KUNSTI – Knowledge Generation for Norwegian Language Technology

Bente Maegaard (1), Jens-Erik Fenstad (2), Lars Ahrenberg (3), Knut Kvale (4), Katarina Mühlenbock (5), Bernt-Erik Heid (6)

(1) CST, University of Copenhagen, Njalsgade 80, DK-2300 Copenhagen, bente@cst.dk
(2) University of Oslo, P.O. Box 1053 Blindern, 0316 Oslo, Norway, jfenstad@math.uio.no
(3) University of Linköping, , SE-58183 Linköping, Sweden, lah@ida.liu.se
(4) Telenor R&D, Snarøyveien 30, N-1331 Fornebu, knut.kvale@telenor.com
(5) Sahlgrenska University Hospital, DART, Kruthusgatan 17, SE-411 04 Göteborg, katarina.muhlenbock@vgregion.se
(6) The Research Concil of Norway, P.O.Box 2700 St.Hanshaugen, 0131 Oslo, Norway, beh@rcn.no,

Abstract

KUNSTI is the Norwegian national language technology programme, running 2001-2006 inclusive. The goal of the programme is to boost Norwegian language technology research. In this paper we describe the background, the objectives, the methodology applied in the management of the programme, the projects selected, and our first conclusions. We also describe national programmes form Sweden, France and Germany and compare objectives and methods.

1. Background

During the 1990's, the commercial potential of language technology became increasingly important. In Europe, several nations recognised that efforts within this area of technology would also be of cultural importance in order to strengthen their own language as a viable alternative to the vast supply of English-based products and services. Norway has around 4.5 mill. inhabitants; a quite small market for products and services that understand Norwegian. Thus, a special effort for Norwegian language technology was requisite if the Norwegian language is to be treated at a reasonable level. At that time Norway had very good research teams, but they were too small and too scattered, so education and training of new researchers and collaboration between the various teams were required.

The Research Council of Norway (RCN) divisions of 1) Culture and Society, 2) Science and Technology and 3) Industry and Energy started in 2000 a co-operation on a study of the research demands in Norwegian language technology (*Språkteknologi i Norge* 2000). Based on this study, the Culture and Society Research Board resolved on 14 September 2000 to initiate a new national basic research programme called: "Knowledge Generation for Norwegian Language Technology" (KUNSTI). With a total budget of 10 mill. EUR the KUNSTI-programme funds eight projects over its program period from 2001 to the end of 2006.

2. Main objectives and priorities

2.1. Basic research, tools and language resources

The main objectives of the KUNSTI Research Programme comprise two closely connected aspects:

1. Strengthening **basic research and skills** in the areas of computational linguistics and speech technology, and areas of relevance to language technology within other fields such as computation, information science, phonetics, linguistics and Norwegian language.
2. Research and development aimed at creating **language technology resources and tools** for spoken and written Norwegian in various forms and, to a lesser extent, also Sámi.

In order to meet these objectives the KUNSTI programme has:

- Promoted increased recruitment, partly in the form of doctorates, partly as post-doctoral fellowships.
- Encouraged publications in international refereed journals and conferences, not only by the project participants themselves, but also collaborative authorships with research workers from foreign sites.
- Paved the way for research projects involving collaboration between research sites, as well as between research and industry.
- Defined two main research areas (see below), and demanded a functional demonstrator each of them.
- Requested a number of language technology tools and computer resources to be developed, having general application beyond the specific projects and even, perhaps, in connection with other types of products than those targeted by the programme objectives.

KUNSTI's purpose was partly a general upgrading of Norwegian language technology research sites and partly results in the form of theoretical insights, prototypes, tools and language resources (LRs) which might be used in developing Norwegian language technology. The target groups were therefore the R&D environments at academic institutions and companies developing language technology. There were also openings for including projects addressing the Sámi language, providing these were otherwise within the scope of KUNSTI.

2.2. Priority research themes

It is an aim that basic and applied projects within KUNSTI should be able to co-operate to the greatest possible extent. This was achieved by relating both types of projects to the same type of language technology application. KUNSTI therefore decided to accord priority to two such areas of application which, at the same time, are sufficiently complex to have relevance to a broad spectrum of research topics:

- Machine translation (MT) and multilingual word processing with emphasis on Norwegian
- Spoken Norwegian dialogue between man and machine.

The themes are loosely interpreted. Neither did the Programme Board exclude research topics falling outside these two areas, for instance projects in connection with information searching in large text/speech databases.

3. Programme management

The management of this programme may have deviated somewhat from the traditional programme management in the Research Council of Norway (RCN).

First, the Board discussed how the main objectives can be achieved, and this led to the secondary objectives, mostly to be seen as means to arrive at the desired result of promoting Norwegian language technology research and application, not as objectives in themselves.

In order to build a firm knowledge base with certain volume over time the KUNSTI Board decided to define one large project within each of the prioritized research areas; MT and spoken dialogues, each with a budget of around 3 mill. EUR. Both projects have a large degree of cooperation between institutions. In addition KUNSTI has supported 6 other projects and a small number of pre-projects. See section 4.

The Board has kept the relation to the projects by nominating a contact person from the Board for each project. This contact person has made site visits to the largest projects, and has had on demand consultations with any project. Finally, the Board organised annual meetings for all project participants. These meetings were two day seminars with presentation of all projects, discussion of common themes, such as semantic representation or language technology evaluation methodologies. At the same time, face-to-face meetings were held with individual projects where relevant.

Traditionally, research projects funded by RCN do not meet the Programme Board once they have been accepted.

4. The KUNSTI projects

The KUNSTI programme has supported 8 projects which are briefly described below

LOGON –Lexicon, Lexical Semantics, Grammar, and Translation for Norwegian

- Is the largest KUNSTI project on the text side
- Co-operation between the universities in Oslo, Bergen and Trondheim.
- Develops an experimental machine translation (MT) system from Norwegian to English.
- Works with Lexical Functional Grammar (LFG), as well as Head-Driven Phrase Structure Grammar (HPSG) – LFG for Norwegian and HPSG for English.
- Uses MRS (Minimal Recursion Semantics) for the semantic description
- The MT system has a fairly traditional architecture based on transfer, but it includes several new ideas. In addition to the symbolic MRS-based transfer, it will be refined with stochastic ranking mechanisms.
- Web page: www.emmtee.net

BRAGE – Speech centric dialog systems

- Is the largest project in the man-machine spoken dialogue field
- Co-operation between three partners: Telenor Research and Development, SINTEF ICT and the Norwegian University of Science and Technology (NTNU) represented by three departments (Electronics and Telecommunication, Computer and Information science, and Language and Communication Studies).
- The overall goal of the project is to develop state-of-the-art dialogue systems for the Norwegian language.
- Is implementing and testing speech only demonstrators based on:
  - Spontaneous speech over telephone;
  - A user friendly "mixed initiative" dialogue;
  - Synthesized speech response;
- Is implementing and testing multimodal demonstrators based on:
  - Composite input ("tap and talk");
  - Composite output ("text, graphics and synthetic speech");
  - A wireless client/server system;
- Web page: www.iet.ntnu.no/projects/brage/

FONEMA - Tools for realistic speech synthesis in Norwegian

- Cooperation between the Institute for Electronics and Telecommunications at the Norwegian University of Science and Technology (NTNU) and Telenor Research and Development.
- The main goal of the project is to establish a framework for speech synthesis with a high degree of naturalness based on unit selection waveform concatenation.
- Main focus is on the TTS back-end, i.e. on prosodic modeling and prediction and on speech generation.
- Web page: www.iet.ntnu.no/projects/fonema/.

BREDT – detecting and processing co-reference

- Work performed at the University of Bergen
- Main focus on machine learning of co-reference.
- Web page: spraktek.aksis.uib.no/projects/bredt/

KunDoc – Knowledge-Based Document Analysis

- Co-operation between CognIT and the University of Bergen
- Web page: www.kundoc.net

KB-N - KnowledgeBank for Norwegian for the economic-administrative domains

- Is aiming to establish a knowledge-bank for economic-administrative domains.
- Test a small part of it in LOGON
- Web page: mora.rente.nhh.no/projects/kbn

TREPIL - Norwegian treebank, a pilot project

- Is aiming at developing methods and tools for building a Norwegian treebank.
- Web page: helmer.hit.uib.no/trepil/

Sámi grammatical analysis

- The aim is to make a parser and disambiguator for Northern and Lule Sámi, to automatically tag a large

text corpus, and make it available to the research community through an adequate search interface
- Web page: gjellatekno.uit.no/

5. The functional demonstrators

The requirement that the large projects should build a functional demonstrator has led to the desired collaboration between the various teams in Norway. The demonstrators are briefly described below.

LOGON

Within the LOGON project the somewhat unconventional approach of combining two different grammatical frameworks for the analysis and generation phases (viz. LFG and HPSG, respectively) has proved to be a good match where different research interests and traditions have met. A functional demonstrator targeting tourism-related publication has been assembled, which already at an early stage of the project contained 100 representative sentences from the corpus together with hand-constructed test suites. A reference corpus of 50,000 words with high-quality literal translation has been selected for the final implementation.

- The system (and hence the demonstrator) rests upon a fairly traditional transfer-based approach, strengthened with several new techniques: Implementation of end-to-end MRS-based translation, allowing the components to defer the resolution of ambiguities and supporting flexibility between the three different phases of translation.
- Two new elements added to the transfer process, namely:
  - the use of typing for hierarchical organization of the transfer rules
  - a chart-like treatment of transfer-level ambiguity.

The initial results are already promising with respect to the feasibility of the approach,. Further on, an integrated part of the entire project will be to measure the success of the demonstrator by end-to-end evaluation.

BRAGE

The BRAGE project focuses on dialogue systems of realistic complexity and user friendliness. Further, user interest in the application is mandatory in order to be able to evaluate (and improve) the dialogue systems. User interest is normally dependent on an access to an updated, full scale application database. Thus an early goal was to a) find one or more applications which fulfill the above requirements and to b) specify the corresponding performance requirements for the different modules. This resulted in the following specifications:

Bus information (BUSTER) in Trondheim was chosen as the first speech only demonstrator.

- The ASR module specification was as follows:
  - Input is spontaneous speech over the telephone;
  - A semantic relevant vocabulary of around 1000 words;
  - A semantic analyzer to identify 5-6 classes;
- The dialogue manager should be able to:
  - Switch between query based, system driven and mixed initiative modes;
  - Accept corrections and multiple requests;
  - Include compact user input verification;
  - Produce compact and informative prompts;
- The development should include:
  - A text-based version;
  - A Wizard-of-Oz (WoZ) version;

As to multimodal demonstrators, no systems for mobile terminals are yet public available. Further, few guidelines exist with respect to user friendliness and tailoring to different applications and user groups. Thus, a major goal within this project is to identify users for which multimodality will be especially beneficial and to gain experience about user friendliness for these groups. A special focus was set on ”inclusive design”, and disabled persons were chosen natural candidates for these experiments. Again, user interest and database access resulted in the following specification:

- Bus information (“Trafikanten”) in the Oslo area was chosen as the first multimodal application.
  - The user interface should consist of:
    - Composite talk and tap;
    - Map based graphics combined with text and TTS
- User friendliness should be evaluated both for abled and disabled persons.

FONEMA

- Develops a Norwegian text-to-speech (TTS) demonstrator which will act both as a research tool and a means for demonstrating TTS quality.

6. Comparison with other national programmes

There are several other national programs within the area of language technology. Some of them, e.g. the French and Dutch/Belgian, are focusing on building infrastructures and language resources in form of databases of annotated speech and text in their own languages. Below is a very brief overview of some activities in other countries.

Sweden

Since 1990 Swedish funding agencies have supported a number of research efforts geared towards language technology. These programs have lasted for up to three years with budgets in the range of 800,000 Euros annually. The last programme ended in 2004.

Some features of KUNSTI, such as the goal to increase cooperation between sites, have been pushed in Sweden also with tangible success. However, compared with the Swedish programmes, KUNSTI is more long-term, has more resources and more focus on international cooperation and visibility.

- At The Royal Institute of Technology (Kungliga Tekniska Högskolan – KTH) a Centre for Speech Technology (CTT) was established in April 1996 as a long-term enterprise for cooperation between Swedish companies, non-commercial organizations

and academic research within the strategically important area of speech technology. External support from the funding agency is however due to end by 2006.

- Web page: http://www.speech.kth.se/ctt/about.html

France

- The **Technolangue** programme
  In France the Technolangue programme has been running over the past 3 years. It has four main objectives: provide language resources, do comparative evaluation of technologies for the French language, promote standardization and finally perform technology watch. KUNSTI started before this programme, and objectives are different. The French approach would not have been suitable for Norway at the time.
- Web page: http://www.technolangue.net/

Germany

- **Verbmobil**
  Germany is very well positioned in Europe wrt. language technology, with more than 60 companies active in the market and more than 80 research labs in 2003 (Lockwood et al. 2003). This is certainly due to the long-lasting support for language technology, e.g. the Verbmobil project at the end of the 1990s. This project coordinated public and private sector investment in research and R&D by bringing together specialists from research and industry in a very ambitious project: mobile speech-to-speech translation.
  KUNSTI has not followed the idea of one single umbrella project; however, the focus on functional demonstrators and on collaboration, both between universities and between universities and industry, is in line with Verbmobil.
- http://verbmobil.dfki.de/overview-us.html

7. Programme results

KUNSTI ends in 2006, and we may already try to evaluate the results. One of the aspects of such an evaluation would be to consider the selection of projects and evaluate the extent to which they fulfill the objectives. We feel that we have a good balance between written and spoken Norwegian. We are also happy that we were able to include a Sámi project, so that a minority language is represented.

One of the assumed resources for KUNSTI was the Norwegian Language Bank (Svendsen et al. 2002), a huge collection of written and spoken data to be used as base material for language technology research and development. However, this Language Bank has not yet been created and the consequence of this is that several projects have had to produce the necessary language resources as part of the project. These resources will be available for further research, so the investment is not lost – to the contrary – but this has diverted time and money away from language technology into resources building.

At a more positive note we can see that the training of young researchers is reaching a good level. 10 PhDs and 2 post docs are being funded by the projects. This is the basic way to increase the number of qualified researchers in Norway, and KUNSTI has reached its goals.

The requirement that the large projects should build a functional demonstrator has led to the desired collaboration between the various teams in Norway. As shown above, the KUNSTI projects have collaboration by teams in different locations and institutions, cross-disciplinary collaboration between different departments in the same university, and collaboration between research and industry.

The programme participants have also been able to establish international cooperation with expert teams in Europe and elsewhere. This international collaboration will have effects long after the end of the KUNSTI programme.

Besides, the scientific results are now beginning to emerge, but it is premature to evaluate these at present.

The KUNSTI Board is satisfied with the preliminary effects of the programme, and plans are being prepared for a continued effort in language technology, in order to capitalize on results and cover other, yet uncovered areas for Norwegian. It is important for Norway and for the Norwegian language to have a strong research basis in language technology. This is true both for human resources and for research and development results that can form the basis for industrial development.

8. Acknowledgement

The authors would like to thank the KUNSTI projects for their contribution in form of project reports.

9. References

KUNSTI – Knowledge Generation for Norwegian language technology, Programme Plan (2001). Research Council of Norway, Oslo.

Lockwood, R., A. Joscelyne (2003): *Benchmarking HLT progress in Europe*, Copenhagen.

Språkteknologi i Norge – eksisterende og påkrevet forskning, (2000) Rapport fra en arbeidsgruppe, Norges Forskningsråd, Oslo.

Svendsen, T., Nordgård, T., Andreassen, L.H., Berg, J.T., Kvale, K., Espeli, T., Johansson, S., Breivik, T., (2002): *Consolidating and Increasing the Availability of Norwegian Human Language Technology Resources*, Oslo