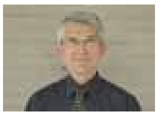
Number 11-12, 2010, September-December

CLARIN seen from the other side of Atlantic



Brian MacWhinney
Carnegie Mellon University
Pittsburgh, USA

I would like to thank the CLARIN group for allowing me to participate, albeit at a distance, in the process of developing the computational infrastructure for language studies. Work now being conducted under the CLARIN banner both complements and supplements work done in North America through the Linguistic Data Consortium (LDC, www.ldc.upenn.edu) and the TalkBank Project (talkbank.org). As the director of the TalkBank Project, I would like to outline ways in which collaboration between CLARIN, LDC, and TalkBank can benefit each project and the research communities we are serving.

It is interesting to compare and contrast the approaches to infrastructure building taken by CLARIN, LDC, and TalkBank. CLARIN has worked to bring together widely divergent projects across more than a dozen European nations. In this vein, the CLARIN mission statement notes that its primary goal is, "to turn existing, fragmented technology and resources into accessible and stable services that any user can share or adapt and repurpose." Traditionally, work in the European tradition has focused on the building of tools and services that can then be deployed across the many European languages. These tools are often developed through work groups, such as CLARIN, that hammer out consistent standards for metadata descriptions, annotation, and analysis tools. Although this process can sometimes seem ponderous, it has the great advantage, particularly in the European context, of developing momentum behind a shared consensus regarding standards. One clear example of this progress is the ongoing development of the IMDI metadata coding system and its web-based implementation through Arbil and the IMDI Browser. (corpus1.mpi.nl). Although this system requires more careful coding than OLAC (www.language-archives.org), it offers greater detail for projects that wish to use a greater array of metadata. Moreover, can be linked in useful ways to further tools for data analysis.

On the North American side, the two major sites of infrastructure development have been LDC and TalkBank. Although these two systems have often collaborated on specific projects (SBCSAE, SCOTUS, AG-Toolkit), they take markedly different approaches to data encoding. LDC has focused its efforts on the archiving and publication of a large number of corpora that use very different data encoding standards. TalkBank, on the other hand, has focused on the development of a database of transcriptions that adhere to a tightly-defined annotation system, called CHAT which is accompanied by a detailed XML specification and validator (talkbank.org/software). The advantage of the approach taken by LDC is that a wide variety of data can be more readily ingested preserving the original format. An advantage of the TalkBank approach is that data can be analyzed more uniformly with a single set of tools. Another advantage of the TalkBank approach is that the standards can allow fields, such as child language (CHILDES), aphasia (AphasiaBank), or second language acquisition (BilingBank), to enforce standards regarding data quality and analysis, sometimes using standardized data collection protocols.

Both LDC and TalkBank have emphasized the importance of providing maximally open access to language data. TalkBank provides totally open access to nearly all of its database. Moreover, TalkBank transcripts linked to media can be played back directly from the browser interface (talkbank.org/browser). The strategy of maximizing open access to materials has played an important role in generating over 4500 published articles based on use of TalkBank data. Happily, this policy of promoting open access is now being forcefully supported by recent initiatives at the ESF, NSF, NIH, and LSA that support and often mandate open, public sharing of research data.

Given recent advances within CLARIN, LDC, and TalkBank, we can begin to see a variety of ways in which these projects could usefully collaborate. Let me provide three concrete examples.

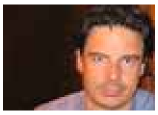
o I have been working with colleagues in Denmark to build up the DK-CLARIN database for Danish spoken and written texts. Because many Danish researchers are particularly interested in Conversation Analysis (CA), we extended the CHAT system to provide a full range of support for CA transcription. To provide a metadata access to this database, we are using the IMDI system. During the process of integration into IMDI, we developed tools that to automatically construct IMDI metadata from CHAT files not just for the Danish corpora, but for all of the TalkBank and CHILDES corpora. Using these tools, we have now incorporated a complete mirrored copy of these corpora inside the MPI repository at (corpus1.mpi.nl).

o We have developed a set of programs designed to convert to and from CHAT and other popular formats such as ELAN, Praat, EXMARaLDA, WaveSurfer, and Transcriber.

o On the higher level of cross-disciplinary cooperation, TalkBank, LDC, and CLARIN have all begun to explore the construction of a shared cyberinfrastructure for the social sciences and humanities. In 2009, I chaired an NSF panel that issued a report on this subject that can be located at (talkbank.org/dreams). This report is very much in line with the new DASISH initiative in which CLARIN will be participating.

The methods and goals of LDC, TalkBank, and CLARIN are moving quickly toward convergence. Hopefully, we can develop methods for building up solid links between these projects, as they seek to integrate themselves into the wider cyberinfrastructure for the sciences. C

Editors' Foreword



Marko Tadić
& Dan Cristea
CLARIN Newsletter editors

Dear readers,
this double issue of CLARIN Newsletter according to the CLARIN project Description of Work was planned to be the last one. It was our intention to have two separate issues, number 11 at the end of 2010 and number 12 that would step into 2011 and thus accomodate the six month prolongation of the project and yet remain within the boundaries of project.

Due to a series of unwanted events, it turned out that our initial editorial plan had to be adjusted, so you have in front of you the double issue 11-12 that covers the planned period by dates, but certainly not by the time of its appearance.

Therefore we decided that CLARIN project, its consortium partners and the community that has built around them in previous three years deserve a proper final and closing issue. We hope that number 13 will not bear any trace of misfortune. On the contrary. We believe that it will pave the way to CLARIN infrastructure in its full form.

But about the next issue you will be able to read in the next editorial. Let's concentrate on this one that is in front of your eyes.

How CLARIN is perceived from the other side of Atlantic you can read in the front page contribution by Brian MacWhinney whom, we hope, we do not have to introduce.

How LT scene, after some time, has become more vigorous again and how research infrastructures with their expected and planned sustainability has become important, you can find out from the Memorandum of Understanding between CLARIN and META-NET. We publish it in its entirety because we believe this gives the long awaited opportunity to LT community to coordinate its efforts at large.

Peter Wittenburg is presenting several possible offsprings of the projects CLARIN and DARIAH within the new call for RI projects. We will certainly hear more about them in the issue to come.

One of the most notable use cases of usage of LT in the humanities, i.e. CLARIN-supported project about the analysis of folk tales is presented by Piroska Lendvai and Thierry Declerck. This topic has been presented at several digital humanities conferences and is rising the interest ever since.

The demonstrative combination of textual and geographical analysis of 17th century manuscript of Romanian Nicolae Milescu's *Iter in Chinam*, is another case that clearly shows how digital humanities depend on LT but also how LT has to be combined with other types of information.

Our middle pages are traditionally oriented to presentation of important events connected to CLARIN. First we have the report by Hetty Winkel from SDH-NEERI 2010 conference that took place in Vienna in October and is to be considered the major event in joint organisation of CLARIN and DARIAH.

It is followed by reports from two LT conferences that embraced Europe from two sides, South-East and North. These are Formal Approaches to South-Slavic and Balkan Languages (FASSBL7) and the fourth Baltic HLT conference. These two show how LT has spread accros the Europe and is growing mature at the regional and not just national level.

The first META-FORUM is presented by Aljoscha Burchardt and Georg Rehm in its full strenght since after the LREC2010 conference, it has been the LT event in Europe that collected the largest number of participants.

Our issues regularly end with reports on the status of LR&T from different European countries. There is no need to drop this practice, so we are bringing you the reports from four countries in this double issue: Israel, Iceland, Turkey and Slovakia. Each of them depicts the different situation and level of development of LT, but what can be noticed is that all these efforts are oriented to a common goal. If we contributed so far to this common goal with our editorial work on the previous issues of CLARIN Newsletter, we certainly hope that we did not disappoint you with this one either.

Enjoy your reading!

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.

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

Memorandum of Understanding between CLARIN and META-NET

An important agreement between two key initiatives in LT community

Introduction

The Network of Excellence "Technologies for the Multilingual European Information Society", henceforth META-NET, established by DG INFSO of the European Commission (EC) under Framework Programme 7, and the "Common Language Resources and Technology Infrastructure", henceforth CLARIN, a consortium established under the Research Infrastructures Programme of the EC, recognise the complementarity of their objectives and declare their intent for multi-level collaboration.

Scope and Objectives of the Participants

CLARIN is a research infrastructure in the ESFRI framework. It constitutes a large-scale pan- European collaborative effort to facilitate research by coordinating and making existing language resources and tools available and readily useable for the Social Sciences and Humanities (SSH) on a sustainable basis. CLARIN offers resources (data and tools) and services to allow computer-aided language processing, addressing one or more of the multiple roles language plays (e.g., carrier of cultural content and knowledge, instrument of communication, component of identity and object of study) in the Humanities and Social Sciences and neighbouring disciplines in the broadest possible sense. By doing this CLARIN will provide an advanced and innovative environment for e-Research and e-Science which will make language resources and technologies (LRT) visible, accessible and interoperable based on portals and standards.

META-NET is a Network of Excellence dedicated to the technological foundations of the European multilingual information society. By developing a long demanded open resource exchange and sharing facility code-named META-SHARE, it will create the basis for developing the necessary technologies and applications for the multitude of European and other relevant languages. By forging an alliance of researchers, technology providers, corporate users, language professionals and other stakeholders and by developing together with these partners a shared vision and a strategic research agenda, META-NET shall prepare an ambitious joint effort needed for realizing Europe's digital single market and information space. By building bridges to neighbouring technology areas, META-NET will approach open research problems in collaboration with other fields such as machine learning, social computing, cognitive systems, knowledge technologies and multimedia content.

Orientation and Target Communities of the Participants

CLARIN is oriented towards the deployment and adaptation of language resources, tools and technologies in order to offer language-related services primarily to the SSH research community (e.g., scholars and researchers in history, anthropology, sociology, linguistics, literary studies, computational linguistics, law, etc.). CLARIN aggregates language data and tools relevant to a wide range of languages, including endangered languages, and to content that has its own properties and processing requirements, be it in the form of text, speech or other modalities. Catering for this community entails offering services to researchers who may have limited IT skills or interest, and therefore require highly automatic processing of data and content. CLARIN anticipates the need of modern research to build interdisciplinary virtual communities working on virtual collections and building virtual workflows to tackle their research questions.

CLARIN is in the process of establishing itself as an ERIC that operates a federation of (interconnected) national CLARIN Centres. They offer language resources and services to the whole European Social Sciences and Humanities communities in a persistent and sustainable way mainly guaranteed by strong repositories. The right to decide about financial and licensing conditions will always remain with the owners, but CLARIN favours and actively promotes a free and open access and open source policy. It builds on existing LRT centres (including repositories, service centres and centres of expertise) and infrastructure initiatives and plays a pioneering role in the emerging European ecosystem of infrastructures, thereby significantly strengthening the European Research Area.

META-NET aims at supporting Human Language Technology (HLT) development, i.e., technologies underpinning language-savvy products and services for the digital information society and a single online market. META-NET focuses primarily on EU languages (official, national but also regional ones) and the languages of EU's major partners, i.e., more on contemporary business and everyday communication language (broadly conceived). All language-enabled communication modalities, i.e., text, speech, still and animated images, sign languages, etc. play an important role. Equally important is the cross-lingual dimension for multilingual Europe.

META-NET aims at creating a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources. Data and tools can be both open and with restricted access rights, free and for-a-fee. This network of repositories, META-SHARE, targets existing but also new and emerging language data, tools and systems required for building and evaluating new technologies, products and services. In this respect, reuse, combination, repurposing and re-engineering of language data and tools, play a crucial role. META-SHARE will eventually be an important component of a language technology marketplace for HLT researchers and developers, language professionals (translators, interpreters, etc.), as well as industrial players, especially SMEs, catering for the full development cycle of HLT, from research through to innovative products and services. To achieve its goals, META-NET intends to play a central role in building an ecosystem with collaborating projects tackling different

Continued on the next page ☞

facets of language resources, technology development and evaluation.

The establishment of such a network may include a joint effort between the European Commission and the Member States through available adequate instruments.

Conclusion

CLARIN and META-NET are two complementary initiatives, serving different communities with different but harmonizable goals. They have identified shared interests and approaches with respect to methods and instruments in the area of language resource infrastructures concerning both technical solutions and organizational models.

Both CLARIN and META-NET aim at describing language resources with appropriate metadata, as well as their sharing, maintenance and reuse. While the fact that the two projects target different user communities may call for different metadata profiles, CLARIN and META-NET envisage an open metadata domain of data and tools, based on interoperable metadata descriptions for categories of data and tools relevant to their joint user communities. The purpose of **interoperable metadata descriptions** is to enable cooperation, the possibility to link data objects and/or their descriptions, the exchange of data of interest to both communities and the way towards **fostering standardisation** and interoperability of data and tools, wherever possible and necessary. Both parties declare their intent to pursue this cooperation and to build upon their deliberation on an agreement concerning existing and emerging standards.

At the time of signature, CLARIN and META-NET intend to collaborate in the setup of language resource repositories, resource identification procedures, metadata infrastructure, as well as legal issues in sharing and distributing language resources.

The parties also agree that while they retain their integrity and pursue their goals, they will explore the possibility to give access to data objects residing in their respective repositories. The parties agree to prepare, by September 2010, an initial list of concrete areas and mutual objectives on which to collaborate, and plans for implementing this.

Signed on behalf of CLARIN
Steven Krauwer, Coordinator of CLARIN

Signed on behalf of META-NET
Hans Uszkoreit, Coordinator of META-NET

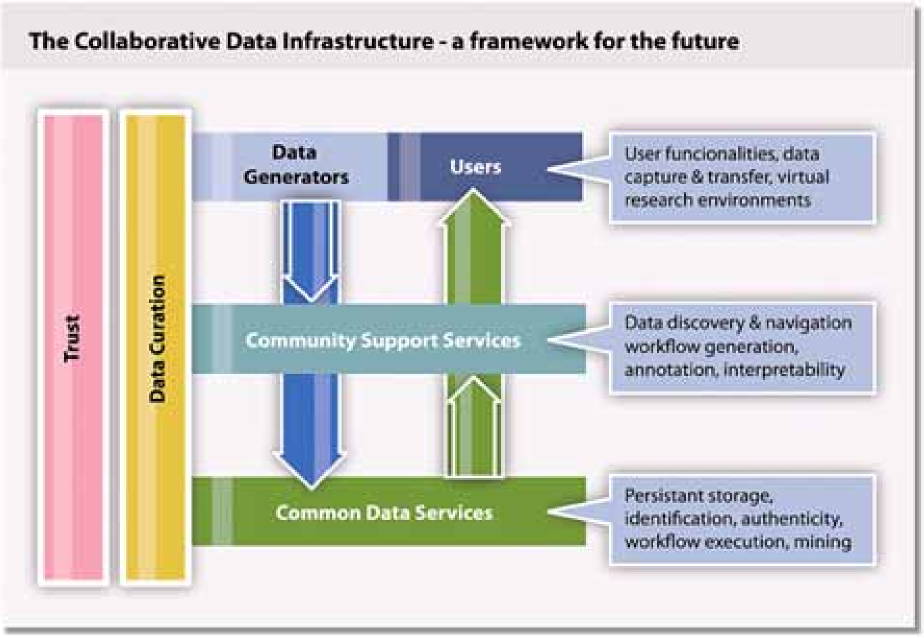
Peter Wittenburg
MPI, Nijmegen

The preparatory phase of the ESFRI research infrastructure projects is close to its end and this also holds for CLARIN. As it looks at this very moment CLARIN seems to be successful in so far that about 18 countries have signaled their interest to sign the memorandum of understanding to participate in the ERIC and there are a few additional countries from which we assume that they will participate, but yet did not finish the national roadmap discussions completely. Of course the CLARIN Executive Board and the colleagues from the national teams have done their homework at several layers. In this article we do not want to explain the CLARIN internal activities – at EU and national level – but want to describe the activities to embed CLARIN in a variety of activities. Yet we cannot say whether all these activities will lead to success, but yet we are rather optimistic.

Data Infrastructure

One CLARIN expert was appointed to become a member of the high level expert group of the EC to work out a strategy for the management of Research Data in the coming decades. The report with the name “Riding the Wave”[1] was handed over recently to commissioner Neelie Croes. One result of the report was the insight that organizing data management and access will be a multilayer task as illustrated in the figure. At the second layer there would be community oriented centers employing experts with decent knowledge about the workflows, formats, semantics and encoding principles of the respective discipline. At the third level we would have common data services which are widely discipline independent and given by large data centers. It is obvious that data curation is a task for all actors starting with the researchers as data creators. The smooth functioning of such a multilayer system can only be guaranteed if mutual trust has been established.

Since CLARIN has a very explicit strategy about data and technology centers as the backbone of a research infrastructure from its beginning, CLARIN was a can-

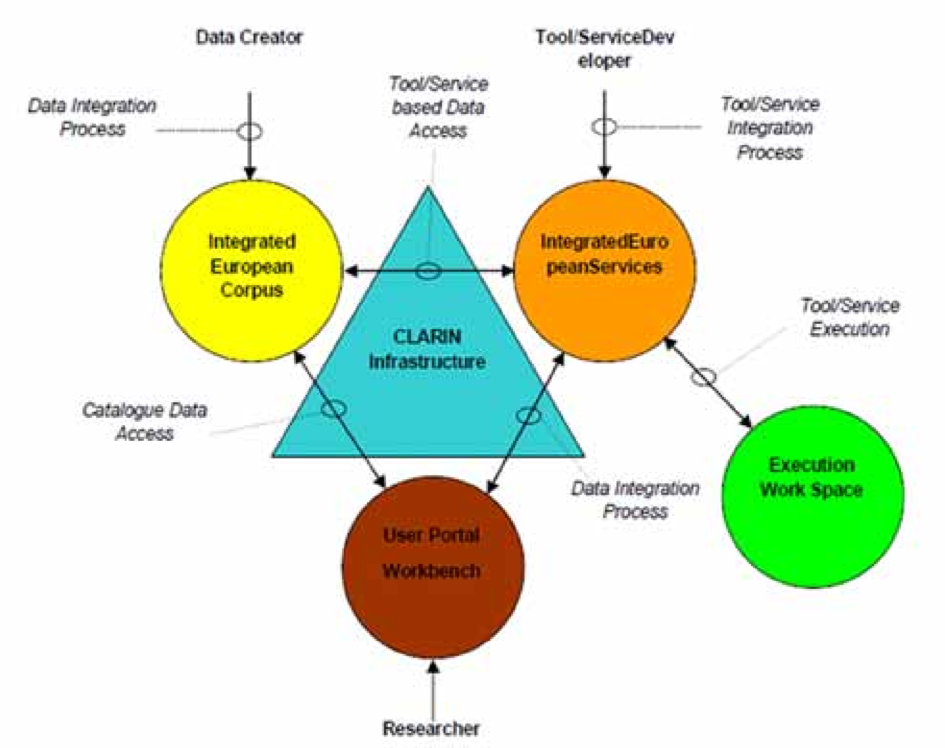

The Collaborative Data Infrastructure - a framework for the future
Trust
Data Curation
Data Generators
Users
User funcionalities, data capture & transfer, virtual research environments
Community Support Services
Data discovery & navigation workflow generation, annotation, interpretability
Common Data Services
Persistant storage, identification, authenticity, workflow execution, mining

Management of data in the collaborative scenario as described by the High Level Expert Group on Scientific Data

Embedding of CLARIN Future

Towards the end of preparatory phase of the ESFRI research infrastructures

didate community to participate in the EUDAT proposal to the EC. EUDAT is a broad consortium existing of strong centers from 15 communities from the various discipline areas on the one hand and 11 of the strongest data centers from various countries on the other hand. By bringing together such a variety of different actors it wants to establish exactly the Collaborative Data Infrastructure as indicated in the figure which is, given the heterogeneity, a challenging task. CLARIN will participate in three ways if the proposal will be accepted: (1) A CLARIN member would act as Scientific Coordinator; (2) It would participate to establish a safe data preservation infrastructure; (3) It would participate to test out a generic web services landscape where large data centers will host the services.



Thus with its participation in EUDAT CLARIN showed its awareness about the need to adopt generic solutions where possible. The following CLARIN centers were accepted to join the consortium: CU Prague, U Tübingen, MPI. Partly the already started collaboration between CLARIN and DEISA[2] in the REPLIX[3] project can be continued. EUDAT will be coordinated by CSC[4] – the Finnish national computer center. Very much related to the EUDAT project is the German Radieschen project which aims at working out a roadmap towards a national data infrastructure. This has already been granted.

Humanities and Social Science Cluster

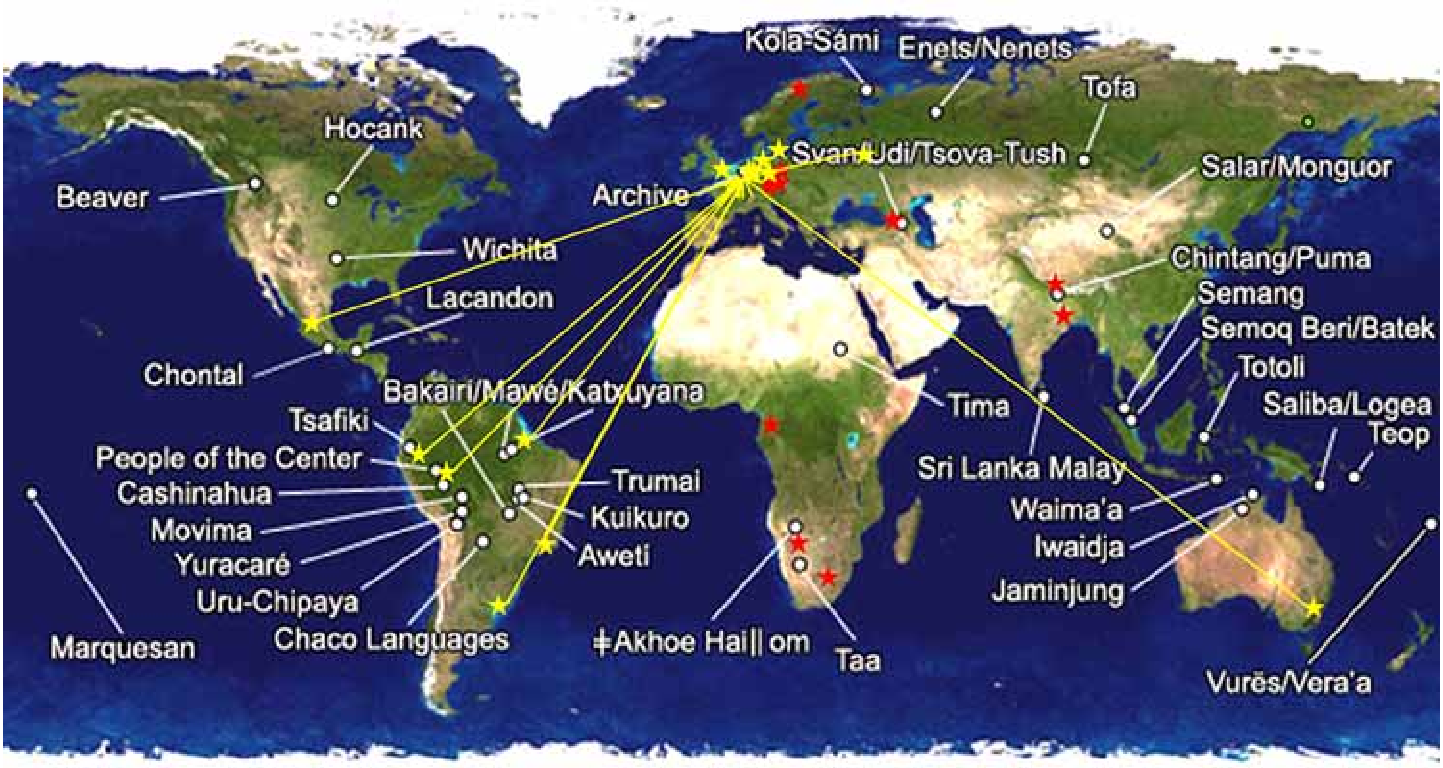
The EC launched a call for a joint activity between the 5 existing humanities and social sciences research infrastructures initiatives (DARIAH[5], CLARIN, CESSDA[6], ESS[7], SHARE[8]) to look for commonalities and to exploit synergies. A board with one representative per initiative was built to work out tasks which are of interest at least for a few of the initiatives and to establish principles for the consortium building. Accepting that DASISH (Data Services Infrastructure for the Social Sciences and Humanities) also will contribute to the eco-system of infrastructures, we finally worked out the following major tasks: (1) getting a deep understanding about data organizations in the disciplines and designing a roadmap towards a higher degree of harmonization; (2) Improve the quality of the social science surveys; (3) offer robust deposit and long-term preservation services; (4) improve metadata quality and the infrastructure integration with respect to AAI (authentication and Authorization Infrastructure), PID (persistent identifiers) and metadata; (5) improve the visibility of all tools and services; (6) develop a cross-disciplinary annotation platform useful for all SSH researchers; (7) improve the legal and ethics situation and (8) carry out lots of activities in education and training.

It was agreed that maximally 5 partners per initiative could join the consortium. CLARIN is represented by UPF Barcelona, U Tartu, U Copenhagen, U Bergen and MPI following a number of published criteria and will participate in all activities except (2) which is only of interest for the social scientists. DASISH will be coordinated by U Gothenburg and MPI will coordinate the research activities.

CLARICLE

The EC launched another call devoted to foster the construction work of CLARIN. Central for CLARICLE is the creation of the Integrated European Corpus and the Integrated European Services making use of the CLARIN research infrastructure. Thus the CLARICLE intention is to fill the infrastructure with useful data and services and to make these available to the users with the help of a user workbench that offers easy access to all resources, services/tools

Continued on the next page ☞



and to other useful infrastructure services. The major challenges were to define the content areas for the Integrated European Corpus and to see who of the CLARIN centers can contribute with accessible resources and services/tools that fulfill the CLARIN requirements widely. CLARICLE also has the intention to act as part of the emerging eco-system of infrastructures and wants to collaborate on execution work spaces with strong data centers.

The following major work dimensions have been identified for CLARICLE:

1) standards,

2) IPR issues;

3) building the integrated corpus;

4) building the integrated services domain;

5) building cross-domain services;

6) adapt the infrastructure;

7) carry out a user survey and launch calls for small project proposals;

8) develop a workbench, annotation and visualization tools.

The integrated corpus will address in particular the needs of the following communities: (a) linguistic variation research based on written, spoken, multimodal and sign language resources; (b) news corpora relevant for historians, social scientists and linguists; (c) recordings and proceedings of parliament sessions and (d) oral history.

The consortium includes 19 CLARIN members from a variety of countries.[9]

INNET

Another small proposal with 4 partners was submitted separately to ask for support for the world-wide activities in documenting and archiving endangered languages. Mostly it is not known even to linguists that from the 6500 languages still spoken every week one is dying and with it an enormous knowledge about linguistic systems, environmental aspects, etc. In the realm of the DOBES project[10] the MPI has set up 13 remote repositories (see stars on the map) at various places world-wide so that a data exchange can be carried out. We have requests for 10 additional repositories and the wish to have regular and close interactions between the various teams.

The proposal thus includes funds for setting up such repositories, for local training courses, for regular workshops and conferences to foster an exchange about methods and technologies and for some educational effort. The intention is to fully adopt all CLARIN requirements in this international exchange. From CLARIN MPI, HAS-RIL and U Cologne (CLARIN D) are involved, since they have a close relationship to the endangered languages topic.

Summary

We can say that CLARIN was very active in planning its construction phase and in bringing services to users, in participating in larger European activities to exploit synergies and in activities to spread CLARIN methods even worldwide. Synchronizing with all participants about the many details and getting all these proposals finished was a very time consuming job during the last half year in 2010. Finally between 25 and 30 CLARIN members are involved in these activities. We could see that for most of these activities the state of the centers and the quality of their services was a crucial criterion. We assume that in future similar calls will be launched, since building research infrastructures is a strategic goal in Europe. Yet we do not know the outcome of the evaluation process. C

References

[1] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[2] http://www.deisa.eu/

[3] http://www.mpi.nl/research/research-projects/the-language-archive/projects/replix-1/replix

[4] http://www.csc.fi/english/pages/parade

[5] http://www.dariah.eu/

[6] http://www.cessda.org/

[7] http://www.europeansocialsurvey.org/

[8] http://www.share-project.org/

[9] During the editing of this issue of CLARIN Newsletter we received the announcement from the evaluation committe that CLARICLE did will not be funded (editor's remark).

[10] http://www.mpi.nl/DOBES .

The CLARIN Folk Tale Use Case at DH-2010 and Elsewhere

Deploying language resources and technology to enable e-Humanities

Piroska Lendvai
HASRIL, Budapest, Hungary
Thierry Declerck
DFKI, Saarbrücken, Germany

In line with one of CLARIN's main goal of deploying language resources and technology to enable e-Humanities, recently a use case has been established, dedicated to the automated linguistic and semantic annotation of folk tales. Emerging from the AMICUS project's research and networking activities (Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts[1]), CLARIN has established special cooperation with D-SPIN[2], the German contribution to CLARIN, whereby the use case is currently under implementation.

The Folk tale use case aims at establishing a combination of linguistic and domain-specific content descriptors from the fields of Literature, Ethnography, and Folklore, offering semantic annotation to describe and model the complex interdependencies between a tale's characters, function roles, and events, and their designated linguistic vehicles. The basic linguistic processing chain is enabled by the WebLicht web services framework (D-SPIN project); its output is being mapped onto TEI[3] and ISO[4] standardized annotation structures. The semantic annotation layer is based on various generic semantic resources (for example a family ontology, a temporal ontology, WordNet[5] and FrameNet[6]) and a specialized annotation schema that originates from a proposal by Vladimir Propp, based on literary and ethnographic studies (Propp 1968). The linguistically and semantically annotated tales are to improve querying and corpus-based research by literature scientists and other Humanities specialists, as well as by laypersons.

Drawing on the connection with the AMICUS project, we established a broader set of cooperation partners, prominently the Swedish School of Library and Information Science — providing expertise on the topic of motifs across various literary genres —, the Universidad Complutense de Madrid — contributing ontological resources like ProppOnto and ProtoPropp, the Fairy-Tale Generator —, and the University of Pittsburgh: creator of the Proppian fairy tale Mark-up Language (PftML). On the basis of their resources, the first step taken within the use case was to develop an extended annotation schema for folk tales – APftML (Augmented Proppian fairy tales Mark-up Language), see (Scheidel & Declerck 2010), which was presented at the AMICUS workshop[7] (Vienna, 21 October 2010), a satellite event to the SDH 2010 conference[8], co-organized by CLARIN and DARIAH.

Digital Humanities 2010: the full auditorium at King's College, London

A first dissemination of our joint work succeeded at LREC 2010 (Lendvai et al. 2010a), where the motivations and research setup were presented to the language technology community. A more focused presentation, including recent developments, has been submitted successfully to the Digital Humanities conference[9] (DH2010) in London (Lendvai et al. 2010b), giving us clear indication that our work in CLARIN is going the right way: the DH conference is the annual international conference for digital scholarship in the humanities. We demonstrated our data, approach, as well as research and implementation methods there within a two-hour poster session, exchanging ideas with colleagues from different research fields. In the Digital Humanities

Continued on the next page ☞

community there is notable interest for obtaining folktales annotated with fine-grained linguistic and semantic annotation, which our use case targets to create. Such systematically grown and automatically assigned markup is currently unavailable to Humanities researchers, except for metadata supporting general textual classification. Our choice of TEI as the basic annotation scheme for textual information, likewise suggested by Laurent Romary and Andreas Witt within a D-SPIN meeting, has also been confirmed, since TEI appeared to be frequently applied and cited. We can thus report on encouraging signals received from both the HLT and DH communities with respect to the use case. An interview conducted at the end of our poster presentation at DH2010, available online, provides a short audio explanation of our work.[10]

References

Scheidel & Declerck (2010) Antonia Scheidel & Thierry Declerck. APftML – Augmented Proppian fairy tale Markup Language. In: Proceedings of the first International AMICUS Workshop.

Lendvai et al. (2010a) Piroska Lendvai, Thierry Declerck, Sándor Darányi, Pablo Gervás, Raquel Hervás, Scott Malec & Federico Peinado. Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010).

Lendvai et al. (2010b) Piroska Lendvai, Thierry Declerck, Sándor Darányi & Scott Malec. Propp Revisited: Integration of Linguistic Markup into Structured Content Descriptors of Tales. In Proceedings of DH 2010.

Propp (1968) Propp, V. A. Morphology of the Folktale. Publications of the American Folklore Society. University of Texas Press, 2nd edition.

Notes

[1] http://amicus.uvt.nl

[2] http://weblicht.sfs.uni-tuebingen.de/englisch/index.shtml

[3] http://www.tei-c.org/index.xml

[4] http://www.iso.org/iso/standards_development/technical_committees/other_bodies/iso_technical_committee.htm?commid=297592

[5] http://wordnet.princeton.edu/

[6] http://framenet.icsi.berkeley.edu/

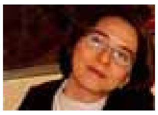
[7] http://amicus.uvt.nl/amicus_ws2010.htm

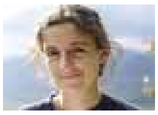
[8] http://www.dariah.eu/index.php?option=com_content&view=article&id=135&Itemid=197

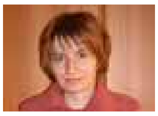
[9] http://dh2010.cch.kcl.ac.uk/

[10] Downloadable at: http://www.arts-humanities.net/audio/interview_thierry_declerck_piroska_lendvai_dh2010.

Nicolae







Daniela Dumbravă
Anamaria Ciucanu
Georgiana Cărăuşu
Faculty of Computer Science,
Alexandru Ioan Cuza
University of Iași

CLARIN represents one of the best European research infrastructures which aim is to make associations between humanistic disciplines and informatics, an increasingly visible tendency in contemporary interdisciplinary studies. For humanists themselves, one of the major challenges is the integrated articulation of the conceptual languages within the various disciplines. Therefore, large scale infrastructures are needed for publishing and sharing data; both informaticians and humanists are intrinsically involved in this process. It seems to us that the main direction of further research projects will be to place or reconstruct knowledge in large scale infrastructures. The project described below – "Nicolae Milescu's *Iter in Chinam* (1676). Visual, computational and encyclopaedic reconstructions" – embodies an enormous set of data, from the 17th century, and dedicated to Northern Asia and Beijing.

General context

Nicolae Milescu, a prominent figure of the European political and intellectual elite in the XVII century, was assigned the task of rigorously exploring the transcontinental travelling routes between Europe and China. This resulted in Milescu's *iter in Chinam* (1676) and his encyclopaedic description of the Northern Asian space, which he presented to the Moscow Parliament in 1678. The purpose of our research is to reconstruct *iter in Chinam* virtually, employing the Google Earth tools and the Natural Language Processing (NLP) technology, and integrating temporal and spatial information (GIS), as well as cartographic and encyclopaedic data. Furthermore, the digitalization of the maps which incorporate *iter in Chinam* will allow us both to analyze them comparatively and to correlate them with Milescu's description of Northern Asia. Multimedia applications will make it possible to uncover the connections between the Asian and the European space which Milescu envisaged in the XVII century, and which remain unknown to the international community, despite their scholarly importance.

The NLP Module

NLP is used in the project to extract temporal information from a corpus of texts which incorporates geographical data. Utilizing shallow-parsing and pattern-matching techniques, spatial information can be extracted from this collection of texts. The NLP module receives as input a plain text and returns a set of annotated versions of the text, one for each type or information (temporal and spatial); spatial expressions will be marked using SpaceML[1] and temporal ones using TimeML[2]. The entire annotation process is automated. Between different versions of annotation several links will be created. Employing these annotations will make it possible to extract automatically all those correlations which are emphasized in Milescu's itinerary, if we consider its spatial and temporal development. Methodologically, one can obtain spatial and temporal annotation with NLP as follows: i) identify the nominalizations of those temporal and spatial expressions which can be extracted from the text (in this case a historical corpus); ii) highlight spatial and temporal relationships between the spatio-temporal entities evident in those expressions. We hypothesize that a significant part of these spatial data and relations between them can be re-elaborated (standardized) and subsequently located into the GIS database.

The Cartography Module (CM)

From this perspective, *iter in Chinam*: A. contains a corpus of texts which lends itself to NLP processing (see *infra*). B. allows for a Google Earth interface, an interactive platform which can select, extract and visually represent the description of this *iter*. Within this

Milescu's Iter in Chinam *(1676)*

Visual, computational and encyclopaedic reconstructions

framework, the project seeks to: i) create an XML document which would emphasize the temporal, spatial, toponimic and etnonimic references and the networks for global communication in the XVII the century; ii) re-create the *iter in Chinam* route according to the characteristics mentioned at points A and B in this paragraph.

The project uses the Google Earth (GE)/Maps (GM) API's for Javascript, integrated in a C# software which stores all the information that can't be found in GE/GM. In practical terms: 1) CM receives a xml file set from the NLP Module, which then changes, depending on its category. XML's should contain data on the space, time and culture described by Milescu in his journal; 2) The changes could be, for example, from a list of places with links in between, processed by the NLP Module, into a KML (XML read by GE) with the places found in the GE database, their coordinates and a KML with the unknown places related to those previously identified. The distance will be measured in *versts* (Russian unit of length; 1 verst = 3500 feet or 1,0668 km). Therefore, the CM is responsible with creating algorithms for detecting the coordinates of the unknown locations, depending on several pieces of information extracted from text, i.e. 17th century cities, distances, directions and narrator orientation. There is an intrinsic link between space and time, as well as between rich historical data. The second set of information will be placed into info balloons, and it will constitute our specialized database. The user can also activate levels of information for buildings, vegetation, info balloons, geographical and ethnographical data. Most of this information and details which may depend on the 3D space will also be stored in a special database, in case the elements do not exist in the GE environment. Parallel activity between the 2D GM scene and the 3D GE view port would be preferred.

Expected benefits of the project

A multimedia application which would incorporate the information obtained and processed following the above mentioned practical stages of this project; a Romanian mini-corpus adnotated to spatiality, based on the aforementioned textual and cartographic sources; the elaboration of articles with an educational role and the promotion of a local media campaign dedicated to *iter in Chinam* and to Nicolae Milescu; last, but not least, the elaboration of a technology of spatio-temporal adnotation of texts and their linking with GIS visual software.

If successful, the innovative technologies developed by this project create the premises of further developing extremely diverse applications, all rooted in textual information, such as: educational: e-learning platforms (level: primary and secondary school, high-school, undergraduate, adult learning); museology: *input* in virtual scenarios for the reconstruction of knowledge simultanously derived from historical, geographical, ethnological, religious, cartographical, etc. information; editorial: the promovation of XVIIth century humanists of several cultures (South-East European, East-Central and Northern Asian).

The work described in this article is partially supported by the grant POSDRU 89/1.5/S/4944.

References

1 BOUTOURA Cryssoula, LIVIERATOS Evangelos, “Some fundamentals for the study of the geometry of early maps by comparative methods”, *e-Perimetron* 1 (1): 60-70; BALLETTI Caterina (2006), “Georeference in the analysis of the geometric content of early maps”, in e-Perimetron 1 (1): 32- 42; JEANSOULIN Robert, PAPINI Odile, PRADE Henri, SCHOCKAERT Steven (2010), *Methods for Handeling Imperfect Spatial Information*, Springer Verlag, Berlin-Heidelberg.

2 Marta Guerrero NIETO, M. J. García RODRIGUEZ, Adolfo U. ZAMBRANA, Willington SIABATO, Miguel A. B. POVEDA (2010), “Incorporating TimeML into a GIS”, IJCLA 1/1-2: 269-283.

Supporting Digital Humanities at Large

An Impression

SDH-NEERI, Vienna, October 19-20, 2010

Hetty Winkel
Utrecht University

In October 2010 CLARIN organized a conference together with our colleagues of the DARIAH project: Supporting the Digital Humanities, (SDH2010). The conference was followed by the second Networking Event for Research Infrastructures: NEERI2010. The local organization was in the hands of our CLARIN colleague Gerhard Budin and his team of the university of Vienna and the Austrian Academy of Sciences. The venue was a centrally located building of the Technical University of Vienna.

The Research Infrastructures panel participants

SDH and NEERI conference formats

SDH2010 consisted of a number of topical sessions where providers and users presented and discussed results, obstacles and opportunities for digitally-supported humanities research. Whereas the focus of SHD2010 was on the types of research made possible by research computing, NEERI2010 addressed the technical, architectural and social challenges of building the infrastructure. NEERI focused on what we share and what we can learn from each other. Examples of such commonalities are architectural issues, communication with users and integration of services and tools.

The conference was opened by Prof Dr Wolfgang Dressler. He took the place of Gerhard Budin, who unfortunately had fallen ill. Prof Dressler is well known to CLARIN, being the Austrian member in our Scientific Board. He then introduced Mrs. Barbara Weitgruber, Head of Scientific Research and International Relations of the Federal Ministry of Science and Research, who welcomed the participants on behalf of the Minister.

The keynote speech was delivered by Prof. Neil Fraistat, Director of the Maryland Institute for Technology in the Humanities (http://mith.umd.edu/) In his speech "Digital Humanities Centers as Cyber infrastructure", Neil took us on a virtual tour through his institute, introducing us to his colleagues and their daily tasks. Thus painting a picture of what a DH centre is all about. A successful center can incubate important research, foster a new generation of scholars, devise creative modes of governance, develop a variety of strategy for funding and build digital collections and tools. "Strong local partnerships" and "Centers as nexus for local and global", were some of his key phrases. There are also pitfalls: the insularity of centers, competition among centers, national boundaries, and cultural divides between language communities. He also stressed the importance of reaching out to the users: how can digital technologies facilitate a broader engagement with different publics and what are the implications for universities in the new digital media, and the implications for the humanities. Neil mentioned centerNet: a network of 100 DH centers in 19 countries, and steering committees in Asia Pacific, Europe, North America, and the UK& Ireland (http://digitalhumanities.org/centernet/). He also referred to CHAIN, the Coalition of Humanities and Arts Infrastructures and Networks. The aim of CHAIN is to support and promote the use of digital technologies in research in the arts and humanities (http://www.arts-humanities.net/chain). An interesting question from the audience was if Neil saw Digital Humanities as becoming part of the main stream of humanities research, or that it might run the risk of standing too much on its own. This was a risk he acknowledged, and we should do our best to prevent this. Another risk factor could be the pressure on young humanist researchers to publish; which would prevent them from working on these innovative research methodologies.

Spreading out to different fields

The programme then continued with parallel sessions in the fields of Archeology, Manuscript Studies, Endangered Languages and Language Variation, and Language Variation and Digital Technology. The second day started with a plenary session on Narrative Psychology, followed by four parallel sessions on Socio-economic history, Musicology, Literary history and the session on Language technologies in humanities studies. The last one was convened by Tamás Váradi, and presented the results of the Call for Humanities projects from CLARIN.

It is impossible to describe in more detail all these sessions here, so I refer to the website where the abstracts of all the presentations can be downloaded (see below).

I would just like to mention two examples. The presentation by Frans Wiering of Utrecht University on Musicology was very interesting and lively. His area is music information retrieval (MIR), and he had some

The keynote speaker Neil Fraistat

The participants of the conference in a proper ambient for digital humanities

nice examples of music to show us the role MIR can play in musicology research. His conclusion is that the benefits for MIR are not solely in supplying tools and data, but in helping the community to formulate its demands and convince them that they need the MIR tools, and more generally, an e-Infrastructure. This is exactly what our CLARIN colleagues in WP3 in the Call for Humanities projects have been dealing with, the results of which were presented in the session chaired by Tamás Váradi. The projects that were supported by CLARIN with expertise and advice, presented themselves in this session. A conclusion of Koenraad de Smedt, the coordinator of the Humanities project, was that it was a real challenge to bridge the gap between research questions and the way to apply the technology. This is an issue that keeps coming back, in this session, but also in the talk of our keynote speaker and in the panel discussion on Wednesday.

What is the message to take home?

The conference was concluded with a panel discussion on the Future of Research Infrastructure and Humanities: "'What is the message to take home"?

Peter Doorn of DARIAH convened this session, and there were representatives from the European Science Foundation (Arianna Ciula), ESFRI SSH (Peter Farago), DG-INFSO (Wim Jansen), HERA (Sean Ryder), DG Research (Lorenza Saracco, also CLARIN project officer), CLARIN (Steven Krauwer), and DARIAH (Laurent Romary).

In this brief report, I can only touch upon some of the issues that were addressed by the panel. Not surprisingly, one of them was about the users: "catch them when they are young", one of the panelists insisted: involve them at an early stage, include it the education at PhD level or sooner. Another issue was about how difficult it is to measure our impact as data infrastructure, and how we have to find ways to do this. Then there is the problem of sustainability, also at governance level. Many countries are prepared to sign an ERIC, but governments and policies change, so this might be a problem if you want stable funding. Open and sustainable access: a key issue in order to make this whole thing work. We should not try and solve this on our own. Wim Jansen mentioned a recent report of the High level group on scientific data: "Riding the Tide", which is calling for joint future thinking and action in the field of integrating data infrastructures. Sean Ryder also mentioned the need for more coordination of the funding agencies. In the European planning for new research programmes, CLARIN and DARIAH could play a role in advising on digital components for new research. Such expertise could be helpful in preparing new research programmes. In the digital e-Humanities you need to work in teams, but this is not something the average humanities researcher is used to. In this respect, Laurent Romary mentioned the example of the musicology project (see above). If we are able to find solutions for historians, archeology, musicologists etc. then we are on the right way. Then how do we commit the users? Most of the users are still language oriented people, we have to reach out to the others: historians, archeologists etc. Neil Fraistat stressed that he does not work with an infrastructure project, if the content is not there. Easy access and simplicity of use is also crucial for RI's. In CLARIN and DARIAH we want to have a knowledge sharing infrastructure, which is at least as important as the technical infrastructure. Dissemination is not enough, it should be community based and made sustainable. This should not have to cost very much. Wim Jansen has a simple approach for the use of the words RI and data: there is RESEARCH infrastructures, and there is research INFRASTRUCTURES. The first one is about producing the data and the cooperation between researchers and the facilities that are produced, and the second one is the technology which is necessary to create and develop the 1st one. RI's like CLARIN and DARIAH are more than just content, but also knowledge sharing, and should also focus on training. Lorenza Saracco again stressed the importance of open access policy.

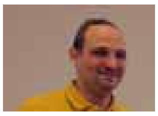
The last conference of CLARIN and DARIAH?

The above is just a brief impression of an interesting and lively conference. Discussions went on in the coffee and lunch breaks, and at the cocktail on Wednesday. And last but not least at the wonderful and animated dinner in the Melker Stiftskeller on Tuesday evening. We would again like to thank Gerard Budin, Claudia Resch, Victoria Weber, Barbara Berger and Daniel Meyrath for their great help in organizing the conference. Also our DARIAH colleagues Peter van Doorn, Milena Piccoli and René van Horik for the smooth collaboration. And of course the other members of the JOCO (Joint Organizing Committee), Martin Wynne, Tamás Váradi, and Matthew Driscoll. And last but not least the convenors of the sessions and their speakers.

Although it was the final conference of the CLARIN and DARIAH projects, we hope and expect that it will be the start for a series of e-Humanities conferences or meetings in the future. All abstracts of the sessions, both of SDH and NEERI can be found at the DARIAH website: http://www.dariah.eu/index.php?option=com_docman&task=cat_view&gid=87&Itemid=200. C

LT in the South-East Europe

FASSBL7, Dubrovnik October 4-6, 2010



Krešimir Šojat
University of Zagreb
Faculty of Humanities and Social Sciences

The Seventh International Conference "Formal Approaches to South Slavic and Balkan Languages" (FASSBL 7) was held from 4 to 6 October 2010 in the beautiful Croatian coastal city of Dubrovnik. The conference was co-organized by the Institute of Linguistics (Faculty of Humanities and Social Sciences, University of Zagreb), Croatian Language Technologies Society, the Department of Computational Linguistics of the Institute of Bulgarian Language "Prof. Lyubomir Andreychin" (Bulgarian Academy of Sciences) and the Norwegian University of Science and Technology. The biannual conference FASSBL aims at bringing together researchers dealing with all aspects concerning formal and computational approaches to South-Slavic and Balkan languages from various institutions from all over the world. This FASSBL is the seventh event in the FASSBL conference series.

Every time getting stronger

This year more than 30 researchers from Austria, Bulgaria, Croatia, France, Norway, Romania, Serbia, USA and Turkey participated at the conference. The conference was divided into 4 sections. The topics covered by the papers extend from morphology (Šnajder & Dalbelo Bašić,), syntax (Comorovski, Radeva-Bork, Vlahova & Atansov, Vučković et al.) and categorial grammar (Mihaliček) to corpus and computational linguistics (Ion et al., Koeva, Koeva et al., Mijić et al., Öztürk et al., Stoyanova, Šojat et al.) as well as machine translation (Iliev). Research infrastructures are traditionally part of the interest of many presenters so CLARIN was mentioned in quite a number of presentations since at least four CLARIN partner sites were represented at the conference.

Lively discussion over LetsMT! project poster

There were 4 invited key-note speakers: Tamás Váradi, Peter Kosta, Karel Oliva and Maria-Luisa Rivero. Let us mention briefly just the first one: Tamás Váradi from The Hungarian Academy of Sciences gave a presentation of the EU sponsored project CESAR which is to begin in the spring of 2011. The project aims at collecting and bringing into the standardised level all relevant resources from partner countries in the project: Poland, Slovakia, Hungary, Croatia, Serbia and Bulgaria (last three being covered by the conference main area of interest). We believe that this project will have a large impact to the availability of language resources and tools that have been collected and developed for quite some time for these languages and that researchers will be able to access them soon over the META-SHARE platform.

The FASSBL 7 conference was also an important event for a number of EU sponsored projects, like CLARIN, ATLAS, Let'sMT!, ACCURAT and others. The members of these project partners were having several papers in the proceedings and gave the presentations. Within the conference the exhibition of European projects dealing with LT and covering respective languages was organised where participants had the opportunity to get acquainted in more details about these projects and discuss their goals and the achieved results.

See you in 2012

The guided city tour and a trip to the most southern Croatian region Konavle were organized for the conference participants, as well as a number of other memorable social events. More information and the proceedings of the FASSBL 7 conference, (edited by Marko Tadić, Mila Dimitrova-Vulchanova and Svetla Koeva) is available at http://hnk.ffzg.hr/fassbl2010. The next FASSBL conference will be held in 2012 at the same venue in the end of September. C

Language Technology in the Baltic

Fourth Baltic HLT, Riga October 7-8, 2010

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

It is a tradition that scientists, developers and users of language technologies from the Baltic region countries meet at the Baltic HLT conference to share new ideas and recent advances in natural language processing, to exchange information and to discuss problems, to find new synergies and to promote initiatives for international cooperation.

The first larger pan-Baltic event on HLT research was the seminar "Language and Technology 2000" held in Riga in 1994. In 2004, ten years later, the first Baltic HTL conference was organized by the Commission of the Official Language of the Chancellery of the President of Latvia. The second conference in 2005 in Tallinn was organized by the Institute of Cybernetics and Institute of Estonian Language, and then the third Baltic HLT conference in Kaunas in 2007 was organized by Vytautas Magnus University and the Institute of Lithuanian language.

This fourth conference took place in Riga on October 7-8, 2010. It was jointly organized by Institute of Mathematics and Computer science (University of Latvia) and Tilde company. More than 70 participants from Estonia, France, Finland, Germany, Latvia, Lithuania and Norway enjoyed presentations covering wide range of topics, including corpus linguistics, machine translation, speech technologies, semantics, and other areas of HLT research.

The morning session of the conference was devoted to CLARIN project and overview of language technologies in Baltic countries. The session started with the invited speech by Steven Krauwer ("CLARIN – how to make it all fit together?") who has been participant and presenter in all Baltic HLT conferences. The presentation provided a brief overview of the current state of CLARIN project and described the steps to be taken, and the strategy to be adopted in order to make CLARIN happen, not as yet another project, but as a sustainable facility for the research community, based on long term commitments from the governments.

Since Baltic countries have actively participated in CLARIN project, the overview presentations of representatives from Estonia, Latvia and Lithuania provided not only an analysis of the current situation in the HLT in Baltic countries, policy, main projects and achievements, but also activities and progress in respect to CLARIN aims.

The afternoon session started with another CLARIN related invited talk by professor Kimmo Koskeniemi: "HFST (Helsinki Finite State Transducer Technology): A new division of labour between software industry and linguists". The presentation proposed an approach where software developers, on the one hand, can easily use language processing modules in their products and linguists, on the other hand, can relatively easy produce such modules for various languages and tasks.

The rest of afternoon was devoted to presentations, demonstrations and posters. Research results, work in progress, and position papers were presented also on the second day of the conference that started with the invited speech by Andreas Eisele: "From corpora to resources and tools – towards a proper treatment of Eastern European languages". The presentation outlined innovative ways how existing techniques can be combined to build the technology required for a proper treatment of Eastern European languages. Examples from EuroMatrix Plus and ACCURAT project as well as other related activities were used to show some important steps towards these goals. Finally, these activities were shown into a larger context of a long-term strategy that will allow to develop the high-quality language technology required for a truly multilingual European society.

The next invited speech by Georg Rehm, "META-NET and META-SHARE: An Overview", provided an overview of META-NET project – Network of Excellence dedicated to fostering the technological foundations of a multilingual European information society. The presentation introduced the architecture and general principles of META-SHARE, which is an action line on creation of an open distributed facility for the sharing and exchange of resources.

The following sessions of presentations were devoted to semantics, machine translation, speech technologies and methods for language processing. The conference ended with panel discussion "How to foster HLT development in Baltic countries".

Steven Krauwer showing CLARIN achievements

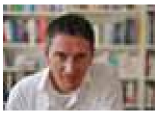
The conference was supported CLARIN, ACCURAT (FP7) and LetsMT! (ICT PSP) projects.

More information can be found on-line: http://www.lumii.lv/hlt2010/.

Challenges for Multilingual Europe

The First META-FORUM

Brussels, November 17, 2010



Aljoscha Burchardt
Georg Rehm
DFKI, Saarbrücken, Germany

Wednesday, November 17, 2010 in Brussels. A cold and foggy autumn day provided the backdrop for META-FORUM 2010, the first large outreach event organised by META-NET, a Network of Excellence forging the Multilingual Europe Technology Alliance (META). META-FORUM 2010 took place in the historic "Theatre" hall of the Hotel Le Plaza and assembled more than 250 participants from a total of 37 countries – a turnout far exceeding our initial expectations. META-FORUM 2010 was the inaugural conference of the META-FORUM series of events. All presentations are available online at http://www.meta-forum.eu.

As the name indicates, META-FORUM was meant to be a gathering point, a hub for diverse communities of interest to meet and discuss developments, problems and opportunities presented by the challenges of a modern and multilingual Europe. In the foyer in front of the Theatre, an exhibition area – META-NET Village – was dedicated to projects the network is collaborating with. Ten projects such as, for example, CLARIN, FLaReNet and PANACEA presented themselves, highlighting their plans of collaborating with META-NET. Software demonstrations from within META-NET proper complemented the picture by showcasing the open resource exchange infrastructure META-SHARE, the Virtual Information Centre and the first results of the Machine Translation research arm of the initiative.

META-FORUM 2010: Highlights of the Programme

The welcome address by Algirdas Saudargas (Member of European Parliament and former foreign minister of Lithuania) conveyed the clear message that language must be handled with the utmost care for it is a strong but also fragile band that ties together communities, social groups, and nations. Complementary, Roberto Cencioni (European Commission, Luxembourg) in his greeting address advised stakeholders from Language Technology research and development that this highly fragmented sector has to join forces to reach critical mass and that it has to improve its credibility and visibility. Hans Uszkoreit (DFKI, Germany), the coordinator of META-NET, took up this topic when he introduced the three lines of action the initiative pursues to reach exactly these strategic goals: building bridges to neighbouring technology fields, designing and implementing META-SHARE, the open resource exchange facility, and building a homogeneous European LT community with a shared vision and strategic research agenda.

The first invited keynote speech presented accessibility and multilingualism as key challenges for the audiovisual industry in the digital age. Yota Georgakopoulou (European Captioning Institute, UK) left no doubt that in the near future there will be an immense economic need within the audiovisual sector to make use of various Language Technology applications. In the session "From Shared Visions to a Strategic Research Agenda" industry representatives and language professionals that had been invited by META-NET into three think tanks – Vision Groups – presented the results of their first two rounds of meetings, which were discussed in a panel afterwards. The discussion is now continued in an online forum at http://www.meta-net.eu/forum and will, together with input and feedback collected through several other means, ultimately lead to the vision paper "European Multilingual Information Society 2020" and a

Hans Uszkoreit presenting projects collaborating with META-NET

Strategic Research Agenda. If you are interested in the needs and visions that have been and will be further discussed in the Vision Groups, please visit the online forum where you can join and contribute to the discussion.

Georg Artelsmair (European Patent Office, Germany) in the second invited keynote speech presented recent developments concerning the machine translation policy at the European Patent Office, which has become a heavy user of multilingual Language Technology. At the end of the first day, Bill Dolan (Microsoft Research, USA) accentuated the need for and also opportunities of close cooperation between industry and the research community, especially the importance of sharing both technology and also data.

The second (half) day of META-FORUM 2010 gave companies that either employ or build Language Technologies the opportunity to present themselves. Seven industry representatives formulated their needs, problems, and wishes for European LT research and development. As an attempt to strengthen the industry community, Jochen Hummel (ESTeam GmbH, Germany) introduced the Language Technology Business Association (LTBA), which is currently in a preparatory phase and which will be launched at the beginning of 2011. John Hendrik Weitzmann (Creative Commons Initiative, Germany) and Prodromos Tsiavos (Creative Commons Initiative, Greece, UK, Norway) addressed problems and solutions with regard to legal issues in basic and industrial research on Language Technology and the sharing of Language Resources through META-SHARE.

In her closing keynote, Swaran Lata (Department of Information Technology, Government of India, New Delhi) provided an overview of numerous multilingual initiatives in India in the past 20 years. Indeed, India, with its 22 languages (and various different scripts), can serve as a blueprint to European initiatives.

META-NET and CLARIN

The META-NET Network of Excellence (http://www.meta-net.eu) is forging the Multilingual Europe Technology Alliance (META) through a concerted effort to building a strong European community around Language Technologies. By working together to provide visionary applications and a Strategic Research Agenda for Language Technology in Europe, META-NET is reaching out to a large and heterogeneous community of stakeholders to help fostering the technological foundations of the European information society. CLARIN and META-NET are two complementary initiatives, serving different communities with different but harmonizable goals. As documented in a common Memorandum of Understanding (available online at http://www.meta-net.eu/collaborations), signed by their coordinators both initiatives have identified mutual interests and approaches with respect to methods and instruments. The next step in the collaboration between the two projects will be the preparation of a detailed list of topics to collaborate on, for example, in the area of metadata descriptions and standards for Language Resources and Language Technologies.

From the overwhelmingly positive feedback we received from the participants and also from our subjective perspective as the organisers of the event, META-FORUM 2010 was a huge success. Even more so if we take into account that META-NET's kick-off meeting only took place in February 2010 – on a very cold and also foggy day in Berlin.

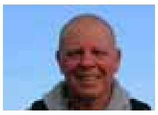
The next edition of META-FORUM will take place in Budapest from June 27-29, 2011.

Contact

Georg Rehm, Network Manager of META-NET, georg.rehm@dfki.de

META FORUM 2010

Language Resources for Hebrew



Alon Itai
*Department of Computer Science,
Israel Institute of Technology,
Haifa, Israel*

Shuly Wintner
*Department of Computer Science,
University of Haifa, Israel*

Introduction

Hebrew is a morphologically rich language whose word formation and inflectional morphology are typically Semitic. The standard Hebrew orthography, "undotted script", is highly ambiguous: Most vowels are omitted, many particles are attached to the word that follows them, and even though the Academy for the Hebrew Language (Gadish, 2001) has issued standards for transcribing Hebrew they are not adhered to. All this results in highly ambiguous texts; the average number of analyses of a word in a running text is 2.64. Consequently, the first step in processing Hebrew is nontrivial and many efforts need to be dedicated to morphological analysis and disambiguation.

In 2003 the Israel Ministry of Science and Technology established a Knowledge Center for processing Hebrew. Its mission was to collect and develop tools for processing Hebrew and make them available to Academia and Industry. Since then the Center, known as MILA, has developed several tools all of which interact with one another š they accept and produce XML data and are available under GPL over the Internet. This note describes the main products of the Center, see (Itai and Wintner, 2008) for additional details.

2 Databases

2.1 Corpora

MILA distributes a number of Hebrew text corpora, obtained from various sources: newspaper articles (*HaAretz*); newswire texts (*Arutz 7*); parliament proceedings (*Knesset*); and a small corpora of *Spoken* Hebrew extracted from CoSIH (Izre'el et al., 2001). Corpus sizes are listed in Table 1.

	HaAretz	Arutz 7	Knesset	Spoken
tokens	11,097	15,107	15,066	93
word forms	306	324	205	12

Table 1: Size of corpora
(in thousands of words)

2.2 Lexicon

Computational lexicons are among the most important resources for NLP. In languages with rich morphology, where the lexicon is expected to provide morphological analyzers with enough infor- mation to enable them to process intricately inflected forms correctly, a careful design of the lexicon is crucial. The MILA Lexicon of Contemporary Hebrew, the broadest-coverage publicly available lexicon of Hebrew, currently consists of about 25,000 entries. Table 2 lists the number of words in the lexicon by main part of speech (POS).

POS	#entries	POS	#entries	POS	#entries
noun	11620	conjunction	82	word prefix	36
verb	4801	numeral	60	quantifier	32
proper name	4769	interjection	59	interrogative	24
adjective	2645	modal	43	copula	24
adverb	466	title	43	negation	5
preposition	128	pronoun	38	existential	4
Total:					24,891

Table 2: Size of the lexicon by part of speech

3 Tools

3.1 Tokenization

Partitioning raw Hebrew data into tokens (words) is slightly more involved than in English due to issues of Hebrew encoding, mixed Hebrew/English, numbers, punctuation etc. We developed a tokenization module which operates on raw data (UTF-8 encoded) and produces an XML corpus. The module is capable of segmenting texts into paragraphs, sentences and tokens.

3.2 Morphological analysis and generation

We initially developed a finite-state morphological analyzer for Hebrew (Yona and Wintner, 2008). This solution, however, turned out to be inefficient (Wintner, 2007), and the analyzer was re-implemented in Java. Analysis is performed by generation: First, all the inflected forms induced by the lexicon (not including prepended prefixes) are generated and stored in a database. Then, analysis is simply a database lookup. At peak performance it is able to analyze 4,000 tokens per second.

3.3 Morphological disambiguation

Identifying the correct morphological analysis of a given word in context is an important and non-trivial task. A single token in Hebrew can actually be a sequence of more than one lexical item. For example, the token *šbth* can be analyzed as *š+b+h+th* "that+in+the+tea", *š+bth* "that+ her daughter"or a single lemma *she took prisoner*.

The task of *segmentation* is partitioning the token into its prefixes and main lemma, while *morphological disambiguation* in addition determines all the morphological features (tense, person, gender, number etc.) Note that morphological disambiguation falls short of *full disambiguation* since it does not distinguish between homographs, e.g., *spr* can be either *sefer* = *book* or *sappar* = *barber*.

MILA distributes three different disambiguation modules. MorphTag (Bar-Haim et al., 2005) is a Hidden Markov Model based tagger. When trained on 4500 annotated sentences, it boasts 97.2% accuracy for segmentation and 90.8% accuracy for POS tagging (Bar-haim et al., 2008). Adler and Elhadad (2006) have developed an *unsupervised* HMM-based method to morphologically disambiguate Hebrew texts. They report results of 92.32% for POS tagging and 88.5% for full morphological disambiguation, i.e., finding the correct lexical entry. Finally, HADAS (Shacham and Wintner, 2007) is a morphological disambiguation module for Hebrew that uses simple classifier for each of the attributes that can contribute to the disambiguation of the analyses produced by the analyzer (e.g., POS, tense, state), and then combines the outcomes of the simple classifiers to produce a consistent ranking which induces a linear order on the analyses. The results are 91.44% accuracy on the full disambiguation task.

These disambiguation modules are fully compatible with the morphological analyzer: they receive as input an XML file which is the output of the analyzer. The output is a file in the same format, in which each analysis is associated with a score, reflecting its likelihood in the context. This facilitates the use of the output in applications which may not commit to a *single* correct analysis in a given context.

References

Meni Adler and Michael Elhadad. An unsupervised morpheme-based hmm for hebrew morphological disambiguation. In Proceedings of the ACL, pp 665-672, Sydney, Australia, July 2006. ACL (http://www.aclweb.org/anthology/P/P06/P06-1084.)

Roy Bar-Haim, Khalil Sima'an, and Yoad Winter. Choosing an optimal architecture for segmentation and POS-tagging of Modern Hebrew. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pp 39-46, Ann Arbor, Michigan, June 2005. ACL. (http://www.aclweb.org/anthology/W/W05/W05-0706.)

Roy Bar-haim, Khalil Sima'an and Yoad Winter. Part-of-speech tagging of Modern Hebrew text. Natural Language Engineering, 14(2):223-251, 2008.

Ronit Gadish (ed.) Klalei ha-Ktiv Hasar ha-Niqqud. Academy for the Hebrew Language, 4th edition, 2001. ISBN 0024-1091. In Hebrew.

Alon Itai and Shuly Wintner. Language resources for Hebrew. Language Resources and Evaluation, 42:75-98, March 2008.

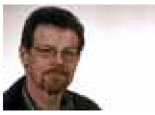
Shlomo Izre'el, Benjamin Hary, and Giora Rahav. Designing CoSIH: The corpus of Spoken Israeli Hebrew. International Journal of Corpus Linguistics, 6:171-197, 2001.

Danny Shacham and Shuly Wintner. Morphological disambiguation of Hebrew: a case study in classifier combination. In Proceedings of EMNLP-CoNLL 2007, Prague, June 2007. ACL.

Shuly Wintner. Finite-state technology as a programming environment. In Alexander Gelbukh (ed.) Proceedings of the CICLing2007, pp 97-106, Berlin and Heidelberg, February 2007. Springer.

Shlomo Yona and Shuly Wintner. A finite-state morphological grammar of Hebrew. Natural Language Engineering, 14(2):173-190, April 2008.

The State of Icelandic LT



Eiríkur Rögnvaldsson
University of Iceland, Reykjavík

Introduction

At the turn of the century, Icelandic language technology (henceforth LT) was virtu-ally non-existent. There was a relatively good spell checker, a not-so-good speech synthesizer, and that was all. There were no programs or even individual courses on language technology or computational linguistics at any Icelandic university, there was no ongoing research in these areas, and no Icelandic software companies were working on language technology.

In 1998, the Minister of Education, Science and Culture appointed a group of experts to investigate the situation in language technology in Iceland and come up with proposals for strengthening the status of Icelandic language technology. The group handed its report to the Minister in April 1999 and in 2000, the Government launched a special Language Technology Program, with the aim of supporting institutions and companies in creating basic resources for Icelandic language technology work. This initiative resulted in several projects which have had profound influence on the field.

Icelandic LT Work 2000-2010

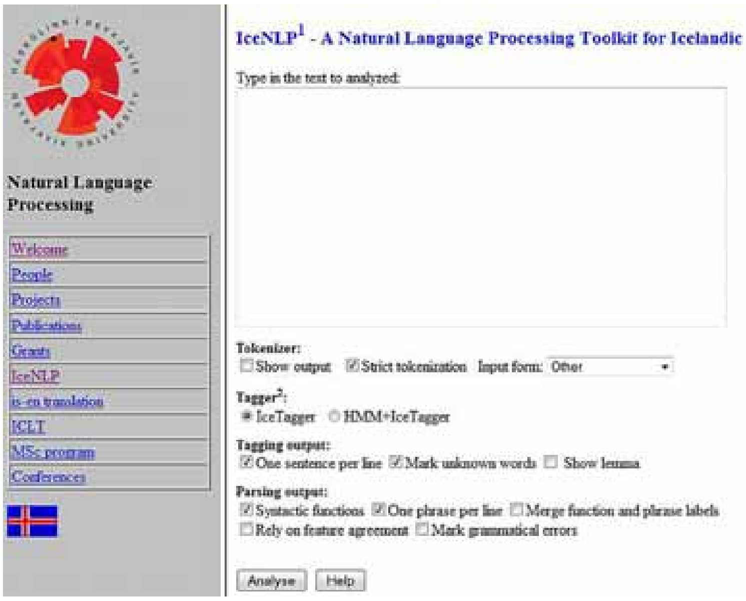
The main direct products of the government-funded LT Program are the following:

- full-form morphological database of Modern Icelandic inflections
- balanced morphosyntactically tagged corpus of 25 million words
- training model for data-driven POS taggers
- text-to-speech system
- speech recognizer
- improved spell checker

Most of these products were developed in cooperation between research institutes and commercial companies. After the LT Program ended six years ago, LT researchers from three research institutes (University of Iceland, Reykjavik University, and the Árni Magnússon Institute for Icelandic Studies), who had been involved in most of the projects funded by the LT Program, decided to join forces in a consortium called the Icelandic Centre for Language Technology (ICLT), in order to follow up on the tasks of the Program. The main roles of the ICLT are to

- serve as an information centre on Icelandic LT by running a website (http://iclt.is)
- encourage cooperation on LT projects between universities, institutions and commercial companies
- organize and coordinate university education in LT
- participate in Nordic, European and international cooperation within LT
- initiate and participate in R&D projects in LT
- keep track on resources and products in the field of Icelandic LT
- hold LT conferences with the participation of researchers, companies and the public
- support the growth of Icelandic LT in all possible manners

Over the past six years, the ICLT researchers have initiated several new projects which have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund. The most important products of these projects are:



IceNLP[1] - A Natural Language Processing Toolkit for Icelandic

Type in the text to analyzed:

Tokenizer:
□ Show output ☑ Strict tokenization Input form: Other

Tagger[2]:
● IceTagger ○ HMM+IceTagger

Tagging output:
☑ One sentence per line ☑ Mark unknown words □ Show lemma

Parsing output:
☑ Syntactic functions ☑ One phrase per line □ Merge function and phrase labels
□ Rely on feature agreement □ Mark grammatical errors

The front page of the IccNLP, a NLP toolkit for Icelandic

- linguistic rule-based tagger, IceTagger
- shallow parser, IceParser
- mixed-method lemmatizer, Lemmald
- (prototype of a) context-sensitive spell checker

In 2009, the ICLT received a relatively large three year Grant of Excellence from the Icelandic Research Fund for the project "Viable Language Technology beyond English – Icelandic as a test case". Within that project, three types of LT resources are being developed:

- database of semantic relations (a pilot WordNet)
- prototype of a shallow-transfer machine translation system
- treebank with a historical dimension

These resources were chosen because they were considered central to current LT work and prerequisites for further research and development in Icelandic LT.

The Prospects of Icelandic LT

Twelve years ago, the LT expert group estimated that it would cost around one billion Icelandic krónas (which then amounted to about ten million Euros) to make Icelandic language technology self-sustained. After that, the free market should be able to take over, since it would have access to public resources that would have been created by the Language Technology Program, and that would be made available on an equal basis to everyone who was going to use these resources in their commercial products.

However, the total budget of the government-funded LT program over its lifespan (2000-2004) was only 133 million Icelandic krónas – that is, 1/8 of the sum that the expert group estimated would be needed. It should therefore come as no surprise that we still have a long way to go. There are only 320,000 people speaking Icelandic, and that is not enough to sustain costly development of new products. At present, no commercial companies are working in the LT area because they don't see it as profitable. It is thus extremely important to continue public support for Icelandic language technology for some time, but given the current financial situation, it does not seem likely that such support will come from the state budget in the near future.

For a small language community and a small research environment like the Icelandic one, it is vital to cooperate, not only on the national level but also internationally. Since 2000, Icelandic researchers and policy makers have taken an active part in Nordic cooperation on language technology. This has been of major importance in establishing the field in Iceland. The Nordic Language Technology Research Programme 2000-2004 was instrumental in this respect.

Iceland has just recently entered the CLARIN consortium, and together with the other Nordic and Baltic countries, Iceland also takes part in the META-NORD project which starts February 1st this year. We sincerely hope that our participation in these projects will help us to develop, standardize and make available several important LT resources and thus contribute to the growth of Icelandic language technology. C

LR&T situation in Turkey

Gülşen Eryiğit
Faculty of Computer and Informatics, Istanbul Technical University (ITU), Turkey

Introduction

Over the last two decades, we see that the interest of the Turkish researchers in language technology research increased noticeably. Although the interest of private sector and funding agencies still remains at its early phase, the ongoing activity in the field is promising.

Most of the important Turkish universities showed their interest in the field by offering courses in language technology. ITU Faculty of Computer and Informatics, as one of the leaders, offers now three specialized graduate courses in the field and many supplementary courses.

The number of language resources such as corpora, treebanks and online dictionaries continue to get increased in number. Also, there exist important NLP tools developed for Turkish.

Turkish Language Resources

Some of the well known data sources for Turkish are as follows:

- online dictionaries of Turkish Language Association (TDK),
- 2M words corpus by Middle East Technical University (METU) and a Discourse Annotation Bank (where METU Turkish Corpus is annotated with respect to connectives, their senses and arguments)
- morphologically and syntactically annotated 5K sentences dependency treebank by METU and Sabancı University,
- Turkish Wordnet (part of BalkaNet Project) by Sabancı University,
- large-scale test collection that contains 408K documents for information retrieval by Bilkent University,
- 20K documents in 8 different categories for text categorization by ITU,
- 423M words web corpus by Boğaziçi University,
- ongoing project on creating a morphologically annotated 50M words Turkish corpus by Mersin University,
- 160K English-Swedish-Turkish Parallel Treebank by Uppsala University,
- 500K English-Turkish parallel sentences, by Koç University,
- transcribed speech corpus by Boğaziçi University.

Turkish NLP Tools

Turkish, being an agglutinative language, poses interesting challenges for the area of language technologies. The models developed for other well-studied languages do not conform to this language for many NLP layers. On the other hand, research in this language serves as reference to many other languages with similar features (morphologically rich, agglutinative, with scarce data resources, …) This makes Turkish very attractive for NLP research not only for Turkish researchers but also at an international level. In recent years, Turkish has been included in many international research projects which aim to develop multilingual systems.

Istanbul Technical University Maslak Campus

The tools which are developed specifically for Turkish automatic language processing dates back to the beginning of 1970s starting with the first morphological analyzer developed at Hacettepe University. Since that time, many other tools have been developed by different institutions. Many tasks still remain as *open research questions and the performance of the tools still needs to get ameliorated*. Some of them may be listed as follows:

- two level morphological analyzer developed at Bilkent and Sabancı Universities, and some replication of the similar approach at Boğaziçi and Koç Universities.
- morphological disambiguators developed at Boğaziçi University and Koç University,
- dependency parser developed at Istanbul Technical University,
- CCG grammar and parser at METU,
- LFG grammar at Sabancı University,
- speech recognition and synthesis tools developed at Boğaziçi University, Tubitak UEKAE MTRD and a private company called SESTEK.

National Funds for LT

The national funding agencies in Turkey are mainly DPT (State Planning Agency) and TUBITAK (The Scientific and Technological Research Council of Turkey). Both of the institutions grant projects on language technology. But a national policy on this subject is still lacking. Thus, the funded projects remain isolated and generally at the level of research projects. We hope that our partnership in CLARIN will speed up the national organization of the Turkish resources and the NLP community in Turkey.

Turkey, being one of the last partners of CLARIN preparatory phase, has to work hard in order to catch up with the ongoing organization. Since Turkey is still not a full member of EU, it seems that it still needs time to overcome the indefiniteness about the legal regulations of ERICs at the national level. DPT works on the issue. Once the preparation phase of Turkey is finalized, we hope that the national roadmap will include CLARIN and provide supports for CLARIN activities in Turkey. C

Slovak language in computer processing

Radovan Garabík
Mária Šimková
Ľ. Štúr Institute of Linguistics,
Slovak Academy of Sciences,
Bratislava, Slovakia

General information

Slovak language belongs to the West Slavic language group, together with Polish, Czech and Sorbian languages. It remained especially close to the Czech language, due to close historical and linguistic ties between the languages (especially in the former common country). Slovak is a typical Slavic language in retaining complex inflectional and derivational morphology, with only minor simplifications compared with neighbouring languages.

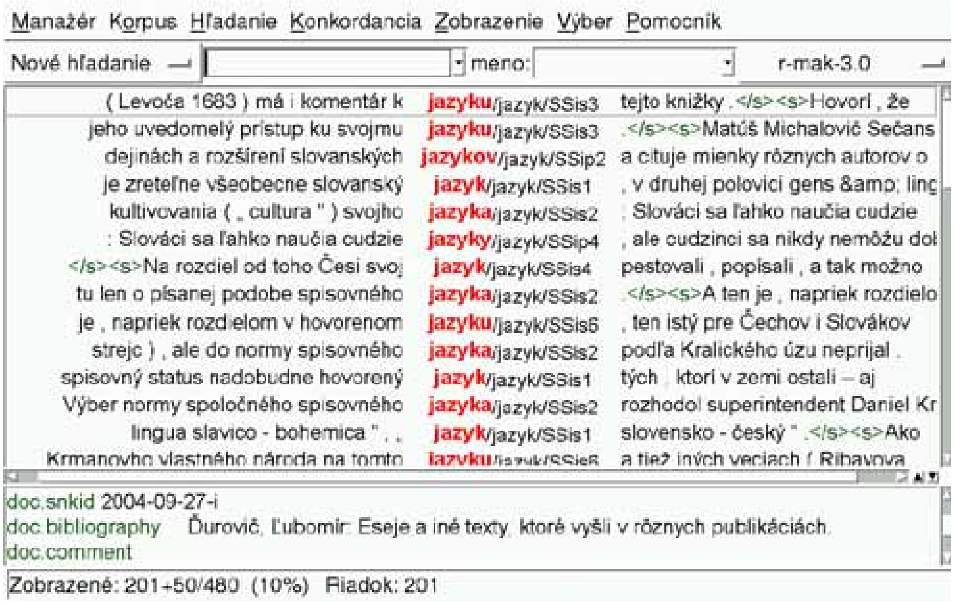
Slovak language is spoken by about 5 million people in Slovakia, but also in some other countries. The most numerous Slovak community is in the USA (1 million people, many of them actually speaking the language), Czech Republic, smaller communities are present in Romania, Hungary, Serbia and other countries.

Slovak orthography is mostly phonemic, with noticeable etymological and morphological features. The language is written in the Latin script, with acute accents marking (phonemic) vowel length, há?eks marking palatals and post-alveolar fricatives. The latest substantial orthography reform has been implemented in 1953, when the language gained contemporary form practically in all of its aspects.

Major linguistic research centres in Slovakia

The *Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava* is the central linguistic institution in Slovakia. Its main area of research is traditional linguistics, with the emphasis on (but not limited to) Slovak language, its history, dialectology, etymology and lexicography, as reflected in a sizeable amount of dictionaries produced. Traditionally, the Institute was connected with the task of regulating the Slovak language, defining its grammar and orthography and producing prescriptive dictionaries. However, in the last decades the main orientation shifted more towards general linguistic research (not limited to the Slovak) and descriptive or bilingual dictionaries and other popular and scientific publications in the field of sociolinguistics, onomastics, language theory, lexicography and others.

Institute of Informatics, Slovak Academy of Sciences, Bratislava is dealing with theoretical and applied research in the field of computer science, information technologies and artificial intelligence. Their NLP tools include information retrieval, knowledge representation speech synthesis and recognition. The main research direction of *Faculty of Arts, University of Prešov* in Prešov is phonetics, phonology, derivatology and morphematic structure of the Slovak. The results of their research are often used in Slovak NLP. An extensive research of child speech is carried out in cooperation with the Faculty of Education, especially in international projects. *University of SS Cyril and Methodius, Trnava* puts special emphasis on German-Slovak confrontational research on collocations, valency and corpus lexicography as well as phraseology and paremiology. *Faculty of Humanities, Matej Bel University Banská Bystrica* is active in sociolinguistics and research in communication and stylistics. Part of their database of recordings of spontaneous speech in the city of Banská Bystrica and surroudings become part of the Corpus of Spoken Slovak. *The Technical University of Košice* is active especially in NLP, computer aided lexicography and ontology research.



Manažér Korpus Hľadanie Konkordancia Zobrazenie Výber Pomocník
Nové hľadanie
meno:
r-mak-3.0

```
      ( Levoča 1683 ) má i komentár k  jazyku/jazyk/SSis3  tejto knižky .</s><s>Hovorí , že
  jeho uvedomelý prístup ku svojmu  jazyku/jazyk/SSis3  .</s><s>Matúš Michalovič Sečans
    dejinách a rozšírení slovanských  jazykov/jazyk/SSip2  a cituje mienky rôznych autorov o
    je zreteľne všeobecne slovanský  jazyk/jazyk/SSis1  , v druhej polovici gens &amp; ling
      kultivovania ( „ cultura " ) svojho  jazyka/jazyk/SSis2  . Slováci sa ľahko naučia cudzie
      : Slováci sa ľahko naučia cudzie  jazyky/jazyk/SSip4  , ale cudzinci sa nikdy nemôžu do
</s><s>Na rozdiel od toho Česi svoj  jazyk/jazyk/SSis4  pestovali , popísali , a tak možno
  tu len o písanej podobe spisovného  jazyka/jazyk/SSis2  .</s><s>A ten je , napriek rozdielo
   je , napriek rozdielom v hovorenom  jazyku/jazyk/SSis6  , ten istý pre Čechov i Slovákov
     strejc ) , ale do normy spisovného  jazyka/jazyk/SSis2  podľa Kralického úzu neprijal .
 spisovný status nadobudne hovorený  jazyk/jazyk/SSis1  tých , ktorí v zemi ostali – aj
 Výber normy spoločného spisovného  jazyka/jazyk/SSis2  rozhodol superintendent Daniel Kr
          lingua slavico - bohemica " , ,  jazyk/jazyk/SSis1  slovensko - český " .</s><s>Ako
 Krmanovho vlastného národa na tomto  jazyku/jazyk/SSis6  a tiež iných veciach ( Ribayova
```

```
doc.snkid 2004-09-27-i
doc.bibliography    Ďurovič, Ľubomír: Eseje a iné texty, ktoré vyšli v rôznych publikáciách.
doc.comment
```

```
Zobrazené: 201+50/480  (10%)   Riadok: 201
```

Web-based interface to Manatte/Bonito platform running Slovak National Corpus

Slovak language in computer processing

The *Slovak National Corpus department* was established as a special project of the Ministry of Culture and Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences in 2002. This marked rapid expansion of Human Language Technology research of the Slovak language, with the Ľ. Štúr Institute of Linguistics becoming the leading research institution in the field. The Institute actively leads R&D in computational linguistics oriented towards contemporary written and spoken Slovak language, covering all aspects of language analysis and processing. Its strong IT background is demonstrated in the number of tools and resources that the department developed during its existence, among the most important are the Slovak National Corpus database and all the necessary associated tools.

The Slovak National Corpus is a big, representative corpus of modern written Slovak (since the 1953 orthography reform). Currently, the whole corpus contains about 780 million tokens. There are several specialised subcorpora (fiction, professional texts, journalistic texts, original Slovak fiction, balanced subcorpus, texts written before 1989). The corpus is automatically lemmatised and morphologically annotated, using its own tagset and morpfology analyser. Access to the corpus is provided free of charge, for non commercial purposes. The corpus legal status is rather unusual, if compared with other language corpora – The Ľ. Štúr Institute of Linguistics did obtain license agreements to use the texts in building the corpus database (for non commercial, research and educational purposes), and the corpus therefore includes the texts with full legal compliance. *New Dictionary of Contemporary Slovak* (*Slovník súčasného slovenského jazyka*) is being compiled with the help of the corpus and is the first Slovak dictionary based predominantly on corpus resources.

The department actively works on several smaller, but no less important projects. *Corpus of Spoken Slovak* is a representative corpus of standard spoken Slovak as spoken throughout Slovakia, consisting of about 160 hours of recordings (1.6 million words). The recordings are manually transcribed on orthographic and phonemic level. Parallel corpora include Slovak-Czech, Slovak-Russian, Slovak-French, Slovak-Bulgarian and Slovak-English corpus. The texts (mostly fiction translations) are automatically sentence-aligned and morphosyntactically tagged. Slovak dependency treebank contains about 50,000 manually syntactically annotated sentences.

The *Slovak Terminology Database* focuses on the field of law, economy and technology, offering about 5000 terminological records that can be classified by circa 20 EUROVOC descriptors corresponding to different soft and hard sciences.

International collaboration

With expansion of data and resources, it is more and more apparent that standardisation and availability of resources, documentation and interoperability are important for further scientific research and technology deployment. Since the beginning, the Slovak National Corpus department is committed to making available all the resources and tools under favourable Open Content and Open Source licensing policies (if other copyright restrictions permit). Collaboration with other partners has proved to be very fruitful and we expect to extend the ties with other CLARIN partners and to improve the interoperability and standard compliance of Slovak NLP resources and tools. 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. Weippl)

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:Jean-Pierre Martens)

University College Ghent (Gent): Faculty of Translation Studies, Language and Translation Technology Team (C:Veronique Hoste)

Katholieke Universiteit Leuven (Leuven): Center for Computational Linguistics (C:Frank Van Eynde); ESAT-PSI/Speech (C:Patrick Wambacq); Language Intelligence & Information Retrieval (C:Marie-Francine Moens)

Katholieke Universiteit Leuven (Leuven – Kortrijk): iTec (Interdisciplinary research on Technology, Education & Communication) (C:Hans Paulussen)

Bulgaria (NCP: Kiril 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 Departement (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 Ćavar)

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 Theocharous)

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 Maegaard) Copenhagen Business School (Copenhagen): Department of International Language Studies and Computational Linguistics (C:Peter Juel Henrichsen)

Dansk Sprognævn – 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 Maegaard)

University of Southern Denmark (Kolding): Faculty of Humanities (C:Johannes Wagner)

Estonia (NCP: Tiit Roosmaa) University of Tartu (Tartu): Institute of Computer Science (C:Tiit Roosmaa)

Finland (NCP: Kimmo Koskenniemi) CSC – the Finnish IT Center for Science (Espoo): (C:Pirjo-Leena Forsström)

Lingsoft Inc. (Helsinki): (C:Juhani 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 Sormunen)

France (NCP: Jean-Marie Pierrel) Centre de ressources pour la documentation de l'oral (Aix-en-Provence) (C:Bernard Bel)

National Center for Scientific Research (CNRS) (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 (ALTIF) (C:Jean-Marie Pierrel)

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 Choukri)

National Center for Scientific Research (CNRS) (Paris): Traitement Électronique des Manuscrits et des Archives (TELMA/DIS) (C:Florence Clavaud)

Université Paris 4 Sorbonne (Paris): Centre de linguistique théorique et appliquée (CELTA) (C:Andre Wlodarczyk)

Université de Strasbourg (Strasbourg): Equipe de recherche LiLPa (Linguistique, Langues, Parole) (C:Amalia Todirascu)

National Center for Scientific Research (CNRS) (Vandoeuvre les Nancy): L'Institut de l'Information Scientifique et Technique (INIST-CNRS) (C:Fabrice Lecocq)

University Paris Est/Paris 12 (Vitry 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:Iryna Gurevych)

TU Dortmund University (Dortmund): Institute for German Language and Literature (C:Michael Beißwenger)

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 Lobin)

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 Tihanyi)

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:Sean 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 Vasiljevs)

University of Latvia (Riga): Institute of Mathematics and Computer Science (C:Inguna Skadina)

Lithuania (NCP: Ruta Marcinekieviciene) Vytautas Magnus University (Kaunas): Center of Computational Linguistics (C:Ruta Marcinkeviciene)

Institute of the Lithuanian Language (Vilnius): (C:Daiva Vaisniene)

Luxembourg (NCP: -) European Language Resources Association (ELRA) (Luxembourg): (C:S.Piperidis/K.Choukri)

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. Bennis)

Univerity 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 Harmsen)

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 Ordelman)

University of Gröningen (Gröningen): 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 Wittenburg)

Radboud University (Nijmegen): Centre for Language and Speech Technology (C:L. Boves / N. Oostdijk); Centre for Language Studies (C:Pieter Muysken)

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 Krauwer)

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:Torbjoerg Breivik)

The National Library of Norway (Oslo) (C:Kristin Bakken)

University of Oslo (Oslo): Department of Linguistics and Nordic Studies, Faculty of Humanities (C:Janne Bondi Johannessen)

University of Tromsø (Tromsø): Det humanistiske fakultet (C:Trond Trosterud)

Norwegian University of Science and Technology (Trondheim): Department of Electronics and Telecommunications (C:Torbjørn Svendsen)

Poland (NCP: Maciej Piasecki) 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 Koseska-Toszewa)

Polish-Japanese Institute of Information Technology (Warsaw): (C:Krzysztof Marasek)

University of Wroclaw (Wroclaw): Instytut Informatyki Stosowanej (C:Maciej Piasecki)

Wroclaw University of Technology (Wroclaw): Institute of Applied Informatics (C:Maciej Piasecki)

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): Pólo de Coimbra (C:Fernando Perdigão)

University of Coimbra (Coimbra): Centro de Esudos 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 (CLUL) (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 Filgueiras)

Romania (NCP: Dan Tufiș) Romanian Academy of Sciences (Bucharest): Research Institute for Artificial Intelligence (C:Dan Tufiș)

University Babes-Bolyai (Cluj-Napoca): Faculty of Mathematics and Computer Science (C:Tatar Doina)

"Al. I. Cuza" University of Iași (Iași): Faculty of Computer Science (C:Dan Cristea)

Romanian Academy of Sciences (Iași): Institute of Computer Science (C:Horia-Nicolai Teodorescu)

University of Pitești (Pitești): Faculty of Letters (C:Mihaela Mitu)

West University of Timișoara (Timișoara): 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:Radovan Garabík)

Slovenia (NCP: Tomaž Erjavec) Alpineon d.o.o. (Ljubljana): (C:Jerneja Zganec 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:Patricio Martínez-Barco)

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:Asuncion Moreno)

Universitat Autònoma de Barcelona (Barcelona): Facultat de Filosofia i Lletres, Dpt. de Filologia Anglesa i de Germanística (C:Ana Fernández Montraveta)

Universitat de Barcelona (Barcelona): Departament de Lingüistica 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): Facultad de Letras, Dept. Filología Española (C:Carlos Subirats)

Girona City Council (Girona): Records Management, Archives and Publications Service (C:Joan Boadas i Raset)

University of Jaén (Jaén): Escuela Politécnica Superior, Departamento de Informática, SINAI group (C:María Teresa Martín Valdivia)

University of the Basque Country (Leioa): Computer Science Faculty, Natural Language Processing Group (C:Arantza Diaz de Ilarraza)

University of Lleida (Lleida): Departament d'Angles i Linguistica (C:Gloria Vázquez)

Autonomous University of Madrid (Madrid): Laboratorio de Lingüística Informática (C:Manuel Alcantara Plá)

University of Málaga (Málaga): Facultad de Filosofía y Letras, Dept. of English, French, and German Philology (C:Antonio Moreno Ortiz)

Universidad Politécnica de Valencia – ITACA (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 Tradución 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 Strömqvist)

KTH Royal Institute of Technology (Stockholm): Department of Speech, Music and Hearing, CSC (C:Rolf Carlson)

Language Council of Sweden (Stockholm): (C:Rickard 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: Gülșen Eryiğit) Istanbul Technical University (Istanbul): Elektrics-Electronics Faculty, Computer Science Department, Natural Language Processing Group (C:Gülșen 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:Briony 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 Langauge (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)

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