Proceedings of the Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, July 2002, pp. 85-90. Association for Computational Linguistics.

# Teaching Computational Linguistics at the University of Tartu: Experience, Perspectives and Challenges

**Mare Koit**
Institute of Computer Science
University of Tartu
koit@ut.ee

**Tiit Roosmaa**
Institute of Computer Science
University of Tartu
roosmaa@ut.ee

**Haldur Õim**
Department of Linguistics
University of Tartu
hoim@psych.ut.ee

## Abstract

The paper gives a review of teaching Computational Linguistics (CL) at the University of Tartu. The current curriculum foresees the possibility of studying CL as an independent 4-year subject in the Faculty of Philosophy on the bachelor stage. In connection with the higher education reform in Estonia, new curricula will be introduced from the next study year where the 3-year bachelor stage will be followed by a 2-year master's stage. It will then be possible to study CL proceeding from two paths: in the Faculty of Philosophy, and additionally also in the Faculty of Mathematics and Computer Science. This way two types of specialists will be trained who will hopefully be able to complement each other in team-work.

## 1 Introduction

Computational Linguistics as a special subject can in Estonia only be studied at the University of Tartu (UT).

Computational linguistics as an independent subject has been taught at UT from the study year 1998/99, various separate courses from the study year 1996/97. The curriculum and various new courses were developed with the support provided by a number of HESP projects. Based on the experience of Western universities, the topics of linguistics, computer science, mathematics and computational linguistics were selected. As the subject was launched in the Department of Estonian and Finno-Ugric Linguistics, Faculty of Philosophy, then linguistic subjects prevail in the curriculum. But the tuition is provided jointly by two faculties: Faculty of Philosophy teaches linguistic subjects, Faculty of Mathematics and Computer Science teaches mathematics and informatics subjects, the computational linguistic subjects have been divided between these two faculties. Language of instruction is Estonian.

Preparing the lectures in new subjects has been mostly accompanied by making available these materials on the web as well. A number of students of computer science and linguistics also listen to several computational linguistics lectures as electives or optional subjects. Therefore we face the problem how to make these subjects interesting and manageable for students with various qualifications.

New curricula will take effect as of the study year 2002/2003 at UT. According to them, the 3 year bachelor study is followed by 2-year master studies which in turn may be followed by 4-year doctoral studies (so far the 4+2+4 year model was used). Bachelor studies will provide the general basic higher education, master studies will lead to a qualification in a specific subject. The new modified curriculum takes into account the experience obtained in the process of teaching so far. In addition to that, the new curriculum for

computer science foresees the (new) possibility to major in language technology. In the following, we will present the structure of the curricula and our experience so far in teaching the subjects of computational linguistics.

## 2 Background to tuition

Tuition can be efficient only when the lecturers themselves are active researchers.

R&D work in computational linguistics and language technology in Estonia is being carried out, in addition to UT, in the Institute of the Estonian Language (IEL is a research institution) and the Institute of Cybernetics of Tallinn Technical University. The IEL focuses on compiling computer lexicons but also on the computer (morphological) processing of Estonian. The topics at the Institute of Cybernetics include the generation of spoken Estonian (together with IEL), the Institute has also started compiling databases necessary for speech recognition.

The R&D at the University of Tartu has focused on the computer analyses of Estonian texts and compiling the text corpora of Estonian underlying that research. The major research topics include

- Formalising the morphology and syntax of the Estonian language
- Formalising the semantics of Estonian (incl compiling a lexico-semantic database)
- Pragmatics: modelling the Estonian (spoken) dialogue.

Based on the results of the research, various linguistic software and resources have been elaborated at UT:

- The morphological analyser and generator of Estonian; the university spin-off language software company Filosoft in turn has used these resources for creating the Estonian spell checker and hyphenator (included in the MS Office package)
- Estonian morphologic disambiguator and syntax analyser
- Various corpora of written Estonian for the period 1890-1990 (total 3 million words), partly morphologically and syntactically tagged
- A corpus of spoken Estonian (300 000 words transliterated), and a corpus of Estonian dialogues based on it (60 000 words).

UT computational linguists have participated in a number of international projects, e.g. GLOSSER, MULTEXT-EAST, TELRI-I, TELRI-II, CONCEDE, EuroWordNet, BABEL, to name some, and carried out numerous projects commissioned by the Estonian Science Foundation and the Estonian Informatics Centre.
The plans for the next years include further development of language software, incl morphological and semantic disambiguator and the syntactic analyser and generator, to study and model the formal structure of dialogues, to expand the tagged corpora. All these outputs can be used in various language technology applications, from aids to the text compiler (e.g. grammar and style checkers in a text editing programme) to machine translation or man-machine dialogue in Estonian.

Previously, in 1960s, UT was engaged in language statistics and automated information retrieval (in 1970s, an information retrieval system for legal texts was developed at UT).

In the second half of 1960s, a special structural linguistics work group (the so-called generative grammar group) was active at the Department of the Estonian Language. It included lecturers, doctoral students and students. Quite a few of the present computational linguists received their first knowledge of CL from that group.

Until 1998, obtaining the CL education was only possible according to individually tailored study plans.

## 3 Curriculum

### 3.1 Earlier experience

While elaborating the CL curriculum and preparing and modifying new courses, we have tried to take into account the experience of other universities.

As demonstrated by a questionnaire carried out in March 1999 in 60 European universities where CL is being taught (de Smedt et al., 1999), three

options are basically used in teaching CL and language technology:

- A minor in philology (dominating)
- A minor in Computer Science
- An independent subject.

The curricula typically include 4 blocks of subjects*:*

1) Linguistics
2) Computer Science
3) Mathematics
4) Computational Linguistics

At UT, as already mentioned, CL is taught as an independent subject in the Faculty of Philosophy. The amount of studies is characterised by the number of credit points (CP) where 1 CP corresponds to 40 hours of work by the student (incl. independent work). The total amount of the 4-year bachelor studies is 160 CP.

In the current curriculum, the amount of CL is 60 CP and it includes the same 4 typical blocks as named above:

Linguistics 20 CP
Computer Science 3 CP
Mathematics 10 CP
Computational Linguistics 27 CP, from them electives 4 CP and the bachelor thesis 12 CP.

As may be seen, the block of Computer Science is very small (presently, it includes only the subjects ”Prolog for linguists”, 2 CP, and ”UNIX for linguists”, 1 CP). In the new curriculum, we have increased the importance of that block.

### 3.2 New curriculum

The necessity for the new curriculum was prompted by the higher education reform in Estonia. The reform was launched in 2000 and is based on various international documents and agreements (Magna Charta Universitatum).

The main objectives of the Estonian higher education reform are:

- to expand the cross-curriculum share of subject areas by widening the opportunities of interdisciplinary studies;
- to simplify the system of university education levels;
- to simplify and expand the opportunities of students from different specialities to continue their studies in other universities (also outside Estonia).

The Bachelor level education will be achieved after completing a 3-year curriculum (nominal study period, 120 CP). The Master level education will be achieved upon completing a 5-year curriculum (nominal study period, 200 CP). The Doctoral level education will be achieved upon completing a 9-year curriculum (nominal study period, 360 CP) and defending the doctoral thesis.

Bachelor studies will provide general theoretical education at the university level. In the first year, the students study subjects shared by the curricula of one broad field. The second year they study following the narrow field module and the third year is devoted to specialised subjects.

Master studies will provide special, professional knowledge and vocational skills.

The ”Estonian and Finno-Ugric linguistics” curriculum that will become operational in the Faculty of Philosophy foresees the possibility of majoring in computational linguistics.

In the bachelor studies one has to take

- two obligatory base modules*:* ”Humanities” (16 CP) and ”Estonian philology” (16 CP);
- obligatory narrow field module ”Estonian and Finno-Ugric linguistics” (16 CP);
- obligatory speciality module ”Computational linguistics” (20 CP, incl a bachelor’s thesis 4 CP);
- two modules of elective subjects (each 16 CP) from the list of the following subjects: ”Estonian language”, ”Estonian language and culture for non-Estonians”, ”Finnish language and culture”, ”Finno-Ugric languages”, ”Hungarian language and culture” and ”General linguistics”;
- electives and optional subjects (20 CP) that may be chosen from any curriculum.

The CL speciality module entails the following subjects: ”Mathematics for computational linguists I”, “Programming”, “Data analysis in humanities”, “Linguistic theories for computational linguists”, each amounting to 4 CP.

Upon completing bachelor studies, the student will receive the Bachelor degree in Estonian and Finno-Ugric linguistics. As a rule, this education will not guarantee entry to the labour market (at least not as a computational linguist) but has to be followed by master studies. It is assumed that at least 75% of students admitted to bachelor studies will continue their master studies.

Master studies curriculum comprises of speciality studies (56 CP), master thesis (20 CP) and optional subjects (4 CP). A prerequisite for starting the master's studies is either a bachelor's degree or education level corresponding to it. A preliminary condition for entrance to the CL speciality is having taken the speciality module during the bachelor studies. Thus every person with a bachelor's degree who has taken the 4 subjects comprising the CL speciality module may enter the CL master's studies.

In master studies, the CL speciality studies will consist of compulsory subjects (22 CP) and electives (34 CP). The compulsory subjects are "Introduction to CL", "Corpus linguistics", "Language technology", "Mathematics for computational linguists II" and the master's seminar. The list of electives is open and will be updated according to requirements and possibilities.

The present list includes subjects from linguistics, computer science and CL.

Linguistics subjects include, for example "Phonology and morphology", „Syntax of Estonian", "Semantics", "Theories of linguistic communication", „Pragmatics". Computer science subjects included in the list are "Artificial Intelligence I and II", "Applied software: Perl", „Databases"; Computational Linguistics subjects include such as "Computational morphology", "Computational lexicology", "Syntactic analyser", etc.

The majority of the subjects are the same that have been and are taught within the existing curriculum but there will also be several new ones, e.g. "Statistical models of natural languages" and "Introduction to speech technology".

The qualification conferred to a graduate will be that of master of Estonian and Finno-Ugric linguistics (computational linguistics). This is a specialist whose computational linguistics education is based on linguistics (this has been the case up to now when studying according to the present curriculum).

## 4 A new opportunity: language technology

In connection with preparing new curricula it became possible to start preparing computational linguists with a different education at UT based on computer science –language technology studies. Proceeding from the new computer science curriculum to come into force in the Faculty of Mathematics and Computer Science it is possible to choose blocks of linguistics and CL subjects that have been named language technology modules (to differentiate them from the CL modules of the curriculum of Estonian and Finno-Ugric linguistics in the Faculty of Philosophy).

Bachelor studies in computer science will provide general knowledge in the classical branches of mathematics and basic knowledge of computer software, hardware, networks and systems, artificial intelligence, software technologies and data protection and a certain amount of practical skills for work in the computer science (incl programming skills). It is possible to choose between theoretical computer science, software systems and language technology. Upon completing bachelor studies, the student will receive the Bachelor degree in Computer Science.

The language technology narrow field module (16 CP) comprises the subjects "Language technology", "Introduction to CL", "Corpus linguistics", "Introduction to general linguistics" and "Database theory".

Master studies in computer science will provide profound knowledge in one area of the computer science enabling the person to carry out development activities in that area; skills to provide consultations; team-work and project management skills. It is possible to major in theoretical computer science, cryptology or language technology.

The qualification conferred to a graduate will be that of master of computer science.

A prerequisite for entrance to master's studies is the bachelor's degree in the computer science (or in a close speciality) and prerequisite subjects amounting to 20 CP (Object-oriented programming, Algorithms and data structures, Introduction to mathematical logics, Elements of discrete mathematics, Algebra I, Data bases).

The didactics of informatics and master seminar are compulsory for all master students (both 4 CP); optional subjects may be chosen for the amount of 4 CP.

Those who major in language technology will have the following compulsory subjects in master studies: Software technology, Automata, languages and translators, Graphs, Theory of databases, Artificial Intelligence I, Computational morphology, Syntax theories and models, Computational lexicology, Semantics, Statistical models of natural languages, total 32 CP. In addition to that, 16 CP of electives from the open list that will be updated according to needs and opportunities as the similar list for computational linguistics

Although these two lists have a quite big overlapping area they are not the same. In case of computational linguistics, linguistic subject will take up a major part of the list; in case of language technology, computer science subjects will replace them (e.g. Methods of logical programming, Methods of functional programming, Systems modelling, Formal languages).

Therefore, a person who has completed this curriculum is an information scientist who has additionally studied linguistic and computational linguistic subjects to such an extent that he/she will have a systematic picture of the tasks of natural language processing and will be able to solve these tasks in co-operation with linguists.

## 5 Problems

The modules of CL and LT contain a number of common subjects that have also to be taught jointly to the students majoring in CL and LT. Preceding from earlier experience it may be assumed that the total number of listeners will not be big but they are with varied preparation. While teaching according to the present curriculum there have been groups including CL students, students from different linguistics fields (Roman-German philology or Estonian as a foreign language) as well as students of computer science. The new curricula have "legalised" the different backgrounds of the students. This will be a challenge to the lecturers: how to present the material in a way that is understandable and manageable for the students, at the same time not being too simple for some students. So far we have overcome the problem by giving different assignments to students with different background. E.g. the students of computer science compile a language processing program while the students of linguistics work with language corpora. In the future these problems will be more acute and require a complex solution. For example, how to compose teams of students with different qualification in such a way that by complementing each other they would be able to solve a joint problem.

Another problem is a small number of students attending special courses. One solution to it could be the web-based tuition, i.e. not only the availability of teaching materials over the web but a special distance learning environment integrating lecture materials, individualised/ personally tailored assignments and check of knowledge. UT has successfully implemented the learning environment WebCT that already includes a number of courses (in the framework of the *e*University project). These do not include any CL courses so far which means that we have to start preparing them by using the already existing web course materials.

## 6 Summary

Our experience in teaching and preparing computational linguists is not extensive yet. Nevertheless, the first CL students have finished their studies and it seems that there will be interested students in the future as well. In addition to that, students of different subjects have attended various CL courses (mainly

"Introduction to CL" and "Language technology"). Students of CL take part in our projects and in our joint seminars thus being exposed from the very beginning to real-life tasks. Therefore we expect them to become qualified labour to universities, research institutions and language software companies.

## References


Konraad de Smedt, Hazel Gardiner, Espen Ore, Tito Orlandi, Harold Short, Jacques Souillot and William Vaughan (eds). 1999. *Computing in Humanities Education. A European Perspective.* SOCRATES/ERASMUS thematic network project on Advanced Computing in the Humanities. University of Bergen.

*Magna Charta Universitatum*
http://www.unige.ch/cre/activities/Magna%20Charta/magna_charta.html (used 26.03.2002)