THE FOLLOWING APPENDICES ARE NOT PART OF THE DESCRIPTION OF WORK. THEY PROVIDE VARIOUS ADDITIONAL INFORMATION THAT WAS REQUESTED BY HTM/ARCHIMEDES OR THAT THE CONSORTIUM FOUND IMPORTANT TO INCLUDE.

A Activities and Results of the TFTs in EXCS 2006-to Date

A.1 Uustalu

Theme No. and Name 0322709s06 *Dependable software and human language technologies*

Duration 2006–2011

Institution Institute of Cybernetics at TUT

Theme leader Tarmo Uustalu

Senior staff Tanel Alumäe (since 1.2007), Hele-Mait Haav, Toomas Kirt (since 1.2008), Mait Harf, Ahto Kalja, Vahur Kotkas, Einar Meister, Jaan Penjam, Jelena Sanko, Olha Shkaravska (until 8.2006), Hellis Tamm, Enn Tyugu, Varmo Vene

Olha Shkaravska was a postdoc from LMU München, she has moved on to Radboud Univ. Nijmegen.

Staff Tanel Alumäe (until 12.2006), Pavel Grigorenko, Kristiina Kindel, Toomas Kirt (10.2006–12.2007), Riina Maigre, Lya Meister, Margus Muskat (until 6.2007), Andres Ojamaa, Ando Saabas, Jaak Simm, Margarita Spitšakova, Andres Toom (since 8.2007)

Goal (from the original proposal) The theme focusses on two important areas of information technology: dependability of software in the context of global computing and user-friendly man-machine interfaces applying human language. The goals are to develop dependable software technologies by improving programming languages and programming language processors and to perfect the Estonian language technology. Accordingly, the project team will develop the composition theory of abstract automata and the semantic theory of effectful and context-dependent notions of computation. It will also develop program logics, program analyses and equivalent type systems for Java virtual machine bytecode, design a corresponding code certificate format and build a certifying Java compiler. It will produce a framework for inducing ontologies. It will further develop the COCOVILA visual programming environment and specialize it for composition of distributed software and reconfiguration. The team will study the variability of the Estonian sound and prosodic system depending on different styles of natural speech (spontaneous speech, dialogue etc). It will investigate models of audiovisual and emotional speech synthesis and create a prototype of an audiovisual speech synthesizer. It will develop statistical models for speech recongnition systems and develop prototypes of man-machine dialogue systems employing speech recognition and synthesis.

Results 2006-date A selecta of works by U. Kaljulaid on semigroups and automata was published (J. Penjam, with J. Peetre), an algorithm for reducing the size of multitape automata was developed, conditions for transition minimality of bideterministic automata were identified (H. Tamm). A categorical analysis based on comonads was developed for the view-update problem of databases was identified (O. Shkaravska). A novel account of tree transducers based on comonads was developed (T. Uustalu, with I. Hasuo, B. Jacobs). A method was given for representing datastructures with cycles and sharing in heterogeneous datatypes via an explicit fixpoint operation (T. Uustalu, V. Vene, with M. Hamana).

Novel type systems for stack-error freedom and secure information flow were developed for a stack-based low-level language. A method was developed for describing program optimizations as type systems with a transformation component (T. Uustalu, A. Saabas). The type-systematic descriptions of program analyses were shown to be applied version of more foundational program logics. A type-systematic account was given for bidirectional program analyses (T. Uustalu, A. Saabas, with M. J. Frade).

A method was devised for ontological annotation of web services and for automatic derivation of composite services (H.-M. Haav et al.). Methods were devised for ontology learning combining web mining and ontology techniques (H.-M. Haav) and for derivation of ontologies from OWL-files (A. Kalja with I. Astrova).

An attribute-based technique was presented for describing the semantics of visual languages of a particular flavour, a logic of higher-order workflows was proposed (E. Tyugu, P. Grigorenko). A monograph was published about algorithms in artificial intelligence. The Cocovila system was enhanced with rich components (A. Ojamaa) and applied to web service composition (R. Maigre). New expertimental software for was developed for computer-aided design (A. Kalja), for modelling and simulation of chain systems and hydraulic systems (M. Harf, with G. Grossschmidt).

The inherent duration and fundamental frequencies of the sounds of Estonian were determined for CVCC (E. Meister). To improve the speech model of the speech recognition prototype, a method for re-evaluating sentence hypotheses was devised (T. Alumäe). The correlation between perceived global accent and pronounciation errors, the role of the duration of a vowel in the perception of its quality and the temporal structure of spontaneous speech were investigated (L. Meister, E. Meister). A method for comparing self-organizing maps for linguistic data was proposed (T. Kirt).

Collaborations, related projects, events 2006-date European projects:

- COST action IC0701 Formal Verification of Object-Oriented Software (16 Dec. 2007/11 March 2008-10 March 2012) (Bernhard Beckert, MC chair; Reiner Hähnle, MC vice-chair; Tarmo Uustalu, MC member for Estonia)
- FP6 IST integrated project 15905 Mobility, Ubiquity, Security for Small Devices, MOBIUS (1 Sept. 2005-31 Aug. 2009) (INRIA, Gilles Barthe, coordinator; IoC, Tarmo Uustalu, partner)
- FP6 IST coordination action 510996 Types for Proofs and Programs, TYPES (1 Sept. 2004-30 April 2008) (Chalmers TU, Bengt Nordström, coordinator; IoC, Tarmo Uustalu, partner)
- FP5 IST thematic network IST-2001-38957 Applied Semantics II, APPSEM II (1 Jan. 2003-30 June 2006) (LMU München, Martin Hofmann, coordinator; IoC, Tarmo Uustalu, partner)

There are many ongoing individual research collaborations with foreign researchers: Hele-Mai Haav - Jørgen Fischer Nilsson (DTU); Ahto Kalja - Hannu Jaakkola (Tampere UT); Einar Meister - Matti Karjalainen, Toomas Altosaar (Helsinki UT), O. Aaltonen (U. of Turku), Stefan Werner (Joensuu U.); Jaan Penjam - Jaak Peetre (Lund U.); Hellis Tamm - Esko Ukkonen (U. of Helsinki); Enn Tyugu - Mihhail Matskin (KTH); Tarmo Uustalu - Thorsten Altenkirch (U. of Nottingham); Luis Pinto, Maria João Frade (U. do Minho); Venanzio Capretta (Radboud U. Nijmegen), Bart Jacobs, Ichiro Hasuo (Radboud U. Nijmegen), Marino Miculan (U. di Udine), Bernd Fischer (U. of Southampton); Varmo Vene - Robin Cockett (U. of Calgary), Makoto Hamana (Tokyo U.), Alberto Pardo (U. de la República, Montevideo), Helmut Seidl (TU München).

Domestic projects:

- HTM target-financed theme SF0322709s06 Dependable Software and Human Language Technologies (Jan. 2006-Dec. 2011) (Tarmo Uustalu, theme leader)
- HTM Estonian CoE in research Centre for Dependable Computing, CDC (Nov. 2002-Dec. 2007) (IoC, coordinator, Jaan Penjam, CoE leader; DCS/TUT, DCE/TUT, DCC/TUT, ICS/UT, TUIT, CybAS, partners)
- ETF grant 7532 Modeling Semantic and Syntactic Dependencies in a Statistical Language Model for Automatic Speech Recognition (Jan. 2008-Dec. 2010, Tanel Alumäe, grant holder)
- ETF grant 7520 Algebraic Automata Theory (Jan. 2008-Dec. 2011) (Hellis Tamm, grant holder)
- ETF grant 7091 Modelling and Simulation of the Dynamics of Technical Chain Systems Described by Multi-Pole Models (Jan. 2007-Dec. 2008) (Mait Harf, grant holder)
- ETF grant 6940 Program Logics, Type Systems and Trustworthy Code Generation (Jan. 2007-Dec. 2010) (Tarmo Uustalu, grant holder)
- ETF grant 6886 Logic-Based Methods for Composition of Distributed Applications (Jan. 2006-Dec. 2009) (Enn Tyugu, grant holder)
- ETF grant 5766 CAD Problem Solving and Technical Systems Modelling Using Distributed Knowledge-Based Systems (Jan. 2004-Dec. 2007) (Ahto Kalja, grant holder)
- ETF grant 5567 Nonclassical Logics and Programming Theory (Jan. 2003-Dec. 2006) (Tarmo Uustalu, grant holder)
- EKKTT project Analysis of Speech and Models of Variability (1 Jan 2006-31 Dec. 2010, Einar Meister, grant holder)
- EKKTT project Resources of Spoken Language and Databases of Speech Technology (1 Jan 2006-31 Dec. 2010, Einar Meister, grant holder)
- EKKTT project Development of Methods for Recognition of Estonian Speech (1 Jan 2006-31 Dec. 2010, Tanel Alumäe, grant holder)
- KM project Simulation Software for Cyberdefence (1 March 2008-31 Dec. 2010) (Vahur Kotkas, grant holder)

- KM project 508/0711 Simulation software for cyber attacks and defence (11 Nov 2007-22 Feb 2008) (Vahur Kotkas, grant holder)
- RAK Measure 2.3 (MKM/EAS) project Development of Infrastructure of the Centre for Dependable Computing, CDC-INFRA (21 Sept. 2005-30 June 2007) (IoC, Jaan Penjam, beneficiary)
- RAK Measure 1.1 (HTM/Innove) project Doctoral School in Information and Communication Technology (1 Sept. 2005-30 June 2008) (TUT, Ennu Rüstern, beneficiary; UT, IoC, CybAS partners)
- RAK Measure 1.1 (HTM/Innove) project Doctoral School in Linguistics and Language Technology (1 Sept. 2005-30 June 2008) (UT, Ennu Rüstern, beneficiary; IoC, partner)
- RAK Measure 1.1 (HTM/Innove) project Establishment of an International Postdoctoral Programme at the Institute of Cybernetics (1 Sept. 2005-30 June 2008) (IoC, Sven Nõmm, beneficiary)

Mait Harf is collaborating with Gunnar Grossschmidt, Ahto Kalja collaborating with Tiit Tiidemann, Irina Astrova. Ahto Kalja is collaborating with Estonian Informatics Centre, contributing specifically to the X-tee project. Enn Tyugu is collaborating with the KV SIVAK R&D unit of the Estonian National Defence College. Jaan Penjam and Hellis Tamm are collaborating with Peeter Normak (Dept. of Informatics, Tallinn U.).

Events organized:

- Summer School 2008 of the NordForsk VISPP network, Kuressaare, 10-16 Aug. 2008 (Einar Meister, local organizer)
- 8th Int. Baltic Conf. on Databases and Information Systems, Baltic DB&IS 2008, Tallinn, 2-5 June 2008 (Hele-Mai Haav, Ahto Kalja, PC cochairs)
- Workshop on Effects and Type Theory, EffTT, Tallinn, 13-14 Dec. 2007 (Tarmo Uustalu, organizer)
- 7th Joint Conf. on Knowledge-Based Software Engineering, JCKBSE 2006, Tallinn, 28-31 Aug. 2006 (Enn Tyugu, PC cochair)
- 8th Int. Conf. on Mathematics of Program Construction, MPC 2006 / 11th Int. Conf. on Algebraic Methodology and Software Technology, AMAST 2006 (incl. Wksh. on Mathematically Structured Functional Programming, MSFP 2006), Kuressaare, 2-8 July 2006 (Tarmo Uustalu, PC chair for MPC, PC cochair for MSFP, Varmo Vene, PC cochair for AMAST)
- 2006, 2007 and 2008 editions of the Estonian Winter School in Computer Science, EWSCS (Palmse)
- biannual Tallinn-Tartu theory days

PhDs defended 2006-date Tanel Alumäe (cosupervised by Einar Meister, Tallinn Univ. of Techn., 2007), Toomas Kirt (Tallinn Univ. of Techn., 2007), Marion Lepmets (cosupervised by Ahto Kalja, Tampere Univ. of Techn., 2007), Meelis Mihkla (supervised by Einar Meister, Univ. of Tartu, 2008)

Target financing 2006 — 2 160 000 EEK; 2007 — 2 454 000 EEK; 2008 — 2 852 400 EEK

A.2 Buldas

Theme No. and Name 0012708s06 *Theoretical and Practical Security of Heterogeneous Information Systems*

Duration 2006–2011

Institution Information Security Research Institute, Cybernetica AS

Theme leader Ahto Buldas

Senior staff Arne Ansper, Margus Freudenthal, Kristo Heero, Sven Laur (since 04.2008), Helger Lipmaa (until 07.2006, since 06.2008), Märt Saarepera, Jan Villemson.

Helger Lipmaa worked at Univ. College London in group at Yvo Desmedt. Sven Laur will defend his PhD degree at Helsinki Univ. of Techn. 25.4.2008.

Staff Dan Bogdanov (since 09.2007), Erki Hermann (since 09.2006), Aivo Jürgenson (since 09.2006), Kristina Kallaste (since 01.2008), Ashok Kumar Kanugula (since 09.2007), Sven Laur (until 03.2008), Martin Pettai, Helen Priisalu (since 09.2006), Jaak Pruulmann-Vengerfeldt, Piret Puus, Uuno Puus, Anu Roos (since 07.2006), Lauri Rätsep (since 09.2006), Rene Tomson.

Goals The overarching goal is to support the development of theories that are able to handle the main security aspects of heterogenous systems, as well as to develop suitable security solutions. In addition to developing new cryptographic protocols and primitives we plan to strengthen the links between security theory and practice by adopting traditional risk analysis methods for measuring security in a quantitative way. Studies are planned in the following main directions:

1. Practice-driven security notions for cryptographic primitives;
2. Constructing secure protocols from simple base primitives;
3. Primitives that are secure in arbitrary environments and configurations;
4. Efficient constructions of cryptographic primitives;
5. Quantitative definitions and evaluation of security;
6. Security in systems with low performance;
7. Developing efficient security management technologies.

Results 2006–date A self-synchronizing authenticated block-cipher mode of operation was proposed, with the ability to resynchronize after the loss of transmission units of sub-block size (Lipmaa, 2006). Privacy-preserving protocols for core data-mining algorithms — Kernel Perceptron and Kernel Adatron — were proposed and private classification protocols and private polynomial kernel computation protocols were given (Laur, Lipmaa, 2006). Practical protocols for data authentication using short manually authenticated out-of-band messages were proposed (Laur, 2006) and extended to group authentication and key agreement (Laur, 2008). An efficient protocol for conditional disclosure of secrets was proposed, where the client's input may be restricted to any set in $\mathbf{NP}/poly$ (Laur, Lipmaa, 2007). The asymptotic behaviour in the mean of a non-commutative rational series, which originates from differential cryptanalysis, was analyzed (Lipmaa 2007). The existence of a two-message argument system for any language in $\mathbf{NP}$ was shown under mild complexity-theoretic assumptions (Lipmaa, 2008). A wide class of commitment schemes were shown to be knowledge-binding, i.e. the list of committed strings that form the published commitment (of fixed size) can be extracted from the code of the committing adversary (Buldas, Laur, 2007). The collision-resistance or even one-wayness of hash functions was shown to be not necessary for the existence of secure hash-function based time-stamping schemes (Buldas, Laur, 2006). The existence of secure time-stamping schemes was shown to not imply the existence of one-way functions at all (Buldas, Jürgenson, 2007). A new game-theoretic framework was developed for analyzing practical security of systems and for deciding about the effectiveness of security measures (Buldas, Saarepera, Villemson, 2006), later extended to deal with fuzzy security estimates (Jürgenson, Villemson, 2007). The same framework was then used to analyze practical security of Estonian elections and the SERVE elections' project in the USA (Buldas, 2007). Counterexamples were provided to a conjecture by Gordon and Loeb that the optimal level of security investment is never higher than about 37% of the protected values (Villemson, 2006). A notion of covering an area was proposed for mobile robot path planning, the set of all minimal covers was described and an efficient algorithm for generating them was constructed (Villemson, 2006). Methods for using the X-Road infrastructure on the international level were proposed (Ansper, Villemson, 2008).

Collaborations, related projects, events 2006-date Cybernetica is a partner in the following EU FP6-FP7 projects:

- AEOLUS (Algorithmic Principles for Building Efficient Overlay Computers), FP6-IST-15964. 09.2005-08.2009.
- BalticTime (Reinforcing eGovernment services in Baltic States throughlegal and accountable Digital Time Stamp), FP6-IST-27751. 01.2006-12.2008.
- HiTS/ISAC (Highway to Security: Interoperability for Situation Awareness and Crisis Management), SEC5-PR-113700, 01.2006-03.2008.
- VirtualLife (Secure, Trusted and Legally Ruled Collaboration Environment in Virtual Life) FP7-ICT-216064. 01.2008-12.2010.

Other international cooperations:

- Sven Laur did his PhD work in Helsinki University of Technology, under the supervision of Kaisa Nyberg.
- Helger Lipmaa has been affiliated with University College London, teaching and co-organizing UCL's MSc programme of Information Security and previously with Helsinki Univ. of Technology.
- Helen Priisalu has close contacts with the data mining research group in Oulu University.

Domestic projects:

- ETF grant 6095 "Universally Composable Security and Its Formalization" (2005-2006, grant holder Peeter Laud)
- ETF grant 6096 "Methods of Game Theory and Risk Analysis in Data Security" (2005-2006, grant holder Jan Villemson)
- ETF grant 6848 "Privacy-Preserving Data-Mining: Cryptographic Methods" (2006-2008, grant holder Helger Lipmaa)
- ETF grant 6944 "Security Proofs for Cryptographic Protocols" (2007-2009, grant holder Peeter Laud)
- ETF grant 7081 "Models of Threat Analysis and Their Applications in Practical Security Evaluation" (2007-2009, grant holder Jan Villemson)
- HTM Estonian CoE in research Centre for Dependable Computing, CDC (Nov. 2002-Dec. 2007) (IoC, coordinator, Jaan Penjam, CoE leader; DCS/TUT, DCE/TUT, DCC/TUT, ICS/UT, TUIT, CybAS, partners)
- RAK Measure 1.1 (HTM/Innove) project Doctoral School in Information and Communication Technology (1 Sept. 2005-30 June 2008) (TUT, Ennu Rüstern, beneficiary; UT, IoC, CybAS partners)
- Cybernetica is a main developer of the Estonian e-voting system (financed by Estonian National Electoral Committee).
- Cybernetica has been a main developer of the secure message exchange system X-Road used as the middleware for making the databases of different state organizations securely available to each other and to the general public (financed by the Department of State Informations Systems).

PhDs defended 2006-date Kristo Heero (cosupervised by Jan Villemson, University of Tartu, 2006), Sven Laur (Helsinki Univ. of Techn., 25.4.2008)

Target financing 2006 — 1 255 000 EEK; 2007 — 1 430 000 EEK; 2008 — 1 800 000 EEK.

A.3 Vilo

Theme No. and Name 0182712s06 *The Methods, Environments and Applications for Solving Large and Complex Computational Problems*

Duration 2006–2011

Institution Dept. of Computer Science, University of Tartu

Theme leader Jaak Vilo

Senior staff Phaedra Agius (since Jan. 2008), Marlon Dumas (since Dec. 2007), Helle Hein, Marina Lepp *née* Issakova (since Sep. 2007), Helle Hein, Härmel Nestra (since Oct. 2006), Ulrich Norbisrath (since Nov. 2006), Rein Prank, Eero Vainikko

Dumas is the group's new professor in software engineering. The position is the first one ever in Estonia's public universities fully financed with privated funds (from Hansabank). The funding was raised by the group. Dumas was selected from among 15 international candidates. His previous position was at Queensland Univ. of Techn. Agius and Norbisrath are postdocs from abroad.

Staff 18 PhD and MSc students

Goal The goal of the research is to develop in an integrated manner novel methods and tools for solving large-scale and complex computational problems on distributed environments like GRID. We will develop methods for formal validation, data security and protection, middleware, as well as algorithms and methods for different applications that require large-scale data analysis. Overall, we will 1) develop data mining, pattern discovery, and machine learning algorithms and tools, 2) continue developing the DOUG solver for solving very large linear equations (Domain Decomposition on Unstructured Grids), 3) develop formal methods and practical approaches for ensuring the correctness, robustness, and data protection of GRID computations, 4) develop end-user interfaces and study user training aspects, and last but not least, 5) apply the developed methods to solve various problems in several application areas, including bioinformatics analysis of gene regulatory networks and gene transcriptional control, computer systems logs analysis, and large database analysis.

Results 2006-date We have developed a new framework for GRID and distributed computing methodologies using peer-to-peer computing and popular instant messaging for the administrative part of setting up the Grid. First successful Monte Carlo simulations and movie rendering applications have been demonstrated in this Friend-to-Friend computing framework (Norbisrath, Vainikko). DOUG development has taken to open source, with University of Bath, and web service interface has been developed (Vainikko). We have developed new bioinformatics tools for mining gene lists for functional relationships g:Profiler (Vilo, Reimand 2007), visualizing high-throughput data KEGGanim (Vilo, Adler), mining large gene networks (graphs based on heterogeneous input data sources) for biologically relevant functional modules GraphWeb (Vilo, Tooming, 2008), a tool Multi-Experiment-Matrix MEM for mining large-scale gene expression data for reconstructing genetic pathways (Vilo, Peterson, Adler, Kolde, Agius), and fast approximate hierarchical clustering of large-scale data HappieClust (Vilo, Kull). Machine learning methods for transcription factor binding site optimization (Tretjakov, Vilo), and factor-gene relationship prediction (Tretjakov, Vilo), are near completion. A visiting summer student Leopold Parts studied the computational identification of conserved micro-RNA genes in 17 fly genomes and published the findings in Genome Research and two fly consortium articles in Nature (2007). Automatic analysis of cryptographic protocols and computational flows has been studied (Laud 2006, 2007), and a tool for an automatic data race analysis of multithreaded C-programs Goblint has been developed (Vojdani, Vene 2006, 2007). Categorical semantics of dynamic programming has been studied (Vene, Kabanov 2006) and different forms of transfinite semantics were developed to avoid the semantic anomaly of program slicing (Nestra 2006, 2007). Computer-aided mathematics learning program T-algebra supporting over 50 types of tasks has been completed and data from first empirical studies collected for analysis (Prank, Lepp). Stochastic oscillators and wavelets have been studied (Hein). Methods for developing process-oriented software systems have been devised and tested (Dumas, 2007). These include methods for transforming high-level business process models into executable code and methods for log analysis of process-oriented software systems to measure business-IT alignment. In-depth evaluation of these methods is ongoing. In parallel, a research stream aimed at evaluating methods for developing forward-compatible presentation components for web applications was started in 2007 (Dumas, Karus).

Collaborations, related projects, events 2006-date European projects:

- COBRED Colon and Breast Cancer Diagnostics (2007-2010) EU FP6 STREP, LSHB-CT-2007- 037730 (coordinator Biosystems International, France)

- ENFIN, Enabling Systems Biology. EU FP6 Network of Excellence (2005-2010) LSHG-CT-2005- 518254 (co-ordinator Ewan Birney, EMBL-EBI, UK)
- ESNATS, Embryonic Stem Cell-Based Alternative Testing Strategies. EU FP7 Collaborative project (2008-2013) (co-ordinator, Univ. Köln)
- ATD, Alternative Transcript Diversity, EU FP6 STREP (2004-2007) LSHG-CT-2003-503329 (co-ordinator Inserm-TAGC, France)
- FunGenES, Functional Genomics of Embryonic Stem Cells, EU FP6 Integrated Project, subcontractor (2006-2007). LSHG-CT-2003-503494, (co-ordinator Jürgen Hescheler, Univ. Köln)
- Baltic GRID (non-funded partner, contributing bioinformatics software for GRID)

Marlon Dumas maintains collaborations with Prof Arthur ter Hofstede (Queensland Univ. of Techn.) with whom he co-supervises 3 PhD students. He also maintains collaborations with the team of Wil van der Aalst (Eindhoven Univ. of Techn.), Marie-Christine Fauvet (Univ. of Grenoble). A PhD student of Prof. Fauvet is undertaking a 4 months stay at University of Tartu. Reciprocally, Marlon Dumas will be visiting professor in Grenoble in summer 2008.

Peeter Laud collaborates with Michael Backes (Univ. des Saarlandes), Eero Vainikko with Robert Scheichl (Univ. of Bath).

Domestic projects:

- HTM Estonian CoE in research Centre for Dependable Computing, CDC (Nov. 2002-Dec. 2007) (IoC, coordinator, Jaan Penjam, CoE leader; DCS/TUT, DCE/TUT, DCC/TUT, ICS/UT, TUIT, CybAS, partners)
- RAK Measure 1.1 (HTM/Innove) project Doctoral School in Information and Communication Technology (1 Sept. 2005-30 June 2008) (TUT, Ennu Rüstern, beneficiary; UT, IoC, CybAS partners)
- Strengthening the ICT MSc programme in Tartu
- ETF grant 7437 Multi-experiment gene expression data matrix analysis (MEM) (2008-2011, Vilo)
- ETF grant 6697 The modeling of dynamical systems by using wavelets and artificial neural networks (2006-2009, Hein)
- ETF grant 6713 Static Analysis of Programs (2006-2009, Vene)
- ETF grant 5766 Re-engineering distributed applications (2008-2011, Vainikko)
- ETF grant 5743 Non-Standard Semantics of Programming Languages (2008-2011, Nestra)
- ETF grant 5724 Bioinformatics of Gene regulation (BiGeR) (2003-2007, Vilo)
- ETF grant 5722 Data Mining Methods and Applications (DMMA) (2003-2006, Vilo)
- Estonian Language Technology: Dictionary informatics (information retrieval) (2005-2008, Vilo)

Events organized: 2006 and 2007 editions of the Estonian Summer School in Computer and Systems Sciences (ESSCaSS, Pedase, Lepanina).

PhDs defended 2006-date Härmel Nestra (supervised by Varmo Vene, Univ. of Tartu, 2006), Marina Lepp (Issakova) (supervised by Rein Prank, Univ. of Tartu, 2007), Maarika Traat (Univ. of Edinburgh, 2006), Phaedra Agius (Rensselaer Polytechnic Institute, Troy, 2007), Ulrich Norbisrath (University of Aachen, 2007)

Target financing 2006 – 1 237 000 EEK; 2007 – 1 405 000 EEK; 2008 – 1 764 000 EEK

A.4 Koit

This theme started only this year. The following is an overview of the past activities of the research staff of the theme.

Theme No. and Name 0180078s08 *Development and implementation of formalisms and efficient algorithms of natural language processing for the Estonian language*

Duration 2008–2013

Institution Dept. of Computer Science, University of Tartu

Theme leader Mare Koit

Senior staff Tiit Hennoste (since Aug. 2008), Päivi Kristiina Jokinen, Neeme Kahusk, Kaarel Kaljurand (since Jan. 2008), Heiki-Jaan Kaalep, Kadri Muischnek, Kaili Müürisep, Tiit Roosmaa, Haldur Õim

Staff Liina Eskor, Olga Gerassimenko, Mark Fišel, Kaarel Kaljurand (until Jan. 2008), Riina Kasterpalu, Helen Nigol, Anton Ragni, Andriela Rääbis, Krista Strandson, Margus Treumuth, Maret Valdisoo, Kaarel Veskis

Overview of the theme The project is a continuation of the earlier target-financed themes (led by professor of general linguistics Haldur Õim) 0180528s98 "The development of computational linguistics tools for Estonian and their application to develop computational resources of the Estonian language" (1998-2002) and 0182541s03 "Computational models and language resources of Estonian: theoretical and practical aspects" (2003-2007) pursued by the research group that unites the researchers and postgraduate students from the Dept. of Computer Science and the Dept. of Estonian and General Linguistics of the University of Tartu. Most of the people who are working on the current project have been the key researchers of one or both of the above research topics, and have also participated in the language technology projects of the national programmes "The Estonian Language and the National Culture" (2002 and 2003), "The Estonian Language and the Nation's Memory" (2004 and 2005) and "Language-Technological Support for the Estonian Language" (since 2006), and in the project eVikings II "Establishment of the Virtual Centre of Excellence for IST RTD in Estonia" (2002-2005) of the EU FP5 IST programme.

Some outcomes of the research group's work (`http://www.cl.ut.ee/`): the collection and annotation of the Estonian text corpora, the corpus of spoken Estonian and the Estonian dialogue corpus, as well as the Estonian-English and English-Estonian parallel corpus, the creation of the database of fixed expressions and the Estonian general thesaurus, that together with references to the English WordNet constitutes the Estonian *wordnet*, being one of the eight *wordnet*-type thesauri that resulted from EuroWordNet-2 project, and is distributed via European Language Resources Association. For the Estonian language, the following models and tools have been developed: a rule- and lexicon-based formal morphology model, as well as the software that implements the model a morphological analyser and generator, a Constraint Grammar based formal syntax model that has been implemented in a surface syntax analyser, and a thesaurus based word meaning disambiguator. We have studied human- human spoken conversation, and worked out the typology of dialogue acts, which is used for the annotation of the dialogue acts in the Estonian dialogue corpus.

The research problems of the current project concentrate on modeling the following four levels of language: morphology, syntax, semantics and pragmatics.

- Morphology

  Problem 1. Changes on the lexical level and modeling them; the coping of the tools for natural language processing with the changes on the lexical level of the actual language use. Goal: developing algorithms for recognition of new words entering the language and words changing their paradigm, as well as identifying the derivational paradigm of these words.

- Syntax

  Problem 2. Fixed expressions as lexical units with their own meaning, government and argument structure. Goal: to study the relationships between the government and argument structure of fixed expressions and the government and argument structure of the 'simple' verb that acts as the nucleus of the given fixed expression. To clarify the possibilities for automatic detection of government.

  Problem 3. The deep syntactic analysis of the sentence. Goal: to find a suitable formalism for the representation of the deep structure of the Estonian sentence, as well as efficient methods both for morphological disambiguation and for the transition to the tree-shaped structure from the flat structure of Constraint Grammar used to date. To adapt the rules of morphological disambiguation for the task of automatic annotation of the Estonian speech corpus. Automatic detection of disfluencies in order to eliminate from syntactic analysis the phrases which do not conform to grammar rules.

- Semantics

  Problem 4. The semantic analysis of the sentence. Goal: developing the conceptual and formal means necessary for constructing the semantic representation of Estonian sentences and discourse.

- Pragmatics

  Problem 5. Dialogue modeling. Goal: to develop a formal model of dialogue that would take into account the general rules of human-human communication, as well as the peculiarities of the Estonian language and culture.

  Problem 6. A language with rich morphology and free word order as the source and/or target language in machine translation. Goal: identify the special needs of a free word order language with rich morphology regarding machine translation, and develop formalisms and methods for successful machine translation from/to such a language.

We base our approach on probabilistic models, and mainly apply machine learning algorithms on corpora of the Estonian language, combining them in appropriate ways with the linguistic rules which have been identified while studying the Estonian language.

All the proposed studies of the Estonian language are unique and unavoidably necessary in order to develop the language technology applications envisaged in "The Strategy of Development of the Estonian Language" (2004-2010) (incl. the speller, term detector, information retrieval, automatic summarisation tools), as such, providing support to the fundamental studies necessary for the goals posed in the national programme "Estonian Language Technology" (EKKTT), and for applications developed in the future.

All the research results will be applicable in the development of new automatic processing software for the Estonian language, as well as for further improvements on the existing software. The main emphasis of the future research will have to move towards developing the formal models of semantics and pragmatics, and towards the shift from the automatic understanding of a single sentence to the understanding of discourse (incl. the dialogue).

Collaboration, related projects Members of group participate in the North-European Association for Language Technology (NEALT), the Special Interest Group on Discourse and Dialogue (SIGdial), the Global Wordnet Association (GWA).

The group is part of the Nordic Graduate School of Language Technology (NGSLT).

There are many ongoing individual research collaborations with foreign researchers: Neeme Kahusk, Haldur Õim, Kaarel Kaljurand – Zurich Univ. (semantic processing); Max Planck Psycholinguistic Institute, Nijmegen (quantitative linguistics), Saarland Univ. (frame semantic tools); Kaili Müürisep, Kadri Muischnek, Tiit Roosmaa, Heiki-Jaan Kaalep – South-Danish Univ. (parsing, treebanks); Växjö and Uppsala universities (treebanks, statistical methods in natural language processing); Stockholm university (parsing, machine translation); Mare Koit – Univ. of Koblenz-Landau, Moscow State Univ. Russian Research Institute of Artificial Intelligence, Taurida Univ. (dialogue and discourse processing); Tiit Hennoste – Univ. of Helsinki (corpora of spoken language).

Domestic projects

- Doctoral School of Linguistics and Language Technology (cooperation with Faculty of Philosophy of the Univ. of Tartu, Institute of Estonian Language, Tallinn Univ of Technology (Institute of Cybernetics)
- ETF grant 7503 Communicative strategies in a communication model: modeling Estonian dialogue on the computer (2008-2011, Koit)
- EKKTT project Collecting and transcribing the corpus of spoken Estonian (2006-2008, Hennoste)
- EKKTT project Corpus query in Keeleveeb (2006-2008, Kaalep)
- EKKTT project Machine translation I (2006-2008, Kaalep)
- EKKTT project Recognition of multi-word verbs in Estonian texts (2006-2008, Kaalep)
- EKKTT project Information dialogue with the computer in Estonian (Koit, 2006-2008)
- EKKTT project Text corpus of Estonian (2006-2008, Muischnek)
- EKKTT project Syntax-based language software and linguistic resources(2006-2008, Roosmaa)
- EKKTT project Semantic analysis of Estonian sentence (2006-2008, Õim)

PhD defended Kaarel Kaljurand (supervised by Norbert Fuchs, Kaili Müürisep, Univ. of Tartu, 10 Jan. 2008)

Target financing 2008 – 1 822 500 EEK.