NO-CLARIN

Norwegian Common Language Resources and Technology Infrastructure

Coordinator: Prof. Koenraad De Smedt, University of Bergen

RCN program *INFRASTRUKTUR*
Støtte til norsk deltagelse i etablering av forskningsinfrastruktur (ESFRI)

1 Vision and scientific goals

The ESFRI-project *Common Language Resources and Technology Infrastructure (CLARIN)* is a large-scale pan-European collaborative effort to make language resources and technology available and readily useable to researchers. The envisaged infrastructure will be integrated via Grid technology to make it interoperable, stable, persistent, accessible and extensible. The infrastructure will act as an eScience mediator offering web services for cataloguing, searching, filtering, processing, reformatting, and visualizing language information in many ways. For more information on CLARIN, see the project website.

The present proposal calls for a national supporting action under the name of NO-CLARIN, with the goal to strengthen Norwegian participation in the preparatory phase of the CLARIN project, in order to prepare Norwegian participation in the construction and exploitation phase of CLARIN starting after 2010. NO-CLARIN will support current Norwegian participation in CLARIN activities, provide a forum for Norwegian language-related research infrastructure that can be related to CLARIN, investigate the possibilities of establishing infrastructure centers connected to the CLARIN grid, and disseminate information on infrastructure status, plans and opportunities to relevant audiences.

2 Scientific and technological environment

Norwegian research in language and text technologies faces a bottleneck, not only with respect to the digitization of language data, but also with respect to the accessibility and reusability of the already digitized data. Many of the present digitized language materials are the result of significant investments and embody great expertise, but often lack documentation, are not easy to access, lack proper support and maintenance, cannot be catalogued due to insufficient metadata, cannot be linked together, are not always adaptable to researcher needs and have generally been underexploited.

CLARIN aims to improve this situation through establishing an infrastructure that provides cataloguing, adequate standardization, better interfaces and linking and faster and more reliable access through Grid-based federation.

It is in the interest of language and humanities scholars in Norway to participate in this pan-European effort, primarily in order to secure the incorporation of language data held in Norway (both for Norwegian and for other languages studied in Norway) in the CLARIN infrastructure, and to secure optimal Norwegian access to, and exploitation of, language materials produced in the geographical area of CLARIN.

3 Description of the Research Infrastructure

As stated in the project description, *"The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organizational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) ..."*

The preparatory phase is paving the way for implementation of the infrastructure by working out a draft agreement between the funding agencies in the participating countries, covering governance, financing, IPR issues, construction and operation. The technical objective is to provide a detailed specification, agreement on data and interoperability standards to be adopted, and a validated prototype covering technical, linguistic and user aspects.

The University of Bergen and its affiliate partner Unifob AKSIS have been assigned tasks in the implementation of Work Packages 3, 6 and 8 in the preparatory phase of the CLARIN project. The resources assigned to these efforts are limited to 9 person months for the whole project duration.

No resources have been allocated in the CLARIN budget to any participation in other Work Packages, nor to the participation of other Norwegian members, nor to any planning, inventory, data handling, networking or dissemination at a national level. These responsibilities are with the national funding agencies, according to the CLARIN workplan, which calls for national complementary actions along these lines.

The currently planned work in NO-CLARIN is aimed at complementing and extending the CLARIN preparatory work through national networking activities, creating a forum for disseminating developments in the preparatory phase of CLARIN, keeping an overview of relevant activities, needs and plans in Norway, securing Norwegian input and positioning towards CLARIN with a view of contributing to CLARIN deliverables that are maximally beneficial for relevant Norwegian research. Extended national participation in CLARIN requires travel funds that allow Norwegian CLARIN members to attend CLARIN meetings. This participation is meant to cover a wider Norwegian influence on design issues in the current preparatory phase and will therefore include participation in as many CLARIN Work Packages as possible. Furthermore, a plan will be drafted for Norwegian participation in the construction phase of the CLARIN infrastructure, including a sketch for a project application and a feasibility study for applying national Grid solutions to Norwegian resources relevant to CLARIN, with a view of assessing Norwegian potential with respect to data and resources for data handling, linking to national solutions for data storage and communication, and participation in the European CLARIN grid.

4 Plan for access and use, data and knowledge management

The expected result of the CLARIN project's construction phase will be an infrastructure based on a network of repositories and service centers where users can deposit and register their resources and that will help turning language technology into usable services.

CLARIN will have a distributed architecture based on 10 to 20 major interoperable centers connected in a Grid. Norway could position itself favorably in this context by running design studies for the use of national grid solutions which will eventually be connected to a European backbone.

5 Impact on research and innovation

Participation in CLARIN will enable researchers in Norway to gain access to digital resources on a new scale of availability and persistence, which will positively affect all researchers in the language sciences and in humanities and social science disciplines working with such materials.

Humanities scholars in the broad sense (e.g. literature, history, philosophy, anthropology, history of art, psycholinguistics) will experience broader and easier access to text and speech materials, archives, historical materials, electronic dictionaries and termbanks and all language related technologies. Computational linguists, theoretical linguists, psycholinguists, applied linguists and language engineers will get access to data to test and optimize language models, run experiments and develop applications.

6 Partners and scientific institutions

NO-CLARIN will be carried out by a consortium led by the University of Bergen and furthermore including the following partners:

1. University of Oslo, member of the CLARIN network (contact person Janne Bondi Johannessen)
2. Norwegian University of Science and Technology (NTNU, contact person Torbjørn Svendsen)
3. University of Tromsø, member of the CLARIN network (contact persons Trond Trosterud and Øystein Vangsnes)
4. Norwegian School of Economics and Business administration (contact person Gisle Andersen)
5. Uni Digital, a department in Uni Research (contact person Eli Hagen)
6. SINTEF ICT (contact person Diana Santos)
7. The National Library (contact person Kristin Bakken)
8. Uninett Sigma / Notur (contact person Jacko Koster)
9. The Language Council (contact person Torbjørg Breivik)

The CLARIN members have a clear interest and good competency in building, managing and utilizing language resources. The academic partners will contribute with scientific competency in relevant fields. The National Library will bring in their experience with very large digitized language collections and data curation, storage and migration. Uninett Sigma will contribute with know-how about eScience infrastructure and the role of NorStore. The Language Council will participate with a perspective on language policies. It has been actively involved in WP7 and attended meetings dealing with IPR matters, accessibility and unique user identification. The Norwegian Language Council takes an active interest in the discussions on developing a BLARK for CLARIN.

7 User groups and international cooperation

The preparatory phase of the CLARIN project is carried out by the CLARIN consortium in cooperation with the members of the CLARIN network all over Europe. Among the members are universities, public and private R&D organizations, national computing centers and international associations like ELRA/ELDA.

NO-CLARIN will cooperate closely with the CLARIN consortium and network. Contacts will be kept with the CLARIN Executive Board and the different CLARIN working groups through meeting activities.

On a national level, NO-CLARIN will link with ongoing actions related to infrastructure concepts, such as INESS, the Norwegian Corpus of Medieval Texts and the workshops on Research Infrastructure for Linguistic Variation Studies (RILiVS), and will address itself to stakeholders in public and private R&D organizations including also IT providers, publishers and media companies.

8 Management plan and localisation

The NO-CLARIN activity will be managed at the UiB, who is the only Norwegian participant in CLARIN, together with its affiliate Unifob AKSIS. UiB has expertise in the field of language resources at the LaMoRe research group which earlier has organized a nation-wide CLARIN-related activity.

NO-CLARIN will contribute to a discussion of suggested involvement of Norway in the next CLARIN phase, depending on foreseeable opportunities and risks. It will contribute to CLARIN plans for the organization of the construction and exploitation of the CLARIN infrastructure and the localization of a center in Norway. It will point out links to existing actions on the national level that involve infrastructure concepts and goals and will sketch possible synergies with the European activities.

9 Time-schedule and deliverables

The preparatory phase of the CLARIN project started on January 1, 2008 and will continue until December 31, 2010. In this period, the current NO-CLARIN proposal is planning the following activities:

1. **A national meeting on research infrastructures for language (Deliverable 1, Spring 2010).** To promote national networking, information dissemination and consensus building, a national meeting will be organized in the spring of 2010, as a more focused follow-up of the meeting in 2008. Its program will be consist of (1) in-depth information on the status and outcomes of the CLARIN project so far, (2) survey and presentation of relevant infrastructure projects in Norway, and (3) consensus building on a plan for language infrastructure in Norway including a link to eInfrastructure (eVita) and integration into the next phase of CLARIN.

2. **National consultation meetings and attendance at CLARIN events.** Norwegian CLARIN members will be given the opportunity to travel to CLARIN meetings at European venues. Smaller meetings will be held to consult stakeholders on specific issues that need clarification and supplemental information.

3. **Plan for Norwegian participation in CLARIN construction phase (Deliverable 2, End of 2010).** Based on the national meeting, the consultations and contacts described above, a plan will be drafted targeted at a project application for the CLARIN construction phase. The plan will include an analysis of needs, status, opportunities and risk related to the infrastructure building after the CLARIN preparatory phase. Aspects considered in this study will

related to scientific orientation, organization, localization, financing, utilization, responsibilities for construction, operation and upgrade and maintenance troughout its entire lifetime. The plan will include a feasibility study for applying national Grid solutions to Norwegian resources relevant to CLARIN, with a view of assessing Norwegian potential with respect to data and resources for data handling, linking to national solutions for data storage and communication, and premises for establishing a national center participating in the European CLARIN grid. A tentative budget and financing plan will be worked out, as well as distribution of responsibilities among competent scientific organizations and other stakeholders. Attention will be given to how the new research infrastructure fits into a long term planning and research strategy of scientific organizations and funding schemes and programs.

10 Budget and funding plan

The financial plan calls for the following expenses:

- National NO-CLARIN meeting costs: travel, accommodation and subsistence for 20 invited participants for two days: NOK 100,000.
- Attendance costs at CLARIN meetings: travel, accommodation and subsistence for approximately 10 trips to European destinations: NOK 80,000.
- Small meeting costs: travel, accommodation and subsistence for 4 trips within Norway: NOK 12,000.
- Salary costs related to consortium participation at meetings, participation in CLARIN working groups and drafting report: 4.5 person months, NOK 309375.[1]
- Coordination of project, preparation of meeting agendas, dissemination and reporting: 0.5 person month, NOK 34,375.
- Administrative support for meeting logistics: 30 hours, NOK 10,465.[2]

The Language Council contributes with NOK 16,545 which has been reserved earlier for CLARIN laison purposes, and also contributes with half a person month (worth NOK 34,375), so a total of NOK 50920. This brings the total requested contribution from RCN to NOK **495,295**.

The Norwegian partners as well as representatives of the CLARIN project have indicated their willingness to contribute further working time towards participation in NO-CLARIN activities as needed.

[1] Based on *forskersats* NOK 825,000 on a yearly basis, monthly cost NOK 68750.

[2] Based on Ltr 45, hourly price 182,3 x 1,4351 = 261.62.