Language Technology
at the University of Tartu

Research and education

- People of 2 faculties are involved:
  - Faculty of Mathematics and Computer Science => Institute of Computer Science => Chair of Language Technology (chair exists since 1-9-2001)
  - Faculty of Philosophy => Department of Estonian and Finno-Ugric Linguistics => Chair of General Linguistics
- Informal research group of computational linguistics
  - Head of the group – professor of general linguistics Haldur Õim

People

- Chair of LT
  - Mare Koit, prof.
  - Tiit Roosmaa, PhD, assoc. prof, vice dean of the Faculty of Mathematics and Computer Science, head of the Institute of Computer Science
  - Heli Uibo, MSc, lecturer/PhD student
  - Kaili Müürisep, PhD, researcher
- Chair of GL
  - Haldur Õim, prof.
  - Renate Pajusalu, assoc. prof.
  - Heiki-Jaan Kaalep, PhD, senior researcher
  - Neeme Kahusk, researcher/PhD student
  - Kadri Muischnek, MA, researcher/PhD student
  - Heili Orav, MA, researcher/PhD student
  - Andriela Rääbis, MA, researcher/PhD student
  - Kadri Vider, MA, researcher/PhD student
  - Tarmo Vaino, network administrator/programmer
  - Urve Talvik, specialist

Faculty
# of Mathematics and Computer Science

- Vice dean Tiit Roosmaa

University of Tartu

- The University of Tartu (UT) is the **oldest and largest** university of Estonia and one of the best-known in Northern and Eastern Europe. From its foundation in 1632 to the first decades of the XX century UT was the only university in the Northern Baltic region.
- **Classical university** - with its eleven faculties, UT is the only classical university of Estonia. It also includes three specialized research institutions: the Institute of Physics, the Estonian Marine Institute and the Institute of Technology. UT has launched its regional program, in the framework of which it has established a network of six UT colleges located each in a different city.

Faculties

Faculty of Theology
Faculty of Law
Faculty of Medicine
Faculty of Philosophy
Faculty of Biology and Geography
Faculty of Economics and Business Administration
Faculty of Education
Faculty of Exercise and Sport Sciences
Faculty of Mathematics and Computer Science
Faculty of Physics and Chemistry
Faculty of Social Sciences

Faculty of Mathematics and Computer Science

Some facts:

- 58 lecturers
- 14 researchers
- 4 institutes
- 826 students (47 PhD students)
- 3 Curriculas on PhD level
- 6 Curriculas on Master’s level
- 4 Curriculas on Bachelor’s level
- 2 Teacher training curricula

Institute of Computer Science

- Chairs
  - Theoretical Computer Science
  - Software Systems
  - Cryptography
  - Distributed Systems
  - Language Technology
- Some facts:
  - 18 lectures
  - 6 researchers
  - 5 chairs
  - 22 PhD students, 54 Master students
  - 1 Curricula on PhD level
  - 2 Curriculas on Master's level
  - 2 Curriculas on Bachelor's level
  - 1 Teacher training curricula

Research and education at the University of Tartu

- Dr. Madis Saluveer, Department of R&D, Head of Development Office

Tartu

Tartu is one of the oldest university towns in Northeast-Europe. It is home to Estonia’s oldest and only classical university, and also a large number of other educational and research institutions. More than 20,000 of its 100,000 inhabitants are university students and over 20,000 study at secondary or vocational schools. Tartu is a town of rapidly evolving innovative entrepreneurship, theatres, parks and scholastic and sporting traditions.

The University Today

- In 2003 the University of Tartu has 11 faculties, 3 research institutes and 5 colleges with more than 70 departments, institutes and clinics. The total number of students is over 18,000, with a teaching staff of 1,250. The University library has joined the electronic information system. The University’s five museums, botanical garden and sports facilities welcome all visitors.
  In all fields of activity the University seeks to maintain the highest standards, paying much attention to developing collaboration and international contacts.

Mission and Vision

- **Mission.** The University of Tartu is a national university uniting different branches of science.
  - The mission of the University of Tartu is to act as the guardian and advocate of a highly educated Estonia through internationally acclaimed research and the provision of research based higher education.
  - The mission of the University of Tartu shall be implemented in co-operation with domestic and foreign partners.
- **Vision.** The University of Tartu is a research university of international repute and the centre of Estonian academic spirit, national culture, scientific language and high-technology innovation.

History

- The University of Tartu was founded as *Academia Dorpatensis (Gustaviana)* **in 1632** by King Gustavus II Adolphus of Sweden.

- In 1802 it was reopened as Die Kaiserliche Universität zu Dorpat by a ukase of the Russian Czar Alexander I.

- In 1919 it became a national university -Tartu Ülikool.

- The University of Tartu has been the alma mater for the entire educational system and for the scientific research in the nation.

International Co-operation

- The University of Tartu has co-operation agreements with 32 universities in 11 countries. The University of Tartu participates in the International Student Exchange Programme ISEP and the SOCRATES programme.
- In addition, it is possible to come to the University of Tartu to teach or study according to university co-operation agreements or on the 120 ERASMUS exchange-agreements.
- The University of Tartu participates in 43 co-operation projects financed by the European Commission with a total budget of 43.2 million Estonian kroons. There are 7 co-operation contracts with other foreign institutions with a total budget of ca one million Estonian kroons.

Research and Development

- The University of Tartu continues to play a central role in Estonian science thanks to its high standards of research and integration between different fields. Its participation in the international programmes and active foreign relations have helped the University and many of its scientists win world renown. Over 3000 publications are published annually. In addition, the University is continually adapting its work to cater for the current needs of society.
- The most remarkable recent research achievements have been in the areas of molecular and cell biology, gene technology, immunology, pharmacology, laser medicine, materials science, laser spectroscopy, biochemistry, environmental technology, computer linguistics, psychology, and semiotics.

Basic Budget

BASIC BUDGET 2003	
1. Income from tuition	Thousands €
1.1. Government-financed tuition	18 250,1
1.2. Governmental allocation for registrars	1 657,4
1.3. Tuition fees	3 782,1
1.4. Tuition fees of Open University	3 295,7
1.5. Tuition fees of continuing education	827,5
1.6. Other paid services	470,2
Total tuition revenue	**28 283,1**
2. Revenues from R&D	
2.1. Targeted funding of research topics	5 709,9
2.2. Targeted funding of infrastructure	1 640,9
2.3. Estonian Science Foundation grants	2 355,4
2.4. Targeted funding of PhD research	552,1
2.5. Research and development contracts	2 754,9
2.6. International contracts and grants	2 528,3
2.7. Other researh revenues	143,5
Total revenue from R&D	**15 684,9**
3. Other revenues	
Total other revenues	1 701,2
Total income of the basic budget	**45 669,2**

Structure of R&D income of UT from 2000 to 2003 (million euro)

	2000	2001	2002	2003	
Source of income					**change 00-03 %**
Target financing	3,523	4,075	4,853	5,712	62,1
ESF	1,84	1,98	2,248	2,356	28,0
Infrastructure	1,112	1,29	1,279	1,641	47,6
PhD students	0,62	0,64	0,57	0,552	-8,1
Foreign contracts	1,147	1,016	1,85	2,529	120,5
Domestic contracts	0,971	1,195	2,024	2,756	183,8
Other	0,818	0,849	0,15	0,143	-82,5
Total R&D income	**10,034**	**11,613**	**12,978**	**15,691**	**56,4**
UT total income (main budget)	**33,127**	**36,753**	**39,773**	**45,688**	**37,9**
Ratio of R&D income from UT total income	**30,29**	**31,60**	**32,63**	**34,34**	**13,4**

FP5 contracts and income 2000-2003

year	number of FP5 contracts signed	Allocated FP5 resources (million euro)	Total income from foreign contracts (million euro)	Income from FP5 contracts (million euro)	Ratio of FP5 income from total contract income
2000	8	1,498	1,147	0,236	20,6
2001	16	1,710	1,012	0,703	69,4
2002	20	2,53	1,851	1,014	54,8
2003	5	0,227	2,529		
Total	49	5,965			

Research group of computational linguistics

- Cooperation with the Institute of Cybernetics at the Tallinn University of Technology and the Institute of Estonian Language
  - 2002 these 3 research units together applied to be a centre of exellence in language technology (head – prof. Haldur Õim) => potential centre
  - 2003/4 language technology development centre (head – Dr. Einar Meister, TUT)
- Members of research group have participated in working out the strategy of development of Estonian language (2004-2010), the language technology part of the state programme “Estonian language and national memory” (2004-2010), the roadmap of Estonian language technology (2004-2010), and are involved in preparation of state programme “Technological support of Estonian language” (2006-2010).
- Main research fields
  - computational morphology of Estonian
  - computational syntax
  - semantics
  - spoken Estonian and dialogue modelling
  - corpora a. o. language resources

Computational morphology
Heiki-Jaan Kaalep (1/2)

- IEL
  - Unification
  - Guessing (stable and unstable inflectional classes)
  - Ülle Viks
- UT
  - 2-level
  - Heli Uibo
- Filosoft http://www.filosoft.ee/index_en.html
  - Unification
  - Lexicon (spelling)
  - Heiki-Jaan Kaalep

Disambiguation
Heiki-Jaan Kaalep (2/2)

- CG
  - Tiina Puolakainen (UT, IEL)
- HMM
  - Heiki-Jaan Kaalep (UT, Filosoft)
- 500,000 word corpus (gold standard)

Computational syntax

- Tiit Roosmaa
- Heli Uibo
- Kadri Muischnek
- Kaili Müürisep

Syntax - Outline

- Projects, funding
- Software: Estonian Constraint Grammar Parser and its applications
- Resources: steps towards Estonian treebank
  - Constraint Grammar corpus
  - Sofie Parallel Treebank
  - Estonian Treebank Arborest

Projects

- Estonian Science Foundation grant No. 3314 “A *formal grammar for the Estonian language*” (1998-2000), total funding 11 600 EUR
- Project “*Syntactically analyzed and disambiguated text corpus*” (2002-2003), funded by Estonian Ministry of Education and Research under the national program “Estonian language and national heritage”, total funding 22 500 EUR
- Project “*Syntax-based language software and the resources needed for its development*” (2004-2008), funded by Ministry of Education and Research, national program “Estonian language and national memory”, in 2004: 16 000 EUR

International cooperation

- Network-type projects funded by NorFA under The Nordic Language Technology Research Programme (2000-2004):
  - *Nordic Treebank Network* (2003-2004), coordinated by Joakim Nivre, Växjö University, joins 15 academic institutions from Sweden, Norway, Denmark, Finland, Estonia and Iceland.
  - *PaNoLa* (Parsing Nordic Languages) follow-up project (Sep-Dec, 2004), coordinated by Eckhard Bick, University of Southern Denmark. The aim of the project is to create VISL teaching treebanks for smaller Nordic languages – Estonian, Faroese, Greenlandic, Icelandic and Sami.

Software: Syntactic Parser for Estonian (EstCGP)

- EstCGP (Estonian Constraint Grammar Shallow Syntactic Parser) is the result of two doctoral dissertations:
  - Kaili Müürisep “Computer Grammar of Estonian: Syntax” (Univ of Tartu, 2000)
  - Tiina Puolakainen “Computer Grammar of Estonian: Morphological Disambiguation” (Univ of Tartu, 2001)
- Current evaluation results of ESTCG:
  - precision 76,4-79,2 %
  - recall 95,5-96,9 %.

Shallow Syntactic Parser: applications

- Noun phrase extraction (K. Müürisep,

T. Puolakainen)

- Automatic summarization (K. Müürisep, students)
- Syntax-based information retrieval

  (K. Kaljurand)
- Grammar check (H. Uibo, students)

Syntactically annotated corpora of Estonian

1. Estonian Constraint Grammar Corpus
   - size – 200 000 running words, ca 15 000 sentences
     - 184 000 words of Estonian original fiction
     - 10 000 words of newspaper texts
     - 6 000 words of legal texts
   - shallow annotation, using Constraint Grammar: a syntactic function is determined for every word-form
   - has been built to train and test EstCGP
   - is being extended semi-automatically
   - planned size by Dec 2004 – 300 000 words
   - website http://math.ut.ee/~heli_u/syntcorpus.html

Estonian Constraint Grammar Corpus

Experiments on EstCGC (K. Kaljurand):

- Conversion of EstCGC to NEGRA export format

http://psych.ut.ee/~kaarel/Programs/Treebank/EstCG2Negra/

- Automatic extraction of syntactic dependency relations

http://psych.ut.ee/~kaarel/Programs/Treebank/DepDict/

Syntactically annotated corpora for Estonian (cont-d)

Two small-scale experimental treebanks:

2. Sofie Parallel Treebank – a Penn-style phrase structure treebank of 100 sentences

3. Arborest – a VISL-style hybrid treebank of 2500 sentences (first 149 sentences manually revised)

Sofie Parallel Treebank

- Sofie Parallel Treebank is a joint effort of the members of **Nordic Treebank Network**
- Material – the 1st chapter of Jostein Gaarder's novel "Sophie's World".
- Currently, the parallel treebank includes Swedish, German, Norwegian, Estonian, Icelandic and two versions of Danish, 20-200 sentences from each language.
- Website of the Sofie Parallel Treebank:

http://omilia.uio.no/sofie

Estonian Treebank Arborest

- Joint work with dr. Eckhard Bick, University of Southern Denmark
- VISL-style (http://beta.visl.sdu.dk) treebank
- Annotated for both **function** (S = subject, P = predicate, O = object, A = adverbial,STA = statement, QUE = question, etc.) and **form** (np, vp, pp, advp, adjp, fcl = finite clause, par = paratagma, etc.)

Arborest

- Automatically generated from a sample of CG-corpus (2500 sentences) with CG→PSG rules
- 149 sentences revised
- 1/3 of sentences correct
- CG→PSG rules are under improvement

Webpage http://corp.hum.sdu.dk/arborest.html

Plans

- To enlarge all three syntactically annotated corpora.
- To improve the CG-to-PSG rules to facilitate the easy semi-automatic way of building an Estonian treebank.
- To investigate, how many semantic information can be derived from the syntactic structure.
- To build a phrase-aligned Estonian-German-Swedish parallel treebank

Semantics

- Haldur Õim
- Heili Orav
- Neeme Kahusk
- Kadri Vider

Semantics – PhD studies

- Kadri Vider – Word Sense Disambiguation of Estonian Verbs According to Lexical-Syntactic Information
- Heili Orav – Semantics of personal traits.
- Neeme Kahusk (PhD student at Tallinn Pedagogical University) –The role of semantic relations in word explanation task demanding quick response

Semantics - Grants

- Target (governmental) financing program
  - Elaboration and implementation of computational linguistics tools for creation of Estonian language resources (SF0180528s98, 01.01.98-31.12.02)
  - Computational models and language resources: for Estonian: theoretical and applicational aspects. (SF0182541s03, 01.01.03-31.12.07)
- Estonian Science Foundation
  - Creation of a Semantic Disambiguator for Estonian (ETF4467, 01.01.00-31.12.02)
  - Concept based resources and processing tools for the Estonian language (ETF5534, 01.01.03-31.12.06)
- Governmental Research Program
  - Human Language Technology: Semantic analysis of Estonian simple sentences

Semantics – current courses of action (1)

- Estonian Wordnet
    - 10,000 synsets, 18,900 word senses
  - WordNet taken as a model
  - EuroWordNet-2 project member 1998-1999
  - Global WordNet Association member

Publications:

- EuroWordNet Technical Reports: Deliverables 2D001, 2D003, 2D006, 2D007, 2D008, 2D010, 2D014, 2D014
- Kadri Vider, Neeme Kahusk, Heili Orav, Haldur Õim, Leho Paldre, 2000. Eesti keele tesaurus (The Estonian Thesaurus) - Publications of the Department of General Linguistics of the University of Tartu, vol. 1. Ed. by T. Hennoste. Tartu, 2000, pp. 127-152.
- H. Orav “Adjectives as semantic problem: wordnet-type thesaurus collection experience” – COMPLEX 2001, Birmingham, UK
- Orav, H. Adjectives in wordnet-type thesaurus: Estonian experience. In Proceedings of the 1st International Global WordNet Conference, Central Institute of Indian Languages, Mysore, India, 2002, pp. 22-25
- Vider, K., Orav, H. Concerning the difference between a conception and its application in the case of the Estonian wordnet Proceedings of the second international wordnet conference. Eds. P.Sojka, K. Pala, P. Smrz, Ch. Fellbaum, P. Vossen. Masaryk University, Brno, 2003, pp. 285-290
- Vider, K., Orav, H. Estonian wordnet and Lexicography. Symposium on Lexicography XI. Proceedings of the Eleventh International Symposium on Lexicography. May 2-4, 2002 at the University of Copenhagen. Ed. by H. Gottlieb, J. E. Mogensen and A. Zettersten. Max Niemeyer Verlag, In press
- Vider, K. Notes about labelling semantic relations in Estonian WordNet. Proceedings of Workshop on Wordnet Structures and Standardisation, and how these Affect Wordnet Applications and Evaluation; Third International Conference on Language Resources and Evaluation (LREC 2002). Ed. by D. N. Christodoulakis, C. Kunze, L. Lemnitzer. ELRA, Las Palmas de Gran Canaria 2002 pp. 56-59

Semantics – current courses of action (2)

- Word Sense Disambiguation
  - SensEval-2 all-words task for Estonian
    - Results: 2 systems, precision & recall 66%
  - Estonian WSD corpus
    - ~100,000 tokens, ~42,000 annotated content words

Publications:

- Kahusk, N. and Vider, K. 2002. Estonian WordNet Benefits from Word Sense Disambiguation. In Proceedings of the 1st International Global WordNet Conference, Central Institute of Indian Languages, Mysore, India pp. 26-31
- Vider, K. and Kaljurand, K. Automatic WSD: Does it make sense of Estonian? - Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse 2001, pp. 159-162
- Kahusk, N., Orav, H., Õim, H. Sensiting inflectionality: Estonian Task for SENSEVAL-2. Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse 2001, pp. 25-28
- Kahusk, Neeme A Lexicographer's Tool for Word Sense Tagging According to WordNet Proceedings of Workshop on Wordnet Structures and Standardisation, and how these Affect Wordnet Applications and Evaluation; Third International Conference on Language Resources and Evaluation (LREC 2002). Ed. by D. N. Christodoulakis, C. Kunze, L. Lemnitzer. ELRA, Las Palmas de Gran Canaria 2002, pp. 1-7
- Kaarel Kaljurand. Word Sense Disambiguation of Estonian with syntactic dependency relations and WordNet. Proceedings of the Ninth ESSLLI Student Session. Ed. by L. Alonso i Alemany and P. Egre. August 2004, Nancy, pp. 128-137

Spoken Estonian and dialogue modelling 1/3

- People
  - Tiit Hennoste (is working at Helsinki University since 1-9-2004)
  - Andriela Rääbis
  - Mare Koit
  - PhD and Master’s students
- Goals
  - to study spoken Estonian, its different registers
  - to collect different kinds of spoken texts into the corpus of spoken Estonian
  - to model human-computer interaction in Estonian

Spoken Estonian and dialogue modelling 2/3

- Corpus of spoken Estonian (started 1997)
  - 490 tapes
  - 1100 transcribed texts (700,000 running words)

- Dialogue corpus (started 2001)
  - spoken dialogues (sub-part of the corpus of spoken Estonian - 400 texts; 100,000 running words)
  - written dialogues collected by the method of Wizard-of-Oz (20 texts, 2500 running words)
  - dialogue acts are annotated in the dialogue corpus – a typology of dialogue acts is worked out
  - theoretical basis of the typology – conversation analysis

Spoken Estonian and dialogue modelling 3/3

- We analyze how various types of dialogue acts are used in a special domain – calls for information (information offices, travel bureaus), and how it depends on Estonian cultural space.
- We are testing machine learning methods for automatic recognition of dialogue acts
- Grants: Estonian Science Foundation, Estonian Ministry of Education and Research
- We have presented our work on “Text, Speech and Dialogue” conference (2003), SIGdial workshops (2003, 2004), LREC 2004 workshop “Compiling and Processing Spoken Language Corpora”, 1st Baltic Conference “Human Language Technologies” (2004) etc.
- International cooperation (previous):
  - Nordic network “Corpus-based research on spoken language” (2000-2004, Tiit Hennoste)
  - Nordic network for researchers in conversation studies (2000-2004)

Language resources1/5

Kadri Muischnek

Corpora

Corpus of Written Estonian 1890-1990
The Mixed Corpus of Estonian:
Balanced corpus (newspaper texts+fiction+science texts)
Morphologically disambiguated corpus
WSD corpus (Word sense disambiguation)
Syntactically annotated corpus

Language technology resources (besides corpora)

Corpus query
Frequency Dictionary
Database of Multi-Word Expressions
Thesaurus
Morphological analyser
Speller of Estonian (HTML)

Language resources 2/5

Corpus of Written Estonian 1890-1990

- corpus of the 1990s (380 000 words newspaper texts + 600 000 words fiction)
- corpus of the 1980s (1 million words, Brown & LOB –style textclasses)
- corpus of the 1970s (170 000 words newspaper texts + 250 000 fiction)
- corpus of the 1960s (200 000 words newspaper texts + 130 000 fiction)
- corpus of the 1950s (240 000 words newspaper texts + 60 000 fiction)
- corpus of the 1930s (120 000 words newspaper texts + 150 000 fiction)
- corpus of the 1910s (180 000 words newspaper texts + 250 000 fiction)
- corpus of the 1900s (170 000 words newspaper texts + 65 000 fiction)
- corpus of the 1890s (190 000 words newspaper texts + 50 000 fiction)

Language resources 3/5

Mixed Corpus of Estonian

Big (in our dreams 200 million words)

non-balanced; contains whole texts, not text samples.

At the moment, the corpus consists of the following:

- Weekly «Eesti Ekspress» (issues 09.08.1996 - 29.11.2001, 7.5 million words)
- daily «Postimees» (issues 27.11.1995 - 10.10.2000, 1760 issues containing 88 600 articles, 32.9 million words)
- weekly «Maaleht» (6 million words coming soon)
- journal «Horisont» (1996 - 2003, 260 000 words)
- journal «Akadeemia» (7,5 million words, coming soon)
- fiction from the year 1995 onwards (4.2 million words)
- PhD dissertations (0.5 million words)
- Parliament transcripts 1995-2001 (13 million words)
- Estonian and European legal documents (ca 1.8 million and 10 million words)

Language resources 4/5

Mixed Corpus contains a balanced subcorpus called

- **The Balanced Corpus**

The aim of this corpus is to enable the comparison of three main textclasses - newspaper, fiction and scientific texts - in written language.

5 million words of newspaper texts

4 million words of fiction (aim: 5 millions)

half million words of scientific texts (aim: 5 millions)

Morphologically Disambiguated Corpus

Fiction	104 000
G. Orwell "1984"	75 800
Newspaper texts	111 000
Legal documents	121 000
journal Horizont	99 000
informative texts	4 000
total	513 000

disambiguated manually by 2 persons

Language resources 5/5

- **Frequency Dictionary**

based on 1 million words (500 000 newspaper texts + 500 000 fiction from the 2. half of the 90ties)

- **Database of Multi-Word Expressions**

based on 6 dictionaries

subpart: Database of Multi-Word Verbs:

data extracted from the dictionaries + collocations extracted from the corpora

Education

Two models of higher education:

- old:
  - – 4 years (Bachelor) +2 years (Master of Arts or Master of Science)

    [+4 years (PhD)]
- new since 2002/2003 (Bologna declaration):
  - – 3+2 [+4]
    - 1 year = 40 credits (AP)
    - 1 credit = 40 work hours (=1,5 ECTS)

PhD studies 1/3

- No speciality of language technology on the PhD level
- The relevant research training is typically carried out under General Linguistics or Computer Science
- The number of PhD student positions has been very limited before 2004 (1-2 in GL, 0-1 in CS)
- Currently, 8 PhD students are specialising in LT (4 in GL, 4 in CS)
  - Individual study plan for every student
    - Obligatory courses 20 AP
    - Optional courses related to the field of specialisation 20 AP
    - PhD thesis 120 AP

PhD studies 2/3

- Optional courses can also be covered by
  - – short courses of visiting professors

    2004 Dr. Graham Wilcock (University of Helsinki) “XML-based document transformations”, Prof. Vadim Stefanyuk (Moscow) “Lisp and artificial intelligence” (supported by Estonian Tiger University)

    2005 February, Prof. Yorick Wilks. **Students of NGSLT are welcome!**
  - – summer schools organised in Tartu
    - 1998 Formal grammars and their applications (8, courses, supported by HESP),
    - 2002 Applications of language technology,
    - 2004 Empirical methods in language technology (2 courses, supported by FW5 programme eVikings II, Estonian Tiger University, and Nordic Treebank Network)
  - – short courses and summer schools abroad (our students have participated in ESSLLI, Finnish GSLT, Swedish GSLT courses, NGSLT, Vilem Mathesius lecture series etc.)

PhD studies 3/3

- 3 PhD theses defended in last 5 years
  - 1999 Heiki-Jaan Kaalep (Creating and use of resources of Estonian in language-technological development work)
  - 2000 Kaili Müürisep (Computational grammar of Estonian: syntax)
  - 2001 Tiina Puolakainen (Computational grammar of Estonian: morphological disambiguation)

Master studies 1/2

- Old model (4+2 years).
    - **Number of tuition free positions is very limited!**

  - Speciality of computational linguistics *on the bachelor level* at the Faculty of Philosophy, started in 1998 (supported by HESP)
    - 6 BA, 1 MA
    - 3 MA students at the moment
  - Some students of general linguistics have been specialised in language technology on the master level
    - 4 MA
  - Some students of computer science are specialising in language technology
    - 8 BSc, 5 MSc since 1999
    - 4 MSc students at the moment

Master studies 2/2

- New model (3+2 years, since 2002/2003)
  - Computational linguistics at the Faculty of Philosophy (3+2) => master of Estonian and finno-ugric linguistics (not MA)
  - Language technology at the Faculty of Mathematics and Computer Science (3+2)=> master of informatics (not MSc)

Course for school children

- Neeme Kahusk and Kadri Vider conducted a training course of computer linguistics in 2002 and 2003 spring term in Hugo Treffner Gymnasium.

PhD studies – personal experience

- Kadri Vider (general linguistics)
- Heli Uibo (computer science)

Different backgrounds

- Kadri –
  - B.A. in Estonian language and literature in 1995
  - M.A. in general linguistics in 1999
  - PhD studies in general linguistics
- Heli –
  - Bachelor's studies in applied mathematics (computer science) 1989-1993
  - M.Sc. in computer science in 1999
  - PhD studies in computer science

PhD courses in CL or LT abroad

Supported by NorFA:

- Graduate School of Language Technology in Finland – 4 students, 3 courses
- Swedish National Graduate School of Language Technolgy – at least 2 students, 3 courses
- the Nordic Graduate School of Language Technology
- Courses in Copenhagen Business School
- Treebank course, a PhD course organized by Nordic Treebank Network (Stockholm University, March 2004) – 2 students

PhD courses in CL or LT abroad

- ESSLLI (European Summer School of Logic, Language and Information)
  - Annual summer school
  - Covers a broad variety of courses ranging from pure linguistics to pure theoretical computer science and logics, through the courses combining these areas (computational linguistics, logic programming, etc.)
  - Participants from University of Tartu (students, whose research topic is within CL or LT):
    - 1998 - 1
    - 1999 – 1
    - 2000 – 3
    - 2001 – 3 (participation of Estonian students supported by NorFA)
    - 2002 – 1
    - 2003 – 2
    - 2004 - 1

- NATO ASI summer school “LT for lesser-studied languages” (Bilkent, Turkey, 2000) – 2 students

PhD courses in CL or LT abroad

- Vilem Mathesius Lecture Series (Charles University, Prague)
  - organized by the Vilem Mathesius Centre for Research and Education in Semiotics and Linguistics
  - 19 lecture series during 1992-2004
  - two intensive weeks with short courses in linguistics and computational linguistics
  - about 20 participants during 1997-2004 from University of Tartu