LREC 2006 Pre-Conference Workshop

Towards a Research Infrastructure for Language Resources

Workshop: 22. May 2006
Magazzini del Cotone Conference Center, Genoa, Italy
Main Conference: 24-26. May 2006

http://www.mpi.nl/lrec/2006

Background

Many teams are working hard on establishing a sound framework for eHumanities where language resources play a fundamental and enabling role both with language as object of research and language as carrier of meaning. The future researcher wants to interact with an integrated and interoperable domain of language resources that is persistent, accessible and extendable. Here, the term "language resources" is used in the more general sense, i.e. they cover data resources (texts of different sorts, annotated multimedia recordings, lexica, grammars, geographical databases etc), tools (aligners, annotators, parsers, taggers, meaning extractors etc) and knowledge sources (metadata, data category registries, relation registries and ontologies). Only a solid and sustainable research infrastructure that transcends national boundaries will help us to realize the researcher's dream. Sustainability is of crucial importance, since researchers will only invest time if they see potential benefits that last.

Many projects have been carried out at national, European and international levels that have helped us to test frameworks, to build up basic technologies, to improve standardization, to create language resource archives and to test new forms of interaction and collaboration. To just mention a few of those initiatives from the domain of language resources (not meant to be exhaustive):

- for standardization work: TEI, EAGLES, ISLE, MILE, ISO TC37/SC4
- for metadata frameworks: DC, IMDI, OLAC, MPEG7, METS
- for schemas: LMF, TIPSTER, EAF, MAF
- for knowledge representation: ISO DCR, GOLD
- for registration, integration and services: INTERA, TELRI, ECHO, DAM-LR, LIRICS

These are all built on strong international backbone network infrastructures, emerging Grid middleware and common standards and frameworks such as XML, RDF and web services. In addition we can refer to national formation processes that will form the pillars for a sustainable international research infrastructure. In Europe for example we can refer to AHDS (UK), DANS (NL), CNRS-eScience (FR) and Max-Planck-Digital-Library (D) as examples for national centers for the humanities.

ESFRI Process

In Europe the issue of pan-European infrastructures to support future eScience scenarios received increasing attention during the last year. This is mainly inspired by the goals of the European Strategy Forum on Research Infrastructures (ESFRI) to establish a priority roadmap for infrastructures. Many disciplines are currently in the process of designing and organizing for research infrastructures that are seen as mature enough to be funded. Based on the experience we

have gained over many years with the language resource community and based on the current existing national infrastructure situation we concluded that the Language Resource and Technology Community is ready to establish such a solid research infrastructure. This is the reason why the CLARIN initiative (Common Language Resources and Technology Infrastructure, http://www.mpi.nl/clarin) was formed, covering institutions from almost all countries in Europe, CLARIN intends to apply for funds in the 7th Framework Program of the EC. It is obvious that Language Resources and Technology have to offer services to the humanities disciplines as well and perhaps even beyond. This is the reason that CLARIN has to synchronize with other initiatives with a broader scope such as EROHS (http://www.portedeurope.org/IMG/pdf/Projet_EROHS-ESFRI.pdf) and DARIAH. Also the European Science Foundation started an initiative focusing on establishing research infrastructures called HERA (http://www.esf.org/esf_genericpage.php).

Language Resource Centers

Language resource centers are the key building blocks for such research infrastructures. They can be digital archives that, by their nature, should be based on principles and technologies that enable accessibility and sustainability such as: (1) Web-accessible metadata standards for resource management and cataloguing (2) Separation of the mutable physical structure from the logical one relevant for researchers; (3) Preservation of bit-stream representations by regular migration to new technology and by distributing them; (4) Facilities to allow interested and qualified researchers to add new data or upload new versions of existing data; (5) Easy and flexible user access to the resources; and (6) Utilization frameworks that take into account the heterogeneity of the resources in terms of linguistic data types, structural differences and differences in linguistic terminology. But there can be other centers that maintain registries of useful components, schemas and tools.

All centers that can play a role here should also share some basic organizational characteristics: (1) they have to be embedded in national research strategies for the humanities; (2) they have to commit themselves to offer stable services and (3) they must be willing and able to act as partners in international scenarios. The latter includes the need to define the organizational, legal and ethical basics of federations. Recently, the partners of the DAM-LR (Distributed Access Management for Language Resources, http://www.mpi.nl/dam-lr) project which is building a federation of archives based on typical Grid components took the initiative to create the Live Archives document (http://www.mpi.nl/dam-lr). It summarizes the principles that should guide the work of Language Resource Archives and received already broad support.

International Networking

The Language Resource and Technology community can also refer to several networks of relevant international collaborations such as TEI, ACL, COCOSDA, DELAMAN, OntoLex, ISO TC37/SC4 and many others guaranteeing that the development of standards and technology is broadly discussed.

Goals

As well as addressing questions as to what the organizational pillars of research infrastructures and the exact identity of federations of language resource centers and archives might be, the workshop will discuss and share information about technologies that can help in setting up and managing large research infrastructures for language resources. All technologies that are important and currently being tested out in European or international projects should be critically discussed to understand their potential and state of maturity. Some time will also be devoted to discussing roadmap issues.

Programme

The workshop offers an interesting programme with a mix of invited and submitted papers. There are contributions from European and international colleagues concentrating on political/organizational and there are more technological oriented papers. The programme will end with an open discussion about the next steps for the CLARIN initiative where also all sorts of related aspects can be discussed. S. Krauwer, T. Varadi and P. Wittenburg who mainly pushed the CLARIN work in collaboration with M. Everaert will be open for all kinds of comments and questions.

After the workshop there will be a closed meeting of all registered CLARIN members. Those who are not yet registered could either talk with one of the three CLARIN coordinators or one of the already registered members about the terms of becoming a member.

Organizers

Peter Wittenburg Max-Planck-Institute for Psycholinguistics, Nijmegen, Netherlands
Remco van Veenendaal Dutch Institute for Lexicology (INL), Leiden, Netherlands
Heidi Johnson AILLA, Texas University, Austin, USA
Linda Barwick PARADISEC, University of Sydney, Australia

Program Committee

Victoria Arranz ELDA, Paris
Linda Barwick Paradisec, U Sydney
Jeannine Beeken TST Center – INL, Leiden
Hans Bennis Meertens Institute, Amsterdam
Steven Bird U Melbourne and U Pennsylvania
Daan Broeder MPI for Psycholinguistics, Nijmegen
Lou Burnard Oxford University Computing Services
Nicoletta Calzolari ILC, Pisa
Khalid Choukri ELDA, Paris
Helen Dry E-Meld, LinguistList, Michigan
Maria Gavrilidou ISLP, Athens
Gary Holton U Alaska, Fairbanks
Michel Jacobson LACITO, Paris
Heidi Johnson AILLA, Austin
Peter van der Kamp Institute for Dutch Lexicology, Leiden
Boyd Michailovsky LACITO, Paris
Richard Moyle AMPM, Auckland
David Nash AIATSIS, Canberra
David Nathan ELAR Archive, SOAS, U London
Nelleke Oostdijk CLS, Nijmegen
Stelios Piperidis ILSP, Athens
Laurent Romary LORIA, Nancy
Florian Schiel BAS, Munich
Gary Simons SIL International, Dallas
Sven Strömqvist Linguistic Department, U Lund
Nicholas Thieberger PARADISEC, Melbourne
Remco van Veenendaal TST Center – INL, Leiden
Peter Wittenburg MPI for Psycholinguistics, Nijmegen
Martin Wynne Oxford Text Archive, UK

Programme

9.00	P. Wittenburg (MPI for Psycholinguistics, Nijmegen)	Welcome and Introduction to Research Infrastructures
9.10	T. Varadi (Hungarian Academy of Sciences, Budapest)	Putting Language Resources Infrastructure to the Test: the ESFRI Challenge
9.35	M. Theofilatou (EC, DG Research Brussels)	Research Infrastructures and FP7
10.10	S. Furui (Tokyo Institute of Technology, Japan)	Research Infrastructures for Systematization and Application of Large-scale Knowledge Resources
10.35	L. Barwick (Sydney University and DELAMAN)	Research Infrastructures – the Australian Perspective
11.10	Coffie break	
11.30	M. Wynne, Sheila Anderson (AHDS, London)	The Arts and Humanities Data Service: research infrastructure in the UK
11.50	N. Calzolari (ILC-CNR, Pisa)	Community Culture in Linguistics – an international perspective
12.10	B. Maegaard (CST, Uiversity of Copenhagen)	Organization Models for RI and existing Infrastructures
12.30	L. Romary, G. Francopoulo et al (ISO TC37/SC4)	The relevance of Standards for RI
13.30	Lunch Break	
14.30	D. Nathan, et al (DAM-LR Project, SOAS London)	Foundation of a Federation of Archives
14.50	D. Broeder et al (DAM-LR Project, MPI Nijmegen)	Integrated Services for the Language Resource Domain
15.10	A. Yli-Jyrä (Helsinki University)	Common Infrastructure for Finite-State Methods and Linguistics Descriptions
15.30	A. Itai (Israel Institute of Technology, Haifa)	Knowledge Center for Processing Hebrew
15.50	O. Streiter et al (?)	Design Features for the Collection and Distribution of basic NLP Resources for the World's Writing Systems
16.30	Coffie Break	
17.00	S. Krauwer, T. Varadi, P. Wittenburg	CLARIN – the next steps
17.30	Panel Discussion	Next Steps Towards a Research Infrastructure for Language Resources
18.30	Closed CLARIN Meeting (only for delegates of member institutions)	
19.30	End	