

Annual Report

Graduate School of Language Technology (GSLT)

2006

1 Overview

The Graduate School of Language Technology (GSLT) is a national graduate school for which Göteborg University (Faculty of Arts) is the coordinating host. General information about the school and its original programme statement can be found on its web page <http://www.gslt.hum.gu.se>. This report covers the calendar year 2006, its fifth year of operation.

Students may currently be registered at any of the following academic institutions:

- University College of Borås
- Chalmers University of Technology
- Göteborg University
- KTH (Royal Institute of Technology)
- Linköping University
- Lund University
- University of Skövde
- Stockholm University
- Uppsala University
- Växjö University

Supervision is also available from SICS (Swedish Institute of Computer Science). We are still open for the addition of further institutions in the future if this should become appropriate.

2 Teaching

The following table gives an overview of the courses offered by GSLT during 2006:

Course	Semester	Level	Credits	Inst	GSLT students	External students
Natural language processing	Spring	1	5	SU	4/4	6/13
Java development for HLT	Spring	2	5	LiU	2/3	3/9
Dialogue systems	Spring	2	5	Liu	4/4	2/3
Lexical semantics	Spring	2	5	UU	3/7	4/15
Machine learning 2	Spring	3	10	VxU	10/13	4/4
Speech technology	Autumn	1	5	KTH	4/4	6/6
Machine translation	Autumn	2	5	UU	3/3	1/3
Treebanks	Autumn	2	5	SU	2/2	2/3

The number of GSLT students and external students in each course is given as a ratio between the number of students that completed the course on time and the number of students that started the course. We note with satisfaction that most courses continue to have a very good quantitative throughput, although exceptions exist, and that course evaluations continue to be mostly positive.

The table below shows the participation of teachers in the different courses. We see that, as in previous years, many courses involve teachers from more than one institution, sometimes with some teacher coming from outside GSLT.

Natural language processing	Martin Volk Lars Borin Torbjörn Lager Barbara Gawronska	SU GU GU HiS
Java development for HLT	Lars Degerstedt	LiU
Lexical semantics	Åke Viberg	UU
Dialogue systems	Lars Ahrenberg Arne Jönsson	LiU
Machine Learning 2	Joakim Nivre Martin Volk Jussi Karlgren Arne Jönsson Lars Borin Oliver Lemon	VxU SU SICS LiU GU University of Edinburgh
Speech technology	Rolf Carlson Björn Granström David House Mats Blomberg Kjell Elenius	KTH KTH KTH KTH KTH
Machine translation	Anna Sågwall Hein Daniel Hardt	UU Copenhagen Business School
Treebanks	Martin Volk Joakim Nivre	SU VxU

3 Seminars

The activities at GSLT intensive weeks and, in particular, the graduate students seminars have been discussed on several occasions over the years. There has been a general agreement that attendance, both from students and faculty, is lower than desired. In the fall of 2005 a questionnaire was distributed among the Ph.D. students on the initiative of Leif Grönqvist and Kenneth Wilhelmsson, and a discussion was held at the last intensive week of that year. This led to a number of changes to the seminar policy starting in 2006, the most important being the following:

- Graduate students are required to give two longer seminars relating to their thesis topic with the main supervisor present. The first one of these should take place after about two years of study, and the second one, about six months before completion of the thesis.
- Calls for thesis seminars should be sent out well in advance.
- In addition to the obligatory thesis seminars, the students may volunteer seminars to present intermediate results or plans.
- A larger share of the GSLT seminars should be given by faculty.

These changes have taken effect during 2006 and have worked out fairly well.

4 Relations to industry

During 2006 we noticed an increased interest from industry in the GSLT graduate students. Contacts were made on an individual basis rather than as the result of an organised effort on part of GSLT, but nevertheless several students have been employed or consulted part time this year.

The invited seminars continued during 2006 covering the following topics: Language technology and new text (Jussi Karlgren, SICS), Automated Question Answering: Template-Based Approach (Erik Sneiders, Askology AB), and The web – what is there today and what may be in the future (Olle Olsson, SICS and the Swedish W3C Office).

GSLT responded to a call by VINNOVA (the Swedish Governmental Agency for Innovation Systems) aiming at forming new structures for cooperation between research centers of excellence, Swedish industry and graduate schools and received a planning grant for a complete application during 2007.

No meeting with the Industrial Advisory Board was held during 2006, though we were in contact with them for the VINNOVA application.

5 PhD Students

Six new students (two women and four men) were admitted in January 2006 with full funding and one student (man) with external funding. Two of the students admitted with full funding were already admitted to the school with external funding. Thus the net gain in number of students to the school was five students.

PhDs with external funding receive their primary funding from their home institution but GSLT pays for expenses involved in taking part in GSLT activities and other expenses (such as the purchase and maintenance of a laptop computer).

This year's admission resulted in the distribution of students over the GSLT participating institutions shown in Table 1.

Institution	Fully funded	Externally funded
Borås	2	1
Chalmers	2	0
Göteborg	8	3
KTH	9	0
Linköping	5	0
Lund	1	2
Skövde	2	0
Stockholm	5	1
Uppsala	4	1
Växjö	2	0

Table 1: Distribution of students over participating institutions

The students are admitted to a range of different degrees at their home institutions. The distribution is shown in Table 2.

Degree subject	2001	2002	2004	2005	2006
Computational Linguistics		2	3	1	3
Computer Science	4	3	1	2	1
Human Computer Interaction	1				
Linguistics	4	7	1	1	1
Natural Language Processing		1	1		
Speech Communication	4	1		2	1

Table 2: Distribution of students over degree subjects

This contrasts with the distribution of first degrees of these students (some students had more than one subject in their first degree) shown in Table 3.

Table 4 shows the percentage of the degree completed at the end of 2006 as reported in individual study plans for 2007. (Note that 100% completion does not necessarily mean that a student has a

Subject of first degree	2001	2002	2004	2005	2006
Cognitive Science	1	2		2	
Computational Linguistics	7	5	2	1	4
Computer Science	3	1		1	1
Foreign languages	3	1	1		
Information Science			1		
Information Technology				1	
Language, Logic and Information			1		1
Language Technology			1		
Library and Information Science		2			
Linguistics	5				
Philosophy		1			
Phonetics	1	1			
Psychology		1			
Physics				1	

Table 3: First degrees of students

finished degree in hand but rather that they have completed the equivalent of four years full-time study towards the degree.) This chart includes students with external funding and some are counted from the year they started their PhD studies previous to joining GSLT.

<i>Percentage degree completed</i>				
Admission to GSLT	≤20%	21-50%	51-80%	81-100%
September 2001				12
September 2002			4	9
January 2004		5	1	
January 2005		10		
January 2006	3	1	1	

Table 4: Percentage of four years full-time study completed

Table 5 shows the number of students who have achieved various percentages of required course credits (60 or 30 credits, depending on the option they are following). This table is based on entry date to a PhD programme, which may be prior to admission to GSLT.

A list of GSLT's PhD students with a short description of their respective thesis project can be found in Appendix B. A list of publications by GSLT's PhD students during 2006 can be found in the appendices following this together with a list of industrial contacts that our students have had during the year.

<i>Percentage course requirements completed</i>				
Admission to PhD in home dept	≤20%	21-50%	51-80%	81-100%
2001 (or earlier)			1	4
2002		1	1	6
2003-2004			1	12
2005		2	1	1
2006		3		

Table 5: Percentage of required course credits completed

6 Completed theses

In June 2006, Magnus Sahlgren successfully defended his thesis *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. The thesis concerns a computational model of word meaning that uses the distributional properties of words to quantify their meaning differences. The thesis was awarded the prize for the most prominent scholarly achievement of 2006 by the Stockholm University Faculty of Humanities. Magnus Sahlgren is employed as full-time researcher at SICS, and works with language technology in a number of national and international research projects.

Genevieve Gorrell successfully defended her thesis in November 2006. The thesis concerns vector space techniques in natural language processing, and focuses on the utility of dimensionality reduction. In particular, an algorithm for singular value decomposition is developed upon. Since graduating, she has worked on the European Union-funded project "Companions", focusing on spoken dialogue systems. She now teaches at the Department of Computer Science at Sheffield University, UK.

Leif Grönqvist defended his thesis *Exploring Latent Semantic Vector Models Enriched With N-grams* in November 2006. The thesis concerns a kind of vector space model called Latent Semantic Vector Models applied to the field of Information Retrieval. Experiments have been performed and evaluated to investigate whether it is possible to improve the models by including N-grams in these normally word-based models. From September 2006 Leif works at a company that develops data mining and visualization applications.

In December 2006, Susanne Schötz successfully defended her PhD thesis *Perception, Analysis and Synthesis of Speaker Age*. The thesis describes six studies of phonetic aspects of age-related variation in speech. The final study developed a research tool which used data-driven formant synthesis and age-weighted linear interpolation to simulate an age between the ages of any two of four female differently aged reference speakers. Evaluation of the tool showed that speaker age may in fact be simulated using formant synthesis. Susanne is now doing research and teaching at the Department of Linguistics and Phonetics, Lund University.

In December 2006, Per-Anders Jande successfully defended the thesis *Modelling Phone-Level Pronunciation in Discourse Context*. The work described in this thesis takes a holistic perspective on pronunciation variation of Swedish and focuses on a method for creating general descriptions of phone-level pronunciation in discourse context. The discourse context is defined by a large set of linguistic attributes ranging from high-level variables such as speaking style, down to the articulatory feature level. Models of phone-level pronunciation in the context of a discourse have been created for the central standard Swedish language variety. Per-Anders is currently engaged in NGOs working with environmental, social equality and human rights issues.

Gustav Öquist defended his thesis *Evaluating Readability on Mobile Devices* in December 2006. The thesis presents findings from five readability studies performed on mobile devices. The dynamic Rapid Serial Visual Presentation (RSVP) format was enhanced with regard to linguistic adaptation and segmentation as well as eye movement modeling. Eye movement tracking was then used to learn more about how the novel presentation formats affected reading compared to other formats on mobile devices. The thesis explains why readability on mobile devices is important, how it may be

evaluated in an efficient and yet reliable manner, and finally, a new predictive text presentation format is pinpointed as the format with greatest potential for improvement. Since January 2007 Gustav works at the Bernadotte eye movement laboratory at Karolinska Institutet and is also involved in a start-up company developing an interactive reading aid tool for dyslectics.

7 Financial Report

The Graduate School had an original budget of 63 MSEK 2001–2007. Göteborg University and the Faculty of Arts have in addition given us permission to budget up to the year 2012 with a budget of 12 MSEK per year.

Table 6 summarises our finances for 2002–06. All amounts are stated in SEK.

	Outgoings 20061231	Outgoings 20051231	Outgoings 20041231	Outgoings 20031231	Outgoings 20021231
Salaries					
PHD	9,266,000	8,472,884	7,368,261	7,757,000	4,216,000
Admin	638,000	542,174	482,732	492,000	359,000
Director	504,000	447,794	435,807	382,000	382,000
	10,408,000	9,462,852	8,286,801	8,631,000	4,957,000
Equipment and services					
Computers	202,000	183,190	387,686	336,000	209,000
Local Comp admin	222,000	261,701	248,164	274,000	148,000
Service, support	13,000	67,606	43,650	57,000	107,000
Copying, mail, etc	128,000	98,789	76,943	91,000	69,000
Equipment	290,000	327,701	257,166	256,000	149,000
	855,000	938,987	1,013,609	1,014,000	682,000
Teaching					
Courses	527,000	423,830	399,178	350,445	438,000
Supervision	907,000	879,890	689,463	917,150	442,000
Local costs	1,097,000	1,071,671	913,380	947,000	478,000
	2,531,000	2,375,391	2,002,021	2,214,595	1,358,000
Travel, meetings, miscellaneous					
Retreat	187,000	165,245	162,903	81,000	144,000
Travel	6,000	19,950	6,206	13,000	81,000
PHD-travel	115,000	121,670	100,011	123,000	14,000
Consultation	0	3,058	44,826	57,000	14,000
Accommodation	42,000	58,145	42,445	47,000	48,000
Meetings	86,000	57,312	76,568	83,000	71,000
Other	4,000	26,997	7,845	15,000	12,000
	440,000	452,378	440,804	419,000	370,000
SUM TOTAL	14,234,000	13,229,608	11,743,235	12,278,595	7,367,000

Table 6: Cost overview 2002-06

8 Computer Systems, Electronic Resources, and Services

During 2006 GSLT upgraded its computer systems to better support courses, administration, and research for its more than 80 users. The most notable hardware upgrade was replacing the RAID system with a new 12 disk-slot Sun StorEdge 3320 SCSI disk array. This new RAID system greatly improved access times for the users' home accounts and is able to hold more resources, e.g. the 16 corpora from the LDC, *Linguistic Data Consortium*, we will acquire in 2007.

As of December 2006 GSLT's server environment contained

- a Sun Fire v440, 4×1GHz UltraSPARC IIIi with 8GB of memory running the Solaris 9 operating system,
- a Sun StorEdge 3320 SCSI Array, 8×300GB 10Krpm Ultra320 SCSI disks, and
- a HP ProLiant DL385, 2×2.4GHz AMD Opteron with 2GB of memory running the RedHat Enterprise Linux ES release 3 operating system.

The client environments contained

- 57 IBM/Lenovo ThinkPad T22/T23/T42/T43/T60 notebooks dual-booting Linux and Windows operating systems,
- 8 Sun Workstation Ultra10/Ultra30 running Solaris 9, and
- 3 Desktop PCs (Dell & CompuStat) running Linux and Windows.

Domains running on GSLT's servers during 2006 were

- `gslt.hum.gu.se` – the Swedish National Graduate School of Language Technology,
- `sprakteknologi.se` – Språkteknologi.se, the National Language Technology Center, and
- `ngslt.org` – the Nordic Graduate School of Language Technology.

Software upgrades of language technology programs and development environments were done continuously over 2006. In the GSLT annual report last year we reported that we were about to introduce a net-based meeting system, *Marratech*, gaining audio/video and a collaborative whiteboard over the older text-based chat system, *IRC*. *Marratech* was used successfully for course meetings in e.g. the NLP and Treebank courses, and we will continue to offer this service in the future.

The computer laboratory has been well visited by GSLT's Ph.D. students, teachers and guests during the intensive weeks. During other weeks, however, we have noticed a decrease in use for this room.

9 GSLT strategy

This year has represented the continuation of a more stable period for GSLT and thus strategy for the future has not been of such importance as in previous years. Our future strategy will become critical again with the last of the currently planned intakes in 2008 when it will become necessary to negotiate for GSLT's existence after 2012. We plan to take up this discussion in the autumn of 2007.

Appendices

A Reports from participating universities

Borås University College

In 2006, the University College of Borås continued to improve its cooperation with researchers in the field of information retrieval, automatic indexing and classification, and language technology. The supervision of our three PhD students at GSLT has been carried out by a Professor and two Associate Professors. These PhD students, in turn, take part in teaching in different courses of the Swedish School of Library and Information Science (SSLIS) curriculum for a coupling of research and education. Lately we have explicitly started utilizing machine learning knowledge in our graduate education as well, and, through GSL, high quality PhD courses have been added to our PhD-programme while research quality has been improved also through collaboration and help with quality assessment using experts from other universities in the program. GSLT as a high quality think tank and a network of professionals whose experience is valuable for SSLIS has remained a priority item on the list of our resources while working on different applications to national and international funding agencies, including the EU 7th Framework Programme. Further, SSLIS has a vested interest in the research topic Focusing on Language for Advanced Information Retrieval, exploring the connection between IR and document classification for libraries as a problem of language representation. We think high on collegial help and language technology resources provided by GSLT partner institutions which enable SSLIS to contribute to the development of the Swedish Library and Information Science field.

Chalmers University of Technology

The department of Computing Science is shared between the University of Göteborg and Chalmers University of technology. The research group in Language Technology has been built up during the last ten years. Our scientific contact with linguists in Sweden has earlier been restricted to the department of linguistics in Göteborg but thanks to the establishment of GSLT we now also have contact with other linguists in Sweden. Our graduate students in Computing can now specialize in language technology and we are impressed by the quality and quantity of organized courses which GSLT has made available to them.

Göteborg University

Language Technology continues to be a priority area in our faculty and GSLT continues to play a central role in this. Collaboration in this area between the departments of linguistics, Swedish language and philosophy continues and also our joint collaboration with Chalmers is continuing on various levels of activity: undergraduate teaching (through the undergraduate programme in language technology), graduate teaching (largely through GSLT) and in research involving joint projects. Some of this research is placed in the GU Dialogue Systems Laboratory located at the IT University at Lindholmen. While language technology is a growth area in our faculty it is still of paramount importance that we train our graduate students in a national context in order to be able to achieve the critical mass of teachers and students and the interdisciplinary spread of teaching and research that is necessary to

make us competitive internationally. GSLT has so far given us full or partial support for ten students, has enabled us to teach courses that we otherwise would not have been able to teach and has also provided our students with a much broader training than would otherwise have been possible. It has also provided significant computational resources for our graduate students and researchers.

KTH, Royal Institute of Technology

The School of Computer Science and Communication at KTH is a fusion of the former departments NADA (Numerical Analysis and Computer Science) and TMH (Speech, Music and Hearing). The new school comprises seven departments: NADA (Numerical Analysis), CVAP, CBN, TCS (computer Science), MDI (Human-Computer Interaction), Media Technology and Graphic Arts and Speech, Music and Hearing including the unit for Language and Communication. Already at the beginning of the two departments NADA and TMH joined GSLT. In total six graduate students got support from GSLT during 2006. Per Anders Jande graduated in 2006. The creation of GSLT has formed a high-quality foundation for education in language technology on a national level. The richness of education offered by the school is very difficult for one single university or institution to offer with limited resources and a small number of students. GSLT has formed an inspiring and supportive network for graduate students and also teachers and an excellent base for cross-fertilization. This will form a new generation of researchers with good national contacts, representing different backgrounds in language technology. After the development stage we have now created a very solid foundation for a true national graduate school in language technology. Thus, it is very promising that we are able to expand the school with new students every year. In addition to the national effort we are now also able to have classes through the NGSLT. The current Speech Technology Course is an example of such a Nordic cooperation where actually a majority of students came from countries outside Sweden. The staff at KTH is also engaged in supervising graduate students outside KTH.

Linköping University

During 2006 one of our students, Gen Gorrell, defended her Ph.D.-thesis, while the employment period for Mustapha Skhiri came to an end. For the other students the GSLT environment and courses have been absolutely essential for their progress. Especially the possibilities to take part in courses in machine learning and statistical methods given by other departments have been essential for their development.

The network provided by GSLT has also lead to an increased cooperation with other departments, especially in the area of machine translation and the construction of parallel corpora. We were also part of a joint application with other departments of GSLT to the Swedish Research Council for language technology infrastructure resources, which received a planning grant.

In the spring semester we conversely contributed to the GSLT course portfolio. Lars Ahrenberg and Arne Jönsson gave a course on Dialogue Systems with participants both from GSLT and NGSLT, while Lars Degerstedt gave his course on Java development for Human Language Technology.

As in previous years, Lars Ahrenberg, was a deputy director for GSLT.

Lund University

The GSLT network has continued to play a very important role for Lund University's education and research within language technology.

During 2006, one of Lund's two associated graduate students in GSLT, Susanne Schötz defended her Ph.D. dissertation, *Perception, analysis and synthesis of speaker age*. Being a part of GSLT, Schötz had the possibility of obtaining cosupervision on the speech technology aspects of her thesis from Rolf Carlson at KTH in Stockholm. As a postdoc, she is engaged in a speech technology research project financed by the Swedish Research Council (Simulect) and led by Gösta Bruce (Lund) and Björn Granström (KTH, Stockholm) two members of the GSLT network.

GSLT has further been valuable for taking the initiative for The Swedish Language Technology Conference in October 2006 where researchers from Lund involved in interdisciplinary research (e.g. Pierre Nugues from the Faculty of Engineering and Jonas Granqvist from the Dept. of Romance languages could ventilate their research findings on the system they developed for evaluating texts of second language learners).

The graduate school and its associated researcher network has further played an important role in extending Lund's involvement in European research projects involving language technology, e.g. the European ESFRI research infrastructure project CLARIN currently being planned and where researchers at the Humanities Laboratory in Lund (Sven Strömqvist, Marcus Uneson and Susanne Schötz) are involved as well as researchers from other centers of language technology in Sweden and other parts of Europe.

University of Skövde

For the University College of Skövde, the possibility to participate in GSLT has been of crucial importance, since the University College has no right to give PhD education on its own. Furthermore, the study programme in Computational Linguistics in Skövde is relatively young (it started 1996). Because of this, collaboration with universities that have longer experience in the field of speech and language technology is of great value. Participation in GSLT has made it possible to employ our first two PhD students, which is a step towards establishing an active and competitive research group.

Without GSLT, we would not have been able to give all necessary courses on PhD level. Now, our PhD students have access to competence represented on several well-known research centers. They also have opportunity to meet other PhD students with different background competence, which, as we hope, will result in a fruitful scientific exchange and cooperation.

Stockholm University

In 2006 Stockholm University has participated in GSLT in various ways. Its three GSLT students have taken various GSLT courses throughout the year while at the same time pursuing their research and teaching undergraduate courses at Stockholm University. In addition, one other Stockholm University PhD student (not financed through GSLT) has also taken GSLT courses. GSLT has enabled them to gain insights into new areas and to work on a broad set of challenging problems.

Involvement in GSLT by student and staff has helped all to get more integrated in the language technology community. The regular meetings during GSLT's intensive weeks, both the talks by the PhD students and the industry seminars, as well as GSLT's retreat and conferences have fostered the exchange of ideas and a sense of striving towards a common goal.

Stockholm University staff have participated in coordinating and teaching GSLT courses and organizing the Swedish Language Technology Conference.

Uppsala University

For Uppsala University participation in GSLT continues to be very important for maintaining the quality of the graduate study program in computational linguistics. In December 2006, Gustav Öquist defended his Ph.D. thesis, entitled *Evaluating Readability on Mobile Devices*, thereby becoming the first graduate from Uppsala University within GSLT. However, since we also received one new student in this year's round of admissions, the total number of students from Uppsala enrolled with GSLT is still four.

In addition to being of central importance for the graduate study program, GSLT also provides a forum for collaboration in research, as evidenced by the ongoing project on building a Swedish BLARK (Basic Language Resource Kit), which is a cooperation between Göteborg University, KTH, Linköping University, and Uppsala University, initiated under the umbrella of GSLT.

Växjö University

For Växjö University participation in GSLT has been and continues to be an absolutely essential component in the build-up of a Ph.D. program in computer science with specialization in language technology. A sign of success for this program is the fact that Ph.D. students from Växjö University have had papers accepted to the ACL conference (the major international conference in language technology, with an acceptance rate of 15–20%) both in 2005 and 2006. Without the collaboration with other universities within GSLT, it would simply not have been possible to maintain a Ph.D. program at this level of quality. It is also worth mentioning that participation in GSLT has also led to collaborations in research that are not directly connected to graduate education.

During 2006, Växjö University has continued to be active in organizing courses within GSLT, being responsible for the level 3 course in machine learning in the spring and teaching within the course on

treebanks in the autumn. Unfortunately, no new students have been admitted to Växjö University in 2006.

B PhD projects

Atelach Alemu Argaw, Department of Computer and Systems Sciences, Stockholm University

Previous degree: M.Sc. in Information Science

Thesis topic: Query Translation and Expansion in Cross Language Information Retrieval

Supervisor: Lars Asker, Associate Professor of Computer Science, Department of Computer and System Sciences, Stockholm University/Royal Institute of Technology (KTH), Stockholm

Assistant supervisor: Jussi Karlgren, PhD, The Human Computer Interaction and Language Engineering Laboratory, SICS.

The research focus is to investigate various ways of translating query terms and query expansion with the intent of boosting IR performance in a cross lingual domain.

The effects of expanding queries before and after translation (through dictionary lookup), and the effects of word sense disambiguation during the lookup will be investigated in detail. The aim here is to find a balance between expansion and reduction of terms in order to come up with a list of keywords with optimal retrieval precision and recall. Statistical collocation measures and machine learning techniques will be used in order to carry out these tasks.

Approaches to increase mono and cross lingual retrieval performance through query expansion and document re-ranking using pseudo relevance feedback will also be investigated.

Jenny Brusk Dept of Linguistics, Göteborg University and Dept of Game Design, Narrative and Time-based Media, Gotland University

Previous degree: M.A. in Computational Linguistics

Thesis topic: Dialogue Management in Virtual Game Worlds

Supervisor: Torbjörn Lager, Prof in General and Computational Linguistics, Dept of Linguistics, Göteborg University

Assistant supervisor: Staffan Björk, Senior Researcher, GAME, Interactive Institute. Guest Lecturer, Dept of Computing Science, Chalmers Technical University.

The aim of this PhD project is to develop a new method for managing game dialogues using statecharts, "a visual formalism for complex systems" (David Harel, 1987). The dialogue manager will be implemented using SCXML, i.e. a new W3C standard which combines XML syntax with statechart semantics. Statecharts are in fact an abstraction of finite state machines, allowing hierarchial as well as orthogonal states. It is thus possible to allow simultaneous processes to run in parallel, which can be exemplified by a situation in which a game character picks up an object while talking. Within the Synergy Project (Torbjörn Lager, Fredrik Kronlid and myself) we are investigating how statecharts and SCXML can be used in dialogue system design as well as for game programming.

Dialogues in games are usually used for providing the player with background story, quests and motivations for performing certain actions in the game, but also as a way to get to know the characters in the game. Most dialogues in games are however canned and thus limited in expressivity and variation. One incentive for this research is to allow more flexible dialogues, through natural language conversations.

A game dialogue manager should for example support

- Story construction
- Characterisation and character development
- Game progression
- Social interaction

We will start by building a dialogue system for the trade domain in games, where a user can buy items in a store and negotiate about the price.

Karin Cavallin Linguistics, Gothenburg University

Previous degree: Master in Computational Linguistics

Thesis topic: Semantic restrictions on syntactic constructions

Supervisor: Robin Cooper, Professor, Linguistics, Gothenburg University

Assistant supervisor:

Questions that have arisen from my research on the existential construction so far are how to deal with the tendency in Swedish to avoid agents in post-verbal position in these constructions. Swedish speakers tend to accept agents in post-verbal position, but they don't seem to produce these sentences (at least not in the corpora used for data collection). We don't want to rule out intuitively grammatically sound constructions, based on their low frequency in corpora. But where do we want to implement these tendencies? In the grammar? In the lexicon? Do we want to implement this kind of restriction at all?

Which conclusions, if any, can we draw from the fact that sparse corpus evidence in Swedish correlates with ungrammatical behaviour in other languages? Can the complete avoidance in other languages and the sparse (basically nonexistent) evidence in Swedish show something about the mental lexicon? Can these observations somehow contribute to the discussion of intuition vs. empiricism as a method of studying syntax and semantics? And what are the consequences for implementing these restrictions in grammars to be used e.g. in translation and multilingual dialogue systems?

In my thesis I want to focus on these questions above. I will also perform an evaluation with regards to which grammatical formalism that best can describe the phenomena, and see how I can describe and explain the phenomena of the existential construction in Germanic languages.

Loredana Cerrato TMH/KTH

Previous degree: Foreign Languages and Literature, with specialization in Linguistics and Phonetics

Thesis topic: Communicative Feedback

Supervisor: David House, ass.prof. TMH/ KTH, Stockholm

Assistant supervisor: Jen Allwod, prof. Dept.of Linguistics Gothenburg University

This thesis deals with human communicative behaviour related to feedback, which is analysed in human-human and human-machine communication. The aim of this study is twofold: give more insight into how humans use communicative behaviour related to feedback and provide valuable data to control facial displays related to visual feedback in synthetic conversational agents. The materials used for the investigations presented in this thesis span from spontaneous conversations video-recorded in real communicative situations, and semi-spontaneous dialogues obtained with different eliciting techniques, such as Map-task and information-seeking scenarios, to a specific corpus of controlled interactive speech collected by means of a motion capture system. When motion capture is used it is possible to register facial displays with a high degree of precision, so to obtain valuable data for the implementation of facial displays in talking heads. A specific coding scheme has been developed, tested and used to annotate human communicative behaviour related to feedback. The annotation has been carried out with the support of different available software packages for audio-visual analysis. One of the final aims of the thesis is to provide data that could be used to control facial displays related to communicative behaviour into synthetic conversational agents.

Susanne Ekeklint, School of Mathematics and Systems Engineering, Växjö University

Previous degree: M.A. in Computational Linguistics, Göteborg University, 2001

Thesis topic: Dependency-Based Semantic Role Labeling

Supervisor: Joakim Nivre, Professor of Computational Linguistics, School of Mathematics and System Engineering, Växjö University and Guest Professor of Computational Linguistics, Department of Linguistic and Philology, Uppsala University

Assistant supervisor: Torbjörn Lager, Professor of General and Computational Linguistics, Department of Linguistics, Göteborg University

The work that will be presented in the thesis deals with new methods for automatic semantic analysis of sentences. The level of semantic annotation that has been considered is of predicate-argument type. This type of annotation can be used in order to improve different natural language processing tasks such as information retrieval, dialog management, translation or summarization. Typically any application that needs to recognize entities answering to question words such as *who*, *when*, *why* can benefit from this type of annotation.

The main goal of this research is to investigate the suitability of dependency-based representations for semantic role labeling. It has already been established that syntactic information is necessary for accurate semantic role labeling. It is however still an open issue which type of syntactic information should be used and how this information should be structured. The majority of published machine driven experiments on semantic role labeling are based on treebanks annotated with phrase structures.

Johan Eklund Swedish School of Library and Information Science, University College of Borås

Previous degree: Master's degree in Library and Information Science

Thesis topic: A comparative study of five ranking algorithms for query expansion

Supervisor: Sándor Darányi, Associate Professor, Swedish School of Library and Information Science, University College of Borås

Assistant supervisor: Sándor Dominich, Associate Professor, Department of Computer Science, University of Pannonia, Hungary

Classification is a central activity in libraries. The available information resources need to be categorized according to structured schemes, such as classification systems, to be accessible to the libraries' users. However, classification is a time-consuming intellectual task and it is an increasingly urgent issue to find useful methods for (at least partly) automating this process. The major research focus of my thesis project is to investigate to what extent one of the most successful classification algorithms, the support vector machine (SVM), can induce a useful classification model given the limited data available in typical library catalogs. The records in the collections available for this project are manually classified according to the SAB classification scheme, which will be used as a point of comparison, and enriched with value-adding data such as reviews and tables of contents. Three major research questions will be investigated: Viewed as a nonlinear optimization problem, how can the training efficiency of SVM for this particular application area be maximized? Which method for multi-class classification gives the overall best performance? Is there any significant difference in classification performance between mono- and multi-disciplinary subject areas?

Markus Forsberg, Computing Science, Chalmers University of Technology

Previous degree: Degree of Licentiate of Engineering *Thesis topic:* Functional Morphology and Lexicon Extraction *Supervisor:* Aarne Ranta, Professor, Computing Science, CTH

Assistant supervisor: Lars Borin, Professor, Institutionen för svenska språket, Göteborgs universitet

The thesis will focus on the tool Functional Morphology, consisting of the following parts: the Functional Morphology system with two case studies: old Swedish and SAL (Swedish Association Lexicon); lexicon extraction via our tool extract.

Functional Morphology is a toolkit for implementing natural language morphology in the functional language Haskell. The main idea behind is simple: instead of working with untyped regular expressions, which is the state of the art of morphology in computational linguistics, we use finite functions over hereditarily finite algebraic data types. The definitions of these data types and functions are the language-dependent part of the morphology. The language-independent part consists of an untyped dictionary format which is used for translation to other morphology formats and synthesis of word forms, and to generate a decorated trie, which is used for analysis.

The tool extract is developed for automatic extraction of lemma-paradigm pairs from raw text data. The tool combines regular expressions containing variables with propositional logic to form search patterns which identify lemmas tagged with their paradigm class.

Eva Forsbom Department of Linguistics and Philology, Uppsala University

Previous degree: B.A. in English and Swedish, Stockholm University, 1990; University Diploma in Language Consultancy, Stockholm University, 1996

Thesis topic: Textlinguistic methods in summarisation and information access

Supervisor: Anna Sångvall Hein, Professor of Computational Linguistics, Department of Linguistics, Uppsala University

Assistant supervisor: Daniel Hardt, Associate Professor, Department of Computational Linguistics, Copenhagen Business School

The purpose of the thesis is to investigate if and how some automated textlinguistic methods can give more relevant hits in information retrieval, and give coherent summaries that are more query and user adapted than those usually given in information systems.

A lexical cohesion analysis is used as a basis for indexing, searching and a short summary in an information system. The analysis is based on a number of knowledge bases containing linguistic or world knowledge, and the result will mainly depend on what knowledge is available.

By combining the lexical cohesion analysis with a Rhetorical Structure Theory analysis, it should be possible to come to terms with some coherence problems in summaries only based on lexical cohesion analysis. At the same time, the less computationally costly lexical cohesion analysis could reduce the number of possible RST analyses, since it also gives an estimate on how closely sentences are related.

Anders Green, Numerical Analysis and Computer Science (NADA), Royal Institute of Technology (KTH), Stockholm.

Previous degree: M.A. in Computational Linguistics, Göteborg University.

Thesis topic: Human-Robot Interaction using Multi-sensory Natural Language Interfaces

Supervisor: Kerstin Severinson Eklundh, Professor, Department of Numerical Analysis and Computer Science, Royal Institute of Technology (KTH), Stockholm.

Assistant supervisor: Henrik Christiansen, Professor, Numerical Analysis and Computer Science (NADA), Royal Institute of Technology (KTH), Stockholm.

How should a Multi-sensory Natural Language User Interface between a human user and an Intelligent Service Robot, be designed in order to support naturalness in human-robot communication, and in particular, communication using speech and gestures in combination? What types of dialogue and what kinds of gestures are required to successfully perform practical tasks using a robot? The view taken in this research is that a Multi-sensory Natural Language User Interface can be thought of as a system that uses several sources of information to analyze the behavior of the user. This information is used by the system to come up with an appropriate response, typically an utterance or an action of the underlying system.

The two main themes of this research are:

1. Establishing principles for the design of usable Multi-sensory natural language user interfaces for Intelligent Service Robots.
2. To develop methods for computational analysis of co-expressive speech and gesture used to instruct Intelligent Service Robots.

Leif Grönqvist, School of Mathematics and Systems Engineering, Växjö University

Previous degree: M.Sc. in Computing Science, Göteborg University, 1994

Thesis topic: Improved Latent Semantic Indexing Models for Information Retrieval

Supervisor: Joakim Nivre, Professor of Computational Linguistics, School of Mathematics and System Engineering, Växjö University

Assistant supervisor: Jussi Karlgren, PhD, The Human Computer Interaction and Language Engineering Laboratory, SICS

The major goal of this investigation is to improve the vector model obtained by LSI, using linguistic information that could be extracted automatically from raw textual data. The kind of improvements in mind for specific applications, are to give a search engine the capability of disambiguate ambiguous words, and to make it possible for a keyword extractor to find not just words, but a list of relevant multi-word units, phrases and words.

The starting point is that multi-word units in the vector model could give the improvements above, since the model as it is includes only single words. How this information should be added is an open question. Two possible ways would be to insert tuples/collocations extracted by some kind of statistics, or to use a shallow dependency parser.

How this should be done is not at all clear, so many experiments will be needed. However, it is important to use extremely fast algorithms. At least one billion words should be possible to prepare in reasonable time (the magnitude of one day), which will limit the possible ways to add these phrases.

The different approaches has to be evaluated using for example a trained vector model and one of the following: a typical IR test suite of queries, documents, and relevance information, texts with lists of manually selected keywords (multiword units included), or a word comprehension test such as TOEFL or ORD in Högskoleprovet.

Mikael Gunnarsson Swedish School of Library and Information Science

Previous degree: Librarian diploma

Thesis topic: Genre Identification

Supervisor: Tor Henriksen, Associate professor, Swedish School of Library and Information Science, Högskolan i Borås

Assistant supervisor: Barbara Gawronska, Associate professor, School of Humanities and Informatics, University of Skövde

The aim is to model genre for document classification with particular respect to the so called logical document structure as a discriminative feature. This is accomplished as a two-stage process of 1) modelling of the document structure and 2) clustering.

Ebba Gustavii Department of Linguistics and Philology, Uppsala University

Previous degree: Master of Philosophy in Language Engineering, 2003

Thesis topic: Automatic translation of productively formed lexical units

Supervisor: Anna Sångvall Hein, Professor in Computational Linguistics, Department of Linguistics and Philology, Uppsala University

Assistant supervisor:

Several aspects of compounds make them particularly difficult to handle in a system performing automatic translation, being it a machine translation system or a system for cross-language information

retrieval. The relation between the parts of a compound is implicit, and thus its interpretation is never wholly compositional. This is, in particular, a problem when translating from a language making frequent use of compounds, to a language generally preferring syntactic constructions instead, since an overt syntactic marker is then to be generated. Related problems concern the choice of morpho-syntactic features of the target construction and its integration into the target sentence. In this thesis I will look into the methods for translating compounds suggested so far, and look for ways to improve them. In particular I will focus on corpus-based methods (as opposed to interpretation-based methods).

Harald Hammarström Department of Computer Science and Engineering, Chalmers University of Technology.

Previous degree: Computer Science, MA 2003, Uppsala University.

Thesis topic: Unsupervised methods in natural language processing.

Supervisor: Prof. Bengt Nordström, Department of Computer Science and Engineering, Chalmers University of Technology. *Assistant supervisor:* Prof. Aarne, Department of Computer Science and Engineering, Chalmers University of Technology.

Prof. Lars Borin, Department of Swedish, Gothenburg University.

So far I have developed some new algorithms for the problem of unsupervised induction of morphology, i.e., input an unlabeled text corpus of a natural language and output some description of how words in that language are declined. I, and most others, try to accomplish this in two phases; segmentation of words and discovery of systematic declension patterns. For the first phase, I have put some new insights on how to use frequencies to segment words (that have a simple base+affix formation) into an efficient and accuracy-competitive algorithm. For the second phase, I have come up with a promising formal criterion for characterizing systematic declension patterns (in the distributional environment typical of a natural language corpus).

In this line of work I still lack a systematic treatment of languages which have many “layers” of declension (as this is not something that can be swept under the carpet), and it still remains to work out exactly how the criterion for characterizing paradigmatic declensions should be exploited. More testing and evaluation are, of course, also needed but I don’t see this as a very big hill to climb. From the data I have at hand now, knowing better now how I can use it, I think I have overestimated the challenges of evaluation earlier.

In the future I wish to continue the above work. There are several directions, such as doing detailed case studies, moving on to parts-of-speech/syntax induction, post-process the morphology-description-output for integration and use in some tool for morphological analysis and generation, notably FM/GF.

Cecilia Hemming, Department of language, University College of Skövde (Department of Linguistics, Göteborg University).

Previous degree: B.A. French and Linguistics, Skövde

Thesis topic: Morphology and semantics of compounds and multi-word terms used to designate technical items

Supervisor: Barbara Gawronska, Associate Professor, Department of Languages, University College Skövde

Assistant supervisor: Joakim Nivre, Professor of Computational Linguistics, School of Mathematics and System Engineering, Växjö University

Technical word formation in both French and Swedish is facilitated by using affixation, compounding and combinations of both. There are production patterns though that seem to be language specific. Contrastive studies can reveal what similarities, divergences and relations there are in the two languages. The result of this linguistic analysis can be used in the development of applications for Machine Translation or Information Extraction.

Anna Hjalmarsson TMH, KTH

Previous degree: Degree of Master of Science in Cognitive Science

Thesis topic: Utterance generation in spoken dialogue systems

Supervisor: Rolf Carlson, TMH, KTH

Assistant supervisor: Joakim Gustafson, Telia Sonera

In conversation humans can choose to say something in a number of different ways but the choice is not arbitrary. Depending on what we want to communicate we consider different linguistic choices available in the current context. A natural part of human conversation is to adapt what we say and how we say it depending on our conversational partners and the dialogue context. This includes syntactic, semantic and lexical variation. A spoken dialogue interface which presents the right information, in the right way, at the right time, to the right user is more efficient and natural to use. However, the main effort within in the research area of spoken dialogue systems has been put on the processes of understanding human speech rather than the processes responsible for generating natural and context aware system utterances. My research focuses on developing spoken utterance generation which to a large extent follows the principles of human-human communication. To generate appropriate and appropriately timed feedback we need better knowledge of how this is done in conversation between humans. An important part of this work is consequently to collect and analyze dialogue data. Analysis of these data will be used to build utterance generation models which can be implemented and empirically tested within the context of a dialogue system.

Hans Hjelm Department of Linguistics, Stockholm university

Previous degree: M.S. in Computational Linguistics, Gothenburg university, 2001

Thesis topic: Ontology learning from parallel texts

Supervisor: Martin Volk, Department of Linguistics, Stockholm university

Assistant supervisor: Joakim Nivre, Professor of Comp. Ling., School of Mathematics and Systems Engineering, Växjö/Guest Prof. of Comp. Ling., Dept. of Linguistics and Philology, Uppsala University

There are many situations where it is of great importance for two agents (human or artificial) to know that they are talking about the same thing. Misunderstandings can be caused by the agents speaking different languages or by them using the same word(s) to mean different things (homonymy/polysemy). To solve this problem, one can make use of an ontology, which, put informally, associates the terms

from a domain with nodes with unique meanings, ordered in a hierarchical structure. Here is another definition of ontology (from Chandrasekaran et al., 1999):

*An ontology holds information about what **categories** exist in the domain, what **properties** they have, and how they are **related** to one another.*

One example where ontologies are put to use: all procurement within the EU now has to be carried out using the Common Procurement Vocabulary, which in fact is an ontology.

The aim of my thesis is to develop a methodology for extracting domain-specific ontologies using information available in parallel corpora. The extracted ontologies would provide the following information:

1. Which terms are relevant to the domain.
2. Which terms are more general and which are more specific (for example, car is more general than SUV).
3. Which terms mean more or less the same thing (for example computer display and computer screen).
4. How a certain term can be expressed in another language.
5. What other relations hold between the terms of the domain (for example, a keyboard is part of a laptop).

References: Chandrasekaran, Balakrishnan, John R. Stephenson and V. Richard Benjamins (1999): *What are Ontologies, and Why Do We Need Them?* In IEEE Intelligent Systems 01/02 1999. 20-26.

Maria Holmqvist Linköpings university

Previous degree: MS in Cognitive Science

Thesis topic: Parallel text processing for Machine translation

Supervisor: Lars Ahrenberg, Professor, Department of computer and information science, Linköpings universitet

Assistant supervisor: Magnus Merkel, Ph.D , Department of computer and information science, Linköpings universitet

Per-Anders Jande, Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm

Previous degrees: B.A. Phonetics, B.A. General Linguistics, B.A. Computational Linguistics, M.A. Computational Linguistics

Thesis topic: Pronunciation variation modelling for Swedish

Supervisor: Rolf Carlson, Professor in Speech Technology, Department of Speech Music and Hearing, Royal Institute of Technology (KTH), Stockholm

Assistant supervisor: to be determined

Although there is a certain degree of individual and random variation in the pronunciation of words in context, the variation is largely systematic within a restricted, relatively homogeneous group of language users. This agreement on systematic variation strategies can be seen as a property of the language variety (e.g. dialect) spoken by the group. The aim for my PhD project is to model this systematic variation inherent to a language variety, with the focus on variation in phone level realisation. The target language variety is central standard Swedish. The methods used for pronunciation variation modelling are data-driven. Spoken language is annotated with various kinds of linguistic and related information and machine learning is used to create models of pronunciation variation from the annotation. The phoneme is the central unit in the pronunciation variation modelling approach employed and the annotation is aimed at describing the discourse context of a phoneme from high-level linguistic variables such as speaking style, down to the articulatory feature level. This multi-variable linguistic context description is then used to predict the realisation of a phoneme in its context. Making information on different linguistic levels available as predictors of phone level pronunciation during machine learning gives a significant reduction of the errors produced by the resulting models. For the data used, models trained on information from multiple linguistic levels on average produce 42.6% less errors on the phone level than models trained on phoneme level information only. Current optimal models produce an average phone error rate of 8.1%, which is an improvement of 63% compared to using canonical phonemic word representations for estimating the phone-level realisation of natural speech.

Richard Johansson Department of Computer Science, Lund University

Previous degree: M.Sc., licentiate

Thesis topic: Automatic semantic analysis and data production

Supervisor: Pierre Nugues, docent, Department of Computer Science, Lund University

Assistant supervisor: Görel Hedin, docent, and Boris Magnusson, professor, Department of Computer Science, Lund University

In general, my research interest lies in automatic methods for semantic analysis of text, and the machine learning methods that are suitable for such tasks. My interest in such tasks stems from my previous work on automatic illustration of narrative texts with multiple events, in particular physical events such as traffic accidents. My research group identified two pivot problems for automatic analysis of such texts: predicate-argument structures ("who does what to whom") and temporal structure (i.e. the ordering of the events in time). Primarily, my own research has focused on the first of these; in particular, I have constructed automatic systems for frame-semantic analysis of predicate-argument relations according to FrameNet. I have also participated in projects that addressed the problem of temporal structure analysis.

Since the automatic systems that have been described in literature all rely (for training of statistical models) on large quantities of annotated data which are not available for small languages such as Swedish, I have also studied and created methods for automatic annotation of corpora, in particular by means of structure transfer from English via parallel texts. My licentiate thesis described a FrameNet-based semantic analyzer for Swedish, which was trained on automatically transferred structures.

My current research focuses primarily on the following key problems:

- New methods and improved mathematical models for transfer of semantic structures from one language to another.
- Methods to measure and improve the quality of the automatically produced data, preferably methods involving a minimal amount of manual labor.
- New methods and models for automatic semantic analysis of text, such as predicate-argument structures.

There are several interesting (and partly unexplored) machine learning aspects of the problems at hand, which are important to my work (in particular for the third point above):

- Prediction of complex structures, where there are strong dependencies between the predicted variables.
- Machine learning with incomplete feedback.
- Machine learning with noisy training data, where the rate of corruption differs between sections of the training data.

Rebecca Jonson Department of Linguistics, Göteborg University

Previous degree: M.A. in Computational Linguistics

Thesis topic: Enhancing Speech Recognition for Dialogue Systems

Supervisor: Robin Cooper, Professor, Linguistics Department, Göteborg University

Assistant supervisor: Rolf Carlson, Professor, Department of Speech, Music and Hearing, KTH

The purpose of the thesis is to explore possibilities of how speech recognition and spoken dialogue systems could be further integrated. A spoken dialogue system has been developed, with the TrindiKit toolkit, which will work as a baseline system. The research will consist in implementing and evaluating different methods that seem plausible for improving the spoken input to the baseline system. An example of such a method would be to filter speech recognition hypotheses according to plausibility in the dialogue information state or the domain. Historically, speech recognition and dialogue systems have been two separate fields and it is just recently that researchers have started to take into account all the information that exists in the dialogue system to assist in the disambiguation task of the speech input. Such research is currently being undertaken in the European TALK project to which this work is connected.

The direct contribution of the research will be a further integration of the Trindikit toolkit with speech recognition modules. However, the general aim of the thesis is to explore possibilities for generic technologies which lead to a closer integration of speech recognition and dialogue management.

Fredrik Kronlid, Department of Linguistics, Göteborg University

Previous degree: M.A. in Computational Linguistics, Göteborg University

Thesis topic: Turn and Dialogue Management in a Multi-Party Setting

Supervisor: Torbjörn Lager, Professor of General and Computational Linguistics, Department of Linguistics, Göteborg University

Assistant Supervisor: Staffan Larsson, Assistant Professor, Department of Linguistics, Göteborg University

The aim of the thesis is to facilitate multi-party dialogue between artificial and human conversational agents by providing accurate and efficient turn and dialogue managers for artificial conversational agents.

The *Synergy* platform, developed for chat, interactive agents and games, is used as a framework for the implementation. *Synergy* uses Harel State Charts as the agent building blocks.

The turn manager design is inspired by the so called SSJ (Sacks, Schegloff, Jefferson) turn-taking model. The dialogue management design builds on work by Staffan Larsson and Jonathan Ginzburg, using and developing the Information State Update/Dialogue Gameboard philosophy.

Monica Lassi, Swedish School of Library and Information Science, University College of Borås and Göteborg University *Previous degree:* Master of Science in Library and Information Science *Thesis topic:* A socio-technical perspective on development of a collaboratory for data collection instruments within Library and Information Science *Supervisor:* Diane Sonnenwald, professor, Swedish School of Library and Information Science, University College of Borås and Göteborg University *Assistant supervisor:* Jussi Karlgren, assistant professor, Swedish Institute of Computer Science

The goal of the thesis is to find out which socio-technical factors that are important for Library and Information Science (LIS) researchers concerning adoption and use of a collaboratory for data collection instruments used within LIS research. I am interested in finding out how LIS researchers choose and evaluate data collection instruments today, and how that process can be supported by a collaboratory by enabling sharing and use data collection instruments through a new medium. Another component of this thesis is to find out what LIS researchers prefer considering the representation of the data collection instruments by metadata in general and a controlled vocabulary in particular. A controlled vocabulary is used to represent the content aspects of a document, what is called a subject representation, in order to provide access points to documents through what they are about. In interdisciplinary areas, such as LIS, there is no single vocabulary that incorporates all concepts of the discipline terms are used in different ways, and one concept may be denoted by many different terms. The main research question is "*Which socio- technical factors affect the design and development and the adoption and use of a collaboratory for data collection instruments within Library and Information Science?*" The main themes include collaboratories, *data collection instruments*, *knowledge architecture (metadata and vocabularies)* and *Library and Information Science*. A grounded theory approach will be used, data including interviews, literature reviews, discipline analysis and a user study evaluating a prototype of a collaboratory for data collection instruments. The dissertation's relation to Language Technology (LT) is at least twofold: firstly, the focus on vocabularies and metadata are connected to semantics, in the case of the dissertation the meaning of concepts of an interdisciplinary discipline. In LT, semantic relations are represented by e.g. ontologies and thesauri, which would in LIS be classified as types of controlled vocabularies. Second, the attempt to enhance scholarly communication in a system for collaboration is connected to Computer Mediated Communication, a field that e.g. studies

how language is used in online environments.

Jonas Lindh Department of Linguistics, Göteborg University

Previous degree: Bachelor of Arts in Linguistics/Phonetics

Thesis topic: Forensic Speaker Identification

Supervisor: Anders Eriksson, Professor, Department of Linguistics, Göteborg University.

Assistant supervisor:

Thesis outline

Current working title: *A Theoretical Basis and Framework for Methods and Measures in Forensic Speaker Identification.*

At the moment I am writing up an article on a major investigation on fundamental frequency. This article is supposed to produce a theoretical basis for methods to measure F0 in forensic casework. Its starting point originates from the modulation theory of speech (MTS) described by Hartmut Traunmüller at Stockholm University (<http://www.ling.su.se/STAFF/hartmut/>). This theoretical starting point is then used to describe both inter- and intravariation for different F0 measures. The baseline from MTS is then suggested as a measure since different liveliness levels of speech influence it less. This is illustrated here http://www.ling.gu.se/~jonas/liveliness_illustration/.

As it is calculated using mean and standard deviation, which are measures severely affected by audio quality, an alternative baseline (A-base) is suggested as outliers or miscalculations influence it less. The A-base is theoretically more or less the same as it is the 5th percentile, situated approximately 1.43 standard deviations below the mean (equal to the baseline).

The measures are tested in 2 experiments where the first one tests robustness, while the second one illustrates the variation of the baseline at different vocal efforts.

Automatically extracted F0 statistics for 109 young males from the Swedia databa se (<http://swedia.ling.umu.se/>) are used to study intervariation for a homogeneous group.

Implications for how this can be used for automatic systems as well as further investigations are suggested.

The next step is to increase the population statistics for relevant measures and to more thoroughly start looking at intravariation for F0 as well as other measures, which can be acoustic, phonetic or phonological.

I am also interested in classification of intravariation. Can different kinds of variation be typical for certain kinds of variation such as sociophonetic, emotional etc. If possible these can be implemented as different parameters in systems to control and compensate for some of the variation.

For more information see my publications on <http://www.ling.gu.se/~jonas/Publications/> or to see references/literature go to my homepage for forensic phonetics <http://www.ling.gu.se>.

se/~jonas/forensic/

Svetoslav Marinov, School of Humanities and Informatics, University College Skövde

Previous degree: M.A in Linguistics, University of Tromsø, Norway

Thesis topic: Dependency parsing for Bulgarian

Supervisor: Barbara Gawronska, Professor of Computational Linguistics, School of Humanities and Informatics, University College Skövde

Assistant supervisor: Joakim Nivre, Professor of Computational Linguistics, School of Mathematics and Systems Engineering, Växjö University

Assistant supervisor: Torbjörn Lager, Professor of General and Computational Linguistics, Department of Linguistics, Göteborg University

Parsing languages that have free word-order is known to be difficult. Dependency grammar seems to be a better candidate than Phrase Structure Grammars for representing phenomena in such languages. Because of its rising popularity, dependency-based treebanks for many languages have been annotated or converted from their phrase structure correspondents. This has facilitated the training of statistical parsers for a wide range of languages (see CoNLL-X Shared Task, 2006).

The thesis relies on the two notions described above and deals with statistical dependency-based parsing for Bulgarian. Based on an automatically converted dependency treebank, MaltParser (Nivre 2005) has been trained for Bulgarian (Marinov and Nivre, 2005). The initial results were promising and even better results were achieved by using the entire treebank. In addition to testing and training MaltParser, I have experimented with DOP for dependency grammars, and will also construct a dependency parser using hand-written rules. This would allow me to compare the three parsing models for Bulgarian. Some of the parsing models could be used in a Machine-translation system (Bulgarian-Swedish) on which I am currently working.

Kristina Nilsson Computational Linguistics Group, Department of Linguistics, Stockholm University

Previous degree: 2003 M.A. in Computational Linguistics, Uppsala University

Thesis topic: Coreference Resolution for Information Access

Supervisor: Martin Volk, Professor of Computational Linguistics, Department of Linguistics, Stockholm University

Assistant supervisor: to be determined.

Coreference Resolution can be defined as the identification of identity between entities in text, both entities identified by named entity recognition and anaphoric references to those entities. Thus, by coreference resolution, information scattered across a text (or if across a collection of texts, by cross-document coreference resolution) can be associated with the entities to which it refers.

Some of the most important application areas for Coreference Resolution can be described as Information Access areas: Information Retrieval and Question Answering, Automatic Summarization, and Information Extraction. The purpose of this project is to investigate whether coreference resolution

can improve on the results of such Information Access systems.

Fredrik Olsson Department of Swedish language, Göteborg University

Previous degree: Ph. Lic.

Thesis topic: Information Access

Supervisor: Lars Borin, Professor, Department of Swedish, Göteborg University

Assistant supervisor: Björn Gambäck, PhD, SICS AB

Working title: A method for semi-automatically defining information extraction templates and acquiring slot-fillers

The thesis project concerns a method for aiding a user in recognizing and organizing his long term need for information wrt to a given domain. Applying the method will result in information extraction templates reflecting the the users information need, as well as initial programs for filling the templates.

The method relies on the assumption that people are generally better at recognizing what is important when they see it, than they are at á priori describing it. This assumption implies that the method under consideration should be interactive (semi-automatic), treating the user as an oracle in possession of all the knowledge needed about the domain. The problem challenged by this project is a multi-facetted one:

- It is an annotation problem; the user needs to mark up the relevant entities present in the texts. The user should only be made to mark the entities judged by the system as too uncertain or troublesome.
- It is a problem of sparse data; user intervention should be kept at a minimum to spare the user from repetitive and tedious tasks that might lead to low quality annotations.
- It is a problem of skewed data; even if given the opportunity, the user is more likely to annotate such structures that correspond to his need for information than those that do not.
- It is a problem of keeping turn-around times in the annotate-train-annotate cycle as short as possible.

In the thesis, I will outline the method addressing the above issues, as well as present the results from a number of empirical experiments focusing on what is believed to be the weakest parts of the method.

Magnus Rosell KTH CSC

Previous degree: M.Sc. Engineering Physics, 2002 Tekn. Lic. Computer Science 2005

Thesis topic: Text Clustering

Supervisor: Prof. Viggo Kann, KTH CSC

Assistant supervisor: Prof. Joakim Nivre, Växjö University and Uppsala University

Text Clustering or automatic grouping of texts is used to divide a set of texts into groups, so called clusters. The goal is to produce clusters such that texts in the same cluster are more similar in content

than texts from different clusters. Many sets of texts are partitioned manually as a matter of routine, in libraries and news papers (the sections of the paper) for instance. These partitions are static and are not always suitable. A new partition may shed new light on a set of texts. The result of a text clustering is dependent on the way the texts are represented. We have investigated how some aspects of the Swedish language affect the result. In connection to this we have also studied evaluation of text clustering. It is very hard to define what a good partition of a set of texts is. Hence it is also very hard to measure. We believe text clustering will become an important tool for exploration and analysis of open text answers in questionnaires. The information in free text answers is almost never used since it is too hard and expensive to do a manual analysis. By using automatic partitions it is easier to find connections and similarities among the answers. We cooperate with the Department of Medical Epidemiology and Biostatistics at the Karolinska Institutet (The Swedish Medical University) to investigate these possibilities.

Magnus Sahlgren, Department of linguistics, Stockholm University and SICS

Previous degree: M.A. in Philosophy and Computational Linguistics

Thesis topic: Distributional Semantics

Supervisor: Jussi Karlgren, PhD, The Human Computer Interaction and Language Engineering Laboratory, SICS

Assistant supervisor: Jens Allwood, Professor of General Linguistics, Department of Linguistics, Göteborg University

My research is focused on how semantic knowledge is acquired and represented in man and machine. In particular, I study the distributional approach to semantic knowledge acquisition, in which semantic information is extracted from cooccurrence statistics. My research concerns both the theoretical underpinnings of distributional semantics, and the practical implementations, and uses of the theories.

On the theoretical side, I am interested in the concept of meaning, and in how meaning resides in language and in the mind. On the practical side, I am interested in the use of vector space models as tools for acquiring and representing the distributional information. Well known models include Latent Semantic Analysis (LSA) / Latent Semantic Indexing (LSI) and Hyperspace Analogue to Language (HAL). As an alternative to these established models, we have at SICS developed a technique called Random Indexing, which is based on Pentti Kanerva's research on sparse distributed representations. The technique is both computationally advantageous - it is scalable, efficient and adaptable - and cognitively justified.

Susanne Schötz, Linguistics and Phonetics, Centre for Languages and Literature, Lund University

Previous degree: M.A. in Phonetics, 2001

Thesis topic: Perception, analysis and synthesis of Speaker Age

Supervisor: Per Lindblad, Associate Professor, Linguistics and Phonetics, Centre for Languages and Literature, Lund University

Assistant supervisor: Gösta Bruce, Professor, Linguistics and Phonetics, Centre for Languages and Literature, Lund University

Assistant supervisor: Rolf Carlson, Professor in Speech Technology, Department of Speech Music and Hearing, Royal Institute of Technology (KTH), Stockholm

My work is focused on finding acoustic and perceptual correlates to speaker age in order to build a model of speaker age based on these correlates. A large number of acoustic features automatically extracted from about 700 natural speakers of different ages are analysed carefully. A small system for analysis by datadriven formant synthesis has been built to systematically test potential cues to and models of speaker age. Listening tests and interviews are used to find out more about cues to speaker age and how they vary between listeners as well as between speakers. If successful, my methods may be applied to other paralinguistic properties of speech as well, including health state, attitudes and emotions. The goal of my research is to help improve the naturalness of synthetic speech.

Markus Saers Department of Linguistics and Philology, Uppsala University

Previous degree: MA in Computational Linguistics

Thesis topic: Machine Translation

Supervisor: Anna Sågvald Hein, professor, dep. Linguistics and Philology, Uppsala University

Assistant supervisor: Joakim Nivre, guest professor, dep. Linguistics and Philology, Uppsala University

The project aims at improving fully automatic translation based on empirical data. At the present, this is carried out by means of statistical machine translation (SMT) and Example-based machine translation (EBMT). The assumption is that these two approaches do not fully exploit the resources available (bilingual corpora and to some extent bilingual treebanks). Other viable resources include mono-lingual resources such as grammars (automatic morphological and/or syntactic analyzers) and pos-taggers, which can potentially increase translation quality.

This description is vague because the project is still vague.

Yvonne Samuelsson Department of Linguistics, Stockholm University

Previous degree: M.A. in Computational Linguistics

Thesis topic: Parallel Treebanks and Machine Translation

Supervisor: Martin Volk, Professor in Computational Linguistics, Department of Linguistics, Stockholm University

Assistant supervisor:

Gabriel Skantze Department for Speech Music and Hearing, KTH

Previous degree: M.A. Cognitive Science

Thesis topic: Error handling in spoken dialogue systems

Supervisor: Rolf Carlson, Professor, TMH, KTH

Assistant supervisor: Arne Jönsson, Professor, IDA, Linköping University

In spoken dialogue systems there are a number of different sources for errors and miscommunication. Many errors arise in the speech recognition process, due to disfluent speech, limited grammar coverage and noise in the environment. The thesis will present and discuss different methods for preventing, detecting and recovering from such errors. This includes the issues of how to select, generate and model natural and efficient grounding and feedback actions (including prosodic realisation), how

to interpret speech recognition results robustly, how to use machine learning for detecting word-level errors, and what to do after complete non-understandings. These methods have been implemented in the dialogue system Higgins, a system for pedestrian navigation, developed as part of the thesis work. The implementation of a complete system facilitates the exploration of how errors are best detected and handled in the different parts of the system. A complete system also facilitates user studies, where the users' behaviours in different error situations are studied.

Mustapha Skhiri, Department of Computer and Information Science, Linköping University

Previous degree: M.Sc. Computer Science

Thesis topic: C Computational Models of facial and head movements in interactive dialogue systems

Supervisor: Bertil Lyberg, Adjunct Professor in Speech Technology, Reader (docent) in Phonetics

Assistant supervisor: Lars Ahrenberg, Professor of Computational Linguistics, Department of Computer and Information Science, Linköping University

Dialogue is an interactive communication of information mainly based on speech, but also visual information such as gesture, facial expression and head movement clearly makes the conversations much smoother and more natural. In my research I will first do an examination of the face and head movements in turntaking situations and also study how these movements are related to the speech signal in human-human and human-computer situation. Then I will make a computational model and implement these movements in a "talking head" and study the perceptual relevance of this in a dialog system that includes visual information.

Håkan Sundblad Department of computer and information science

Previous degree: Masters in cognitive science

Thesis topic: Question answering

Supervisor: Arne Jönsson, Professor, Dept. of computer and information science, Linöping university

Assistant supervisor: Magnus Merkel, Assistant professor, Dept. of computer and information science, Linöping university

The thesis investigates the question classification process in question answering systems. Focus is on taxonomies and methods for classification.

Per Weijnitz, Department of Linguistics, Uppsala University

Previous degree: M.A. of Philosophy in Language Engineering

Thesis topic: Hybrid methods in machine translation

Supervisor: Anna Sågvall Hein, Professor of Computational Linguistics, Department of Linguistics, Uppsala University

Assistant supervisor: to be determined

Grammar based systems will typically fail to process data that are not covered by its grammars. Rule based machine translation systems are no exception. An MT system could recover from such failures by processing ungrammatical (in the technical sense) input data using complementary, not necessarily grammar based, resources and methods. I am investigating how this could be accomplished and what

these resources and methods could be.

Marcus Uneson, Linguistics and Phonetics, Centre for Languages and Literature, Lund University
Previous degree: B.A. in Musicology, M.A. in Phonetics
Thesis topic: Data-driven induction of phonological rules
Supervisor: Sven Strömquist, Professor, Linguistics and Phonetics, Centre for Languages and Literature, Lund University
Assistant supervisor: Pierre Nugues, Associate Professor, Department of Computer Science, Lund University

My areas of interest fall within computational phonology, dealing with questions such as: Given a pair of words represented as strings in some phonetic (or at least alphabetic) transcription, how different are they? Given two languages represented as a set of such pairs, how different are they? How can one transform one member of a pair into the other? Borrowing methods from computational biology, I'd like to explore ways of inducing such rules automatically.

Automatically learned phonological rules may be of interest for instance for pronunciation modelling and historical linguistics. Similarly, phonetically or phonologically motivated string comparison measures have applications in very diverse areas, including dialectometry, speech technology, historical linguistics, and identification of confusable drug names.

Jessica Villing Department of Linguistics, Göteborg university
Previous degree: M.A. Computational Linguistics 2005
Thesis topic: Multimodal In-Vehicle Dialogue Systems
Supervisor: Robin Cooper, professor, Department of Linguistics, Göteborg university
Assistant supervisor: Staffan Larsson, dr, Department of Linguistics, Göteborg university

The aim of the thesis is to investigate methods for improving the human-machine interaction in in-vehicle multimodal dialogue systems. The in-vehicle environment makes heavy demands on the usability of the dialogue system. Unlike most other environments where a dialogue system is used (for example phone applications for ticket booking) the user has to pay full attention to something other than the system, namely the traffic situation. The interaction with the dialogue system is a secondary task, which must be taken under consideration. My focus will be on the cognitive load of the driver. I want to investigate methods for measuring the cognitive load, and try to distinguish between different types of cognitive load to be able to adjust the dialogue depending on what kind of cognitive load the driver is experiencing. If the driver is stressed due to a heavy traffic situations the dialogue system should be able to pause the dialogue until the driver is ready to resume. On the other hand, if the driver is experiencing high cognitive load due to a difficult task, maybe the system should change dialogue strategy to make the task simpler.

Kenneth Wilhelmsson Institutionen för lingvistik, Göteborgs universitet
Previous degree: MA in computational linguistics

Thesis topic: Heuristic Analysis of Free Text Using Diderichsen's Sentence Schema

Supervisor: Robin Cooper, Professor, Institutionen för lingvistik, Göteborgs universitet

Assistant supervisor: Dimitrios Kokkinakis, Research Assistant, Section of Lexicology and NLP (Språkdata), Göteborgs universitet

The thesis aims at evaluating the possibility of using a schema-based approach as an NLP component for Swedish syntax analysis. The methodology includes heuristic linguistic rules that are triggered and applied depending on the analysis up to each point. The first and most crucial task is correct identification of primary finite verbs (*'huvudsatsernas finita verb'*). This is followed by identification of other bounded (non-recursive) constituents on the primary level – non-finite verbs, particles, reflexive objects and (most) sentence adverbs. Identification of primary conjunctions (between main clauses or primary finite verb phrases), final period or similar, and what is in *SAG* (1999) referred to as *förfält* is also identified here.

The remaining constituents of the level of analysis that is called *primär satslösning* are in most cases unbounded (recursive): primary subjects, objects/predicatives and other adverbials. A new kind of chunking: *rank-based chunking* is examined for the task of separating adjacent recursive constituents (mainly NPs and PPs).

For questions such as whether a PP is an attribute of an NP, or, an adverbial, two handcrafted valency lexica for Swedish are employed. These come from *Nationalencyklopediens ordbok* and *Lexin: Svenska ord*. They include approximately 7000 verbs, 7000 nouns and 1800 adjectives when preprocessed and made machine-readable together with a dedicated base form look-up functionality.

The approach is thus depending on Diderichsen's sentence schema, syntactic valency information from two lexica, a number of collections of particular word types (e.g. sentence adverbs, auxiliary verbs and personal names), and a large set of heuristic rules that often are directly motivated by traditional grammar. – There is, however, no generative grammar component (such as a regular grammar or a CFG), although some processes such as the chunking is somewhat reminiscent of this. The lack of a separate grammar component makes the method applicable for programming in most imperative programming languages. It represents a parsing method similar to manual traditional functional analysis of Swedish.

For the development of this implemented method, *Stockholm Umeå Corpus* has been used to provide text with the best possible correctness in tagging and to represent current published Swedish text. One or two part-of-speech taggers are also built.

Many different NLP applications seem directly dependent of a working functional preprocessing like this, e.g. automatic paraphrasing systems and question answering systems.

Pontus Wärnestål Institution of Computer and Information Science, Linköping University

Previous degree: Ph. Lic.

Thesis topic: Dialogue Management in Conversational Recommender Systems

Supervisor: Arne Jönsson, Professor, Inst. Of Computer Science, Linköping University

Assistant supervisor: Lars Degerstedt, Ph. D. Inst. Of Computer Science, Linköping University

The overall objective for the thesis is to develop mechanisms for modeling high-quality user preferences to be used in the process of reducing complexity in search for information, and to support domain exploration to help users develop and enhance their own understanding of the domain and their domain preferences. This in turn can improve the preference elicitation and thereby the recommendation prediction accuracy. The work is based on previous projects (see list of publications) and can be extended by extending the conversational model (dialogue strategy or policy) and the user model (especially the user preference model).

The goal is to create a theory/model of user-system recommendation dialogue interaction that eliminates complexity, lets the user focus on his/her tasks, and supports the user in learning about the domain and cultivate her preferences in a personalized way. This is achieved by focusing on the conversational model (1), the user model (2), and to a lesser extent the domain model. The search model is not the main focus of this work but may be affected by 1-2. The theory/model is implemented in the music recommender system AcornSong.

Read more: <http://www.ida.liu.se/~ponjo/Research/>

Preben Wik Dept. of Speech, Music, and Hearing, CSC, KTH

Previous degree: Candidatus Philologiae in Language, Logic and Information, 2002, University of Oslo, Norway

Thesis topic: The Virtual Language Tutor – Design and Implementation Issues

Supervisor: Björn Granström, Professor, Department of Speech Music and Hearing, Royal Institute of Technology (KTH), Stockholm

Assistant supervisor: Olov Engwall, Department of Speech Music and Hearing, Royal Institute of Technology (KTH), Stockholm

The goal of my PhD is to design and implement the framework for a new type of language learning, using a virtual language tutor as a supplement to the traditional teacher. The aim is to make a universal language tutor, with placeholders for language specific modules, and user specific applications. The process will be twofold. Top-down, looking at the overall architecture of the system, and bottom-up doing an incremental build of software components needed for the implementation of such a system. The implementation will – at least initially, focus on Swedish for adult immigrants. The system will be put in use as part of a language course that is being developed at KTH. A database with speech-data from people with different language background will be created as a result of this. An iterative process of evaluation, redesign, and implementation of the system will be performed, based on the results of the course, evaluation of the speech- database, and feedback from the students using the tutor.

Gustav Öquist, Department of Linguistics, Uppsala University

Previous degree: M.A. in Computational Linguistics, Uppsala University (planned 2005)

Thesis topic: Improving Readability on Small Screens

Supervisor: Anna Sångvall Hein, Professor of Computational Linguistics, Department of Linguistics, Uppsala University

Assistant supervisor: Jan Ygge, Professor of Ophthalmology, Department of Clinical Neuroscience, Karolinska Institutet

Assistant supervisor: Mikael Goldstein, Associate Professor of Psychology, Migoli HB

The aim with the thesis is to bring forward efficient and usable text presentation formats for small screens. To learn about this, we have performed usability evaluations on mobile devices where we have compared traditional text presentation to different variations of a dynamic text presentation format called Rapid Serial Visual Presentation (RSVP). In the RSVP format, the text is presented as chunks of words in rapid succession at a single visual location, a technique that often is claimed to reduce eye movements and increase reading speed. Our findings confirm that the RSVP format is efficient on a mobile device, but we also found it to impose a greater cognitive load on the reader. However, by using linguistic adaptation of the text presentation pace, we were able to reduce task load for most factors. In order to improve the adaptation further, and get more data on what actually happens while reading, we developed a system for measuring eye movements while reading on mobile devices. The results from the two eye movement studies performed so far show that the RSVP format may eliminate eye movements and mostly does reduce them. However, the reduction does not seem to reduce cognitive load. In fact, it rather seems to increase cognitive load. These empirical findings disprove the theoretical assumption of the RSVP format, that suppressing eye movement reduces cognitive load. Instead, we propose that dynamic text presentation formats should try to stimulate eye movements similar to how we are accustomed to read on paper in order to improve readability on mobile devices.

Lilja Øvreid, Spåkdata (NLP-unit), Department of Swedish, Göteborg University

Previous degree: B.A. in Language, Logic and Information, B.A. in Language, Logic and Information (major), Linguistics (minor) and English (minor)

M.A. in Language, Logic and Information, Oslo University, Norway, 2003

Thesis topic: Disambiguating animacy: Acquisition of animacy information for syntactic disambiguation
Supervisor: Elisabet Engdahl, Department of Swedish, Göteborg Univ. Department of Swedish Language, Göteborg University

Assistant supervisor: Joakim Nivre, Uppsala and Växjö Universities

Animacy is a referential property of nouns which has been claimed to figure as an influencing factor in a range of grammatical phenomena in various languages. It is closely correlated with other linguistic dimensions such as agentivity, topicality etc. Even so, little work has been done on automatically acquiring such information.

The purpose of this thesis is to investigate automatic acquisition of animacy information and evaluate the usefulness of this information first and foremost in syntactic parsing of Scandinavian. For acquisition of animacy information, a range of experiments varying the specification of the learning task and algorithm as well as the level of analysis of input data will be conducted. The focus will be on corpus-based methods. In particular, lexical acquisition both at a type (lemma) and token level will be performed in an effort to combine information found at the syntax-semantics interface with more contextually based co-occurrence measures.

Extrinsic evaluation of the usefulness of animacy information will mainly focus on its use in syntactic disambiguation in Scandinavian, however, other possibilities might also be investigated if there is time (coreference resolution, semantic role labeling etc.).

C Licentiate theses

Johansson, Richard (2006) *Natural Language Processing Methods for Automatic Illustration of Text*, Licentiate thesis, Department of Computer Science, Lund University.

Svanfeldt, Gunilla (2006) *Expressiveness in virtual talking faces*, Licentiate thesis, Department of Speech, Music and Hearing, KTH.

D PhD theses (with abstracts)

Gorrell, Genevieve (2006) *Generalized Hebbian Algorithm for Dimensionality Reduction in Natural Language Processing*, PhD Thesis, Linköping University, Department of Computer and Information Science.

Abstract

The current surge of interest in search and comparison tasks in natural language processing has brought with it a focus on vector space approaches and vector space dimensionality reduction techniques. Presenting data as points in hyperspace provides opportunities to use a variety of well-developed tools pertinent to this representation. Dimensionality reduction allows data to be compressed and generalised. Eigen decomposition and related algorithms are one category of approaches to dimensionality reduction, providing a principled way to reduce data dimensionality that has time and again shown itself capable of enabling access to powerful generalisations in the data. Issues with the approach, however, include computational complexity and limitations on the size of dataset that can reasonably be processed in this way. Large datasets are a persistent feature of natural language processing tasks. This thesis focuses on two main questions. Firstly, in what ways can eigen decomposition and related techniques be extended to larger datasets? Secondly, this having been achieved, of what value is the resulting approach to information retrieval and to statistical language modelling at the n-gram level? The applicability of eigen decomposition is shown to be extendable through the use of an extant algorithm; the Generalized Hebbian Algorithm (GHA), and the novel extension of this algorithm to paired data; the Asymmetric Generalized Hebbian Algorithm (AGHA). Several original extensions to these algorithms are also presented, improving their applicability in various domains. The applicability of GHA to Latent Semantic Analysis-style tasks is investigated. Finally, AGHA is used to investigate the value of singular value decomposition, an eigen decomposition variant, to n-gram language modelling. A sizeable perplexity reduction is demonstrated.

Grönqvist, Leif (2006) *Exploring Latent Semantic Vector Models Enriched With N-grams*, PhD Thesis, Växjö University, School of Mathematics and Systems Engineering.

Abstract

This thesis deals with a kind of vector space model called *Latent Semantic Vector Model*, or LSVM, calculated by the technique *Latent Semantic Indexing*. An LSVM can be used for many things, but I have mainly looked at one direct application: document retrieval. What we can gain from an LSVM is the possibility of searching for content rather than specific keywords. Using an LSVM in a document retrieval system has been shown to improve the quality of the returned document lists, which makes it easier for the user to find the information he or she wants. The problem attacked in this thesis is that an LSVM in the normal case contains just single words, while the terms one searches for in many cases are multi-word expressions.

LSVMs have been trained with various parameter settings for training data, vocabulary, matrix size, context size, and last but not least, different ways to include multi-word expressions directly into the models. The aim has been to determine how the performance of an LSVM

changes when we go from a word-based model to a model containing both words and multi-word expressions. To be able to measure the changes, two evaluation methods have been used: synonym tests and document retrieval. Synonym testing has been performed for Swedish and document retrieval for both Swedish and English. The results are improved when multi-word expressions are added for the synonym test task, but change for the worse for document retrieval. For English, the latter change is not significant.

This work has also resulted in two new resources, well suited for evaluation of various models: the evaluation set SweHP560, containing 560 Swedish synonym test queries from *Högskoleprovet*, and the new metrics RankEff and WRS for document retrieval evaluation, which handle the problem of an incomplete gold standard in a better way than existing metrics like MAP and bpref.

Jande, Per-Anders (2006) *Modelling Phone-Level Pronunciation in Discourse Context*, PhD thesis, TMH, Speech, Music and Hearing, KTH.

Abstract

Analytic knowledge about the systematic variation in a language has an important place in the description of the language. Such knowledge is interesting e.g. in the language teaching domain, as a background for various types of linguistic studies, and in the development of more dynamic speech technology applications. In previous studies, the effects of single variables or relatively small groups of related variables on the pronunciation of words have been studied separately. The work described in this thesis takes a holistic perspective on pronunciation variation and focuses on a method for creating general descriptions of phone-level pronunciation in discourse context. The discourse context is defined by a large set of linguistic attributes ranging from high-level variables such as speaking style, down to the articulatory feature level. Models of phone-level pronunciation in the context of a discourse have been created for the central standard Swedish language variety. The models are represented in the form of decision trees, which are readable for both machines and humans. A data-driven approach was taken for the pronunciation modelling task, and the work involved the annotation of recorded speech with linguistic and related information. The decision tree models were induced from the annotation. An important part of the work on pronunciation modelling was also the development of a pronunciation lexicon for Swedish. In a cross-validation experiment, several sets of pronunciation models were created with access to different parts of the attributes in the annotation. The prediction accuracy of pronunciation models could be improved by 42.2% by making information from layers above the phoneme level accessible during model training. Optimal models were obtained when attributes from all layers of annotation were used. The goal for the models was to produce pronunciation representations representative for the language variety and not necessarily for the individual speakers, on whose speech the models were trained. In the cross-validation experiments, model-produced phone strings were compared to key phonetic transcripts of actual speech, and the phone error rate was defined as the share of discrepancies between the respective phone strings. Thus, the phone error rate is the sum of actual errors and discrepancies resulting from desired adaptations from a speaker-specific pronunciation to a pronunciation reflecting general traits of the language variety. The optimal models gave an average phone error rate of 8.2%.

Sahlgren, M. (2006) *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*, Ph.D. dissertation, Department of Linguistics, Stockholm University.

Abstract

The word-space model is a computational model of word meaning that utilizes the distributional patterns of words collected over large text data to represent semantic similarity between words in terms of spatial proximity. The model has been used for over a decade, and has demonstrated its mettle in numerous experiments and applications. It is now on the verge of moving from research environments to practical deployment in commercial systems. Although extensively used and intensively investigated, our theoretical understanding of the word-space model remains unclear. The question this dissertation attempts to answer is: what kind of semantic information does the word-space model acquire and represent?

The answer is derived through an identification and discussion of the three main theoretical cornerstones of the word-space model: the geometric metaphor of meaning, the distributional methodology, and the structuralist meaning theory. It is argued that the word-space model acquires and represents two different types of relations between words - syntagmatic or paradigmatic relations - depending on how the distributional patterns of words are used to accumulate word spaces. The difference between syntagmatic and paradigmatic word spaces is empirically demonstrated in a number of experiments, including comparisons with thesaurus entries, association norms, a synonym test, a list of antonym pairs, and a record of part-of-speech assignments.

Schötz, Susanne (2006) *Perception, Analysis and Synthesis of Speaker Age*, Ph.D. thesis, Department of Linguistics and Phonetics, Lund University.

Abstract

Speaker age is an important paralinguistic feature in speech which has to be considered in the study of phonetic variation. Knowledge about this feature may be used to improve speech technology applications, e.g. automatic speech recognition and speech synthesis. The present thesis describes six studies of several phonetic aspects of age-related variation in speech.

As the speech production mechanism changes from young adulthood to old age, speech is affected in numerous ways. Human perception of speaker age is based on cues such as pitch, speech rate and voice quality, and is fairly accurate. However, it is still unclear which cues are the most important ones. The first study included in this thesis investigated the role of F_0 and speech rate (word duration) in age perception. It was found that while these cues may be less important than spectral ones (e.g. formant frequencies), they still correlate with chronological as well as perceived age. In the second study, two stimulus types of various lengths were compared. Results indicated that while longer stimulus duration (regardless of speech type) seems to improve the age estimation of females, spontaneous speech (regardless of duration) appears to contain more important cues for perception of male speaker age.

In the next two studies, several automatic estimators of speaker age were built, none of which reached the same accuracy as humans. Important features in machine perception of age were

also investigated. It was found that prosodic features seem to be more important in the estimation of female age, while spectral features (e.g. F_2) appear to be more important for male age.

Although several acoustic correlates of speaker age are known, their relative importance has not yet been established. The next study analysed 161 features, automatically extracted from segments in six words produced by 527 speakers. Normalised means were used to ensure that the features could be compared directly. The most important acoustic correlates of speaker age were identified to be speech rate (segment duration) and intensity range. However, F0 and some spectral measures (e.g. F_1 and F_2) may also, if used in combination with other features, be important correlates of age.

Synthetic speech may sound more natural if speaker age is included as a parameter. The final study developed a research tool which used data-driven formant synthesis and age-weighted linear interpolation to simulate an age between the ages of any two of four female differently aged reference speakers. Evaluation of the tool showed that speaker age may in fact be simulated using formant synthesis. The tool will be used in further studies of analysis by synthesis of speaker age.

Öquist, G. (2006) *Evaluating Readability on Mobile Devices*, Ph.D. thesis, Department of Linguistics and Philology, Uppsala University, Acta Universitatis Upsaliensis. Studia Linguistica Upsalien-sia 4.

Abstract

The thesis presents findings from five readability studies performed on mobile devices. The dynamic Rapid Serial Visual Presentation (RSVP) format has been enhanced with regard to linguistic adaptation and segmentation as well as eye movement modeling. The novel formats have been evaluated against other common presentation formats including Paging, Scrolling, and Leading in latin-square balanced repeated-measurement studies with 12-16 subjects. Apart from monitoring Reading speed, Comprehension, and Task load (NASA-TLX), Eye movement tracking has been used to learn more about how the presentation formats affects reading.

The Page format generally offered best readability. Reading on a mobile phone decreased reading speed by 10% compared to reading on a Personal Digital Assistant (PDA), an interesting finding given that the display area of the mobile phone was 50% smaller. Scrolling, the most commonly used presentation format on mobile devices today, proved inferior to both Paging and RSVP. Leading, the most widely known dynamic format, caused very unnatural eye movements for reading. This seems to have increased task load, but not affected reading speed to a similar extent. The RSVP format displaying one word at time was found to reduce eye movements significantly, but contrary to common claims, this resulted in decreased reading speed and increased task load. In the last study, Predictive Text Presentation (PTP) was introduced. The format is based on RSVP and combines linguistic chunking and adaptation with eye movement modeling to achieve a reading experience that can rival traditional text presentation.

It is explained why readability on mobile devices is important, how it may be evaluated in an efficient and yet reliable manner, and PTP is pinpointed as the format with greatest potential for improvement. The methodology used in the evaluations and the shortcomings of the studies are discussed. Finally, a hyper-graeco-latin-square experimental design is proposed for future evaluations.

E Conference publications by GSLT's graduate students

- Ahlgren, Per and Leif Grönqvist (2006) Retrieval evaluation with incomplete relevance data: A comparative study of three measures (poster abstract), in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 872–873.
- Ahlgren, Per and Leif Grönqvist (2006) Measuring retrieval effectiveness with incomplete relevance data, in *Current Research in Information Sciences and Technologies: Multidisciplinary approaches to global information systems (InSciT2006 proceedings)*, Guerrero-Bote, Vicente, pp. 74–78.
- Alemu Argaw, Atelach and Lars Asker (2006) Amharic-English Information Retrieval, in *Working Notes of CLEF 2006, Alicante, Spain. September 2006*.
- Alemu Argaw, Atelach and Lars Asker (2006) Increased Retrieval Performance using Word Sense Discrimination, in *Proceedings of The 13th Conference on Natural Language Processing (TALN 2006)*, Leuven, Belgium. April 2006.
- Alemu Argaw, Atelach, Lars Asker, Rickard Cöster, Jussi Karlgren and Magnus Sahlgren (2006) Dictionary-based Amharic-French Information Retrieval Lecture Notes in Computer Science: Accessing Multilingual Information Repositories, Springer Berlin/Heidelberg.
- Asker, Lars, Atelach Alemu Argaw, Björn Gambäck, and Magnus Sahlgren (2006) Applying Machine Learning to Amharic Text Classification, in *Proceedings of The 5th World Congress of African Linguistics (WOCAL)*, Addis Abeba, Ethiopia. August 2006.
- Berglund, Anders, Richard Johansson, and Pierre Nugues (2006) Extraction of Temporal Information from Texts in Swedish, in *Proceedings of LREC-2006, The fifth international conference on Language Resources and Evaluation*.
- Berglund, Anders, Richard Johansson, and Pierre Nugues (2006) A Machine Learning Approach to Extract Temporal Information from Texts in Swedish and Generate Animated 3D Scenes, in *Proceedings of EACL-2006, 11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Carlson, R., J. Edlund, D. House, M. Heldner, A. Hjalmarsson, and G. Skantze (2006). Towards human-like behaviour in spoken dialog systems, in *Proceedings of Swedish Language Technology Conference (SLTC 2006)*. Gothenburg, Sweden.
- Chanev, A., K., Simov, P. Osenova, and S. Marinov (2006) Dependency conversion and parsing of the BulTreeBank. *Proceedings of LREC Workshop, Genoa, Italy, May 2006*.
- Forsberg, M., H. Hammarström and A. Ranta (2006) Lexicon Extraction from Raw Text Data, in *Advances in Natural Language Processing: Proceedings of the 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006*, LNCS 4139, ed. by T. Salakoski, F. Ginter, S. Pyysalo and T. Pahikkala, Springer, pp. 488–499.
- Goldstein, M., Öquist, G., and Lewald, I. (2006) Evaluation of PreCodia, a Computerized Reading Aid for Readers Suffering from Dyslexia, in *Proceedings of Human Factors in Telecommunication 2006 (Sophia-Antipolis, France)*, IGI Group, Brighton, MA., pp. 127–134.

- Green, Anders and Helge Hüttenrauch (2006) Making a Case for Spatial Prompting in Human-Robot Communication, in *Proceedings of Multimodal Corpora: From Multimodal Behaviour theories to usable models, workshop at the Fifth international conference on Language Resources and Evaluation, LREC2006, Genova, May 22-27*.
- Green, Anders, Helge Hüttenrauch and Elin Anna Topp (2006) Measuring Up as an Intelligent Robot – On the Use of High-Fidelity Simulations for Human-Robot Interaction Research, in *Proceedings of PerMIS '06, Performance Metrics for Intelligent Systems, The 2006 Performance Metrics for Intelligent Systems Workshop, Gaithersburg, MD, August 21-23, 2006*.
- Green, Anders, Helge Hüttenrauch, Elin Anna Topp and Kerstin Severinson Eklundh (2006) Developing a Contextualized Multimodal Corpus for Human-Robot Interaction, in *Proceedings of Fifth international conference on Language Resources and Evaluation, LREC2006, Genova, May 22-27*.
- Green, Anders, Britta Wrede, Kerstin Severinson Eklund and Shuyin Li (2006) Integrating Miscommunication Analysis in Natural Language Interface Design for a Service Robot, in *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS2006, October 9 - 15, 2006, Beijing, China*, pp. 4678–4683.
- Gunnarsson, Mikael (2006) Genre Identification, in *Proceedings of The Document Academy Meeting, Berkeley, CA, October 13-15, 2006*.
- Hammarström, H. (2006) Poor Man's Stemming: Unsupervised Recognition of Same-Stem Words, in *Information Retrieval Technology: Proceedings of the Third Asia Information retrieval Symposium, AIRS 2006, Singapore, October 2006*, LNCS 4182, ed. by Hwee Tou Ng, Mun-Kew Leong, Min-Yen Kan and Donghong Ji, Springer, pp. 323–337.
- Hammarström, H. (2006) A Naive Theory of Morphology and an Algorithm for Extraction, in *Proceedings of SIGPHON 2006: Eighth Meeting of the the ACL Special Interest Group on Computational Phonology, 8 June 2006, New York City, USA*, ed. by R. Wicentowski and G. Kondrakm, Association for Computational Linguistics, pp. 79–88.
- Hammarström, H. (2006) A New Algorithm for Unsupervised Induction of Concatenative Morphology, in *Finite State Methods in Natural Language Processing: 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005: Revised Papers*, LNCS 4002, Springer, pp. 288–289.
- Hjelm, Hans (2006) Extraction of Cross Language Term Correspondences, in *Proceedings of LREC 2006*.
- Hjelm, Hans and Christoph Schwarz (2006) LiSa - Morphological Analysis for Information Retrieval, in *Proceedings of the 15th NODALIDA conference, Joensuu 2005* ed. by S. Werner, Ling@JoY : University of Joensuu electronic publications in linguistics and language technology 1.
- Jande, Per-Anders (2006) Integrating Linguistic Information from Multiple Sources in Lexicon Development and Spoken Language Annotation, in *Proceedings of the LREC workshop on merging and layering linguistic information*, pp. 1–8.

- Jande, Per-Anders (2006) Modelling Pronunciation in Discourse Context, in *Proceedings of Fonetik*, ed. by Gilbert Ambrazaitis and Susanne Schötz, pp. 69–72.
- Johansson, Richard and Pierre Nugues (2006) A FrameNet-based Semantic Role Labeler for Swedish, in *Proceedings of Coling/ACL-2006*.
- Johansson, Richard and Pierre Nugues (2006) Construction of a FrameNet Labeler for Swedish Text, in *Proceedings of LREC-2006, The fifth international conference on Language Resources and Evaluation*.
- Johansson, Richard and Pierre Nugues (2006) Investigating Multilingual Dependency Parsing in *Proceedings of the Tenth Conference on Computational Natural Language Learning (CONLL-X)*.
- Johansson, Richard and Pierre Nugues (2006) Automatic Annotation for All Semantic Layers in FrameNet, in *Proceedings of EACL-2006, 11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Jonson, Rebecca (2006) Generating statistical language models from interpretation grammars in dialogue systems, in *Proceedings of 11th Conference of the European Association of Computational Linguistics (EACL). Trento, Italy*.
- Jonson, Rebecca (2006) Dialogue Context-Based Re-ranking of ASR Hypotheses, in *Proceedings of Workshop on Spoken Language Technology (SLT), Aruba, Dec 10-13 (2006)*.
- Kronlid, F. (2006) Turn Taking for Artificial Conversational Agents. (Eds.) *Cooperative Information Agents X*, LNAI 4149, ed. by M. Klusch, T. Rovatsos and T.R. Payne, Springer-Verlag, pp. 81–95.
- Lindh, J. (2006) A Case Study of /r/ in the Västgöta Dialect, in *Papers from FONETIK 2006, Working Papers, 52*, Department of Linguistics and Phonetics, Lund University, pp. 85–88.
- Lindh, J. (2006) Preliminary Descriptive F0-statistics for Young Male Speakers, in *Papers from FONETIK 2006, Working Papers, 52*, Department of Linguistics and Phonetics, Lund University, pp. 89–92.
- Nilsson, Kristina, Hans Hjelm and Henrik Oxhammar (2006) SUIs – cross-language ontology-driven information retrieval in a restricted domain, in *Proceedings of the 15th NODALIDA conference, Joensuu 2005* ed. by S. Werner, Ling@JoY : University of Joensuu electronic publications in linguistics and language technology 1.
- Nilsson, Kristina and Aisha Malmgren (2006) Towards automatic recognition of product names: an exploratory study of brand names in economic texts, in *Proceedings of the 15th NODALIDA conference, Joensuu 2005* ed. by S. Werner, Ling@JoY : University of Joensuu electronic publications in linguistics and language technology 1.
- Nivre, J., J. Hall, J. Nilsson, G. Eryigit, and S. Marinov (2006) Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines, in *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL). NY, June, 2006*.
- Olsson, F., J. Karlgren, P. Hansen, M. Svensson, R. Cöster, and M. Sahlgren (2006) Consensus and Opinions; Quality and Churn, in *Proceedings of the Workshop on Computational Approaches*

- to Analyzing Weblogs, AAAI Spring Symposium 2006, March 27-29, Stanford University, Palo Alto, California, USA.*
- Sahlgren, M. (2006) Towards pertinent evaluation methodologies for word-space models, in *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006, May 24-26, Genoa, Italy.*
- Samuelsson, Yvonne and Martin Volk (2006) Phrase Alignment in Parallel Treebanks, in *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories, TLT 2006, Prague, Czech Republic.*
- Skantze, G., J. Edlund, and R. Carlson (2006) Talking with Higgins: Research challenges in a spoken dialogue system, in *Proceedings of Perception and Interactive Technologies* ed. by E. André, L. Dybkjaer, W. Minker, H. Neumann, and M. Weber, Springer, pp. 193–196
- Skantze, G., D. House, and J. Edlund, (2006) Grounding and prosody in dialog, in *Working Papers 52: Proceedings of Fonetik 2006*, Lund, Sweden: Lund University, Centre for Languages and Literature, Dept. of Linguistics and Phonetics, pp. 117–120.
- Skantze, G., D. House, and J. Edlund (2006). User responses to prosodic variation on fragmentary grounding utterances in dialogue, in *Proceedings of Interspeech 2006 ICSLP. Pittsburgh PA, USA.*
- Villing, J. and S. Larsson (2006) Dico - A Multimodal In-vehicle Dialogue System, in *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial'06, Sem-Dial 10), Potsdam, Germany*, pp 187–188.
- Volk, Martin, Sofia Gustafson-Capková, Joakim Lundborg, Torsten Marek, Yvonne Samuelsson and Frida Tidström (2006) XML-based Phrase Alignment in Parallel Treebanks, in *Proc. of EACL Workshop on Multi-dimensional Markup in Natural Language Processing.*
- Wallers, Å., J. Edlund, and G. Skantze (2006) The effects of prosodic features on the interpretation of synthesised backchannels, in *Proceedings of Perception and Interactive Technologies* ed. by E. André, L. Dybkjaer, W. Minker, H. Neumann, and M. Weber, Springer, pp. 183–187.
- Öquist, G. (2006) Multimodal Interaction with Mobile Devices: Outline of a Semiotic Framework for Theory and Practice, in *Proceedings of Wireless Networks and Systems 2006 (Setubal, Portugal)*, INSTICC Press, Setubal, pp. 276–283.
- Øvrelid, Lilja (2006) Towards robust animacy classification using morphosyntactic distributional features, in *Proceedings of EACL 2006 Student Research Workshop, Trento, Italy, 3.-7. April, 2006.*

F Presentations not associated with a publication

- Borin, L. and M. Forsberg. Mannen är faderns mormor: Svenskt associationslexikon i utökad elektronisk form, presented at SLING 2006.
- Brusk, J. and M. Eladhari (2006) Playing the Character. Seminar on Playing Roles, March 30 – 31 2006, Tampere University.
- Eklund, J. (2006). Ergebnisse des Projektes LIVA. Presentation given at the 29. Österreichischer Bibliothekartag 2006 in Bregenz, Austria.
- Forsbom, Eva (2006) A Swedish Base Vocabulary Pool, Swedish Language Technology Conference, Göteborg.
- Hjelm, Hans (2006) Ontology Learning from Parallel Texts – presentation held at TNC (Terminologiceentrum).
- Jonson, Rebecca (2006) Context-based Re-ranking and Confidence Level Classification of N-Best Hypotheses, SLTC, Göteborg.
- Lindh, J. (2006) Preliminary F0 Statistics and Forensic Phonetics, IAFPA 2006, Gothenburg.
- Lindh, J. (2006) F0 Statistics, Robustness and Measures – Implications for Forensic Speaker Identification. SLTC-2006, Gothenburg.
- Samuelsson, Yvonne (2006) Experiences from building a German-English-Swedish Parallel Treebank, International Symposium on Parallel Treebanks, Stockholm, Sweden.
- Wilhelmsson, K. (2006) Identification of Primary Finite Verbs in Swedish Text, poster presentation at SLTC 2006.
- Øvrelid, Lilja (2006) Exploring the distribution of animacy: experiments on Norwegian. Presentation at Quantitative Investigations in Theoretical Linguistics 2 (QITL-2), Osnabrück, June 1st-2nd 2006.

G Other publications

Hammarström, H. (2006) Review of *Lithuanian Romani*, Languages of the World/Materials 452, by Anton Tenser, Lincom, 2005, LINGUIST LIST 17.1511.

Hammarström, H. (2006) Review of *Some Aspects of the Grammar of Zina Kotoko*, LINCOM Studies in African Linguistics 54, ed. by Bodil Kappel, David Odden, and Anders Holmberg, Lincom, 2002, LINGUIST LIST 17.818.

Sahlgren, M. (2006) Concept-Based Text Representations for Categorization Problems. ERCIM News, No.64, January 2006.

H Industry contacts by GSLT graduate students

Leif Grönqvist Began employment full time at Spotfire AB three months before defending thesis.

Hans Hjelm Worked on implementing some linguistic analysis (both statistical and rule based) in an intranet search tool for IntraFind Software AG (www.intrafind.de), a company based in Munich, Germany.

Richard Johansson Consultancy to Artificialife, Montreal, on design and implementation of NLP tools and knowledge representation.

Rebecca Jonson Contacts with Speech Technology department at Telefonica I+D, Spain.

Jessica Villing Work in the Vinnova funded DICO project, a cooperation between GU, Volvo Technology, Volvo Trucks, Volvo Cars, TeliaSonera and KTH. The project aims to demonstrate how state-of-the-art spoken language technology can enable access to in-vehicle features.

Gustav Öquist is co-founder of Precodia, a subsidiary of Stockholms DyslexiCentrum AB. The company is developing an interactive reading aid tool for dyslectics which is based on language technology.

Lilja Øvreliid Employed part-time at Meltwater News (www.meltwater.com), a media monitoring company, to apply NLP-techniques to document search and analysis.

