The Icelandic Centre for Language Technology (ICLT) Annual Report September 2009–August 2010

Sigrún Helgadóttir, Árni Magnússon Institute for Icelandic Studies
Hrafn Loftsson, Reykjavik University
Eiríkur Rögnvaldsson, University of Iceland

1 Introduction

The Icelandic Centre for Language Technology (ICLT) was established on June 15th, 2005. The Centre has its roots in the LangTec Project of the Ministry of Education, Science, and Culture, which lasted from 2000 to 2004. ICLT is run jointly by the *Institute of Linguistics* at the University of Iceland (UI), the *School of Computer Science* at Reykjavik University (RU) and the *Department of Lexicography* at The Árni Magnússon Institute for Icelandic Studies (AMI).

This annual report is meant to give the Language Technology (LT) community in Iceland and elsewhere, and our sponsors and funding agencies, an overview of the activities of ICLT for the operating year September 2009–August 2010.

ICLT serves as a platform for cooperation between the participating institutions. It serves its role by:

- serving as an information centre on Icelandic LT by running the websites `http://www.iclt.is` and `http://www.maltaeknisetur.is`
- encouraging cooperation on LT projects between universities, institutions and private companies
- organising and coordinating university education in LT
- taking part in Nordic, European and international cooperation in the field of LT
- initiating and taking part in research projects in LT
- initiating and taking part in practical projects in LT
- keeping track of resources and products in the field of LT
- holding a bi-annual LT conference with the participation of researchers, companies and the public
- supporting the growth of Icelandic LT in all possible manners

2 Board members

Each member institution appoints one member of the board. The board members for the last operating year were:

- Professor Eiríkur Rögnvaldsson, appointed by the *Institute of Linguistics* at the UI (chairman)
- Dr. Hrafn Loftsson, appointed by the *School of Computer Science* at RU
- Sigrún Helgadóttir, MSc., appointed by the *Department of Lexicography* at the AMI

3 Current researchers and research areas

During the operating year, the following researchers from the member institutions have taken direct part in the activities of ICLT: Anna B. Nikulásdóttir (PhD student, UI), Eiríkur Rögnvaldsson (Professor, UI), Matthew J. Whelpton (Senior Lecturer, UI); Hannes H. Vilhjálmsson (Associate Professor, RU), Hrafn Loftsson (Assistant Professor, RU), Kristín Bjarnadóttir (Research Lecturer, AMI) and Sigrún Helgadóttir (Project Manager, AMI).

During the last few years, these individuals have carried out research in the following areas of LT: context-sensitive spell checking, corpus construction and annotation, lemmatisation, morphology, part-of-speech tagging, shallow parsing, extraction of semantic relations, machine translation, speech recognition, and interactive virtual environments.

As a result of these research projects, a number of BLARK (Basic Language Resource Kit) units have been created for the Icelandic language. Part of the BLARK has been made open source and is currently available at `http://icenlp.sourceforge.net/`.

ICLT had the following visiting researchers during the operating year:

- Joel C. Wallenberg. NSF post-doctoral fellow, UI.
- Francis M. Tyers. PhD student, University of Alicante.

4 Funding

ICLT does not receive any direct funding for its activities. On the other hand, the participating institutions indirectly support the activities of ICLT in the form of facilities for seminars and students, overhead cost in research projects, etc.

During the operating year, the members of ICLT received the following grants:

- Project grant: *An advanced search in Icelandic corpora on the web.* Project manager: Sigrún Helgadóttir (AMI). Co-applicants: Eiríkur Rögnvaldsson (UI), Ásta Svavarsdóttir (AMI). Duration: 3 months. Grant: 0.42 million ISK. Sponsor: The Icelandic Student Innovation Fund
- Project grant: *The correction of OCR text.* Project manager: Kristín Bjarnadóttir (AMI). Co-applicants: Sigrún Helgadóttir (AMI), Ásta Svavarsdóttir(AMI). Duration: 2 months. Grant: 0.94 million ISK. Sponsor: Directorate of Labour in Iceland.
- Project grant: *Digitized books from the 16th to the 19th century.* Project manager: Eríkur Rögnvaldsson (UI). Duration: 2 months. Grant: 0.94 million ISK. Sponsor: Directorate of Labour in Iceland.
- Project grant: *Digitized letters and manuscripts from the 16th to the 19th century.* Project manager: Eríkur Rögnvaldsson (UI). Duration: 2 months. Grant: 0,56 ISK. Sponsor: Directorate of Labour in Iceland.
- Project grant: *A gold standard for Icelandic.* Project manager: Eiríkur Rögnvaldsson (UI). Co-applicants: Hrafn Loftsson (RU), Sigrún Helgadóttir (AMI). Duration: 3 months. Grant: 0.42 million ISK. Sponsor: The Icelandic Student Innovation Fund
- Project grant: *An improved shallow parser for Icelandic text.* Project manager: Hrafn Loftsson (RU). Co-applicants: Eiríkur Rögnvaldsson (UI), Jón Eðvald Vignisson, CLARA. Duration: 3 months. Grant: 0.28 million ISK. Sponsor: The Icelandic Student Innovation Fund
- Project grant: *Icelandic diachronic treebank.* Project manager: Eiríkur Rögnvaldsson (UI). Duration: 2 years. Grant: 0.8 million ISK. Sponsor: University of Iceland Research Fund.

In addition, the following grants awarded to ICLT researchers before September 2009 were still running during the operating year:

- Grant of Excellence: *Viable Language Technology Beyond English – Icelandic as a Test Case*. Project manager: Eiríkur Rögnvaldsson (UI). Co-applicants: Hrafn Loftsson (RU), Matthew J. Whelpton (UI), Kristín Bjarnadóttir (AMI), Anthony Kroch and Joel Wallenberg (University of Pennsylvania), Mikel L. Forcada (Universitat d'Alacant). Duration: 3 years. Grant: 43.5 million ISK. Sponsor: The Icelandic Research Fund.
- Project grant: *A tagged Icelandic corpus*. Project manager: Sigrún Helgadóttir (AMI). Duration: 4 years. Grant: 18.5 million ISK. Sponsor: The Ministry of Education, Science and Culture.

5 Activities

5.1 Service

The following lists participation of the members of ICLT in program committees or reviewing for conferences/journals/grant agencies:

- Kristín Bjarnadóttir: Programme Committee member for *IceTAL 2010 – 7^{th} International Conference on Natural Language Processing*, Reykjavik, Iceland. 2010.
- Sigrún Helgadóttir: Programme Committee member for *IceTAL 2010*.
- Hrafn Loftsson: Programme Committee Chair for *IceTAL 2010*.
- Hrafn Loftsson: Programme Committee member for 7^{th} *SaLTMiL Workshop on "Creation and use of basic lexical resources for less-resourced languages"*, LREC 2010, Valetta, Malta.
- Hrafn Loftsson: Programme Committee member for *FreeRBMT09 – First International Workshop on Free/Open- Source Rule-Based Machine Translation*, Alicante, Spain.
- Eiríkur Rögnvaldsson: Programme Committee member for *IceTAL 2010*.
- Hannes H. Vilhjálmsson: Programme Committee member for *IceTAL 2010*.
- Matthew Whelpton: Programme Committee member for *IceTAL 2010*.

5.2 Membership in international organisations

ICLT is a member of CLARIN (Common Language Resource and Technology Infrastructure; `http://www.clarin.eu/`) and ELRA (European Language Resources Association; `http://www.elra.info/`), and a supporting member of NEALT (Northern European Association for Language Technology; `http://omilia.uio.no/nealt/`).

5.3 International collaboration

During the operating year, ICLT cooperated in the following international projects:

- "Viable Language Technology Beyond English – Icelandic as a Test Case". The main objective is to develop scientific LT methods that are suited for less-resourced languages (`http://iceblark.wordpress.com/`). The project consists of three work packages: A semantic network with semantic mining, a shallow-transfer machine translation system, and the development of parsing strategies and a treebank. International collaborators are from the University of Pennsylvania (USA), and Universitat d'Alacant (Spain).

- "META-NORD". The aim is to establish an open linguistic infrastructure in the Baltic and Nordic countries. An application was submitted to the European *Information and Communication Technologies Policy Support Programme*. International collaborators are from Tilde SIA (Latvia), Københavns Universitet (Denmark), Tartu Ulikool (Estonia), Universitetet i Bergen (Norway), Helsingin yliopisto (Finland), Institute of Lithuanian Language (Lithuania), and University of Gothenburg (Sweden).

5.4 ICLT seminar series

During the operating year, ICLT continued its LT seminar series. Table 1 shows the talks given in the series.

Date	Venue	Title	Lecturer
25/05/2010	RU	NLP-post-processing of OCR-output of business cards	Bettina Harriehausen-Mühlbauer University of Applied Sciences, Darmstadt

Table 1: Talks given during the operating year in the ICLT seminar series.

5.5 LT conference

On April 15^{th} 2010, ICLT organised its third bi-annual LT conference. The conference was held at RU and attended by academics, people from industry, and others interested in LT. The following talks were given:

- *Towards a lexical infrastructure for Swedish language technology.* Lars Borin (Gothenburg University).
- 12^{th} *Century Homilies: The Cutting Edge in Parsing.* Joel Wallenberg, Einar Freyr Sigurðsson, and Anton Karl Ingason (UI).
- *Icelandic machine translation: Recent progress.* Martha Dís Brandt (RU) and Francis Tyers (University of Alicante).
- *Merkingarbrunnur: merkingarupplýsingar með hjálp tölfræðiaðferða.* Matthew Whelpton and Anna Nikulásdóttir (UI).
- *Að skilja umræðu á netinu – hagnýting íslenskra málgreiningartóla.* Jón Eðvald Vignisson (CLARA).
- *Beygingarlýsing íslensks nútímamáls og verðlaunasamkeppnin "Þú átt orðið".* Kristín Bjarnadóttir (AMI), Hjalmar Gislason (Data Market), Borgar Þorsteinsson, Stefán Ingi Valdimarsson and Tihomir Rumenov Rangelov.

5.6 IceTAL 2010 – 7^{th} International Conference on Natural Language Processing

ICLT organised IceTAL 2010, an international conference on NLP, which was held at RU, August 16-18 2010. IceTAL was the seventh in the series of the TAL conferences, following GoTAL 2008 (Gothenburg, Sweden), FinTAL 2006 (Turku, Finland), EsTAL 2004 (Alicante, Spain), PorTAL 2002 (Faro, Portugal), VexTAL 1999 (Venice, Italy) and FracTAL 1997 (Besançon, France).

Our program committee (PC) consisted of 45 recognized researchers and professionals in the field of NLP from Belgium, China, Cuba, Denmark, Finland, France, Germany, Iceland, Italy, Japan, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, and the United States.

We called for submissions both from academia and the industry on any topic that is of interest to the NLP community, particularly encouraging research emphasizing multidisciplinary aspects of NLP and the interplay between linguistics, computer science and application domains such as biomedicine, communication systems, public services, and educational technology.

As a response, we received 91 submissions from authors representing 37 countries in Europe, Asia, Africa, Australia, and the Americas. Each submission was reviewed by three PC members or external reviewers designated by the PC members. The reviewing process led to the selection of 43 papers (30 full papers and 13 short papers) which were presented at the IceTAL conference.

In addition, two invited talks were given at IceTAL 2010:

- "Reliving the History: The Beginnings of Statistical Machine Translation and Languages with Rich Morphology". Jan Hajic, Charles University, Prague, Czech Republic.
- "Harmonizing WordNet and FrameNet". Christiane D. Fellbaum, Princeton University, Princeton, USA.

Registered participants at IceTAL were about 60. The preparation for this conference took up signifiant time of the activities of ICLT during the last 12 months. Further information about IceTAL can be found at the conference website `http:\\icetal.ru.is`.

5.7 Masters program in LT

In fall 2007, ICLT started its Masters program in LT. The program, which is run jointly by RU and the UI, is a two year interdisciplinary program (120 ECTS credits).

Students must either have a BA degree in the humanities (languages and linguistics) or a BS degree in computer science or related subjects (such as electrical and computer engineering). In the past, students have had the possibility of taking courses at foreign universities participating in the Nordic Graduate School of Language Technology (NGSLT). Unfortunately, the NGSLT activities ended in September 2009 and the future of ICLT's masters program is therefore uncertain.

5.8 Courses

The following courses, which are part of the joint LT program, were taught by members of ICLT during the operating year:

- Natural Language Processing. Fall 2009. Teachers: Hrafn Loftsson and Hannes Högni Vilhjálmsson (RU).
- Parsing and parsing methods. Spring 2010. Teacher: Joel C. Wallenberg (UI).

5.9 Supervision of students

During the operating year, members of ICLT supervised the following BSc/MSc/PhD students working on LT theses or projects:

- Anna B. Nikulásdóttir (UI), a Phd student in LT. Thesis: *A Semantic Database for Icelandic Language Technology.* Supervisor: Matthew Whelpton.
- Martha Dís Brandt (RU), an MSc student in LT. Thesis: *Developing a shallow-transfer translation system using existing open-source tools.* Supervisor: Hrafn Loftsson.
- Svavar K. Lúthersson (RU), a BSc student in Computer Science. Final project: *Tagging and parsing a large corpus.* Supervisor: Hrafn Loftsson.
- Jökull H. Yngvason (RU), a BSc student in Computer Science. Independent study: *Automatic extraction of verb subcategorization frames for Icelandic.* Supervisor: Hrafn Loftsson.
- Ragnar L. Sigurðsson (RU), a BSc student in Computer Science. The Icelandic Student Innovation Fund: *An Improved Shallow Parser for Icelandic.* Supervisor: Hrafn Loftsson.

- Guðmundur Ö. Leifsson (UI), a BSc student in Computer Science. The Icelandic Student Innovation Fund: *An advanced search in Icelandic corpora on the web.* Supervisor: Sigrún Helgadóttir.
- Kristján F. Sigurðsson (UI), a BA student in Icelandic. The Icelandic Student Innovation Fund: *A gold standard for Icelandic.* Supervisor: Eiríkur Rögnvaldsson.
- Jón F. Daðason (UI), a MSc student in Computer Science. Directorate of Labour in Iceland: *The correction of OCR text.* Supervisor: Kristín Bjarnadóttir.
- Kristján Rúnarsson (Utrecht Conservatory), a BMus student. Directorate of Labour in Iceland: *The correction of OCR text.* Supervisor: Kristín Bjarnadóttir.

  In addition to the students mentioned above, Anton Karl Ingason (UI), an MA student in Language Technology, and Einar Freyr Sigurðsson (UI), an MA student in Icelandic linguistics, worked as research students in the Treebank work package of the project *Viable Language Technology Beyond English – Icelandic as a Test Case.*

6 Publications

6.1 Peer reviewed papers

In the period covered by this report, researchers in ICLT published the following peer reviewed papers in the field of LT:

- Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir and Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the* 7^{th} *International Conference on Language Resources and Evaluation, LREC 2010.* Valetta, Malta.
- Anna Björk Nikulásdóttir and Matthew Whelpton. 2010. Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the* 7^{th} *International Conference on Language Resources and Evaluation, LREC 2010.* Valetta, Malta.

6.2 Editorial work

- Hrafn Loftsson, Eiríkur Rögnvaldsson and Sigrún Helgadóttir (Eds). 2010. Advances in Natural Language Processing. 7^{th} International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 2010, Proceedings. © 2010 Springer.

7 Forthcoming activities

During the next twelve months, we plan to continue our work with the aim of achieving the objectives stated in section 1.

Due to the reasons stated in Section 5.7, we will not be able to admit new students into our LT Masters program. We will, however, try to seek collaboration with foreign universities, with regard to LT education, by other means.

8 Summary

In addition to the regular activities of ICLT, like organising a seminar series, running a LT Masters program, supervising students, organising the bi-annual LT conference, and working on research projects, the main activity of ICLT during the operating year was the planning and organisation of the IceTAL 2010 conference.

The conference was a success as witnessed by the high quality of accepted papers and the numerous positive comments acknowledged by the participants. IceTAL has definitely helped ICLT to become better known in the global LT community.

Further information is available from the ICLT web page at `http://www.iclt.is`.