28 February 2011

# BAULT Application Report

## RESPONSIBLE PERSON OF THE RC:

*Last name:* Koskenniemi
*First name:* Kimmo
*E-mail:* kimmo.koskenniemi@helsinki.fi

## NAME AND ACRONYM OF THE RC:

*Name of the participating RC:* Building and use of language technology
*Acronym of the participating RC:* BAULT

## KEY FOCUS AREAS:

*Selected key focus area of UH:* 8. Kieli ja kulttuuri - Language and culture
*Comments for selecting/not selecting the key focus area:* More specifically, the Faculty of Arts has selected three major fields of research: (1) Cultural and linguistic diversity, (2) Language and interaction and (3) Language technology and corpus linguistics. The current RC fits to the field (3).

## FOCUS AND QUALITY OF THE RC'S RESEARCH:

The aim of the research community (RC) is (1) to make language resources (i.e. materials and tools) seamlessly available for the researchers, (2) develop and facilitate new types of language oriented research and (3) develop new multilingual computational methods for processing language materials.

The RC focuses centrally in the "Language and culture" area in the strategy of the University of Helsinki, and more specifically in "Language technology and corpus linguistics" which is one of the three main focus fields of the Faculty of Arts.

The group includes the FIN-CLARIN project which is listed in the national roadmap of the 20 new infrastructures to be built. FIN-CLARIN is a component in the European CLARIN (Common Language Resource and Technology Infrastructure) of the ESFRI roadmap.

FIN-CLARIN and similar services are used for linguistic research when studying language structure, its use and variation and when developing language technological applications such as spellers, parsers and machine translation.

CLARIN is committed in solving three problems for the user: (1) to find relevant materials when they exist, (2) provide a seamless way to get the necessary permissions to use the materials and (3) to enable the use of different materials together and use them with existing tools. The problems are solved by providing metadata for searching, an authentication and authorization framework for the permissions and standardization to guarantee the compatibility of materials and tools. Both written and spoken texts are stored and made available through CLARIN.

In particular, the RC is widely known for the leading edge research on finite-state computational morphology and surface syntactic parsing. Google Scholar shows more than 3000 references to publications by the Finnish researchers in this area. The recent Helsinki Finite State Technology (HFST) project has combined several international research groups to cooperate under the HFST platform.

The availability of language materials and tools for utilizing them has already changed the nature of language-oriented research. Using databases consisting of tens of thousands millions of words, one can instantly resolve problems which would have been impossible in the past. With adequate collections of texts, one can study the actual grammatical complexity in huge texts of different languages in order to show universal regularities and limits. One can also study the emergence of words and grammatical patterns and see how they are used, how their use varies according to areas and how they change in the course of time.

For Russian linguistics, the Integrum service offering access to practically all Post-Soviet newspapers and periodical is available for the RC and is heavily used. The corpora can benefit research in very different ways but only it researchers have the skills to use them. An example based on Russian material is given here. In Russian there is a structure which is a contamination of the passive and active voices. It is quite a normal construction among others in printed media but rather rare. The structure has been studied by about ten scholars but only on the basis of some 30 instances. When the researchers of the RC created a special way to find phrases with this construction, they could create a massive set of 3000 examples. On a thorough analysis of the examples, the researchers totally reshaped the description of this phenomenon. The paper on this topic published in Russian Linguistics was during half a year the most read article of the journal. Corpora can be used not only in studies on the lexicon or grammar, as it is usually done, but they can also give new information on people's communicative behavior. So, by using media text corpora, the same research group compiled a classification of reasons why people pretend to understand or not to understand. CLARIN aims at providing access to similar amounts of texts for our national and other European languages.

FIN-CLARIN has also produced annotated text collections (tree banks) and synonym thesauri (word nets) which are necessary in many branches of research, including research within the RC and research in other disciplines such as information retrieval and computer science. The initial but substantial versions of these were produced on a tight schedule by subcontracting the work to external parties. The tree bank contains some 17,000 sentences and the Finnish WordNet some 180,000 words. The results are now freely available on the net. The development of Finnish language technology was severely restricted because such resources were missing.

The RC has a long history of digitizing Oriental and African language materials, actually dating back to the 1960ies. In particular, the digital cuneiform corpora of the State Archives of Assyria by Simo Parpola, the digital corpus Indus script texts and its computer analysis by Asko Parpola and Swahili corpora and morphological and syntactic analysers by Arvi Hurskainen have been internationally noted and recognized to be at the top of their fields.

The RC works mainly at the Department of Modern Languages where the famous VARIENG research group is located. It is one of the forerunners of English corpora. The close contact with them makes it possible to expand cooperation in corpus knowledge outside the RC itself.

The RC works in cooperation also with the Finnish, Swedish and Finno-Ugric studies in getting all relevant content accessible for researchers through the infrastructure. In particular, dialectal Finnish Swedish conversation with transcriptions were collected.

Another area of activities is the multilingual language technology by Lauri Carlson where automatic and computer aided translation and terminology is the object of the studies and the theme of national and European projects.

*Ways to strengthen the focus and improve the quality of the RC's research:* The cooperation with the researchers of domestic and other languages and FIN-CLARIN can be improved, which would result in the widening of the archives of language materials. More materials would become visible and reachable for the scholars. A sufficiently large critical mass of users and language resources is essential.

In addition to the concrete building of the infrastructure, the training and support for using it is essential.

It would be very important that Finland becomes a member of the European CLARIN ERIC (which seems likely as the Ministry has signed the Memorandum of Understanding and will participate in the negotiations). CLARIN ERIC would imply longer term continuity and would upgrade CSC into a European CLARIN class A centre.

## PRACTISES AND QUALITY OF DOCTORAL TRAINING:

Practically all scholars of the RC participate (or have participated) as supervisors or students in the Langnet national PhD graduate school. Another national graduate school, the Finnish Graduate School of Language Technology, KIT-GS (2002-2009) was participated in by many members of the RC and was merged with Langnet in 2009 and became one of its thematic programs.

Recruitment of PhD students is accomplished according to Langnet procedures, i.e. open calls distributed through mailing lists and other media to reach prospective domestic and foreign students and a two-stage selection by a ranking of the supervisor pool of the subprogram and by final selection by the Board of Langnet. In KIT-GS, the Board made the selection of PhD students.

Many PhD students have a secondary supervisor in addition to the main supervisor. Langnet is multidisciplinary and most of the Finnish universities participate in it. In the Language technology program of Langnet, there are several students from the information technology of Aalto university and these form a group together with the students with a more humanistic background. This was also the practice in the KIT-GS where we had e.g. some students of information research from the University of Tampere and cognitive neurolinguistics from Turku.

The Nordic Graduate School of Language Technology (NGSLT) was funded in 2004-2008 by the Nordic Council of Ministries. NGSLT enabled the entry of the young PhD students in the international research community by using leading edge scholars as teachers in researcher courses common with the KIT-GS. The students who were to become future scholars and teachers got a wide view of the emerging research topics.

PhDs of language technology have careers both inside and outside the academic circles. A majority of them are currently or have been working in commercial companies developing relevant methods. Linguistic PhDs are typically more interested in academic careers.

At the University of Helsinki, all doctoral and master's level students of Russian are taught to use sophisticated language databases such as Integrum and Russian National Corpus. There are plans to extend this practice more widely in the Langnet graduate school.

RC's strengths and challenges related to the practises and quality of doctoral training, and the actions planned for their development: The students and the supervisors are multidisciplinary but they have very different backgrounds which is both a strength (or an opportunity) and a weakness (or a threat). Students with a humanistic background would need support in using appropriate methods, and the required training and support is not readily available. The problem with many linguistic students would not be fixed through a short course, because they would often need more extensive background knowledge in mathematics and statistics. Providing suitable courses would be possible but it takes time and requires courses and teaching material which are not available at present. The problem is recognized by Langnet and the RC. New courses for fixing this gap are planned.

## SOCIETAL IMPACT OF RESEARCH AND DOCTORAL TRAINING:

The language technological tools and modules for local languages are essential for enabling the use of those languages in all facets of normal life. Otherwise there is a threat that local languages (Finnish, the three Sámi languages spoken in Finland) will lose ground and prestige and be more likely to be replaced by world languages.

The META-NET where the RC is a member, addresses the commercial and societal use of language resources and technology. The RC is a major partner in the META-NORD project of the EU FP7 and its task is to gather connections and form a community also for non-academic parties. This complements the goals of the CLARIN efforts which address mainly the academic needs.

FIN-CLARIN builds some resources by subcontracting them from the commercial enterprises based on competing offers. Some results, e.g. the open source HFST tools are used by commercial companies. In this way, information about methods and resources flows in all directions.

National Library digitizes huge amounts of written materials for their own purposes of archiving. The RC cooperates with them in order to enable the research use of the materials. Part of the material is old and in the public domain. The RC has started negotiations with an organization of authors and publishers, Kopiosto, in order to allow the restricted research use of even copyrighted materials using so called agreement licenses.

Ways to strengthen the societal impact of the RC's research and doctoral training: Tieto ja viestintäteollisuuden tutkimus TIVIT Oy is a Finnish Strategic Centre for Science, Technology and Innovation (SHOK) which represents the IT industry. TIVIT has indicated its support for the FIN-CLARIN-CONTENT project funded by the Academy (2011-2015). The RC will tighten its cooperation with TIVIT in the near future.

The RC is planning plans to strengthen the cooperation with the National Library (Kansalliskirjasto) in reusing the digital materials there and in enhancing the library services using tools and methods developed by the RC.

Producing the missing language resources and tools for Finnish is a part of the so called "Basic Language Resource Kit" or BLARK which is considered essential for the survival of any language in the world which becomes more dependent of technology.

## INTERNATIONAL AND NATIONAL (INCL. INTERSECTORAL) RESEARCH COLLABORATION AND RESEARCHER MOBILITY:

- National Langnet graduate school covers all branches of linguistics, including the RC, and our scholars act as supervisors in Langnet
- EU Marie Curie project CLARA specifically promotes researcher mobility. The RC has had two incoming positions, each for two years.
- During the NGSLT funding in 2004-2008, PhD students had more international contacts, especially with other students in the Nordic countries.

RC's strengths and challenges related to research collaboration and researcher mobility, and the actions planned for their development: Strong international contacts through past and present EU, Nordic and other international projects make the RC visible among international research institutions. Visitors appreciate the wide expertise of UH but Finland is not the first choice for a foreigner to stay for longer periods.

## OPERATIONAL CONDITIONS:

The language infrastructure is built for improving the language research and making it more efficient and productive. Limited facilities exist now. We think that building the new national and international infrastructure has a certain priority at this stage and more content and coverage should be provided as soon as possible. At the same time, basic research using the existing infrastructure is vital for guiding the building new facilities and services.

Teachers have a full set of courses to deliver, project researchers and PhD students participate in the teaching with a small share which appears to be beneficial both for the career development of the staff and for the breadth of topics offered for the students.

RC's strengths and challenges related to operational conditions, and the actions planned for their development: The circumstances for building the CLARIN infrastructure are now favourable, the network of user institutions has been built, the IT expertise is available, the participating Finnish institutions have expressed their desire to cooperate and contribute by depositing their language resources in the common infrastructure. The funding for FIN-CLARIN is available until the end of 2012. According to the plans on the national roadmap of infrastructures, the building of the FIN-CLARIN needs about the same level of annual funding until the end of 2020. The infrastructure becomes usable in steps. The first stages of the infrastructure, namely the authentication and authorization system were opened in February 2011.

Once the national and European CLARIN infrastructure is fully up and running, it enables a new kind of research which is more productive and of higher quality because the research results will be open for repeating and checking by the rest of the research community.

## LEADERSHIP AND MANAGEMENT IN THE RC:

For historical and practical reasons, Kimmo Koskenniemi is the principal investigator of several projects such as the EU CLARIN, FIN-CLARIN, CLARA and META-NORD. Krister Lindén is the actual research director participating in the boards of the projects, hiring personnel and directing all activities.

FIN-CLARIN itself is divided into five teams with specific tasks and leaders (the word bank, the treebank, the language bank, the applications and the theory). FIN-CLARIN represents several institutions: University of Helsinki, University of Tampere, University of Eastern Finland, University of Oulu, University of Jyväskylä, University of Turku and Aalto university, and the Research institute for languages in Finland (Kotus) and the CSC. FIN-CLARIN has a board which consists of a member from each participating institution.

VARIENG is directed by professor Terttu Nevalainen and it is not part of this RC in the current research evaluation. Only professor Nevalainen has a shared presence also in this RC.

Slavists, orientalists and other linguists form their own groups. Prof. Mustajoki leads the Slavists, professor Asko Parpola the Indologists, professor Simo Parpola the Assyriologists etc. Most of these are principal investigators themselves and participate in other RCs as well.

Fred Karlsson has led two typologically and corpus linguistically oriented projects on morphological and syntactic complexity. One of the main results is that there are clear quantitative and qualitative restrictions on all types of nested and tail recursion.

RC's strengths and challenges related to leadership and management, and the actions planned for developing the processes: The formation of CLARIN-ERIC will be critical because it will greatly affect the operations of the RC essentially. The ministry will decide the future organization leadership, management and funding of the national CLARIN consortium. It would be desirable that the future organization be mostly consistent with the present one.

The arrangement also decides the division of labour between the University of Helsinki hosting the FIN-CLARIN and the CSC.

## EXTERNAL COMPETITIVE FUNDING OF THE RC (in euros):

AF: 4,405,000
TEKES: 1,725,000
EU: 1,388,000
ERC: 0

FOUNDATIONS
Names of the foundations:
Amount of funding: 0

OTHER INTERNATIONAL FUNDING
Names of the funding organizations:
- Nordic Council of Ministers/NorFA
- Forsmark Nuclear Plant/Nordic Nuclear Safety Research
Total amount of funding: 167,000

OTHER NATIONAL FUNDING
Names of the funding organizations:
- Ministry of Education (and Culture)
- Nokia Oyj
Total amount of funding: 6,766,000

## RC'S STRATEGIC ACTION PLAN FOR 2011 - 2013:

- Participation in and the full utilization of the initial core of the European language resource infrastructure CLARIN, opening European language archives to Finnish scholars through seamless access.
- Coordinating and building the national Finnish language resource infrastructure FIN-CLARIN.
- Training more scholars to use the language resource infrastructure.
- Enabling and advancing repeatable and verifiable language research.
- Starting a project possibly within the programmes of TIVIT in order to upgrade CLARIN so that users may access and search large collections of materials even if parts of them reside in different centres, and to seamlessly combine tools available at different centres for processing such materials.
- VARIENG subject specializations in corpus compilation, annotation and metadata design will benefit the Building and Use of Language Technology (BAULT) and the FIN-CLARIN initiatives. Teaching collaboration within these frameworks is also expected to intensify.

## CONTRIBUTION OF THE RC MEMBERS IN COMPILATION OF THE STAGE 2 MATERIAL:

All members participated in compiling the material. Facts and initiatives collected and earlier versions of the submission text were distributed to members for comments.  The final revision of the text was submitted.