# CLARIN: Norwegian and Nordic perspectives

**Koenraad De Smedt**
University of Bergen and Unifob AKSIS

## Abstract

This position paper adresses the question whether there is a need for Nordic cooperation on building a language infrastructure for the Humanities, given the existing European cooperation in the CLARIN project and a number of national initiatives. It will be argued that the Nordic level is not superfluous, but in fact it seems the most efficient and appropriate level for cooperation, based on size, common culture, cooperation record and existing frameworks.

## 1 Status of coordination of language resources in Norway

Not since the Norwegian Computing Centre for the Humanities, established in Bergen in 1972, was discontinued as a centre with national responsabilities in the 1990s, has Norway had a centralized coordination of digital resources for the Humanities. A plethora of activities has taken place in the past four decades, resulting in a wealth of digital resources and technologies, many of very high quality, but as a whole rather disparate and not easy to access for exploitation.

The KUNSTI research program (2001–2007) provided an important stimulus to language technology research in Norway, but this program was not targeted at unifying existing language resources into a usable whole. Since 1999, there have been numerous surveys, studies, reports and plans aimed at building a comprehensive Norwegian language bank, but this goal as such has not received adequate funding so far. A Norwegian government proposition in 2008 stated that the Norwegian HLT Resource Collection shall be established on Jan. 1, 2009, but although this activity has received considerable moral support from the government, the Language Council and the universities, it has so far not received substantial funding. Besides, the latter initiative is targeted at language technology development and not at a comprehensive language infrastructure for the Humanities.

## 2 The *Infrastruktur* program

Recently, Norway's strategic investment in research infrastructures was accelerated by initiatives abroad, especially by Europe's engagement in scientific infrastructures since the 2000 Strasbourg Conference on Research Infrastructures, leading to the first ESFRI roadmap in 2006 and the projects in the Capacities program since 2007.

In early 2008, a strategy document was published, *Verktøy for forskning: Nasjonal strategi for forskningsinfrastruktur (2008–2017)*, that envisaged the establishment of a special research fund of NOK 20 billion with a yearly yield of NOK 800 million,[1] 75% of which would be channeled through the Research Council of Norway (RCN) and 25% of through R&D institutions.

Further recommendations, including continued participation in NDGF (Nordic Data Grid Facility), have come through the strategy document *Nasjonal strategi for eInfrastruktur*, which outlined in particular the electronic platforms necessary for digital infrastructures.

The first call for proposals in research infrastructures under the program *Infrastruktur* was published in early 2009 by the Research Council of Norway (RCN), with an initial overall budget framework of approximately NOK 400 million for the initial announcement. This call, with a deadline of April 22, is open to all scientific disciplines and encompasses several categories of research infrastructure described in specific calls, among

[1] Actual alotted amounts are dependent on the national budget. The current national budget partly complies this proposition, and a government fund with a start capital of NOK 4 billion will be established.

which the category most relevant to CLARIN is *Scientific databases and collections*. This program currently seems the best option for building up some of the language resources and technologies that will make up Norway's contribution to CLARIN, although heavy competition can be expected from all scientific disciplines, witnessed by the fact that in the pre-proposal round, the call was oversubscribed by a factor of 25.

On the one hand, a number of proposals for rather specific large-scale language resources are being prepared for this call, such as a database for speech and dialect data, one for syntactically and semantically annotated corpora, etc. It is expected that these infrastructure projects, if funded, will strive to be compatible with CLARIN, but it does not currently seem guaranteed that the results will in fact be incorporated in CLARIN. There is not even a plan for joining these resources in a single national infrastructure for language resources.

On the other hand, a coordination project entitled NO-CLARIN is submitted that ensures national networking and liaison of national activities to the CLARIN effort. NO-CLARIN will promote networking between actors and stakeholders in Norway through events and other communication. It will also run case studies and pilots to investigate the possible establishment of a Norwegian CLARIN center, while Nordic cooperation in this area will also remain a possibility. NO-CLARIN builds on previous coordination activities in late 2008, in particular a national seminar with 36 participants which also included several representatives from other Nordic countries. Current support schemes at RCN only cover networking and preparatory activities under the preparatory phase of ESFRI projects, while schemes for national support under the next phase of ESFRI projects are not yet available.

## 3 Nordic cooperation on language technology

The Nordic countries have a good record of cooperation and mutual understanding, partly thanks to regular cooperations in higher education, researcher training and research projects, partly stimulated by specific programs, in particular the recent Nordic language technology program (2000-2004, extended to 2005). The networking activities stimulated by this program did not only focus on specific research fields, but also included a coordinated documentation activity (NorDokNet) and an outreach to the Baltic countries in 2005. An extension and consolidation of these cooperation and networking efforts was attempted through bids for a Nordic Center of Excellence (2005), a Nordic documentation effort with industry through Nordisk InnovationsCenter (2005), and a Joint Nordic Use of Infrastructures (2007), but all three bids were unsuccessful.

However, in 2006 the Northern European Association for Language Technology (NEALT) was founded and established good publication channels. Furthermore, the Nordic language councils have a good tradition of cooperation that also encompasses the stimulation of language technology applications for the Nordic languages. As part of this cooperation, a working group on *Språkvård och språkteknologi i Norden* was established and a report *SpråkVis — Språkteknologisk vismansrapport* was ordered. It is in this spirit of Nordic cooperation and language appreciation that further joint work on language resources and technologies seems feasible.

## 4 Nordic initiatives in e-infrastructures

The Nordic countries have a number of instruments promoting research cooperation. In particular, The Nordic Council of Ministers provide funding of common actions in education and research through programmes and actions administered by NordForsk. One recent NordForsk initiative is The Nordic eScience Initiative, which may bear relevance to CLARIN. Its task is to *"... describe what Nordic level functions and services would be beneficial for coupling digital resources using Grid technology, including computational resources, data repositories and key research instruments. The proposed functions and services should, by federating resources and competences, add value to Nordic research communities and to the NGIs. Furthermore, the proposal may propose a joint Nordic framework for resource provisioning and sharing/aggregating national resources. The Nordic centers/metacenters have already made significant progress in this direction."* From this description, it appears that this task could be a good match for reaching the goals of CLARIN at a Nordic level.

## 5 Perspectives for Nordic cooperation on CLARIN

As mentioned above, there has been an unsuccessful attempt to obtain funding for a Joint Nordic Use of Infrastructures, but with careful planning, joint Nordic activities may still be realized. I believe that Nordic cooperation is beneficial because Nordic projects in this area will have the most efficient dimension. Nordic countries have good expertise, but since research groups are small, it is only through pooling that a critical mass will be reached. On the one hand, even at a national level, the research capacity in the area of language resources and technologies of a country like Norway is quite limited. On the other hand, full interaction between 23 countries at a European level is quite complex and requires enormous management resources. In contrast, Nordic cooperative projects would be of a manageable size, but at the same time they embody a sufficient economy of scale.

Research infrastructures are expensive to establish and run. CLARIN is currently estimated to cost EUR 23.2 million in the construction phase. While the data throughput on the CLARIN grid is expected to be smaller than typical amounts, for instance, in particle physics or climate research, language data is more heterogeneous and structured, such that curation of language data, as well as search in annotated data, is more complicated and expensive than for the huge amounts of data that is produced by the Large Hadron Collider experiments. CLARIN will therefore be a distributed facility relying on networked centers with special expertise at specific centers.

There will be also a need for physical platforms with large media for datastorage and supercomputers that perform searches in databases with good response time.[2] It is inevitable that demands will be placed on cost-efficiency; such demands are already being made in the Norwegian *Infrastruktur* program. In this context, the benefits of cooperation and the necessity to operate swiftly and efficiently make it natural to consider extending national networking efforts once again to a Nordic level, perhaps in the following ways:

1. Communication forums and meetings ought to be established to exchange and discuss common experiences, proposals and solutions, for instance through Nordic workshops on language infrastructure research and through invitations of other Nordic partners to national seminars.

2. A laison ought to be established between the Nordic partners in CLARIN and relevant Nordic actors in e-Infrastructure, including the eNoria Task Force on Sustainable Nordic Grid Collaboration, and NDGF, with the view of exchanging information between linguistic and technical communities.

3. The linking of national language infrastructure centers in a Nordic grid solution ought to be investigated and tried out in case studies and experiments. Such a grid might in the first instance be easier to achieve on a Nordic scale than on a full European scale.

4. Financing possibilities in order to support some of the above actions ought te be looked at on a Nordic level, perhaps also on national and European levels.

## 6 Conclusion

The main reasons for working on a Nordic level are the following. First, relevant actors at the Nordic level know each other, have a record of cooperation, and share a common culture (including a research culture). Second, there is an important 'Goldilocks' argument of finding the right size: whereas research communities in most of the Nordic countries are too small, and the European community may be a bit too big, the Nordic community seems just the right size. Third, there are existing Nordic cooperative initiatives in eScience that may serve as a frame, platform or jumping board, whichever metaphor one prefers. The best thing to hope for is that research and infrastructure activities on the various levels (local, national, regional and European) will not be in the way for each other, but will complement each other in the spirit of *subsidiarity*, in the sense that activities should be managed on the level where it is most efficient to do so.

## 7 Links

1. `http://www.clarin.eu`

2. `http://www.spraakbanken.uib.no/utredninger.page`

[2] Treebank searches, for instance, may involve arbitrarily complex graph traversals that place heavy demands on CPU power and memory.

3. http://www.regjeringen.no/nb/dep/kkd/dok/regpubl/stmeld/2007-2008/stmeld-nr-35-2007-2008-.html?id=519923

4. http://cordis.europa.eu/esfri/

5. http://link.uib.no/?vhuj

6. http://www.rcn.no

7. http://www.ndgf.org/

8. http://link.uib.no/?JAD3

9. http://www.cst.dk/nordoknet/

10. http://omilia.uio.no/nealt/

11. https://kitwiki.csc.fi/twiki/bin/view/Main/LTExpertPanelReport

12. http://www.nordforsk.org

13. http://www.nordforsk.org/text.cfm?id=499