; Sheffield; Wim Peters

University of Surrey; Guildford; Lee Gillam

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

Language Technologies Unit, Bangor University; Bangor; Briony Williams

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

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