

# GSLT Joint Advisory Group

Report 2007

## **Executive Summary**

The GSLT is currently creating a nation-wide community to provide graduate students in Sweden with comprehensive training in Language Technologies. At the current time there are a total of 47 students registered in one of the participating universities (of these 14 have 100% non-GSLT funding and 2 have 50% non-GSLT funding) and 11 doctoral theses have been successfully examined.

The graduate school continues to address a real national need as evidenced by the ability to attract excellent students and the active participation of researchers of international stature. In addition to the primary goals of providing a graduate education, GSLT has created extensive collaborations between researchers at the participating institutions.

The advisory group members appreciated the open and constructive atmosphere at the 2007 meeting and congratulated the graduate school for the progress thus far in meeting targets, and commending in particular for the quality of the admissions procedure.

The group is pleased that the merits of the school have been recognized, and encourages the endeavors of the graduate school to ensure funding beyond 2012 and so that new students will be able to join the program in 2009. The shared opinion of the advisory groups is that the graduate school should continue to focus on what it does well, that is ensuring a top-level graduate education in Language Technologies, but that it might consider extensions in the scope of its program to include recent developments in human-machine communication and access to multimedia information, and to develop extended ties to industries exploiting language technologies.

## **1 Introduction**

A joint meeting of the GSLT international advisory and the Swedish advisory groups was held on September 11, 2007 at Göteborg University. The participants were the members of international advisory group: Lori Lamel (CNRS-LIMSI, Orsay), John

Nerbonne (University of Groningen) and Stephen Pulman (Oxford University); the members of the Swedish advisory group: Johan Boye (Telia Sonera), Eva Ejerhed (Newsmachine), and Jesper Högberg (Hewlett Packard). The GSLT was represented by Lars Ahrenberg (Linköping), Robin Cooper (Göteborg), Joakim Nivre (Växjö) and as well as student representatives Håkan Burden and Jonas Lindh.

The meeting was held in an informal manner, with active participation of all attendees. The agenda addressed three main points: responses/changes resulting from the previous report of the meeting in October 2004; discussion of the upcoming review of the Graduate School; and the future of GSLT and languages technologies (LT) in general. The meeting was characterized by a frank and open exchange of ideas on graduate training in language technology in general and on opportunities for the GSLT in particular.

This report summarizes the discussions of the meeting and makes some recommendations to GSLT.

## **2 Actions from last time**

At the 2004 advisory board meeting some suggestions were made concerning the Level 3 courses (advanced level courses), to address difficulties the GSLT had encountered (primarily deciding what the level 3 course should be and getting enough critical mass of participation for them). Some of these suggestions have resulted in program modifications such as courses involving joint student projects, courses involving common shared tasks, or courses studying and reporting on hot research topics. Some of these have been quite successful (KTH Dialog course). Our perception, however, is that these courses are still a bit problematic (in particular concerning the level of participation) and later on in this report some new suggestions are made (see next Section).

Another suggestion from 2004 was to include some research highlights on web site. This should require only a minimal effort as the web pages should be proposed by the students, with the agreement of their supervisors.

The annual Swedish Language Technology conferences offer an opportunity to disseminate information to potential industrial partners. However, our impression is that while these conferences are popular with the research community they attract less industrial interest than had been hoped for.

Recommendations were also put forward for the Annual Report and general reporting practices which were largely taken into account for the 2005 and 2006 annual reports. In order to clearly show the added value of GSLT a suggestion was made to compare performance to (some) similar programs elsewhere in Sweden. The advisory boards realize that it may be impossible to compare graduate training in language technology inside vs. outside the GSLT, since all Swedish universities are participating in the

GSLT, but we suggest that a comparison to programs in Linguistics or in Computer Science would be insightful.

At the previous joint advisory board meeting there was a discussion of the need for centralized computing and data resources at GSLT. This has been developed and the students expressed satisfaction with the current situation. Tools are being shared and students have access to online data. Most students also use facilities at their local institutions, but the central facilities are heavily used. The GSLT was accepted as an LDC member which gives students access to all of the LDC corpora. This is important since many of the LDC corpora are widely used and allow for international comparisons.

There was also a brief discussion of changes in Sweden from the Bologna reform (a subject of relatively extensive discussion in 2004). Contrary to expectations in 2004, Sweden has maintained its four year undergraduate educational framework (in particular the “Magister”), alongside the Bologna Masters, making moot the issue of whether the graduate education scheme (Ph.D.) would need to be revised as a consequence. This discussion naturally led to the topic of whether or not GSLT should extend the program to offer a Masters program in Language Technology. This would need to be implemented in a manner analogous to the PhD program since only the individual universities can grant the Masters degrees.

Another related issue was the possibility of increasing the number of students admitted in a Licentiate degree program providing a high-level formation in LT that may interest industry more than PhDs. The advisory boards noted this as an interesting option for students and research groups interested in issues very close to application.

Several attempts have been made to add more cross fertilization (between speech and language technology) in the curriculum as had been suggested. This still seems to be a bit limited from the faculty viewpoint but maybe less so from the students. Maybe, as a result of the interdisciplinary environment, the students are doing better at this than the faculty, and higher levels of cross fertilization can be expected as the PhDs advance to faculty positions. It is noticeable that several thesis titles suggest exactly the kind of interdisciplinarity work that we hoped for. There have also been some contacts with other departments at the participating universities, but generally with people who already consider themselves in the LT field.

### **3 Future opportunities**

To reiterate the view of the advisory board, our impression is that things are going well and that the GSLT participants would like to continue to ensure the improved coordination among Swedish research universities.

As a response to the worry that the GSLT is engaging too little industrial interest, the advisory board offers the following suggestion: We urge the GSLT to require or

at least encourage groups of students to develop sample web-based applications that demonstrate the potential and value of LTs, and to make these publicly available. Such an activity could meet the level 3 course requirements by involving practical work in application development. One of the objectives of such projects would be to introduce LTs to a wider audience including the general public and industry. This sort of project could also serve as an element of a portfolio that students might compile during their graduate studies which presumably would be useful when applying for future employment or research grants. The advisory board emphasizes that students should be encouraged to see such development exercises as a proving ground for small commercial ventures which might be launched even without substantial financial backing. One possibility would be to interest business school students in participating in such projects. By encouraging students to develop their ideas in this way, the GSLT might stimulate LT application without needing to open the doors of the largest Swedish industries. The board notes that the particular needs of the larger industries may not mesh with many of the LT application ideas even though these are realistic and of potential (modest) commercial interest. Naturally this should not be taken as advice to ignore interesting connections to the larger companies, only to look beyond them.

In addition, the advisory board noted that this suggestion might allow faculty members with application ideas to have them realized in prototype applications that might be developed by students and subsequently extended by others in subsequent years.

The advisory board is also enthusiastic about contributions LT is making in areas that are not directly exploitable commercially, but which are nonetheless valuable scientifically and socially. The board suggests that the GSLT consider whether modest efforts in these directions might not be valuable.

Two further suggestions arose from email exchanges among board members which could strengthen student awareness of how their research relates to the (commercial) state-of-the-art. The first is for level three courses to include student reviews of commercially available LT products (based on easily available documentation only) and which would be presented as a seminar in very much the same way as narrow research problems are studied. The second idea is to invite companies manufacturing LT products to give seminars on their view of current and long term challenges, thereby giving students an opportunity to learn the concerns and directions of companies who already know the trade. The latter also increases the student-industry contact broadening their contact network.

One area concerns access to cultural heritage. One notable effort has been that of using speech recognition software to enable indexing and search functionality of the enormous SHOAH archive in the project "Multilingual Access to Spoken Language Archives" (MALACH, <http://malach.umiacs.umd.edu/>), comprised of 115,000 hours of video in dozens of languages. LT techniques are also being experimented with in connection with provided access to handwritten archives (Schomaker [http://www.ai.rug.nl/alice/nwo-catch-scratch/index\\_english.html](http://www.ai.rug.nl/alice/nwo-catch-scratch/index_english.html)). The progress in LT has been facilitated by advances in computational and storage capacity which has led to many recent similar efforts around the world, of-

ten financed by national agencies. Access to huge multimedia corpora is not possible without resorting to the use of automated processing, and LTs provide many crucial components.

Many scientific and scholarly areas are using LT to include large amounts of data in scientific analysis, e.g. confronting descriptive and theoretical linguistics with large amounts of data, which, however, require some analysis (intelligent search) in order to be useful. We also see LT applied in areas such as aphasiology and psycholinguistics (improving, e.g. frequency estimates), and in dialectology and historical linguistics (both in model construction, but also in improving access to large data reserves). More adventurously, there are efforts to involve LT in analyses of political speech, in the history of ideas, and gauging sentiment in selections from large text archives.

## 4 Other topics

*Annual reports* The advisory board appreciates the efforts in the annual reports to directly address previous concerns and recommendations.

*Impact of GSLT:* One of the topics was how to measure the impact of the GSLT on research and education in Sweden. The perception of the board members is that Swedish researchers are more present and visible in international instances. It would be of interest to keep track of concrete information such as participation in international conferences and workshops, publications, organizational program committees, session chairs, keynote/invited speakers, etc. and to the extent possible compare this to the state-of-affairs prior to the formation of GSLT. Another indication is the employment of GSLT graduates. Where are they currently employed (post-docs, research/faculty positions, industry)? It was suggested that the school use those that have chosen industry as liaisons.

*Postdoctoral program:* This subject was once again discussed and the general consensus was unanimous that GSLT should focus their effort on the graduate education and not on post-doctoral formation. The GSLT should aim at creating a base of researchers that go on to industry and academia. Some of these needs may be addressed by increasing the number of Licentiate students.

The advisory board also noted that the GSLT participants have also achieved an unusual level of coordination among themselves, based on a recognition of the effective division of expertise in Swedish LT that enables them to concentrate on individual foci of research in a way that would not be feasible without the coordination of the GSLT. The recognition of the effective division of expertise has been necessitated by the admissions procedure, in which students are assigned to the most suitable advisors. The board remarks that the resulting “division of labor” in LT is likely to continue to serve the researchers and the Swedish LT field well.

## 5 Conclusions

In summary, the joint advisory board extended their congratulations for progress thus far in meeting targets (despite the usual complications of programs getting started, young careers, and parental leaves). The number of graduates is progressing with 7 thus far and 13 foreseen over the next year. There is extensive and effective collaboration in teaching and thesis supervision, as well as more collaboration and improved scientific research by both faculty and students. The board also expressed their view that it is strategically important to continue to contribute to the development of technologies for the Swedish language and train advanced researchers to generate new ideas. It is important to maintain the momentum gained thus far, and the board strongly encourages the school to continue its efforts building upon what has already been accomplished, potentially expanding scope of activities to encompass recent advances in human-computer interaction and access to multimedia information.

The advisory board reminded the GSLT of its enormous potential value to Swedish businesses, to public agencies and to the Swedish people. The applications of language and speech technology are many and growing in number. We are seeing regular improvement in applications such as spelling checkers, hyphenation routines, and grammar checkers; text-to-speech synthesis used to read text aloud to the blind; in dictation software, especially for victims of repetitive strain injuries; language control systems that enforce standards of documentation in the aerospace industry; information search and information extraction; and in named entity recognition, producing names of people, places and organizations (for indexing, for clipping services and for information retrieval).

New applications are seen on a regular basis. For example, language technology is being used in the automatic summarization of documents, including multiple documents. It is also finding use in the automatic evaluation of educational testing, including the essay examinations administered *inter alia* by the Education Testing Service. It is being used to detect potentially confusable names in connection with drug approval at the U.S. Food and Drug Agency and, similarly, by marketing firms testing whether proposed names might have the wrong connotations in some languages. It is used in transliteration software (to render names from non-Roman alphabets and in computer-assisted instruction software, not only that aimed at language learning, but also that aimed at education professionals in finding their way among opaque technical and legal documents.

There other new applications ideas emerging regularly, but it is important to appreciate that the applications will never perform optimally without intelligent input from Swedish language experts, knowledgeable about the techniques, their potential shortcomings, and the proper means of assessing their suitability for novel application ideas. This is the opportunity which the GSLT is poised to seize.

With respect to similar international efforts, the board regards the GSLT program as an excellent model for graduate education that maximizes the use of resources—

especially the resource of scarce technical expertise. This allows the research faculty, staff and students at individual participating departments to specialize more effectively (because they can rely on expertise elsewhere), guaranteeing better matches between faculty supervision and students interests, and it allows universities to react more flexibly in recruiting experts at their respective departments (again, because they can rely on some expertise beyond their local department), and thereby providing the universities a more robust organization in the face of shifting personnel. The increased networking within Sweden is beneficial for students and researchers alike.

The joint advisory board strongly recommends that GSLT staff and hosting institutions continue their investment, which has proved to be a win-win situation for all.