The Icelandic Centre for Language Technology (ICLT) Annual Report September 2011–August 2012

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 2011–August 2012.

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

The following researchers from the member institutions are affiliated with 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 (Associate Professor, RU), Jón Guðnason (Assistant Professor, RU), Kristín Bjarnadóttir (Research Lecturer, AMI), Kristín M. Jóhannsdóttir (UI) 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 synthesis, interactive virtual environments and computer-assisted language learning.

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 for LT projects:

- Project grant: *Fjölnir fyrir hvern mann* [The Journal Fjölnir (1835-1847) made accessible]. Project manager: Kristín Bjarnadóttir, AMI. Participants: Jón Friðrik Daðason and Kristján Rúnarsson. Duration: 3 months. Grant: 1,020,000 ISK. Sponsor: The Icelandic Student Innovation Fund.
- Project grant: *Software for manual correction of scanned texts.* Applicant: Guðrún Kvaran, AMI. Co-applicant and project leader: Sigrún Helgadóttir, AMI. Grant: 1,100,000 ISK. Sponsor: The University of Iceland Research Fund.

In addition, the following grants awarded to ICLT researchers before September 2011 were still running during the operating year:

- Project grant: *Baltic and Nordic Parts of the European Open Linguistic Infrastructure (META-NORD).* Project manager: Tilde SIA (Latvia). Co-applicants: Københavns Universitet (Denmark), Tartu Ulikool (Estonia), Universitetet i Bergen (Norway), Helsingin yliopisto (Finland), Háskóli Íslands (ICLT; Iceland), Institute of Lithuanian Language (Lithuania), Göteborgs Universitet (Sweden). Duration: 2 years. Grant: 2,250,000 EUR. Sponsor: The ICT Policy Support Program.
- Project grant: *A System Architecture for Intelligent CALL.* Project manager: Hrafn Loftsson, Reykjavik University. Co-applicants: Lars Borin, University of Gothenburg; Birna Arnbjörnsdóttir, University of Iceland. Duration: 2 years. Grant: 38,500 EUR. Sponsor: NordPlus Sprog.

5 Activities

5.1 Service

During the operating year, the members of ICLT participated in editorial boards, program committees or reviewing for conferences/journals/grant agencies in the field of LT:

- Sigrún Helgadóttir: Programme Committee member for *CHAT 2012 – The 2^{nd} Workshop on the Creation, Harmonization and Application of Terminology Resources.* Madrid, Spain.
- Hrafn Loftsson: Editorial Board member for the *Northern European Journal of Language Technology.*
- Hrafn Loftsson: Reviewer for the Icelandic journal *Gripla.*

- Hrafn Loftsson: Programme Committee member for *SaLTMiL-AfLaT Workshop on "Language technology for normalisation of less-resourced languages", LREC 2012*, Istanbul, Turkey.
- Hrafn Loftsson: Programme Committee member for *FreeRBMT12 – Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, Gothenburg, Sweden.
- Hrafn Loftsson: Programme Committee member for *JapTAL 2012 – 8^{th} International Conference on Natural Language Processing*, Kanazawa, Japan.
- Eiríkur Rögnvaldsson: Programme Committee member for *The Fifth International Conference on Human Language Technologies*, Tartu, Estonia.
- Eiríkur Rögnvaldsson: Programme Committee member for *LThist2012, First International Workshop on Language Technology for Historical Text(s)*, Vienna, Austria.
- Eiríkur Rögnvaldsson: Programme Committee member for *META-RESEARCH Workshop on Advanced Treebanking, LREC 2012*, Istanbul, Turkey.

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:

- "META-NORD" (`http://vefir.hi.is/metanord/`). The aim is to establish an open linguistic infrastructure in the Baltic and Nordic countries. International collaborators: 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).
- "A System Architecture for Intelligent CALL". The aim is to design and develop open system architecture for supporting ICALL systems. International collaborator: University of Gothenburg (Sweden).
- "Almannarómur" (`http://almannaromur.hr.is/`). The aim is to collect recordings of spoken Icelandic, and make them open and available for research and develoment. Collaborator: Google Inc.
- "Íslenskur talgervill" (an Icelandic text-to-speech engine). The aim is to develop a new text-to-speech engine for Icelandic, whose quality is comparable to the best systems in foreign languages. Collaborators: The Icelandic organization for the blind and visually impaired and Ivona Software (Poland).

5.4 ICLT seminar series

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

5.5 The LT Conference day

On April 27^{th} 2012, ICLT held its 4^{th} bi-annual LT conference day. This time, it was organized in collaboration with "Íslensk málnefnd" (The Icelandic Language Council) and META-NORD. The name of the conference was "Máltækni fyrir alla" (Language Technology for everyone) and it refers, at the one hand, to the policy of the Icelandic Language Council that LT should be accessible to everyone, and, on the other

Date	Venue	Title	Lecturer
08/11/2011	UI	Multimodal conversation analysis of institutionalized political TV interview	Sigrún M. Ammendrup, Reykjavik University
06/12/2011	RU	Tagging Icelandic: The development of recent years	Hrafn Loftsson, Reykjavik University
31/01/2012	UI	MerkOr: A semantic net for Icelandic Language Technology	Anna Björk Nikulásdóttir, University of Iceland
29/03/2012	UI	An example of corpus-driven quantitative approaches to the study of linguistic variation in English	Javier Pérez-Guerra, University of Vigo

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

hand, to the fact that the conference should be of interest to everyone bearing an interest in LT. The conference day was held at the UI and attended by academics, people from industry, and others interested in LT.

The following talks were given:

- *Íslensk máltækni í evrópsku samhengi – META-NORD og META-NET* [Icelandic LT in an European perspective – META-NORD and META-NET]. Eiríkur Rögnvaldsson, UI.
- *Del eller dø?* [Share or die?]. Sabine Kirchmeier-Andersen, Danish Language Council.
- *Íslenska er málið: Tölvur og íslensk málstefna* [Computers and the Icelandic Language Policy]. Haraldur Bernharðsson, Icelandic Language Council.
- *Gagnasöfn frá sjónarhóli notandans* [Language Resources from a user's perspective]. Kristín M. Jóhannsdóttir, UI.
- *Talgervlar og tungumál sem fáir tala* [Text-to-speech systems and languages spoken by few]. Kristinn H. Einarsson, The Icelandic organization of the visually impaired.
- *Almannarómur – Söfnun á íslensku talmáli fyrir talgreiningu* [Almannarómur – collecting Icelandic spoken data for speech recognition]. Jón Guðnason, RU.
- *Samhengisháð ritvilluvörn* [Context-sensitive spelling correction]. Jón F. Daðason, UI.
- *"Hér er ég, um ég, frá ég" – Mikilvægi fallbeygingar í leitarvélum* [The importance of inflection in search engines]. Jón E. Vignisson, CLARA.

5.6 Other talks

During the operating year, the members of ICLT gave the following other talks on LT issues:

- *Orðin í Markaðri íslenskri málheild og íslenskur orðaforði* [The words of the Tagged Icelandic Corpus and the vocabulary of Icelandic]. Kristín Bjarnadóttir, AMI. Presentation at the Linguistic Society of Iceland, UI, January 20^{th}, 2012.
- *Málheild sem hluti af orðabókarlýsingu* [Tagged Corpus as a part of a lexicographic description]. Sigrún Helgadóttir, AMI. Presentation at the seminar "Aðgengi að orðaforðanum" [Accessing vocabulary] at the annual conference of the humanities ("Hugvísindaþing"), UI, March 10^{th}, 2012.
- *Glíman við orðmyndirnar* [Tackling word forms]. Kristín Bjarnadóttir, AMI. Presentation at the seminar "Aðgengi að orðaforðanum" [Accessing vocabulary] at the annual conference of the humanities ("Hugvísindaþing"), UI, March 10^{th}, 2012.

- *Sögulegur íslenskur trjábanki og nýting hans* [An Historical Icelandic Treebank and its utilization]. Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson and Joel Wallenberg. Talk presented at the seminar "Gagnagrunnar í málfræði" [Databases in linguistics] at the annual conference of the humanities ("Hugvísindaþing"), UI, March 10^{th}, 2012.

- *Gagnagrunnur um sagnflokka og táknun rökliða* [The Verb Class and Argument Structure Database]. Jóhannes Gísli Jónsson and Matthew Whelpton, UI. Talk presented at the seminar "Gagnagrunnar í málfræði" [Databases in linguistics] at the annual conference of the humanities ("Hugvísindaþing"), UI, March 10^{th}, 2012.

- *Hvert á að sækja orðaforðann?* [Where should we get vocabulary from?]. Kristín Bjarnadóttir, AMI. Presentation at the seminar "Íslenska sem viðfangsmál í íslensk-erlendum orðabókum. Sjónarmið við aðferðir og öflun, val og framsetningu efnis" [Icelandic as the target language in bilingual Icelandic dictionaries. Criteria for methods, collection, selection and presentation]. Organized by the journal *Orð og tunga*, AMI, May 4^{th}, 2012.

5.7 Supervision of students

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

- Anna B. Nikulásdóttir (UI), a Phd student in LT. Thesis: *A Semantic Database for Icelandic Language Technology*. Supervisor: Matthew Whelpton.

- Guðmundur Örn Leifsson (UI), an MSc student in Computer Science. Thesis: *A System Architecture for Intelligent Computer-Assisted Language Learning*. Supervisor: Hrafn Loftsson.

- Jón Friðrik Daðason (UI), an MSc student in Computer Science. Thesis: *Context-sensitive spelling correction of OCR text*. Supervisors: Sven Þ. Sigurðsson and Kristín Bjarnadóttir.

- Ólafur Waage (RU), a BSc student in Computer Science. Independent study: *Continued development of Apertium-IceNLP: A rule-based Icelandic to English machine translation system.*.

5.8 MSc/PhD Committes

During the operating year, members of ICLT took part in committee work for the following MSc/PhD students working on theses in the field of LT:

- Anna B. Nikulásdóttir (UI), a Phd student in LT. Thesis: *A Semantic Database for Icelandic Language Technology*. Committe members: Eiríkur Rögnvaldsson, Hrafn Loftsson.

- Jón Friðrik Daðason (UI), an MSc student in Computer Science. Thesis: *Context-sensitive spelling correction of OCR text*. Committe member: Hrafn Loftsson.

6 Publications

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

- Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *Proceedings of the SaLTMiL-AfLaT Workshop on "Language technology for normalisation of less-resourced languages", 8^{th} International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey.

- Jón Guðnason, Oddur Kjartansson, Jökull Jóhannsson, Elín Carstensdóttir, Hannes H. Vilhjálmsson, Hrafn Loftsson, Sigrún Helgadóttir, Kristín Jóhannsdóttir and Eiríkur Rögnvaldsson. 2012. Almannaromur: An Open Icelandic Speech Corpus. In *Proceedings of the Third International Workshop on Spoken Language Technologies for Under-resourced languages (SLTU 2012)*. Cape Town, South Africa.
- Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir and Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). In *Proceedings of the SaLTMiL-AfLaT Workshop on "Language technology for normalisation of less-resourced languages"*, 8^{th} *International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey.
- Hrafn Loftsson, Sigrún Helgadóttir and Eiríkur Rögnvaldsson. 2011. Using a morphological database to increase the accuracy in PoS tagging. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria.
- Anna B. Nikulásdóttir. 2012. Tölvutækur merkingarbrunnur fyrir íslenska máltækni – Grunnur lagður að því að tölvur skilji merkingu í íslenskum textum. *Orð og tunga*, **14**:19-38.
- Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the* 8^{th} *International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey.
- Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson and Joel Wallenberg. 2011. Creating a Dual-Purpose Treebank. In *Proceedings of the ACRH Workshop, Heidelberg. Journal for Language Technology and Computational Linguistics*, **26(2)**:141-152.
- Bolette S. Pedersen, Lars Borin, Markus Forsberg, Krister Lindén, Heili Orav and Eiríkur Rögnvaldsson. 2012. Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD. In *Proceedings of the Global Wordnet Conference*, Matsue, Japan.
- Inguna Skadina, Andrejs Vasiljevs, Lars Borin, Koenraad De Smedt, Krister Lindén and Eiríkur Rögnvaldsson. 2011. META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries. In *Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm*. Chiang Mai, Thailand.
- Andrejs Vasiljevs, Markus Forsberg, Tatiana Gornostay, Dorte H. Hansen, Kristín M. Jóhannsdóttir, Krister Lindén, Gunn I. Lyse, Lene Offersgaard, Ville Oksanen, Sussi Olsen, Bolette S. Pedersen, Eiríkur Rögnvaldsson, Roberts Rozis, Inguna Skadina og Koenraad de Smedt. 2012. Creation of an Open Shared Language Resource Repository in the Nordic and Baltic Countries. In *Proceedings of the* 8^{th} *International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey.
- Matthew Whelpton. 2012. From human-oriented dictionaries to computer-oriented lexical resources – trying to pin down words. *Orð og tunga*, **14**: 1-18.

7 Summary

During the operating year, the members of ICLT received two new grants for research and development in the field of LT. Two projects were funded with grants from previous years, and, in addition, ICLT took part in new international projects without any special funding.

One of the main obstacles to further development of LT in Iceland is the lack of funding. Some European countries have established long-lasting specially funded programs for research and development of LT, as well as for LT education, e.g. Estonia, Finland and Sweden. As mentioned in Section 1, the Icelandic Ministry of Education, Science and Culture established the LangTec Project which lasted from 2000 to 2004. This project was short, but laid the foundation of LT in Iceland.

Now it is imperative that the authorities establish a new program. Without such a program, it will be difficult to make the Icelandic language function successfully in the Information Society.

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