KUNSTI – Knowledge Generation for Norwegian Language Technology

An Interim Report from The Programme Board

Norges forskningsråd

The Research Council of Norway
P.O.Box 2700 St. Hanshaugen
N–0131 OSLO
Telephone: +47 22 03 70 00
Telefax: +47 22 03 70 01
bibliotek@forskningsradet.no
www.forskningsradet.no/english

The report can be ordered at:
www.forskningsradet.no/publikasjoner

or green number telefax: +47 800 83 001

Number of copies: on demand

Oslo, February 2006

ISBN 82-12-02271-4 (trykksak)
ISBN 82-12-02272-2 (pdf)

0. Preamble

This report describes the situation of the KUNSTI programme in the spring of 2005. Results obtained since then may be found for each project, following the link given in the section.

The report is produced by the Programme Board, which has the following members:

- Managing Director, Professor **Bente Maegaard,** *chair person*
  Center for Language technology, Copenhagen, Denmark
- Professor **Lars Ahrenberg**
  Institute for Computer Science (Institutionen för datavetenskap - IDA), University of Linköping, Sweden
- Professor **Jens Erik Fenstad**
  Institute for Mathematics, University of Oslo
- Senior Researcher **Knut Kvale**
  Telenor FoU/R&D, Oslo
- Director **Katarina Mühlenbock**
  DART - The Centre for AAC and computer access in west Sweden for children, young people and adults with disabilities, Gothenburg

1. Introduction

Language technology is technology incorporating information about natural language with the aim to perform linguistic and cognitive processes, including imitate aspects of natural language skills, such as automatic dictation, speech-based dialogue systems, automatic language control, machine translation and information searching in large data sets, both text and speech. Language technology is founded on several different subjects and research fields, from both technology and the humanities. During the 90's, the commercial potential of language technology became increasingly important in Norway. In addition, Norwegian efforts within this area are also of cultural importance in order to strengthen Norwegian as a viable alternative to the vast supply of English-language products.

The Research Council divisions of Culture and Society, Science and Technology and Industry and Energy started in 2000 a co-operation on a study of the research demands in Norwegian language technology. Based on this study, the Culture and Society Research Board resolved on 14 September 2000 to initiate a new programme for language technology. This programme was to be run as an internal co-operation in the Research Council between the three divisions mentioned above. The Programme is running for 6 years, from 2001 to 2006 inclusive.

A total of eight projects are now funded by the programme. They all started in 2002-03 and are now beyond mid-point. The Programme Board expects that a report on the structure of the projects and the results obtained so far is of interest at this point, in order to provide input for the discussion about possible follow-up activities.

To place the individual reports in a proper context we have included some general information on the original programme objectives. There is also a section on the management of the programme with special emphasis on the close interaction between the Board and the individual projects throughout the project period.

2. Programme objectives

The structure and the content of the Research Programme are highly inspired by the study report mentioned above (*Språkteknologi i Norge – eksisterende og påkrevet forskning*, 2000).

2.1 Main objectives

The objectives of the 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 computer science, information science, phonetics, linguistics and Norwegian language.

2. Research and development aimed at creating language resources and tools for spoken and written Norwegian in various forms and, to a lesser extent, also Saami.

Whilst the activities under item 1 are typically oriented towards development of theories and methods, activities under item 2 will to a greater extent generate concrete results in the form of systems and linguistic data. However, individual projects often incorporate activities under both headings.

Language technology is distinguished by its pronounced multi-disciplinary basis. The contacts it generates across traditional departmental divides (for instance between the humanities and technology) are not of a superficial or enforced character, but dip deep into the respective subjects' central fields of interest. The important new insight arising at the new interfaces makes the multi-disciplinary aspect not only a matter of consequence for a variety of applications, but also an issue of basic research.

At the same time, not all research in the relevant subjects has been covered by the Programme, which is limited to projects explicitly accommodating language technology issues, i.e. issues concerned with the automatic processability of natural language and which, whether they be basic or applied, should lead towards insight which can be implemented.

2.2 Secondary objectives

An important objective of the Programme was to promote increased recruitment, partly in the form of doctorates, partly as post-doctoral fellowships.

Publication in international refereed journals has been encouraged, preferably based on collaborative authorship with research workers from foreign sites.

In accordance with the Programme's primary objectives, the intention was for the majority of the Programme's activity to centre on collaboration entailing definite added value, and there was a wish to see projects involving collaboration between research sites, as well as between research and industry. This was regarded as a necessity if the Programme was to contribute to the creation of new expertise in language technology.

Within each of the Programme's two main areas (see section 2.4) one of the results should be a prototype or a functional demonstrator.

Another objective was that within the scope of the larger collaborative projects a number of language technology tools and computer resources should 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.

2.3 Target groups

Participants in the Programme were expected to be:

1. Research sites and researchers within the disciplines upon which language technology is based
2. Other sites working with innovative, research-based development of language technology applications.

The Programme's purpose was partly a general upgrading of Norwegian language technology research sites and partly results in the form of theoretical insights, prototypes and resources 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 the Programme.

2.4 Priority research themes

It was an aim that basic and applied projects within the Programme should be able to co-operate to the greatest possible extent. The aim was to achieve this by relating both types of project to the same type of language technology product. The Programme therefore accorded priority to two such areas of application which, at the same time, were sufficiently complex to have relevance to a broad spectrum of research topics:

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

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

The Programme encouraged a variety of forms of project co-operation in connection with the priority themes.

3. Programme management

The management of this programme may have deviated somewhat from the traditional programme management in the Norwegian Research Council.

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.

Secondly, when project applications were received, and had been reviewed through the traditional Research Council reviewing process, the Programme Board discussed those that had a positive review with the applicants. During these discussions, the Programme Board expressed its visions to the projects, and certain changes were made by the projects. In particular the discussions with the two largest projects were very fruitful. In the past, projects were either accepted or rejected, but this Board also influenced the projects.

The Programme has supported a small number of pre-projects and 8 projects. The two largest projects are in the fields of MT and spoken dialogue as foreseen. These two are collaboration projects, and they are pretty large (around 3 mill. EUR each). All projects are described in 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 do not meet the Programme Board once they have been accepted.

4. KUNSTI projects: Reports on work in progress

The project managers have been asked to provide a short progress report on the individual projects. Each report should clearly describe the aim of the particular project, the methodology and the project structure, and the results obtained so far. They were asked to write for their professional peers. This means that not every part of the reports is equally understandable to everyone, but the hope is that the reports taken together are of interest to the international research community as an example of an applied research program in natural language technology.

4.1 LOGON

Introduction. LOGON is the largest KUNSTI project on the text side and includes three universities: Oslo, Bergen and NTNU, Trondheim. It is a response to the call for proposals which asked for a large project including several groups working on an integrated effort resulting in a functional demonstrator reflecting a multitude of tasks and techniques. The choice fell on the development of an experimental machine translation (MT) system from Norwegian to English. The system has a fairly traditional architecture based on transfer, but it includes several new ideas. In addition to the symbolic transfer, it will be refined with stochastic ranking mechanisms.

The basic underlying assumptions of the project are: (i) Translation is at its core a semantic activity—a main concern for any translation is to preserve interpretation in context, or at least the meaning of the source expressions. (ii) Deep linguistic methods, based on full grammatical analysis, have reached a level of sophistication and broad coverage where they, combined with faster algorithms and computers, make real world applications feasible. (iii) Empirical studies of language use and statistical preferences based on these studies are necessary in language technological applications. Only when these techniques are combined with proper theoretical models will we see their full potential.

Semantics in machine translation. We have chosen semantic representations as the level of transfer. The specific meaning representation language used is Minimal Recursion Semantics (MRS). This is one of the logically based meaning representation languages within computational semantics which facilitates underspecification of scope relations. As the MRSs are implemented in a typed formalism with inheritance they also enable generalization over classes of predicates, and thereby enable MT components to defer the resolution of ambiguity. We follow the main trend within the MRS tradition and apply an event based (Neo-Davidsonian) logical language.

Grammars, generation and analysis. To begin at the far end, realization of post-transfer MRSs in LOGON builds on the pre-existing LinGO English Resource Grammar (ERG) and LKB generator. ERG is an HPSG-based grammar developed over several years to serve various project applications and domains, starting with VerbMobil. Within LOGON it has been further developed to the hiking domain, and currently has a 21,000 lexeme lexicon and 95 per cent coverage on our initial development corpus. ERG was already equipped with MRS before the start of LOGON. ERG is written to work with the LKB system which is a tool for parsing and generation with type-based grammars. Yet, out-of-the-box LKB generator performance was not satisfactory and one task within LOGON has been to enhance it.

On the analysis side we chose to use an LFG-based grammar for Norwegian, NorGram. This is the most developed computational deep grammar for Norwegian, currently with more than 80,000 lexemes in the lexicon, and it is under active development of one of the involved groups. There is also a theoretical motivation. We want to consider the transfer representations at a semantic level independent from particular grammar formalisms. While MRS was originally developed in an HPSG setting, we have shown that it can also be integrated with other grammar formalisms. NorGram assigns the usual LFG representations c-structure (PS tree) and f-structure (expressing grammatical relations like subject and object) to sentences. The LFG architecture allows the projection of new representations by co-description, and the MRS-structure is projected off the f-structure in this way. For the LFG based analysis the XLE system developed at PARC is used.

Transfer, system integration and regression testing. Unlike in parsing and generation frameworks, there is less established common wisdom in terms of (semantic) transfer formalisms and algorithms. LOGON follows many of the main VerbMobil ideas—transfer as a resource-sensitive rewrite process, where rules replace MRS fragments (SL to TL) in a step-wise manner—but adds two innovative elements to the transfer component, viz., (i) the use of typing for hierarchical organization of transfer rules and (ii) a chart-like treatment of transfer-level ambiguity. The transfer machine is implemented in an LKB type of formalism**.**

The three core components (analysis, transfer, generation) are implemented as separate processes managed by a central controller which passes intermediate results through the translation pipeline. All component communication is in terms of sets of MRSs and, thus, can easily be managed in a distributed and (potentially) parallel client–server set-up.

The project involves researchers at three different sites, the parallel development of two different grammars and a transfer module, and a code repository of its core demonstrator with around 650 megabytes of software and linguistic resources. To keep track of this, the competence and profiling technology and the [incr tsdb()] tool (which was originally developed for parsing with HPSG implementations and specifically the ERG) has been modified and extended for use in the MT scenario. This makes it possible to assess progress and keep track of a multitude of central system measures like coverage, ambiguity and speed over successive system revisions.

Empirical and statistical layers. MT research has been dominated by statistical methods for the past decade or so, and from that point of view our approach based on deep linguistic analysis may seem old fashioned. But like a growing number of colleagues, we doubt the long-term value of pure statistical approaches. Even though such systems perform as well as symbolic systems on quantitative performance measures, they are also known to sometimes deliver totally unacceptable output. At this point in history these systems seem to have reached the "ceiling" and are not able to circumvent the problems without qualitative enhancements. Also within the statistical community one has started to search for ways to integrate more linguistic information.

Our approach is to start "from the other side" and add statistical techniques on top of the deep system. The goal of the deep approach is to restrict the output in the target language to be a well formed sentence and to, as far as possible, preserve the meaning of the input sentence. But the deep approach works on the competence level and will in many cases deliver a large number of alternative analyses. To select between them, performance oriented methods based on empirical observations have to be employed. We have so far experimented with a ranking of the realized translations against a language model based on the BNC with promising results. We have also experimented with ranking of the generation step. We will proceed with experiments with stochastic ranking for parse selection and for the transfer step, ultimately selecting the most probable translation of an input string.

Evaluation and resources. The translation domain is chosen to be hiking descriptions, both because they have a potential applied value and because they contain challenging constructions not always in the focus of linguistic theory. The initial demonstrator was developed on the basis of 100 representative sentences picked from texts in the domain, together with hand-constructed test suites containing basic words and constructions in the two languages. For the remaining project period we have selected a 50,000 word corpus with high-quality literal translations to serve as the reference corpus.

The project reuses and further develops several language technological resources for Norwegian. One overall goal for KUNSTI is to develop reusable resources for Norwegian. LOGON aims at, besides the Norwegian grammar, (further) developing a tokenizer, a morphological analyzer, and, in particular, a computational lexical data base which may serve as the inter locus for different applications.

An integrated part of the project is end to end evaluation. This will serve both as a research task and as a way to measure the success of the demonstrator.

Doctoral projects and training aspects. A main goal for the whole KUNSTI program is the training aspect, building competence in language technology in Norway. LOGON has employed four doctoral students. Their main task is the work on their thesis project. As this is supposed to be independent work, the doctoral projects are not necessary components in the downstream LOGON demonstrator. They are rather independent modules which might contribute to the quality of the demonstrator when integrated. One project concentrates on the statistical ranking of the results of the generation and translation system. Another project considers how the addition of soft constraints to the system may enhance performance. One project is geared towards semantic distinctions in transfer and how these can be learned from a translation corpus. The last project pursues developing HPSG for certain Norwegian constructions.

The LOGON project has also given stipends to several master's students who have worked on relevant master's projects. These projects include the analysis and translation of numeric expressions, a semantic and translational analysis of locative expressions, evaluation, translation of compounds, collocation extraction, and lexicon acquisition for Norwegian.

International collaboration. The project maintains active international collaboration. Several international recognized scholars in the field have attended the project meetings, and, so far, we have had one international guest researcher each year for three months. The collaboration includes, through the Delph-In-network (related to HPSG, ERG and LKB): Stanford, Saarbrücken, Cambridge and Sussex Universities and NTT, Japan. The LFG and ParGram-network include collaboration with Stanford and, in particular, the Palo Alto Research Center (PARC).

Project assessment. Half way through the project we have developed the first demonstrator, in fact several consecutive versions of it. Even though the coverage is limited, in particular in the transfer step where we started from scratch, the initial results are already promising with respect to the feasibility of the approach, in particular the way symbolic and statistical layers can be integrated. The challenge for the next period will be first to extend the coverage considerably to cover as much as possible of the new 50 000 word reference corpus. Then the focus will be shift to the precision of the system and how the statistical methods can be employed at all levels to guide the selection of the best output.

Further references. For more information on the LOGON-project, see the project web page: www.emmtee.net/.

4.2 BRAGE – speech centric dialog systems

Introduction. BRAGE is the largest project in the man-machine spoken dialogue field. The project is a 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).

Internationally, speech centric dialogue systems based on small clients (mobile phone, PDA) have shown to be a viable alternative for information retrieval when a user has no access to a PC/internet. Further, such a system can be the only alternative for large groups of disabled. A speech centric dialogue system consists of several modules: an automatic speech recognizer (ASR), a semantic analyzer, a dialogue manager including a reasoning part, a prompt generator, a text-to-speech synthesizer (TTS) and an application database. Each module has normally both generic and task specific parts. Multimodal system variants have user interfaces which include text and graphics as well as speech.

The motivation for the BRAGE project was the realization that dialogue systems for the Norwegian language had a significant lower performance than state-of-the-art internationally. This applies also with respect to complexity; i.e. for which kind of tasks such systems could be implemented. In order to do research on speech based dialogue systems, the complementary and multidisciplinary knowledge of the above partners constitutes a natural starting point for the project.

The overall goal of the project is to develop state-of-the-art dialogue systems for the Norwegian language. This logically results in the following concrete project goals:

- Implementation and testing of speech only demonstrators based on:
    - Spontaneous speech over telephone;
    - A user friendly “mixed initiative” dialogue;
    - Synthesized speech response;
- Implementation and testing of multimodal demonstrators based on:
    - Composite input (”tap and talk”);
    - A wireless client/server system;
    - Composite output (“text, graphics and synthetic speech”);
- A research activity focused on:
    - Relevance to dialogue systems;
    - 4 Ph.D. dissertations;
    - Acceptance of publications in refereed international conferences/journals;
    - International and national cooperation;

Methodology and project structure. The focus in the project is 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 benefitial 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.

The basic research within BRAGE is focused towards dialogue systems, the different modules and their co-operation. For the ASR-module, the main research focus is on robustness against various kinds of backgrounds and transmission channel noise, and how to deal with different dialects and speaking styles e.g. spontaneous speech versus ”read aloud text”. Dialogue strategies and formalisms are a research area in itself. In order not to be restricted by the current ASR-performance, it was decided to include a research activity on text-based versions of complex and even problem-based dialogue tasks.

As an integrated part of the research activity, four PhD-candidates are financed by the project. International refereed publications are a natural measurement of research quality. One aims at reaching at least 25 publications as a result of the project research.

Collaboration is important with respect to research quality. Internationally we will collaborate with acknowledged institutions both in USA, Japan and within EU. Nationally, the project partners have a long tradition for cooperation.

Results achieved. The text-based version of BUSTER was developed during 2003. In parallel the WoZ-version was developed and testing was continued in 2004. The final operating demonstrator was delayed due to inadequate ASR-performance both on commercial and in-house recognizers. However, during spring 2005 the in-house recognizer was substantially improved, and a full-scale version of the demonstrator will be ready for public testing in the autumn of 2005. Testing indicates that our current in-house recognizer represent state-of-the-art with respect to spontaneous speech in Norwegian.

In spring 2005 initially work has been done with respect to developing a second speech only demonstrator. The chosen application concerns telephone switchboard services at the university (NTNU and SINTEF).

A first version of the multimodal demonstrator for Trafikanten was developed during 2003. During 2004 the demonstrator was improved with respect to both technical performance and user friendliness. The demonstrator was tested both by abled and disabled persons. In spring 2005 focus has been set on usability for dyslectics and aphasics.

Publications. By the end of spring 2005 a total of 18 publications have been presented at different international conferences with referee. In addition we will present at least 3 papers during the rest of 2005. Further, 4 Ph.D.-students are working with their dissertations on different topics as robust speech recognition and verification, spontaneous speech characterization and dialogue strategies.

International collaboration. Internationally collaboration is established with ATR (Kyoto), ISK (Tokyo), Georgia Tech. (Atlanta), IBM (New York) and COST action 278 (EU). As to the first three institutions, the collaboration has the form of long term research visits by people attached to the project. Further, co-operation is established towards Nordic institutions like KTH, HUT and AUC. We also plan to establish contact with Linkøping University during the autumn of 2005.

Further references. For more information on the BRAGE-project, see the project web page: www.iet.ntnu.no/projects/brage/.

4.3 FONEMA - Tools for realistic speech synthesis in Norwegian

Introduction. The FONEMA project (Greek for 'speech sound, utterance') is a cooperation between the Institute for Electronics and Telecommunications at the Norwegian University of Science and Technology (NTNU) and Telenor Research and Development. The project period is from 2003 to mid-2007.

The motivation for the FONEMA project was the realization that current text-to-speech synthesis (TTS) for the Norwegian language an inferior quality compared with TTS for other languages, and that development costs for new synthetic voices based on state-of-the-art technology are too high. Concatenative speech synthesis based on unit selection is the current state-of-the-art technology for TTS, and is capable of producing superior naturalness compared to other methods. This technique opens interesting possibilities such as cloning of speaking styles and personalized voices.

Unit selection waveform synthesis demands labor and cost intensive processing relatively large speech databases, which results in high costs when producing new voices, or new dialects or accents. The aim of the project is to develop methods and tools for data driven waveform synthesis, emphasizing methods for automating the process of creating databases for new speakers. The project includes methods for specification of linguistic content, annotation and processing of the speech recordings, in addition to the development of an improved front-end with automatic annotation of intonation. The project will contribute to an important strengthening of competence in speech technology and linguistics, and the results can act as a basis for commercial players intending to develop state-of-the art speech synthesis for Norwegian.

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. The framework will include:

- A Linguistic model for Norwegian prosody for TTS
- Procedures for establishing speech databases with a speaking style suitable for different applications, including
    - methods for defining a manuscript for speech recordings ensuring satisfactory phonemic and prosodic coverage;
    - procedures for recording, digitalization and data organization;
    - methods for automatic phonemic and prosodic annotation;
    - methods and procedures for building an efficient database
- A general synthesis module for production of natural sounding speech based on an existing front end and unit selection synthesis.

The project will also produce a TTS demonstrator which will act both as a research tool and a means for demonstrating TTS quality.

Methodology and project structure. The focus of the project is on the TTS back-end, i.e. on prosodic modeling and prediction and on speech generation. These modules are of utmost importance for naturalness in synthetic speech. The front end linguistic processing is based on an existing front end, developed by Telenor R&D for use in their Talsmann diphone synthesizer. Only smaller modifications on this front end have been necessary to make it fit into the new framework.

Previous work by Telenor R&D has shown that an existing phonological model for Norwegian intonation (the Trondheim model, Fretheim 1992) can be successfully applied to diphone synthesis. This model forms the basis for prosodic modelling for database annotation and can be applied to prosodic prediction in the TTS engine.

Recent development in TTS research has demonstrated a convergence of methodologies applied in automatic speech recognition (ASR), pattern classification and speech synthesis. In particular the increased focus on data driven methods for speech generation in TTS has necessitated the use of statistical methods well known in speech recognition. The project group has a strong background in ASR and statistical methods which will form the basis for the development of annotation tools and for research the unit selection synthesis.

The project is organized in six sub-projects which interact closely. The subprojects are

- Linguistic model, including prosody model; tools for phonemic analysis and tools for prosodic analysis.
- Design methodology for unit selection databases
- Reference database, i.e. the design, recording and organization of a unit selection database for evaluation of annotation tools and for development of unit selection synthesis algorithms

- Annotation tools, i.e. the development of automatic tools for phonemic and prosodic annotation
- Unit selection methodology, including research on cost functions and search methods for unit selection synthesis.
- Demonstrator, the development of a TTS system for research and for demonstrations.

A PhD student is financed by the project, working on cost functions and concatenation methodology. In addition a PhD student financed by NTNU is attached to the project.

Results achieved. The prosodic model for Norwegian intonation has been simplified to form the basis for the envisioned needs of unit selection synthesis. A system for automatic prosodic labelling that predicts and positions prosodic features (syllable stress) based on lexical and linguistic information as well as acoustic parameters extracted from the recorded speech signal has been developed. Experimental results show that the method is capable of classifying syllables as unaccented or accented with high accuracy (>90% correct).

An XML-structure that incorporates all phonetic and prosodic information produced by the TTS front-end into a uniform description has been developed. The existing linguistic front end has been adapted to conform to this standard.

A baseline system for automatic phonemic labelling has been developed. Improvements to the system employing voicing information in order to better determine boundaries between phones with different voicing status have been implemented, showing promising results. The voicing information is used to better determine boundaries between phones with different voicing status. Currently, work is under way to enable the system to select the correct pronunciation in sentences where the pronunciation cannot be uniquely predicted.

A major endeavor in 2004 has been the specification, recording and labeling of a reference database. The intended primary use of the reference database is evaluation and further development of automatic annotation tools. In addition, the database will be used for experiments and optimization of unit selection procedures. Tools for manuscript design have been developed, along with tools and protocols for recordings. 2000 sentences have been recorded for each of two professional speakers, corresponding to approximately 2 hours for each speaker. 10% of the recordings have been manually annotated, both phonemically and prosodically. Experiences from the production of the database, as well as use of the database for unit selection experiments will be invaluable input when specifying a production base to be used for the demonstrator.

An important observation from work on the reference database has been that in a speech corpus there will always be discrepancies between the expected realization of a written sentence and the actual pronunciation. We have proposed a method based on utterance verification for automatically detecting sentences with such discrepancies. This will simplify the quality control of large databases for unit selection speech synthesis.

In unit selection speech synthesis it is important that the cost functions employed in the database search matches human perception. In a perception experiment, we have studied different alternatives to calculating the concatenation cost with the aim of finding spectrally based measures that correlate well with human perception of discontinuities.

Further references. For more information on the FONEMA-project, see the project web page: www.iet.ntnu.no/projects/fonema/.

4.4 BREDT – detecting and processing co-reference

Introduction. Detection of co-reference is an important task in natural language processing. It helps to create cohesion and coherence in a text, and it is essential for finding out what a text is about. In text-to-speech, the correct detection of co-reference is a significant factor for separating new and given information. Typically, new information should have more prominence, although stressing old information may have the effect of contrasting this information with other known information. It is also indispensable for machine translation, where the correct translation of a pronoun is intimately connected to finding the antecedent of the pronoun.

Correct reference detection is a notoriously difficult task, with very modest success rates. This is partly due to ambiguity and vagueness in language use. Obviously, it is also dependent on background knowledge that is not available directly from the text at hand. However, the construction of an ontology that is complete enough to be useful for arbitrary text is a task that has proved extremely difficult.

Methodology and project structure. In BREDT there is a focused on machine learning of co-reference. This task can be started with fairly moderate resources, but a substantial research effort has to be devoted to developing annotation guidelines, and building up a fair size database of correct examples of co-reference.

Since the start of BREDT in 2003, an annotation manual has been developed, and used to annotate a small amount of co-reference pairs. This limited data set has been used to develop a demonstrator, which can find an antecedent for some of the most common pronouns. During this work, various ways of weighting the evidence from the database have been developed, and work is now in progress of improving both coverage and performance of the demonstrator.

Links to other KUNSTI projects. BREDT may deliver valuable resources to many other projects in the KUNSTI-programme: KUNDOC (document analysis), LOGON (MT), TREPIL (corpus resources), FONEMA and BRAGE (speech technology), and it may be adapted to, and used in, projects modeling special domain knowledge, such as KB-N and KUNDOC. A further research direction is to implement more languages. Since language independent representations are used, this ought to be a feasible task given a tagger that can provide the relevant underlying features.

International collaboration. The group has published widely in international conferences on topics related to discourse annotation and machine learning. A demonstrator, and its documentation, is continuously updated.

Further references. For more information on the BREDT project, see the project web page: spraktek.aksis.uib.no/projects/bredt/.

4.5 KunDoc – Knowledge-Based Document Analysis

Introduction. KunDoc is a research project conducted by CognIT as and the University of Bergen, Norway. The project is funded by the Norwegian research council, within the KUNSTI framework. KunDoc aims at developing a new method for analysing Norwegian text. The project seeks to recognise co-reference relations (e.g. "Clinton" and "ex-president") in order to build up knowledge representations and discourse structures of natural language texts.

The detection of Coreference chains in texts is closely related to the task of anaphora resolution. Current research in anaphora resolution can be divided into methods that are mainly based on heuristics (traditional methods), statistics-based methods, such as machine learning and some hybrid approaches that try to combine both.

Methodology and project structure. KunDoc aims at using knowledge representations (ontologies) for the resolution of references that usually are hard to find by the mentioned approaches since they depend on the availability of world knowledge. Knowledge-based disambiguation of text is not new in itself (some methods such as frames or scripts date back to the early days of Artificial Intelligence). However, most knowledge-based approaches have been dismissed due to the lack of processing power, memory and availability of knowledge. In the mean time, research in the fields of knowledge representation, information extraction and initiatives within the Semantic Net has provided both for tools and methodologies for acquisition, storage and re-use of knowledge.

The KunDoc approach is cross-disciplinary, combining methods from Computational Linguistics with Knowledge-based (AI-oriented) methods. The central questions KunDoc seeks to answer are:

- What kind of ontological knowledge is needed in order to support coreference chaining?
- How can this knowledge be acquired and used?

In order to answer these questions the KunDoc work plan implements the following tasks:

1. Collection of data
2. Implementation of parser
3. Statistic analysis and learning of ontologies
4. Adaptation of existing Coreference techniques
5. Use of ontologies for Coreference
6. Extraction of discourse structures
7. Testing and evaluation
8. Documentation

Work Outline. The plan for the project basically consists of three phases. In the first phase, after the collection of the document bases, we will look into the knowledge intensive methods for anaphora resolution. Methods such as Wilks' preference semantics or Carbonnell's multi strategy approach will be evaluated and – if suitable – adopted since they employ world knowledge in order to disambiguate referents.

In the next step, ontology will be built describing the domain of the text corpus. The basic concepts of the initial domain will be modelled manually, the refinement will be carried out (semi-)automatically in the course of the project.
While modelling the ontology, methods for extracting the relevant knowledge for disambiguating texts will be developed.

In the third step the inference rules that transfer the domain knowledge into sectional restrictions will be developed in order to resolve unclear reference in new documents.

Further references. For more information on the KunDoc-project, see the project web page: www.kundoc.net/.

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

Introduction. KB-N is a 3-year project aiming to establish a knowledge-bank for economic-administrative domains. The concept of a text-based knowledge-bank builds on the underlying assumption that domain-focal special knowledge is embedded in text produced typically by domain experts for documentary, argumentative, didactic or general communicative purposes. It further assumes that the essential knowledge content is embedded in relatively language independent concepts and manifested through relatively language specific terminology (in casu English and Norwegian used in economic-administrative domains), and that such terminology is stratified with respect to domain specificity ranging from general shared terms down to a small set of domain-focal terms**.**

Modules and functions. KB-N represents the culmination of efforts to develop, refine and integrate computational strategies and NLP tools for

- corpus design and analysis
- automatic and semi-automatic extraction, representation, and retrieval of terminology
- dynamic linkage of terminology and its authentic textual manifestation (co-text)
- dynamic thesaurus creation
- dynamic concordance display of authentic collocational and phraseological evidence

Methodology. (i). *Exploiting Parallel Text*. Relevant domain text is extracted from formal institutional channels, introductory textbooks, research articles and popularized publications across 30-odd subdomains. Texts are scanned or downloaded from relevant sources, routinely XML-coded and POS-tagged.

Where strictly parallel texts are available the two language versions are aligned for semi-automatic equivalence mining. An intuitive semi-hierarchic classification of essential subdomains of economics and business administration forms the basis of storage and retrieval of terms as well as texts. Contact is maintained with KunDoc, another KUNSTI project, for the purpose of refining our concept-based classification system.

(ii). *Term extraction*. Automatic term extraction from Norwegian text (unlike English) is very much in its infancy and is being developed from scratch through an associated project to establish and refine a Term Candidate Filter. The filter draws on algorithms for Named Entity Recognition. The KB-N Term Candidate Filter is continually being refined and will be presented at TKE 2005.

(iii). *Concept systematization*. Getting from automatically generated term candidates to solid terms is a critical stage where domain expert knowledge intersects with terminological principles, and man/machine interaction must lean heavily on the former, especially in identifying "missing" concepts. The same holds for the establishing of conceptual hierarchies: As indicated above concepts are viewed as gateways to the domain knowledge, and the way they are organized and structured requires deep understanding of the specific knowledge content of the domain. The KB-N system provides essential support in allowing dynamic conceptual hierarchies to be developed, changed and revised alongside the terminological analysis.

(iv). *Term Entry*. Once a limited set of domain focal terms has been automatically extracted from the text and the resulting Term Candidate List has been interactively pruned, a semi-automatic module handles the actual entry of relevant terminological material into the corresponding term record, along with essential contexts and collocations. As the term base construction progresses, required definitions of central concepts will be added. Each term record preserves information linking it to the actual textual occurrence of the term in question. Since it is concept-based the term-bank will easily accommodate the addition of more languages (German, French and Spanish are obvious candidates).

Each term record is being structured with a view to making it accessible from an automatic translation system which is being developed as part of the KUNSTI program under the project title LOGON. Allowing such an MT system access to Norwegian source terms as well as their equivalent English target terms during a stage of its preprocessing routines prior to actual dictionary look-up will allow the system to insert a domain-relevant English term equivalent without interference from general language vocabulary. We are approaching a stage of maturity in both LOGON and KB-N which will allow a pilot implementation of this idea.

Achievements to date. We are continually updating a working Demonstrator version of the integrated KB-N software suite for handling corpus based concordancing, automatic term extraction, semi-automatic term selection, thesaurus building on-the-fly etc. The online operation of the Demonstrator has been shown publicly on several occasions, and has also been tested in an advanced teaching context at NHH during the spring of 2005.

Applications . The theoretically most interesting use of the KB-N Termbank will be in the context of Norwegian-to-English automatic translation of the relevant types of domain text already referred to. A range of other applications are envisaged for the knowledge bank. It is designed as a web-enabled resource available for systematic terminology registration and look-up, textbook authoring, as well as e-learning. Here KB-N will be closely integrated with an established e-learning system to provide interactive study support for students of economic-administrative subjects.

Further references. For more information on the KB-N-project, see the project web page: mora.rente.nhh.no/projects/kbn/.

4.7 TREPIL - Norwegian treebank, a pilot project

Introduction.The TREPIL project is aimed at developing methods and tools for building a Norwegian treebank. The pilot project will produce a design and specification, together with a well motivated methodology and well tested tools, in preparation of a subsequent, large project that will build a Norwegian treebank of a suitable size. TREPIL will not deliver a full size treebank, but a proof of concept in the form of a prototype treebank.

Methodology. Considering the fact that manual annotation is labor intensive and prone to inconsistency, the approach is based on semi-automatic annotation with a deep wide-coverage parser, for which the NorGram grammar is chosen. This LFG grammar, developed in the NorGram and ParGram projects, is being extended and coupled to a new tool being developed for discriminant-based manual disambiguation. The analyses to be stored in the treebank represent structures at triple strata: constituent structure, functional structures and semantic MRS-structures.

Project structure. The project cooperates with the Palo Alto Research Center (PARC), which provides and maintains the XLE parse tool. The project also cooperates closely with the LOGON-project, which has an interest in treebanking for machine translation, and with the Nordic Treebank Network (NTN) which promotes treebanking in the Nordic countries.

Results achieved. A white paper, co-authored with colleagues from NTN, has been published. At the LFG'05 conference, a poster presentation and demonstration of the first version of the discriminant based tool will be given**.**

Further references. For more information on the TREPIL-project, see the project web page: helmer.hit.uib.no/trepil/

4.8 Sámi language technology project

Introduction. The aim of the project 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.

Methodology and project structure. The project uses the so-called finite-state framework. The parsers are parsers are written as transducers, with two-level rules for the morphophonology. Suprasegmental morphology is triggered via quasi-affixes introduced with the segmental morphology in the suffix lexica. The disambiguation is done by constraint grammar, i.e., by a set of contextual rules removing inappropriate readings from the sentence in question**.** As a part of the disambiguation process, we have also mapped and disambiguated grammatical functions, such as subject and object, but also potentially hierarchical information such as left-modifier-of, etc.

There are two computational linguists and one (half-time) programmer working on the project. In addition to that, there is a twin project, run by the Sámi Parliament, aimed at building a spell checker for Northern and Lule Sámi. The two projects share the source code for the transducer, and collect a common corpus of texts. They differ only on the use of the transducers: The university project uses them as input to the disambiguator, and includes the texts in the search interface, whereas the Sámi Parliament project makes the transducers the base component of a normative spell checker, and integrates it in various desktop applications.

Applications. The Sámi project cannot be funded by selling products in a market. There simply are too few Sámi users. Contrary to many language technology projects, we are then able to offer our source code as open source. The importance of this for other language communities lies not primarily in the linguistic rules themselves, but in the infrastructure we have made for the project as a whole, including directory structure, Makefile setup, etc.

Results achieved. So far, most effort has been concentrated upon Northern Sámi. There are still lacunas in the preprocessors and lexica, and the project is able to recognise only approximately 95-97 % of new word form tokens and 90-93% of new wordform types in unseen, running text. The disambiguation results are better; it is possible to disambiguate the grammatical analysis with a precision of 94% for the morphology and 93% for the syntax, and with a recall of 99 %, with an ambiguity of 1.056 (i.e. 1056 readings per 1000 words). The morphological tag set used is a fine-grained, fully specified, set. With a POS tag set of the size normally used, the result would have been far better. As for Lule Sámi, one still does not have a full lexicon, but the morphological transducer gives good coverage of the inflectional grammar.
The file infrastructure has already been used as starting point for making parsers for Greenlandic, Komi and Udmurt, with promising results. The project participates in the Nordic PaNoLa II-project, in essence making a treebank for use in interactive pedagogical programs. At present the transformation from linear to hierarchical structure is done semi-manually, when making a phrase structure grammar on the basis of our disambiguated output we will benefit from the work done within the TREPIL project.

Further references. For more information on project, see the web page: gjellatekno.uit.no/

5. Programme Results

The programme is going into its last year, 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 fulfil 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 the programme was the Norwegian Language Bank, 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. Ten PhDs are being funded by the projects. This is the basic way to increase the number of qualified researchers in Norway, and the programme has reached its objectives.

The requirement that the large projects should build a functional demonstrator has led to the desired collaboration between the various teams in Norway. Here are some examples: In the LOGON project researchers from Oslo, Bergen and Trondheim collaborate intensely to make the demonstrator work. In BRAGE, researchers from different departments in Trondheim collaborate. So, we have collaboration by teams in different locations and institutions, but also collaboration which is cross-disciplinary between different departments in the same university. Finally, in KunDoc we have the collaboration between research and industry – and between Oslo and Bergen.

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 programme.

Besides, the scientific results are now beginning to emerge, but it is premature to evaluate these on the basis of the 2005 results.

The Programme Board is satisfied with the effects of the programme, and plans are being prepared for a continued effort in language technology, in order to capitalise on results and cover other, yet uncovered areas. 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.

6. Acknowledgement

The Programme Board would like to thank the KUNSTI projects for their contributions of project reports.

7. References

Språkteknologi i Norge – eksisterende og påkrevet forskning. Rapport fra en arbeidsgruppe.
The Research Council of Norway. Oslo 2000

KUNSTI – Knowledge Generation for Norwegian Language Technology. Programme Plan.
The Research Council of Norway. Oslo 2001

Home pages for the KUNSTI projects:
BRAGE: http://www.iet.ntnu.no/projects/brage/
BREDT: http://spraktek.aksis.uib.no/projects/bredt
FONEMA: http://www.iet.ntnu.no/projects/fonema/
KB-N: http://www.nhh.no/spr/sff/kbn/
KUNDOC: http://www.kundoc.net/
LOGON: http://www.emmtee.net/
SAMI LANGUAGE TECHNOLOGY: http://giellatekno.uit.no/
TREPIL: http://helmer.hit.uib.no/trepil/

ISBN 82-12-02271-4 (trykksak)
ISBN 82-12-02272-2 (pdf)