Strategic plan for the development of research in Language Technology at the University of Gothenburg

A long range plan
with particular reference to the current research period 2009–12

Amended version, 2009-06-10

Contents

1 Language technology

Language technology (LT) consists of a wide spectrum of technologies that enable computers to handle both spoken and written natural language. It lies within the framework of information and communication technology (ICT) and is a strongly interdisciplinary field. Fundamental research in LT largely involves the development of algorithms for language processing and mathematically rigorous descriptions of various linguistic phenomena and methods for the automatic acquisition of linguistic knowledge from empirical language data. It also involves ascertaining the mathematical and logical properties of such descriptions and algorithms. Linguistic data resources play an important role in modern LT research, a significant part of which is fundamentally empirical in nature.

LT, that is the processing of natural language, has been a dream since the early days of computing when there was a great deal of investment in machine translation in the 1950s which was subsequently perceived not to have met expectations. It became an established international discipline during the 1980s with leading academic centres in the US, Europe and Japan. However, at this stage of development it was still very limited in nature, dealing with small fragments of natural language with carefully designed hand-written programs. This severely limited the usefulness of the research as a general technology and it is significant that at that stage of development the term in use to describe the field was not language technology but *computational linguistics*.

During the 1990s there was a revolution in the field which imported statistical and machine learning techniques. This has led to a technology that can handle large amounts of linguistic data and provide robust systems that cover enough of the variety of human language to provide systems that are commercially viable or clearly have commercial potential in the future. It was during the 90s that the term "language technology" first came into use and LT proper came into existence.

The first decade of the 21st century has seen a huge development in the use of language technology commercially, for example in the use of automatic dialogue systems for simple tasks such as call routing or travel information and the use of linguistic analysis is search engines such as Google. However, since 2000 there has been increasing awareness among researchers that the "statistical revolution" of the 90s has its limitations and that we need to combine the kind of detailed analyses associated with the hand-crafted theory-based approaches of the 80s with the new found power of statistical and machine-learning algorithms and large data processing capabilities. At the same time, since 2000 there has been increasing awareness that we need to be able develop computational methods for dealing with the meaning or content of language, the field of study known as semantics. This is reflected in a range of developments such as the semantic web, encoding of semantic information on linguistic data resources and the development and

increasing use of semantics based resources such as WordNet and FrameNet.[1]

These developments require a significant investment in linguistic data resources – digitized text and speech collections, refined by the addition (manual or automatic) of linguistic annotations, and databases of lexical and grammatical knowledge. In addition to this there is a need for general-purpose or domain-specific lexical and grammatical tool resources, that is, tools which enable the use of lexical and grammatical knowledge about words, syntactic, morphological and phonological structure and semantic interpretation in LT applications. In fact, much energy in the LT community is today directed at ensuring the permanence and reusability of language resources, since it is increasingly acknowledged that the realization of language resources for even a single language are truly large-scale, very expensive undertakings. In the VR planning project *An infrastructure for Swedish language technology* 2007–2008 (VR dnr 2006-6763), a national consortium consisting of seven partner institutions and coordinated by Gothenburg (Språkbanken) estimated the cost for building a Swedish BLARK (Basic LAnguage Resource Kit) and a Swedish national corpus to be on the order of 130 MSEK (at the then current indirect cost rate of 35%), a figure that is on a par with similar estimates or actual costs in other countries. The consortium prepared a proposal for funding of a national basic infrastructure for Swedish LT, which is currently being processed by VR.

Figure 1 represents the interface between LT research and potential users of LT. At the core is LT theory, including basic research in linguistics, logic and computer science. Theory is used as a basis for the implementation and building of LT artifacts, such as dialogue systems, corpus processing and exploitation tools as well as web services.

All LT research is informed by the close interplay between theory and practical applications. The best way to test a computational theory is almost always to implement it as a computer program. As such, the work in LT encompasses theoretical research themes on the one hand, and application areas on the other – i.e., areas that function as testing grounds for theory. This indicates a great strength of the field: even the most abstract and theoretical LT research is directed towards practical goals and concrete applications with ultimate commercial and social potential. While many of the computer applications created to test research hypotheses are far from being marketable products, in recent years an increasing number of university prototypes have either inspired or provided the basis for commercial systems. This trend is in keeping with the fact that over the last decade, LT has been recognized not only as a promising area of academic research, but as an important growth industry as well. This is documented, for example, on websites dedicated to information about LT for the general public

[1] See <http://wordnet.princeton.edu/> and <http://framenet.icsi.berkeley.edu/>.

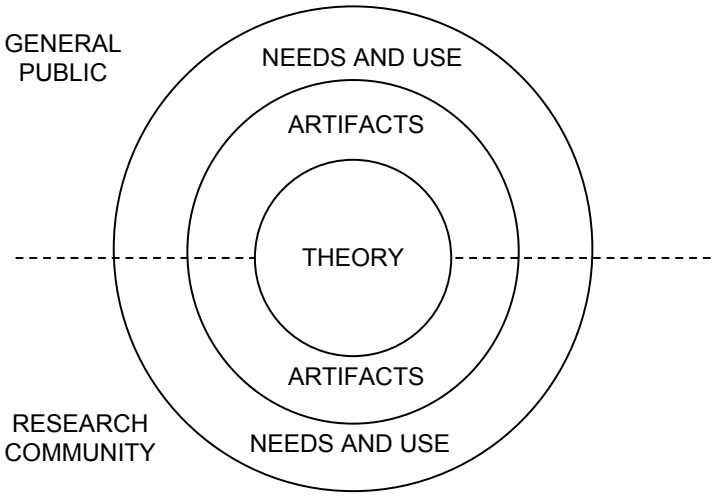


Figure 1: LT research

such as <http://www.lt-world.org/> and <http://sprakteknologi.se/>, the latter a collaboration between the University of Gothenburg and the Swedish Language Council (Språkrådet).

Language technology: the future.

In a fast moving and young field like language technology it is difficult to predict what will happen even in the near future. However, there are a number of trends pointing to future developments where we feel that Gothenburg has an opportunity to make a significant contribution. These are:

wide coverage systems The ability to deal with a wide coverage of data, e.g. the processing of free text, will continue to develop. Whereas many of the techniques used now are shallow statistically based techniques, these will in the future be integrated with more exact rule-based techniques based on implemented grammars of various kinds.

linguistic resources and tools The importance of linguistic resources and tools for their enhancement and analysis has been recognized within LT. The largest development of these resources is, however, within English and other "large" languages. LT has become increasingly multilingual but there is still a long way to go if the "smaller" languages are to enjoy technological equality.

semantics We use language to communicate information. A great deal of what has been successful in LT so far has to do with the form of language rather

than it contents (i.e. the semantics of language). While there has been a great deal of theoretical work on natural language semantics (much of it based on advanced techniques developed in logic during the twentieth century) it is only recently that we have gained some understanding of how semantics might be dealt with in the kind of large-scale applications which current LT is interested in. This is a growth area which from an LT perspective is still in its infancy and the challenge for the future is to find a way of matching the kind of shallow statistical techniques which are available with the deep logic-based analyses that have been developed in theoretical linguistics and linguistic philosophy.

machine learning Machine learning (ML) plays an important role in many areas of LT ranging from speech recognition to machine translation. The kind of features that ML systems are trained on, however, are often shallow and not based on linguistic theory. There are good computational reasons for this. The kind of complex deep analyses that linguistic theories provide, yield hypothesis spaces which are simply too large to be computationally tractable. This has led to separate cultures based on ML and linguistic theory. However, in the last few years there there have appeared a few pieces of work pointing to important connections between the ML-based research and theoretical research. A new window of opportunity seems to be opening which will enable the two approaches to work together with a potential for providing new insights into how language is adapted to deal with new content. A central question is how conversational agents adapt to each other's language and a challenge is how to get machines to learn language through interacting with their human users.

expansion beyond LT The importance of large linguistic data resources and tools for enhancing and analyzing them has already been recognized within the field of LT. However, a development that is taking place now is that this is in the process of being recognized by increasing numbers of other fields from medicine to the humanities. In the future this will offer significant opportunities for LT to strengthen its base and become recognized as providing resources and tools that are central both to academic research and future generations of speech and language enabled products.

2 Language technology research in Gothenburg

The development since the start of the new millenium towards incorporating more theory-based and meaning oriented technology into the statistical techniques developed in the 90s means that Gothenburg, like a few other centres in the world such as the Bay Area (including Stanford and Berkeley), Edinburgh and Saarbrücken, is particularly well placed to make a significant contribution. It has a long

tradition of empirical work combined with a tradition in logic (both philosophical and computational) and the semantic analysis of natural language. Above all, perhaps, there is a concentration of leading scholars in various areas who have shown a commitment over the past ten years or so to working together on language technology in a multidisciplinary environment. This commitment manifests itself in a number of joint publications and a joint LT seminar series, among other things.

An overview of current research in LT in Gothenburg is presented in figure 2. This figure shows the diversity of topics and multi-faceted nature of LT research in Gothenburg, which is well advanced in many areas, especially of course in relation to LT for the Swedish language.[2]

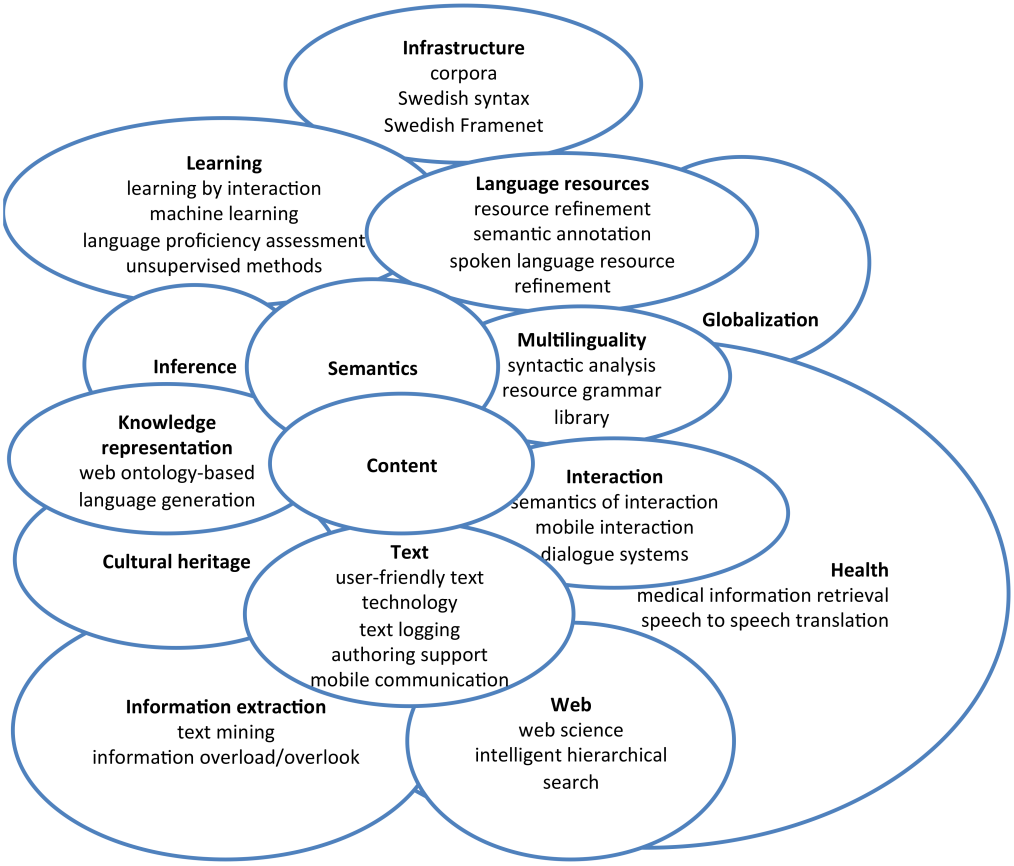


Figure 2: Current LT research in Gothenburg

At the centre of many of the numerous research activities listed in figure 2 are

[2]Throughout this document, we have deliberately omitted references to publications, for reasons of space. We can produce publication lists on demand, or GUP may be consulted for the group's publications.

content and *semantics*, which are increasingly becoming a central concern for a lot of our work ranging from information extraction to translation and multilingual applications. The unsolved research problems of language technology today are mainly those that deal with understanding of language, i.e., semantic and pragmatic processing of language, both spoken and written. Thus, work on content and semantics will form the main thrust of the plan presented in this proposal.

LT research is conducted mainly by collaborating groups from three departments[3] representing five academic subjects: computational linguistics, natural language processing, computer science, general linguistics and theoretical philosophy. We are also in the process of increasing involvement by colleagues from other disciplines, who are interested in using LT as a research tool in their own area of specialization (eScience), or who are interested in LT as one of several options for human-computer interaction (interaction design).

Much of the LT research that we conduct in Gothenburg can be seen as falling into three overlapping and interacting areas, each of which is distributed across the departments where LT research is pursued. These are: *text technology and linguistic data resources*, *grammar technology and linguistic theory* and *dialogue technology and spoken interaction*. The relationship among the three is such that grammar, broadly construed to include semantics and theoretical linguistic analysis, feeds both text technology and dialogue technology, which in turn inform and guide the work on grammar technology and linguistic theory.

Recognizing this general tripartite nature of LT research in Gothenburg, we now propose to also formally organize our work along these lines, turning the three loosely organized research groups into three "labs", the *Text Technology Lab*, the *Grammar Technology Lab* and the *Dialogue Technology Lab*. We will present more detail about our ongoing and planned research using this perspective below in section 4.

3 Plans for strengthening the area

The plan for strengthening the area comprises the following elements:

- the establishment of interdisciplinary thematic research groups ("labs"), as outlined in the previous section and detailed in section 4 below, administratively placed within an interdisciplinary Centre for Language Technology (CLT). One of the advantages of organizing our work in the way we are proposing is that we will be able to track developments in the three areas in a way not previously possible and make strategic decisions based on this

[3](1) Department of Computer Science and Engineering, Faculty of the IT University, University of Gothenburg and Chalmers University of Technology; (2) Department of Philosophy, Linguistics and Theory of Science, and (3) Department of Swedish Language, both the latter in the Faculty of Arts, University of Gothenburg.

information. An important practical aspect of this organization will be the creation of permanent research engineer posts (two full-time equivalents) to support the computational implementation work in all three labs;

- the creation (and maintenance) of a "CLT toolkit", a set of mutually and externally compatible state-of-the-art open-source LT tools and accompanying linguistic resources, which we expect will be of great help in speeding up LT prototyping efforts and generally in avoiding duplication of effort in our research. The toolkit will in part be based on existing tools that we have developed such as the Grammatical Framework (GF) in the grammar group and the dialogue system toolkit TrindiKit in the dialogue group as well as corpus enhancement and analysis tools developed within the text group. Part of our effort will be devoted to increasing compatibility between these various software tools so that they can be used together;
- the continuation of the CLT seminar series (featuring both local and guest researchers). This series was started with a one-time allotment of strategic funds from the University in 2007 and is ongoing, but after the initial funding ended, increasingly rarely with external guests;
- a yearly CLT workshop including guest researchers and guest students;
- postdoctoral research fellowships;
- PhD student fellowships. This will be especially important if GSLT (see section 5) is not able to admit new PhD students, although there will not be sufficient funds in this budget to replace GSLT's contribution to LT in Gothenburg;
- active forming of collaborations with researchers in non-LT disciplines at the University of Gothenburg and Chalmers University of Technology, especially collaborations on eScience and interaction design;
- collaborative projects with international labs, including extended visits by research personnel (see appendix B).

By attracting recent graduates from leading graduate schools to our postdoctoral fellowships and research positions, CLT will benefit from rapid sharing of new ideas and technologies. The ability to provide funding for visiting senior researchers as well as research time for local researchers is essential to the establishment of a creative and productive research centre. CLT will provide an appropriate infrastructure, for example, for the organization of workshops and conferences in the area. Work within CLT will be organized so as to encourage dynamic interactions between fundamental research projects in LT and the development of prototype applications and systems.

4 Current and future projects

Here, we present an overview of current and planned activities in the three labs, indicating existing and planned external collaborations with academia, industry and the public sector. The following table shows current external funding and submitted project proposals per lab:

	Text	**Dialogue**	**Grammar**
Current	14 MSEK	7 MSEK	5 MSEK
Submitted	36 MSEK	13 MSEK	38 MSEK

Note that this is a snapshot of the situation as of early 2009, rather than a picture of accumulated or average external funding over time, which accounts for the apparent uneven distribution of funds across the labs. In addition, two large collaborative proposals coordinated by the text technology group are under consideration by the Swedish Research Council:

1. a proposal by the University of Gothenburg for an initiative in the strategic research area eScience, where LT takes the central place (65 MSEK over three years, including OH) (see section 4.1.2 and p. 18 below);
2. a proposal by a seven-member national consortium for a Swedish LT infrastructure, consisting of a Swedish national corpus and a BLARK (Basic LAnguage Resource Kit) (130 MSEK over 7 years, including the then current 35% OH).

In table 1 we illustrate the interdisciplinary nature of the labs, by showing how some of the different research topics we are engaged in distribute over traditional academic subjects – represented by the Department of Swedish Language, the Department of Philosophy, Linguistics and Theory of Science (FLoV) and the Department of Computer Science and Engineering – and the research labs in this proposal.

The approximate number of refereed publications produced by the three groups over the last five years (2004–2009, as documented in GUP) are as follows:

Text technology group	60
Grammar technology group	100
Dialogue technology group	60

Below we give a brief account of some of the main research themes in each of the labs.

4.1 The text technology lab

The focus of the fundamental research in the text technology lab is on LT data resources and tools for working with large amounts of text (understood in the widest sense: transcribed speech is also to be considered text here).

	Text	Dialogue	Grammar
Swedish Language	lexical resources, corpus collection, corpus tools, syntactic resources, morphological resources, semantic resources, machine learning	corpus collection	language resources, parsers, syntactic theory, semantic theory
FLoV	corpus collection, corpus tools, machine learning, web science	dialogue system development, dialogue theory, coordination and learning in dialogue, communication studies	parsers, computational syntax, computational semantics, semantic theory, formal languages, logic
Computer Science	parsers, morphological analysis, translation, web science	grammar resources for dialogue systems, dialogue theory	semantics, syntactic resources, machine learning, formal languages, logic

Table 1: The interdisciplinary aspect of the proposal

4.1.1 LT resources and tools

Since 1975, the University of Gothenburg's Språkbanken (Swedish Language Bank <http://spraakbanken.gu.se>) has made Swedish and other language resources available online to both the research community and the general public. Språkbanken possesses a unique combination of competences in the areas of Swedish text corpora, parallel text corpora, Swedish computational lexicons and LT tools for the processing, annotation and presentation of text corpora. These competences are made even more effective by being coupled with the kind of stable organization required for sustained large-scale corpus processing and presentation. Over the years Språkbanken's corpora and lexicons have been widely used for research, teaching and other related purposes. In particular, a good number of PhD theses in Sweden and Finland have used Språkbanken as a data source.

4.1.2 eScience

A prominent application area of the research in the lab is *eScience*, understood here as the use of LT-based methodologies and tools for working effectively with large amounts of digitized text as primary research data, as in many humanities and social sciences disciplines. Much of the LT research in the lab and in Gothenburg in general can potentially be applied for this kind of eScience.

Corpus linguistics – eLinguistics

The language sciences have long made use of LT as a research tool in the form of *corpus linguistics*, whereby large digital text collections (*corpora*) are made searchable via linguistically relevant search criteria. During the 1960s Gothenburg became a pioneer in Swedish corpus linguistics by compiling Press-65, one of the first large electronic text corpora in a language other than English. The approximately 200 MW corpora now available online in Språkbanken have since the mid-1960s been used in research in computer-based lexicography and lexicology. Many of the recent large Swedish published dictionaries have been worked out by lexicographers at Gothenburg. A notable recent focus in our research are corpus-based studies of learner (second-language) Swedish and LT-informed corpus-based applications in the area of computer-assisted language learning.

The Department of Philosophy, Linguistics and Theory of Science maintains an extensive set of spoken language corpora, which have served as the basis for numerous studies of spoken language phenomena. They have also been used for comparing the properties of written and spoken language. Approximately 60% of the corpora are video recorded, making them an interesting data source for the study of multimodal communication. Along these lines, several studies have been conducted on the relationship between gestures and spoken language that have strongly influenced current research on Embodied Communication at the Zentrum für Interdisziplinäre Forschung at Bielefeld University, Germany.

Also at the Department of Philosophy, Linguistics and Theory of Science is the SweDia 2000 dialect database. The database consists of recorded speech material from 107 Swedish dialects (1200 speakers, 750 hours recording time). The recordings were made as a joint effort by the departments of linguistics at Umeå, Stockholm and Lund universities in 1998–2002. It has long been used as an eScience resource resulting in more than 70 publications. There are several fundamental scientific questions that may be addressed in a fruitful way by an analysis based on data of the kind contained in SweDia. The SweDia database is now hosted by the department, and work is in progress to develop the database further as an eScience resource.

Life sciences

Research on applying LT in the life sciences domain has been a focus of research at the Department of Swedish Language during recent years. The research is based on authentic text material, both clinical narratives and scientific medical content. In the medical setting, for instance, vast amounts of health-related data are constantly collected. These data constitute a valuable source of primary research material; however, to empower clinical researchers to locate and make highly efficient use of the knowledge that is encoded therein, the material must be better integrated and linked via effective automated processing. Such advanced capabilities would facilitate the construction of hypotheses based upon novel associations between extracted information, the undertaking of retrospective studies based upon patient narratives, drug discoveries, the early detection of adverse drug events, the improvement of searching and browsing capabilities, and the like. It would also provide a more focused and effective means of searching and collecting patient-related information from unstructured text – a process that is currently restricted by language-inherent complexity, variation and ambiguity.

Gothenburg has an established network of academics and professionals, both in Sweden[4] and in Europe,[5] that are actively interested in issues related to this field. Recently, the Swedish National Board of Health and Welfare (SoS) and Gothenburg's LT researchers have begun to collaborate on a one-year SoS-funded research project to transform a large archive of scientific medical text material into a format that is suitable for text mining. This material is annotated with a systematically organized computer processable collection of medical terminology known as SNOMED CT (the Systematized Nomenclature of Medicine – Clinical Terms). Plans are also in the works to establish a competence center for *clinical language technology* that will supply clinical researchers in academia and the pharmaceutical industry with LT tools and services that are customised to meet different individual needs.

Web science

An important aspect of LT is the way in which it connects to other, more visible, technologies. In particular, we have reason to believe that LT will become increasingly important to web technology, making the web seem smarter, more interactive, more conversational, and more sensitive to the information needs of people with a first (and perhaps only) language not so widely used.

[4]Sahlgrenska University Hospital, the AstraZeneca pharmaceutical company, the Swedish National Board of Health and Welfare, the Swedish Association of the Pharmaceutical Industry, DSV/KTH-Stockholm University.

[5]The National Centre of Text Mining, Manchester, UK and the NCSR Demokritos, Athens, Greece.

The web is currently largely made up of linked documents, often text documents. Language technology may add value to the web by extracting some form of meaning from the human-readable text of the pages, based on rules or statistics – meaning that can make search more effective, or help building the semantic web, a web made up of linked data.

In line with the general interest in LT-based eScience, the Department of Philosophy, Linguistics and Theory of Science has recently formed a group approaching the emerging field of Web Science from an LT perspective.

4.1.3 Planned research activities

Fundamental research

The planned fundamental LT research in the lab will have as its centerpiece the creation of a full-scale computational lexical resource for modern Swedish – and to some extent earlier forms of Swedish, primarily the 19th century language – with rich semantic, syntactic and morphological information ("Swedish FrameNet++"). We see this as a necessary core component for the development of methodology and tools for automatic semantic annotation of text. Further, this work will allow us to explore various corpus-based lexicon creation methodologies, mixing the manual – which we master almost to perfection, but which is prohibitively labor-intensive – with experimental resource-economical machine learning approaches. This research will be conducted in collaboration with the grammar technology lab.

Applications

The main application areas that we wish to initiate or develop further in the lab are the following:

- language resources and language technology tools used for eScience and Web Science, in particular in linguistics and in medicine and other life sciences (as outlined above), but also in history and literature (see under *Expansion beyond LT* on p. 18 below)
- text generation in the cultural heritage domain (in collaboration with the grammar technology lab and the Gothenburg City Museum)

4.2 The grammar technology lab

4.2.1 Grammar engineering

LT research at the Department of Computer Science (CS) started in 1997 with a Logic and Language seminar joint between CS (Bengt Nordström) and Linguistics (Robin Cooper). The interest on the CS side was mainly within the Programming Logic group, which had been working on Type Theory and the foundations of logic and programming languages since the late 1970s. The LT group was formed in 2001, at the initiative of Nordström together with Aarne Ranta, who had joined CS in 1999.

The Grammatical Framework (GF) was first created in 1998 when Ranta worked at Xerox Research Centre Europe; its purpose was to build systems that work on a language-neutral semantic content (expressed in Type Theory) with views in different natural languages. In GF, it is possible to translate semantic content between languages without loss of meaning. In addition to this, Type Theory also provides a model for human-computer interaction, which was later worked out into a mathematical model for dialogue systems as well.

A specific mission of the LT group at CS has been to make LT accessible for programmers without linguistic training. We see that LT in itself is seldom interesting for end users – but LT as a part of a larger system can add a lot of value to it. As a proof of this, the LT group has had joint projects with many other research groups at CS:

- Programming Logic: type-theoretical semantics, user interfaces for theorem provers.
- Functional Programming: tools for language design and implementation.
- Formal Methods: authoring tools for software specifications, automated theorem proving in semantics.
- Software Engineering: grammars as software libraries.

4.2.2 Authoring support

The development of parsing techniques for the study of syntactic deviations in text as a support for language development of school children is the focus of research by Sylvana Sofkova Hashemi. The main goal of her research includes the development and adaptation of different tools and language resources that will support better the writing development and writing education in school, the exploration of keystroke logging tools as a didactic support and for language checking. She is also interested to study collective and collaborative writing in web 2.0 tools and the impact on language development and education.

4.2.3 Syntactic theory

Work on syntactic theory and particularly on the syntax of Swedish is carried out mainly in the Department of Swedish Language (Elisabet Engdahl). There has been a focus on the syntax of information structure and the use of syntactic structure in dialogue. There is considerable work on the use of corpus materials and the appropriate way to enhance them with syntactic information as well as work on current grammatical theories such as Head Driven Phrase Structure Grammar, Dependency Grammar and Optimality Theory.

4.2.4 Semantics

Work on semantics is distributed among philosophers and linguists at the Department of Philosophy, Linguistics and Theory of Science (Robin Cooper, Björn Haglund, Dag Westerståhl) as well as the Department of Computer Science (Aarne Ranta). In addition, work specifically on lexical semantics is being conducted in the Department of Swedish Language. There is a concentration of work going back many years on formal semantics using logical techniques of various kinds (classical model theory, type theory, situation theory and type theory with records). There is a particular concentration on generalized quantifiers (quantifier expressions in natural language other than those found in classical logic), and other phenomena that go beyond standard first order logic (such as intensionality). There has also been a considerable amount of work on compositionality and dynamic semantics, including the development of semantic theories to account for learning and interaction.

4.2.5 Planned research activities

Fundamental research

One focus of fundamental research over the next few years is the development of wide coverage grammar applications that can handle free text. This applies both to syntactic and semantic coverage. A second focus, related the first, is the development of algorithms for systems that can learn incrementally from linguistic input.

Applications

The main application areas that we wish to initiate or develop further in the lab are the following:

- the extension of the GF Resource Grammar Library from 12 to 30 languages, including the 23 official EU languages; this will be done by coordinating an international open-source effort, including a series of summer schools;
- the development of a large-coverage parsing grammar for some languages of the Resource Grammar Library, starting with Swedish; this will be done in collaboration with Språkbanken and the Text Lab;
- the development of tools for multilingual on-line translation based on Semantic Web ontologies and the Resource Grammar Library; this will be done in international collaboration, possibly within a European consortium (we coordinated an FP7 proposal MOLTO, "Multilingual On-Line Translation", submitted in April 2009);
- research on combining statistical and grammar-based translation methods, partly based on the use of word alignment data generated from multilingual GF grammars;

- research on robust parsing, based partly on statistical methods and partly on fuzzy matching between strings and abstract syntax trees, with results to be evaluated in the mentioned large-coverage grammars and also in systems built in the dialogue lab;
- the development of software engineering tools to support grammar application programming; one of the research engineers to be employed will play a key role in developing e.g. an IDE (Integrated Development Environment);
- authoring support for non-adult speakers of Swedish;
- general computational semantic tools giving deep analyses of constructions which are not covered by first order logic.

4.3 The dialogue technology lab

4.3.1 Dialogue systems

Spoken dialogue (between humans) has been an important research topic in Gothenburg linguistics since the establishment of linguistics as a separate discipline in the 1970s. In the late 1990s Gothenburg began to profile itself internationally as a centre for dialogue systems research in the context of a series of EU-funded projects (through the work of Robin Cooper, Staffan Larsson and their co-workers). Today, Gothenburg is a main player in the international dialogue systems research community, and (together with Edinburgh and Saarbrücken) one of the main European research centres in the field. Since the early 00s, dialogue systems research in Gothenburg is largely coordinated within the Dialogue Systems Lab, which conducts interdisciplinary research on the theory and practice of dialogue systems. Research in the dialogue lab has had significant impact on the field of dialogue systems research both in the form of theoretical contributions such as the Information State Update approach to dialogue management and Issue-based dialogue management, as well as software in the form of the TrindiKit toolkit for dialogue systems R&D and the GoDiS dialogue system. The latter is currently being commercialised by spin-off company Talkamatic, in cooperation with Volvo Cars.

4.3.2 Applied speech technology

The phonetics section of the department of Philosophy, Linguistics and Theory of Science has a small but very active research group in the area of applied speech technology, and more specifically forensic phonetics. (Anders Eriksson, Jonas Lindh) The two major technologies are automatic speech recognition and automatic speaker recognition (in forensic case work the term now used by us is forensic speaker comparison) and the merge in between those.

4.3.3 Mobile communication studies

Researchers at the department of Applied IT are currently in the process of gathering previous research efforts around mobile communication into one coherent

research agenda. They explore aspects of human communication mediated by mobile technology. This involves the collection and detailed analysis of both written and spoken communication on and through mobile devices. The aim is to advance our knowledge about communication in mobile activities. Ylva Hård af Segerstad and Alexandra Weilenmann have for the last 10 years worked with different aspects of mobile communication.

Hård af Segerstad has focused on text-based interaction through mobile phones (text messaging, or SMS). Data from text messaging has been gathered in two sets of corpora. The technology used for text input (e.g. predictive text technology) is one of the factors which influence language use and communication patterns in text messaging.

Weilenmann has focused on interaction on and through mobile devices, and particularly on oral communication. Conversational and ethnographic data has been collected in a number of fields, both leisure and work related.

4.3.4 Planned research activities

Fundamental research

An important next step in our work on dialogue is to develop theories of how humans coordinate their language and learn new language through interaction. We aim to investigate the possibilities of adapting machine learning techniques (currently oriented towards training on large datasets) to incremental learning on the base of input gained from dialogic interaction. One of the promises of this research, as well as of related research in the area of in-vehicle dialogue systems, is the development of adaptive dialogue systems which put less cognitive stress on their users. Another goal is to make a connection between our work on speaker comparison and dialogue systems, so that, for example, systems can recognize users by their speech from among a small number of users registered with the system. A further goal is to develop methods and tools for capturing and analysing linguistic interaction involving language technologies (dialogue systems and mobile communication technologies).

Applications

The main application areas that we wish to initiate or develop further in the lab are the following:

- dialogue systems that can adjust their language to users
- dialogue systems that can learn language from users
- systems that can differentiate among speakers by examining phonetic and possibly other properties of their language
- tools and methods for analysing mobile communication and dialogue system interaction

Summary

As a summary we take up the five LT development areas which we introduced on p. 3 and indicate how our work in the labs addresses them. Some of the planned work continues threads which have already been developed in Gothenburg. Some of it represents innovative work in areas that are new to us.

wide coverage systems We are no strangers to dealing with large amounts of linguistic data and there is a long tradition in the work associated with the text technology lab of both data collection and data analysis. However, the work associated with the grammar technology lab has been largely rule-based and focussed on detail rather than wide coverage. While we have started scaling up this work, notably in the collaborative effort on functional morphology using work from both labs, there are still many challenges to be met. We believe, however, that the future of the field lies in the application of the kinds of detailed theory-based methods used in previous work associated with the grammar technology lab to the broad coverage statistical methods for shallow analysis in use today.

linguistic resources and tools While we have considerable experience in the development of linguistic resources and tools, there is still a great need of development in this area in respect of Swedish and the other languages of Sweden. While we have made a significant start on Swedish there is still a lot of new work to be done and other languages in Sweden are in even greater need of resources.

semantics There is a tradition of detailed theoretical work on semantics in Gothenburg some of which has been applied to small scale rule-based application. In the proposed CLT organization this work will be associated with the grammar technology lab. There are, however, important aspects of semantics, particularly lexical semantics, which are associated with the work of the text technology lab and the dialogue technology technology lab. New techniques are emerging for connecting this kind of work to the shallow statistically-based techniques currently used in the kind of semantics which has been developed in LT over the past 3–5 years. There are still important conceptual challenges to be met in relating these two strands of work and Gothenburg is perhaps one of a handful of research centres in the world where there is sufficient expertise and interest in order to carry out the investigation. One of the keys, we suspect, is the relating of machine learning techniques to semantic analysis.

machine learning Work associated with the dialogue technology lab has recently begun to move in the direction of using ML techniques to create dialogue systems that will learn new language from interaction with users. This represents a new departure for our research and will involve a departure from

the standard ways in which ML techniques have been developed. ML standardly involves "batch learning", that is, the presentation of large amounts of data to a machine equipped with a learning algorithm which as a result acquires a system capable of performing a certain task. This kind of learning is in contrast to human learning and the kind of learning needed for dialogue systems which learn from interaction. The new kind of learning is "incremental learning" where a small amount of data (even a single datum) will provide the machine with a hypothesis which it can immediately test by interacting with the user. User feedback will provide evidence as to whether the learner's hypothesis was correct or not. A measure of success in incremental learning might mean that machines could learn more efficiently and would not need to be fed with large amounts of collected data. This is innovative work which represents a new direction for us (and for most other work in ML).

expansion beyond LT LT has now reached a stage of maturity that it can present itself as a tool for the use of other disciplines. The recent initiative in eScience coming from work associated with the text technology lab shows that this development is feasible (independent of whether the particular current initiative secures funding). Our experience in the past has been that it is a non-trivial matter to get people from other disciplines to understand what LT researchers actually do. Language is something that is normally taken for granted and it is difficult for people outside the discipline to understand what the basic problems are in language processing. However, our initial work is beginning to pay dividends we feel and we could be about to embark on a period of new collaboration and understanding of the relationship of our research to other disciplines. A cross-disciplinary seminar series (*Thesis Antithesis Prosthesis: Posthuman perspectives on language technologies*; see <http://www.ling.gu.se/projekt/tap>) was funded by the Faculty of Arts from 2005 to 2008 and attempted to extend the interface of language technology (in a wide sense) to the humanities in general. In the eScience initiative, we have now been able to involve a number of relevant "early adopters" in various fields outside LT and Linguistics, such as Literature, History, Political Science, Education and Medicine, in some cases reflecting already existing collaborative projects or existing advanced plans for such projects. This represents a new opportunity which we are eager to pursue in the future. At the present time, one very important platform for doing this – at least on the European level – is the ESFRI CLARIN network <http://www.clarin.eu>, where the text technology lab is represented (both by Språkbanken and the SweDia group).

5 Relationship to teaching

Since LT is a young and fast developing field where technology and methods can change rapidly from one year to the next, it is important that there be a close connection between research and teaching. This has meant that our teaching often involves the presentation of new research and students get an opportunity to become involved in work related to our current projects. It also means that our courses are revised from year to year depending on input from students, our industrial contacts and research developments. LT researchers at the university collaborate on degree programmes at all three levels: undergraduate, masters and PhD. There is a joint undergraduate programme in LT, coordinated by the Department of Philosophy, Linguistics and Theory of Science with contributions from the Department of Swedish Language and the Department of Computer Science. LT figures as a specialization in the masters degree in Computer Science which is run by the Department of Computer Science. The Department of Philosophy, Linguistics and Theory of Science coordinates the national Graduate School of Language Technology (GSLT) which, since its beginning in 2001, has provided the bulk of our funding for PhD students in all three departments (in addition to providing funding to departments in other Swedish universities who provide graduate training in LT). The most recent GSLT intake was in 2008 and subsequent intakes depend on the school securing future funding. GSLT's Academic Board (which includes three of the senior scientists involved in this proposal) is actively concerned with raising awareness of the GSLT's crucial role in a national strategy for PhD level training in LT. (See Appendix C.) GSLT is currently actively pursuing possibilities for national training in LT at the masters level (see Appendix D). An important component in this is making LT courses available using web distribution (possibly including iTunes U now that it is available to Swedish universities) and distance learning technology and GSLT is currently working on making its courses available in this way.

6 Dissemination strategy

The main dissemination strategies are: national, including Scandinavian, and international conferences, submission to peer-reviewed journals, end-user and public events, press releases, an information/open day workshop and the CLT and språkteknologi.se websites. In addition, our resources and tools will be distributed according to an open source policy whenever possible. The dissemination strategy will be continuously reviewed and refined until the end of the envisaged three year period. All members of the various CLT groups will be actively involved in this.

6.1 Objectives and instruments

The general objectives of the dissemination strategy plan are:

- to exploit the existing CLT infrastructure (tools, show cases, demos).
- to promote and publicize more the already developed CLT tools, encouraging as many (e)scientists as possible to start a dialog with CLT and participate in CLT activities (academic dissemination)
- to exploit the links of the CLT members with other researchers in the area of eScience
- to identify new targets and develop new partnerships for all the applications related to CLT group members
- to intensify cooperation with industry partners for implementation in products and services enabling the transition of academic thinking into industry thinking
- to engage relevant policy makers both at the local (university level) as well in the industrial and public sectors to CLT activities
- to form an advisory board with professionals from the academia (at the national and international level), industry and society

The dissemination strategy will be continuously refined until the end of the three year period taking into account the technical progress, as well as the inputs of the potential users of our tools. We emphasize the role of the project websites which we expect to become a focus of attention among researchers in other disciplines and information seekers. Furthermore, we consider open-source software resources and open-source linguistic resources as an important factor in our dissemination plan. Thus, in the context of the dissemination strategy the following instruments will be used:

- enrichment of the CLT website – creating web awareness about CLT and LT in general
- attending/organizing/hosting forums, events, seminars and workshops
- demonstrations of applications
- general and individual meetings with target groups, both with the research community and the general public

- press releases
- information leaflets (in English and Swedish) to be distributed at various events
- writing academic and technical papers for national and international LT conferences including focused ones such as in the area of medical informatics
- papers in scientific journals
- visits by CLT members to relevant LT labs and groups
- inviting prominent members of the international LT community for short or longer term visits

We consider also the advisory board (see appendix A) as another instrument to increase the dissemination of our infrastructure and visions. We anticipate that the advisory board members will provide valuable feedback to CLT but also present and discuss their views outside the CLT group.

6.2 Scheduled activities

The scheduled activities are divided into local, national and international. The local ones include: CLT members' participation at the International Science Festival ("Vetenskapsfestivalen") with at least one event per year during 2009–2012; the participation (preferably with students) at the Student and Knowledge Fair "Kunskap & Framtid" (in November every year); the creation and distribution of a flyer describing aspects of Language Technology, in particular education possibilities, success stories, LT applications; participation in the university's Annual Report at least twice between 2009-2012 and also participation at the "Universitets-TV" and create material for the I-Tunes U and YouTube. Moreover, we plan to develop on-line and user-friendly demos and tutorials and organize annual information and open day workshops, starting by the end of this year. Web publishing in English at the university, faculty and department level will also be prioritised.

At the national level we plan to write papers for scientific publications in Swedish journals, for instance there is a scheduled paper to be submitted at the Journal of the Swedish Medical Association *Läkartidningen*; also popular science publications in Swedish magazines: e.g. *Computer Sweden* and *Språktidningen* are foreseen. We plan to participate in the Swedish Society of Medicine fair "Läkarstämma" that will take place in Stockholm at the end of November 2009 with a poster and an oral presentation.

At the international level, members of the CLT groups have already submitted papers (or plan to submit) papers to a number of scientific LT conferences and workshops, and a book on GF is scheduled for review by the end of the year. Moreover, a GF Resource Grammar Summer School is scheduled for mid-summer 2009 and a Swedish SEMEVAL-2 (SEMantic EVALuation) task is currently scheduled for the end of 2009 – beginning of 2010.

7 Success indicators

General indicators by which to gauge the success of the research program outlined here are the following:

- project proposals (approved, submitted, "ready to go")
- initiated and completed PhD projects
- postdoc projects
- guest researcher visits
- publications
- collaborations with industry, international labs
- collaborations with disciplines outside the LT environment
- external funding
- tool and resource usage (among ourselves; by other groups, as measured by downloads and references in research papers)

In particular it will be important for us to keep track of our development in respect of the five LT development areas we have identified in this plan:

- wide coverage systems
- linguistic resources and tools
- semantics
- machine learning
- expansion beyond LT

7.1 Milestones for 2010

For the first evaluation in the summer of 2010, additional important indicators are

- that the labs are up and running
- successful appointment of research support staff (project administration officer and research engineers)
- successful initiation and evaluation of internal (including postdoc) research projects
- internal conference
- GF summer school
- SemEval

7.2 Milestones for 2011

- evaluation of internal research projects
- successful initiation and evaluation of joint ventures with international laboratories
- internal conference
- evaluation of impact on and support for PhD training
- implementation of advice from the advisory board

7.3 Milestones for 2012

- evaluation of internal research projects and joint ventures with international laboratories
- evaluation of PhD training progress
- successful initiation and evaluation of industry related projects
- internal conference
- strategic plan 2013–2017

A Organization

The research activities described in this proposal will be located to the three labs – the Text Technology Lab, the Dialogue Technology Lab and the Grammar Technology Lab – but conducted within the wider context of CLT (Centre for Language Technology), where seminars, internal conferences and other events are organized on a regular basis, and which will maintain a central web site with information about language technology R&D in Gothenburg.

CLT organization

CLT will be organized as a centre ("centrumbildning") with administrative placement at the Department of Swedish Language.

The centre will have a small permanent staff, including a director (scientific coordinator) (10% of full-time; initially Lars Borin), a project administration officer (50%; to be appointed), and research engineers (200%; to be appointed). The technical staff will not be physically located in one place. However, in order to achieve the aims of harmonizing our various LT tools and applications, of developing standards and best practices for their deployment, and generally of furthering cross-lab synergies, it is important that the technical staff in CLT and the three labs come together regularly on an organized basis for effective knowledge and competence transfer among the labs.

The coordination and planning of the activities of CLT will be executed by a scientific board consisting of 10 members: the director and three representatives from each of the three labs. One of their first tasks will be to make a more detailed budget for the centre. The director will chair the board and be responsible for the day-to-day administration of the centre.

We plan in addition to appoint an advisory board, consisting of nationally and internationally renowned researchers in LT and industry representatives.

Lab organization

Each lab will have two co-directors, coming from different disciplines, in order to ensure the nature of the labs as interdisciplinary research platforms. The proposed lab co-directors are as follows:

Text Technology Lab

Lars Borin, professor of natural language processing, Språkbanken/NLP, Dept of Swedish Language

Aarne Ranta, professor of computer science, LT research group, Dept of Computer Science and Engineering

Grammar Technology Lab

Aarne Ranta, professor of computer science, LT research group, Dept of Computer Science and Engineering

Elisabet Engdahl, professor of Swedish language, Dept of Swedish Language

Dialogue Technology Lab

Robin Cooper, professor of computational linguistics, Dept of Philosophy, Linguistics and Theory of Science

Bengt Nordström, professor of computing science, LT research group, Dept of Computer Science and Engineering

B Application for funding

We divide our application into three levels corresponding to the three levels of funding:

1. Laboratory facilities and junior research support
2. Junior research support and joint ventures with other international laboratories
3. Research training and relations with industry

Level 1: Laboratory facilities and junior research support – 3 MSEK

The lowest level of funding (3 MSEK/year) is a fairly modest amount of money, and we plan to concentrate it on enhancing our fundamental research infrastructure. Our most urgent need at this time is for the kind of basic infrastructural support which is hard to fund through individual external projects. This includes *research engineers* (shared equally among the three laboratories) and *administrative assistance* in managing projects and project applications. On the 3 MSEK level of funding, we can also employ one fte junior researcher (post-doc, assistant professor or lecturer). This may correspond to more than one individual, depending on other available research funding, teaching requirements, etc.

Here and below the budget item "research support" includes travel money and other research support expenses, including funds for inviting visiting researchers/ guest lecturers. On the 3 MSEK level of funding it also includes the cost of arranging an annual internal workshop.

Director (10%)	0.1 MSEK
Project administration officer (50%)	0.3 MSEK
Research engineers (200%)	1.2 MSEK
Equipment, service contracts, software licenses	0.4 MSEK
Junior research support (100%)	0.7 MSEK
Research support	0.3 MSEK

Level 2: Junior research support and joint ventures with other international laboratories – 2 MSEK

On the second level of funding, we would support two additional fte junior researchers, the aim being to have one fte per lab (although, as mentioned above, this may in fact involve more individuals).

We would also reserve funds (included under the item "research support" in the following table) for developing long-term relationships with other national and international laboratories such as Stanford, King's College London, Edinburgh, Oslo, KTH, Lund, again broadly within the planned research topics described in section 4 above. The relationship would involve joint ongoing projects and exchange of personnel.

In keeping with the spirit of this national and international collaboration, we will use part of the funding at this level to set up an advisory board to CLT, consisting of international and national researchers and members from industry (see Appendix A).

Junior research support (200%)	1.4 MSEK
Advisory board	50 kSEK
Research support	0.55 MSEK

Level 3: Research training and relations with industry – 2 MSEK
We will use money for doctoral positions (doktorandanställningar) in order to involve PhD students in our laboratories' activities. This level of funding will be sufficient to allow us to admit a total of three PhD students for the time period covered by this budget, i.e., less than one per year on average. Thus, we will be able to compensate only in small part for the fact that GSLT funding has not been secured for the future (typically, out of the five PhD students admitted every year to GSLT, one or two would be placed at Gothenburg).

We would also extend our internal research support to include development of projects which could attract collaboration with industry.

Doctoral positions (300%)	1.5 MSEK
Industry oriented research support	0.5 MSEK

C GSLT's document on a national strategy for higher education in language technology

Nationell strategi för högre utbildning i språkteknologi

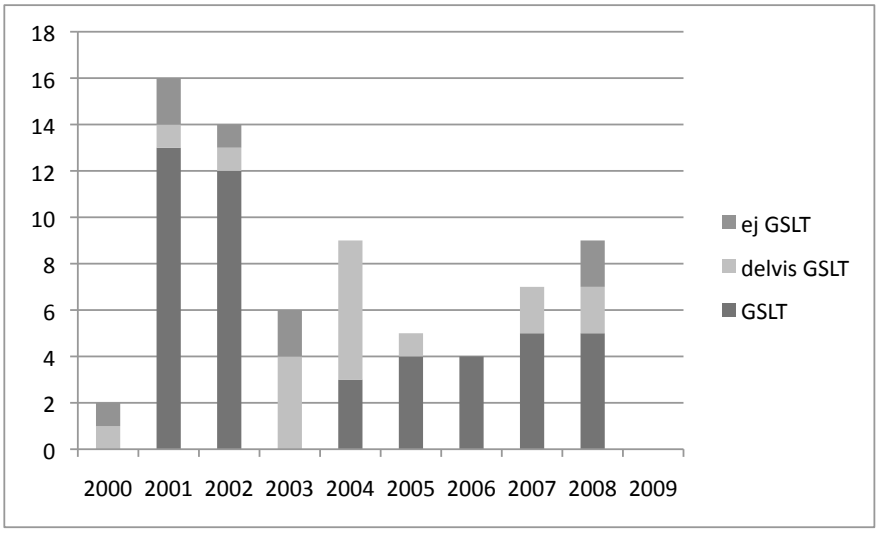
Nationella forskarskolan i språkteknologi (GSLT)

29 april 2009

I Sverige bedrivs forskning och utbildning inom språkteknologi på flera lärosäten, vilka sedan 2001 samarbetat i den nationella forskarskolan i språkteknologi, GSLT. Finansieringen för GSLT kommer från den dåvarande regeringens satsningar på sexton utvalda forskarskolor med Göteborgs universitet som mottagare och värduniversitet. Göteborgs universitet har förlängt användningen av den permanenta anslagsökningen för forskarskolans ändamål fram till och med 2012. Det finns när detta skrivs inget beslut om dessa medels användning därefter, varför GSLT inte längre tar in nya doktorander. (Den sista antagningen skedde 1 januari 2008.)

Under den tid som GSLT funnits har utvecklingen på språkteknologins område varit stark. Förbättrad teknik för taligenkänning och talsyntes har lett till ökad praktisk användning inom områden som kundtjänster, undervisning och tillgänglighet. Förbättrade metoder för analys av innehåll i text tillämpas inom omvärldsbevakning och sökning, och framsteg på översättningsteknologins område har möjliggjort en snabbare utveckling av översättningssystem för nya språkpar och samtidigt lett till kraftfullare datorstöd för lokalisering och översättning av programvara och dokumentation. Nya tekniker bygger på en kombination av specialistkompetenser som förut saknat gemensamma problem och mötesplatser.

Denna utveckling är påtaglig även i Sverige. En nyligen genomförd studie vid Uppsala universitet räknade till 47 språkteknologiska företag i Sverige varav många startats under de senaste åren. Flera av dessa företag har GSLT-doktorer



Figur 1: Intagning till forskarutbildning med språkteknologisk inriktning, 2000–09

anställda eller i ledningen. Av 19 färdiga doktorer från GSLT är i dag 11 st verksamma inom universitetsvärlden (varav 3 i utlandet) och 8 st i näringslivet.

GSLT har under perioden 2001–2008 svarat för merparten av doktorandfinansieringen inom språkteknologi. Det är därför angeläget att formulera en nationell strategi för den högre utbildningen och forskningen inom området. Figur 1 visar antagningen av doktorander med språkteknologisk inriktning under perioden 2000-2009 fördelat på antagningar som finansieras av GSLT och sådana som finansieras med andra medel. Forskarstuderande inom språkteknologi har varierande bakgrund, förutom språkteknologi förekommer datavetenskap, informatik, lingvistik, kognitionsvetenskap, filosofi och språkämnen.

Det nationella samarbetet inom GSLT har av både doktorander och forskare bedömts bidra till en kraftig kvalitetsförbättring av forskarutbildningen i Sverige. Detta vidimeras i Högskoleverkets rapport ”Utvärdering av 16 nationella forskarskolor”(Rapport 2008:16 R) som också framhåller (sid. 108) att GSLT utvecklat forskarskolekonceptet och “på ett ganska unikt sätt, åstadkommit en nationell kraftsamling inom sitt område”. För att kunna bibehålla och utveckla kvalitén på nationell nivå med nya generationer av forskare behövs en kritisk massa av doktorander som antas till forskarutbildning och bildar nya nätverk.

Vi bedömer att det krävs minst fem nya forskarstuderande per år i Sverige för att kunna hålla utbildningen internationellt konkurrenskraftig med regelbunden kurs- och seminarieverksamhet.

Vi anser vidare att en nationell strategi för forskarutbildningen bör kombineras med en motsvarande strategi för utbildningen på avancerad nivå (master- och magisterutbildning.). Målet bör vara att åtminstone 25 studerande årligen påbörjar en sådan utbildning med specialisering i språkteknologi. Dessa studenter, som kan ha sitt huvudämne i datavetenskap, lingvistik, fonetik, kognitionsvetenskap eller annat relevant ämnesområde, kan vara fördelade på olika masterutbildningar på de deltagande universiteten, men via det nationella språkteknologiska nätverket, ta del av specialiseringskurser på hög nivå.

GSLT:s ledningsgrupp

Robin Cooper, ordf
Professor i datalingvistik
Göteborgs universitet

Lars Ahrenberg
Professor i datorlingvistik
Linköpings universitet

Lars Borin
Professor i språkvetenskaplig databehandling
Göteborgs universitet

Rolf Carlson
Professor i talteknologi
KTH

Sándor Darányi
Universitetslektor i biblioteks- och informationsvetenskap
Högskolan i Borås

Joakim Nivre
Professor i datorlingvistik
Uppsala universitet, Växjö universitet

Bengt Nordström
Professor i datavetenskap
Chalmers Tekniska Högskola

Pierre Nugues
Docent i datavetenskap
Lunds universitet

D GSLT's proposal for a national masters school

National Masters School in Language Technology*

GSLT

May 19, 2009

Introduction

Sweden is a relatively small country and competence in language technology is spread over a number of academic institutions which have been gathered together in GSLT, the national graduate school of language technology. GSLT would now like to propose that a nationally coordinated "masters school" in language technology would offer the possibility of combining the advanced training competencies of various institutions, thereby creating a critical mass of students taken from a broad range of subjects. This will increase the possibility of attracting both national and international students to a greater degree than is possible with our current national graduate programme where commitment to a four year research degree makes it difficult to admit students who are unknown to us or who have not yet be able to show their research potential in this field through an undergraduate degree. Our proposal is that each collaborating site would have its own masters programme in language technology which would then draw on the resources, mainly courses, of the national school. In this respect the organization would be similar to that of PhD studies in GSLT where each student is registered and examined at her home organization but takes part in courses and other activities organized by the national school. It does not seem feasible that several universities in Sweden could individually attract the numbers of students needed to make a viable masters programme in this specialist area or provide the breadth that is necessary for an internationally competetive programme. We believe that a national masters school would solve this problem and allow individual universities to exploit their specialist

*Contact people: Lars Ahrenberg (Linköping), Robin Cooper (Göteborg), Olov Engwall (KTH), Torbjörn Lager (Göteborg), Beáta Megyesi (Uppsala), Aarne Ranta (CTH)

profiles within the area while sharing national resources in a way that would make smaller numbers of students registered in the local masters programmes an economically feasible proposition.

Opening the Masters courses to students from other Bachelor's programmes (and countries) is expected to attract more students from a wider range of backgrounds than is currently possible with the current degree structure.

Overview

Name of the school The name of the school shall be Masters School in Natural Language Technology (Mastersskola i Språkteknologi).

Content and goals Natural Language is becoming more and more important in IT applications. Some key areas are:

- extracting information from documents on the web
- producing documentation to products and systems in a multilingual environment
- localization of software to different languages and cultures
- speech-based communication with computers and also small devices (e.g. embedded in the car or home)

The goal is to provide a combination of courses that make it possible for students with an appropriate background to specialize in natural language technology. The specialization can be profiled towards computational linguistics or computer science, and towards a career in industry or academic research. There will, however, be a significant number of common, obligatory courses.

Number of credits The school will support two year programmes which carry 120 ECTS credits.

Targeted students Courses will be taught in English and aimed at the international student community. The programme will be open to students with

a background in language technology/computational linguistics, computer science, cognitive science and, with appropriate prerequisites, general linguistics, modern (or ancient) languages, philosophy and mathematics.

Detailed description

Degree structure and course content

A typical curriculum that the school would support has four semesters of study each of which corresponds to 30 ECTS credits. The first three semesters will normally have four courses each carrying 7.5 ECTS credits. The final semester will be devoted to the writing of a thesis (30 ECTS credits). Students are required to take 90 ECTS course credits (45 ECTS credits at level 2, 15 at level 3) and 30 ECTS credits of thesis work. A typical recommended study programme would be:

Semester 1:
Natural Language Processing (national, level 1)
A natural language oriented programming course (local/national, level 1)
Advanced methods course (local, level 1)
Linguistic resources (national, level 2)

Semester 2:
Speech Technology (national,level 1)
Statistical Methods (national, level 2)
two more level 2 courses

Semester 3:
Two level 2 courses à 7.5 ECTS credits
One level 3 course à 15 ECTS credits

Semester 4:
Thesis (30 ECTS credits)

Prerequisites for the courses

Students with an undergraduate degree in language technology, computational linguistics, cognitive science or computing science would be eligible to apply for level 1 courses without further prerequisites.

Students from general linguistics, languages, philosophy, mathematics and natural sciences are eligible , with the additional prerequisite of 15 ECTS credits in programming and 7.5 ECTS credits logic and/or discrete mathematics.

Examination and forms of study

Teaching will be largely web-based using technology such as Marratech, Google applications and Second Life, and will, in addition to traditional lectures delivered on the web, consist of "hands-on" lab-based exercises and projects with a problem-based orientation. The examination of courses will be a mixture of traditional examinations, course papers, practical lab exercises and programming projects. There will also be opportunities for students to meet physically, normally at least for initial and final sessions of courses.

Connection to research

All of the departments involved in this school are actively involved in externally funded research at both the national and European level and regularly collaborate with each other on such projects. Good masters students will be encouraged to relate their work to these projects.

Market demand for graduates

Language Technology is a rapidly growing area which creates a demand for experts internationally in both industry and research.

Recruitment base

GSLT has an international recruitment basis, and aspirants for GSLT might well be interested in our Masters programme. There are several GSLT applicants each year who are not admitted into the PhD programme because there is only a limited number of places available even though they are well qualified for advanced work. There is a further number of applicants each year who show promise but who have too little training or for whom we have too little information (in particular for foreign students) for us to be able to commit to a PhD. Such students could well be admitted to the Masters programme with a view to continuing to the PhD.

Another potential source of masters students comes to us through our work on NGSLT. For the introductory GSLT courses the number of students coming from NGSLT is often far greater than those from GSLT. Often 10–20 students from around the Nordic area (including the Baltic States and NW Russia) attend GSLT courses. Some of these could be interested in a masters degree.

As a rough estimate, we could expect to get 15–25 students every year.

Organization and budget plan

The school will be administered by a board (*ledningsgrupp*) which includes at least one representative from each participating university. One of the members of the programme's board will be director of the programme and 20% of the director's time will be devoted to running the programme and advising students on general matters relating to the programme, which courses to take etc. The director will also be provided with administrative and systems assistance to support course and admissions administration, maintenance of the programme's website and systems administration associated with course software and the like (perhaps 20% of full-time for each of these functions). Funds for teaching will be administrated centrally in the school and distributed to those sites conducting the teaching. A possible model for this is that each participating university contributes those funds which would be assigned to the masters students they have registered (possibly a fixed sum per student per year so that each participating university contributes the same per year per student). Students should have their travel and living expenses covered when taking part in the programme's activities and also be provided with a laptop computer. If national funding is not available for this, a fixed sum per student per year should be contributed by the participating universities to a central

fund which is disbursed to students who need to travel. This will ensure that none of the participating sites are financially disadvantaged by geographical location.

The following budget is a rough estimate of approximate costs that could be involved in such a school.

Annual cost when the Masters school is fully operative assuming an yearly intake of 25 students

Cost of instruction and instructors' travel and accommodation expenses					
per course	80,000 SEK	number of courses	10	Total	800,000 SEK
Student travel and accommodation					
per student	4,500 SEK	number of students	25	Total	112,500 SEK
Supervision of masters thesis paid locally					
Administration					
Director, administrator, computer support, board travel					500,000 SEK
					1,412,500 SEK

Concrete plans for immediate future

Uppsala is committed to starting a masters degree in language technology in the autumn of 2010. This degree relies on the existence of national courses. A proposal for what should be place in Uppsala by the autumn of 2010 is given in Figure 1 showing planned dependencies on offerings of national courses.

GSLT has so far planned national course offerings up to Spring 2011 and they have been designed to harmonize with the plans for the Uppsala masters programme. Fig. 2 shows the planned national courses for Autumn 2010 and Spring 2011.

Considerations on the form of courses and course materials

We aim to produce courses of excellence which can be used not only by Swedish masters programmes but also by programmes in other Nordic countries and elsewhere. (Considerable interest has already been expressed by our colleagues involved in the Nordic Graduate School of Language Technology on collaborating with this.) The greatest practical problem we face is the travel involved for

<table>
<tr><td>National Master in LT</td><td>Weeks</td><td>BA in language technology
----
BA in general linguistics/modern languages/ancient languages AND 15 ECTS in programming and 7.5 ECTS in logic/discrete mathematics</td><td>BA in computer science</td></tr>
<tr><td rowspan="4">Semester 1</td><td>W 36-40</td><td rowspan="3">Computer Science [B] (Local)</td><td>Linguistics I 7.5 [B] (Local)</td></tr>
<tr><td>W 40-44</td><td>Phonetics 7.5 [B] (Local)</td></tr>
<tr><td>W 44-48</td><td>Grammar 7.5 [B] (Local)</td></tr>
<tr><td>W 36-03</td><td colspan="2">Natural Language Processing 7.5 [A] (GU)</td></tr>
<tr><td rowspan="3">Semester 2</td><td rowspan="2">W 04-13</td><td colspan="2">Speech Technology 7.5 [A] (KTH)</td></tr>
<tr><td colspan="2">Statistical Methods 7.5 [A] (UU)</td></tr>
<tr><td>W 14-22</td><td colspan="2">Advanced Method Course 7.5 [A] (Local)
Project Work 7.5 [A] (Local)</td></tr>
<tr><td rowspan="3">Semester 3</td><td rowspan="2">W 36-44</td><td colspan="2">Language Resources 7.5 [A] (GU)</td></tr>
<tr><td colspan="2">Machine Learning 7.5 [A] (UU)</td></tr>
<tr><td>W 46-03</td><td colspan="2">Student's choice 15 (Local)</td></tr>
<tr><td>Semester 4</td><td>W 03-22</td><td colspan="2">Thesis Work 30 [A] (Local)</td></tr>
</table>

B – Basic course
A – Advanced course

Local – Given by the university where the student is registered
Courses in blue – Given by the responsible university (GU/KTH/UU) on the national level (existing GSLT course)

Figure 1: Proposal for Uppsala's programme starting autumn 2010

Semester	**Level 1**	**Level 2**
Autumn 2010	NLP	Linguistic Resources Machine Translation Natural Language Generation Java Development for HLT
Spring 2011	Speech Technology	Statistical Methods Speech Synthesis Treebanks Lexical Semantics

Figure 2: Planned national courses 2010–11

students if we follow the GSLT model of courses with intensive periods of teaching in Gothenburg or elsewhere. It is also not so clear that intensive teaching periods are the most suitable form of delivery for masters students. We therefore aim to produce course materials that are flexible with respect to the extent that traditional classroom teaching is employed. We aim at the following:

- Course materials should be in English
- They will be made available on a website
- They will include a downloadable video presenting the course (perhaps a filming of a classroom instance of the course). In addition to being available on the national website, the video will also be published on iTunes U to increase visibility for the course.
- They will include a complete set of exercises, projects etc. for the course
- Also a complete set of course notes, overhead presentations used and bibliographical references
- The course should be in a form so that in principle it could be delivered by a local teacher (e.g. an advanced graduate student) who is not necessarily an expert in the particular subject matter of the course.
- The course website should name an expert or several experts (e.g. the course coordinator) willing to consult with local teachers if questions about the material should arise.
- Course materials should be reviewed by the expert at three yearly intervals to determine whether changes should be made to keep the course up-to-date.

In this way we hope to maximize the use of the expert knowledge we have without creating an impossible teaching situation for the experts.

E Amendments and additions to previous version

The following amendments and additions relate to the six points raised by the external reviewers:

GSLT There is additional text on p. 19, which describes in more detail our strategy relating to the uncertainty of GSLT's future. A strategy document written by GSLT's Academic Board has been added as Appendix C. GSLT's proposal for a national masters school has been included as Appendix D.

Cooperation outside LT The fifth future development area described on p. 4 and p. 17 concerns the use of LT resources and tools in other academic research as well as in future product development.

Future strengths In the summary starting on p. 17 we show how our planned work relates to five future development areas in LT (first introduced on p. 3) and indicate in broad terms what is novel about the work and what is new for us.

Grammar Lab On p. 14, the activities planned in the grammar lab for the period 2009–2012 have been specified in more detail.

Vision On p. 3 we have added an account of five future development areas in LT which we believe that Gothenburg LT can make a significant contribution to. We have also indicated that these development areas should be tracked in evaluation by being included among our success indicators (p. 22).

New blood Originally, we interpreted the instructions given to us by the University to mean that activities aimed at attracting additional – external – funding should be prioritized in the plan. With this in mind, we defined a budget where "seed money" and visiting researchers were given prominent place. Considering that the external reviewers view this as a significant weakness of the plan – they devote more text to this point than the other five together – we have revised the budget in Appendix B accordingly.