Nordic co-operation in building the language resource infrastructures

Kimmo Koskenniemi
University of Helsinki
Finland
kimmo.koskenniemi@helsinki.fi

Antti Arppe
University of Helsinki
Finland
antti.arppe@helsinki.fi

Abstract

This paper attempts to identify worthwhile goals when building Nordic language resource infrastructures and the relevant parties who should participate their planning and construction. Finally, some actions are suggested which could move us closer to the goals which have been set.

1 Background

We have a long tradition of Nordic co-operation within language technology (Koskenniemi et al. 2007), including a long series of NODALIDA conferences, the Nordic Research Program 2001-2004, NGSLT, and we now have the NEALT organization which hosts special interest groups such as the SigInfra dedicated to research infrastructures for language resources. Similar co-operation has also been practiced in linguistics, e.g. the NordForsk summer schools and the Scandinavian Conference of Linguistics (SCL).

The *European Common Language Resource and Technology Infrastructure (CLARIN)* infrastructure entered its EC funded preparatory phase 2008-2010 and is creating frameworks according to which the operational CLARIN could be built. All Nordic and Baltic countries are participating CLARIN in various roles.

In Finland, FIN-CLARIN, a consortium of research institutions involved in linguistics and language technology has been formed in 2007 to strive towards CLARIN objectives at a national level. Currently, FIN-CLARIN encompasses the Universities of Helsinki, Joensuu, Jyväskylä, Oulu, and Tampere, the Research Institute for the Languages of Finland (KOTUS/FOCIS), and CSC – IT Center for Science, but the consortium remains open to all other Finnish academic organizations with an involvement in linguistic research or having language resources and technologies available for such research.

As the first step, the FIN-CLARIN consortium members have conducted in 2008 a survey of linguistic research resources and tools that exist within their organizations. In all, 76 distinct collections of resources have been identified in this survey, for which the key descriptive data, identifying the resource, its content, location, and access requirements are available at the FIN-CLARIN website as well as the general *ad hoc* registry maintained by CLARIN[1].

As a second step, the FIN-CLARIN consortium has commissioned from CSC – IT Center for Science a White Paper concerning the various possibilities for setting up a Finnish national Authorization and Authentication Infrastructure (AAI) for language resources, as well as a proposal covering the requirements specifications and actual construction plan for implementing such an infrastructure in Finland. Such an AA infrastructure is the technical bedrock which allows for the potential use of a language resource at any of the participating Finnish organizations according to the Single-Sign-On (SSO) principle, i.e. requiring a user's identification only at one's own Finnish home organization. In practice, this now completed development plan realizes the technical framework of the envisioned CLARIN infrastructure within Finland, and is planned to be fully conformant with the pan-European CLARIN AAI, the kernel of which is planned to be operational already in 2009. As the third step, the FIN-CLARIN consortium has commissioned from CSC the actual construction of this AAI in Finland within 2009.

2 Nordic goals

One important goal of Nordic research infrastructures for language resources is obviously to make language and lexical materials accessible

[1] see *http://www.clarin.eu/view_resources*

and usable for all those who need them for research, teaching, language planning or similar purposes. The access and use of existing materials should be facilitated, new materials should be created, and measures should be taken in order to secure maximally free availability of the future materials already when the materials will be created.

Just within the Nordic countries, the CLARIN infrastructure should allow for researchers interested in e.g. the overall state of the Swedish language, i.e. Swedish spoken and written both in Sweden and in Finland, to easily access the language resources currently physically located at several institutions, first and foremost *Språkbanken* (The Swedish Language Bank) in Göteborg, Sweden, CSC – IT Center for Science, Finland, the Department of Scandinavian languages and literature at the University of Helsinki, and the Research Centre of the Languages of Finland, regardless of what their home organization currently is. Likewise, the CLARIN infrastructure should allow for researchers in e.g. the Department of Finno-Ugrian Studies at the University of Helsinki to have easy access to the substantial Sámi resources at the University of Tromsø. In addition to such ease of access, the CLARIN infrastructure aims to provide user-friendly interfaces to aggregate such scattered resources as single virtual corpora, and to conduct the most common search and concordancing operations for researchers lacking extensive skills in language technology and programming, which would be necessary to work by themselves directly with the source format of the resources.

The resources for CLARIN or national language resource infrastructures are limited. In order to proceed fast and get the appropriate high quality services available, the Nordic participants now have an opportunity to get more by smart division of labour and by co-ordination, making the most of the current individual strengths of all the parties.

This paper also discusses how the Nordic countries could better integrate themselves in the European CLARIN which is, of course, the best, if not the only way to offer the Nordic researchers the access to materials and tools in other EU countries.

3 Actors

It is important to get the relevant parties involved, including but not restricted to:

- researchers in various disciplines such as linguistics, language technology, or machine learning who need linguistic materials in their research and who sometimes produce new materials,
- researchers in other disciplines who in fact essentially work with linguistic data, e.g. historians, sociologists, or theologians, just to mention a few fields,
- funders of research projects who can require allowing free access, and compliance with standard formats as new materials are produced as a result of the projects,
- specialists in language planning or language cultivation (*språkvård*), who utilize the materials in their work and compile new dictionaries, norms for language users, and compile new corpus materials,
- commercial parties such as publishers and broadcasting companies who own or possess written and spoken materials, as well as language technology companies who need written or spoken corpus materials and create language technology tools using these materials,
- libraries, museums, and some commercial companies such as Google and Microsoft Corporation which may have huge archives of materials and which are involved in digitizing and storing these archives,
- organizations of authors and journalists, as well as the organizations which process the copyright fees of authors and performers, and
- experts in copyright legislation.

There is an obvious need for attracting relevant parties to the work because relevant materials exist and are controlled by them. In addition, risks will increase if those parties are not motivated and co-operative.

At first sight, some of these parties might appear to have conflicting interests. It would be nice for the researchers if they could use all published materials on an open access basis. This might, however, conflict with the legitimate commercial interests of the publisher if they intend to print and sell copies of such a work. We think that there may still be workable compro-

mises where the commercial publisher can feel comfortable and safe at the same time as the researcher can use the texts and other language materials fairly freely. In order to find and establish such practices, one definitely needs contacts, discussions, and negotiations, and in the long run, relatively permanent, organized forums through which such activities take place. Importantly, establishing relations of trust between the various actors requires extensive engagement and time.

4 Organizing Nordic co-operation

Probably the best and only truly operational basis for Nordic co-operation with language resource infrastructures would be based on *national infrastructure consortiums* which are anyway needed in the CLARIN framework. They will be the essential *primary* parties in applying for national funding and in setting priorities for tasks and steps in building resources and the infrastructure.

The European CLARIN will neither build nor fund the national or regional CLARIN centres, and the European CLARIN will not build the materials for national languages. These tasks have to be funded and carried out nationally, and most likely through some national consortium which represents the most relevant parties.

SigInfra of NEALT is a special interest group dedicated for the advancement of Nordic co-operation in language resource infrastructures. SigInfra cannot, however, assume alone much of the responsibilities of building the national infrastructures. But SigInfra, together with national consortia, definitely can make the building of CLARIN compatible resources and centres much more successful.

In a nutshell, the organization could consist of national language resource consortia and a board consisting of one or two representatives nominated by each consortium.

5 Forms of co-operation

Let us suppose that there is a national consortium in each country which is building a national infrastructure for language resources. If so, that would provide an excellent basis for Nordic co-operation aiming at the integration of the national infrastructures into mutually compatible CLARIN nodes. Simply put, a board consisting of representatives from those consortia would plan, co-ordinate, and synchronize the common activities. The national consortia would then carry out the actual tasks which have been agreed upon.

The board could e.g.

- co-ordinate the collecting of certain information by the participating member consortia (such as an inventory of national digital text, speech and lexical materials),
- co-ordinate the application for any national funding and the implementation of the (successful) funding decisions, and store and make the results available as needed,
- initiate discussions and possible negotiations concerning the optimal selection of institutions and centres for various CLARIN service centres, along with the co-operation and division of labour between present or future CLARIN service provider centres,
- discuss and provide recommendations on types and levels of CLARIN metadata describing the language materials,
- discuss and co-ordinate producing, enhancing and sharing of software tools to become parts of CLARIN resources or services,
- apply for Nordic funding for arranging meetings about Nordic language resource infrastructures,

The board would have no resources and practically no funding of its own. All work would be carried out with the funding of national consortia and by their staff. Therefore, the adequately funded national research infrastructure consortia are crucial.

6 Expected results of the co-operation

There are many kinds of small or important results or benefits that could be achieved with Nordic co-operation.

One achievable goal would be that through cooperation, the CLARIN infrastructure in the Nordic countries could become operational earlier than if the countries would act uncoordinated. A common effort might have a better opportunity of getting adequate local funding. The cooperation might also help national efforts to find better practices and avoid (repeating) poor design and miscalculations, and to learn from the experiences of organizations which have had a

chance already to try out the construction of some service. For instance, the forthcoming Finnish experiences in setting up a national AAI for language resources could perhaps be utilized by other Nordic national consortia.

Another, equally important goal would be that the implementation of a good functional Nordic CLARIN might be less expensive to build. This could result in from the division of labour where partners concentrate their efforts in components where they have special expertise, and reuse parts which others have created, or simply benefit from the prior experiences of other partners.

There is a shortage of qualified technical people with the necessary skills to implement the technical infrastructure. Some computing centres at universities and national research institutions may have such personnel, but those centres may already be involved in a range of support activities serving many scientific fields. CLARIN is not the only research infrastructure within the European Union. Once we are able to secure such human resources in some organization in a Nordic country and, in addition, establish a good working relationship with such an organization to cater to CLARIN needs, we might as well make the most of such capacity throughout all the Nordic and Baltic countries.

The technological environment, in which CLARIN operates, is dynamic, and our regional infrastructure must prepare to adjust itself even after it has been constructed. For instance, cooperation among the existing national authentication (identity) federations requires a relatively extensive network of mutual agreements. It is possible that such regional federations might in a few years time be replaced by a single pan-European identity federation. Nevertheless, in the meantime we have to settle with what is possible or exists now. Solutions first adopted and the organizations initially providing services may thus change. CLARIN is a distributed research infrastructure which allows and requires the moderate duplication of resources and services which, in turn enables gradual development and improvement of the services.

Present archives of digital language materials are somewhat scattered. The acquisition of the material and management of permissions for their use necessarily involves many institutions. On the other hand, the data processing of language materials is mostly modest. Even collections containing some 10^{12} words of text are technically quite manageable, so that the processing and searching of such masses is not a real problem. But, managing standardized and high quality data security, state of the art authentication and authorization and metadata harvesting might consume a significant portion of the dedicated personnel resources at some relevant centres. Maybe, we could do with fewer centres (maybe even with a single one) to provide certain services. This should, of course, be accomplished so that end-users in all Nordic and Baltic countries can receive an equally high level of support and services regardless of their affiliation.

7 Conclusions

We urge that Nordic organizations with linguistic resources and tools formally establish a national CLARIN consortium for each Nordic and Baltic country, unless one already exists. If and when such already exist, we encourage that the national consortia be extended, if necessary, to include all relevant national organizations. These organizations should apply for European CLARIN membership. Parallel to this, we propose that the Nordic and Baltic national consortia formally establish a forum or an organ for cooperation and agree upon principles which guide this co-operation. It is our firm belief that such cooperation and coordination of Nordic CLARIN activities will be of substantial benefit to all involved parties.

References

Kimmo Koskenniemi, Krister Lindén and Torbjørn Nordgård (editors). 2007. *Expert Panel Report: The Nordic Countries, A Leading Region in Language Technology*. Publications, No. 44, Department of General Linguistics, University of Helsinki.