

The Icelandic Centre for Language Technology
(ICLT)
Annual Report
September 2007–August 2008

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

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 2007–August 2008.

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 a website <http://www.tungutaekni.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: Eiríkur Rögnvaldsson (UI), Hannes Högni Vilhjálmsón and Hrafn Loftsson (RU), Kristín Bjarnadóttir and Sigrún Helgadóttir (AMI). Furthermore, Anton Karl Ingason, an undergraduate student in Icelandic language at the UI, worked closely with researchers in ICLT on many of the activities.

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, 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 (e.g. see <http://nlp.ru.is>).

4 Funding

ICTL 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: *Developing a shallow-transfer translation system using existing tools*. Project manager: Hrafn Loftsson (RU). Collaborators: Eiríkur Rögnvaldsson (UI) and Mikel L. Foracada (University d’Alicant). Duration: 2 years. Grant: 2.0 million ISK. Sponsor: RU Development Fund.

In addition, the following grants awarded to ICLT researchers before September 2007 are still running:

- 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.
- Project grant: *Context-sensitive spell checking*. Project manager: Eiríkur Rögnvaldsson (UI). Collaborators: Hrafn Loftsson (RU) and Sigrún Helgadóttir (AMI). Duration: 2 years. Grant: 3.1 million ISK. Sponsor: The Icelandic Research Fund.
- Project grant: *Improved tagging accuracy of Icelandic Text*. Project manager: Hrafn Loftsson (RU). Collaborators: Eiríkur Rögnvaldsson (UI) and Sigrún Helgadóttir (AMI). Duration: 2 years. Grant: 3.1 million ISK. Sponsor: The Icelandic Research Fund.

We remark that these grants, however, can only be used to support project specific activities, and not for activities related to the Centre as such.

5 Activities

5.1 Service

Eiríkur Rögnvaldsson is active in the international research community in the field of LT. The following lists international service activities contributed by him during the operating year:

- Vice-President, NEALT (North-European Association for Language Technology)
- Member, Language Technology Working Group (Nordic Language Council)
- Board member, NGSLT (Nordic Graduate School of Language Technology)

During the operating year, ICLT wrote a draft of a chapter about LT and software translations with regard to *Íslensk málstefna* [The Icelandic language strategy] compiled by *Íslensk málnefnd* [The Icelandic Language Committee] (<http://www.islenskan.is/>).

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/>).

Both RU and UI are members of NGSLT.

5.3 International collaboration

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

- “Change and variation in Icelandic syntax”. A workshop with the aim of devising one or more collaborative research projects on the topic for which research funding could be sought in the US, Iceland and perhaps Europe. Main organisers were Anthony Kroch (University of Pennsylvania) and Sigríður Sigurjónsdóttir (UI). Participants were from USA, Iceland, UK, and the Faroe Islands.
- An application to NordForsk (main applicant: Kimmo Koskeniemi, University of Helsinki): “Nordic Language Resource Infrastructure - preparatory phase”.

5.4 ICLT seminar series

During the operating year, ICLT continued its LT seminar series. Each talk was attended by 20-45 people. Table 1 shows the talks given in the series.

Date	Venue	Title	Lecturer
06/11/2007	RU	Automatic thesaurus extraction for Icelandic	Frank A. B. Cassata, RU
11/12/2007	UI	BIN (Morphological Database of Inflections) and its children	Hjálmar Gíslason, Síminn ehf. Kristín Bjarnadóttir, AMI
01/08/2008	RU	The Webspeaker (Vefþulan)	Helga Waage, Hexia.net
07/03/2008	RU	Language technology for Sámi, Faroese and Greenlandic: A presentation of ongoing work in Tromsø	Trond Trosterud, University of Tromsø
07/03/2008	RU	Grammar-based language technology as an answer to the challenges facing Icelandic and other circumpolar languages	Trond Trosterud, University of Tromsø
20/05/2008	RU	Electronic Grammar Checking in Natural Languages	Bettina Harriehausen-Mühlbauer, TH Darmstadt
26/06/2008	AMI	You shall know a word by the company it keeps	Lou Burnard, Oxford University

Table 1: Talks given during the operating year

Furthermore, on March 7th 2008, ICLT, together with *Íslensk málnefnd* organised a symposium on Language Technology and Software Translations. The event took place at RU and was attended by about 60 people.

5.5 LT conference

On April 18th 2008, ICLT organised its second LT conference. The conference was held at the UI and attended by academicians, people from industry, and others interested in LT. The following talks were given:

- *Cells and Culture: Disambiguation of Biomedical Documents*. Mark Stevenson, University of Sheffield.
- *Lemmald: A new lemmatiser for Icelandic*. Anton Karl Ingason, UI.
- *Representing Function and Behaviour in Multimodal Communication*. Hannes Högni Vilhjálmsson, RU.
- *Open-Source Machine Translation for Icelandic: the Apertium Platform as an Opportunity*. Mikel L. Forcada, Universitat d'Alicant.
- *Context-Sensitive Spell Checking*. Anton Karl Ingason and Skúli Bernhard Jóhannsson, UI.

5.6 Masters program in LT

In Fall 2007, ICLT started its new 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). Furthermore, students have the possibility of taking courses at foreign universities, especially those participating in the Nordic Graduate School of Language Technology (NGSLT).

Two students were admitted to the program in Fall 2007.

5.7 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 2007. Teachers: Hrafn Loftsson and Hannes Högni Vilhjálmsson.
- Linguistics in Language Technology. Fall 2007. Teacher: Eiríkur Rögnvaldsson.

In addition, ICLT organised an NGSLT course on Information Extraction in Spring 2008. The course, which was held at RU, was attended by about 35 students from Estonia, Finland, Iceland, Latvia, Lithuania, Norway, Russia and Sweden.

The instructor was Dr. Mark Stevenson, University of Sheffield, and Hrafn Loftsson was the main organiser.

5.8 Supervision of students

During the operating year, the members of ICLT supervised the following BSc/MSc students:

- Ida Kramarczyk (RU), an MSc student in LT. MSc project: *Improved Tagging Accuracy of Icelandic text*, funded by the The Icelandic Research Fund. Supervisor: Hrafn Loftsson.
- Verena Henrich and Timo Reuter (RU/UAS Darmstadt), MSc exchange students in Computer Science. Project: *CombiTagger: A System for Developing Combined Taggers*. Supervisor: Hrafn Loftsson.

6 Publications

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

- Ingason, A. K. and Helgadóttir, S. and Loftsson, H. and Rögnvaldsson, E. 2008. A Mixed Method Lemmatization Algorithm Using Hierachy of Linguistic Identities (HOLI). In *B. Nordström and A. Raante (eds.), Advances in Natural Language Processing, 6th International Conference on NLP, GoTAL 2008, Proceedings*. Gothenburg, Sweden. *Advances in Natural Language Processing*, pp. 205–216 (Lecture Notes in Computer Science, Vol. 5221.) Springer, Berlin.
- Loftsson, H. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, **31(1)**:47–72.
- Loftsson, H. and Rögnvaldsson, E. 2008. Linguistic richness and technical aspects of an incremental finite-state parser. In *Proceedings of 'Partial Parsing 2008', workshop at the 6th International Conference on Language Resources and Evaluation, LREC 2008*. Marrakech, Morocco.
- Rögnvaldsson, E. 2008. Icelandic Language Technology Ten Years Later. In *Proceedings of 'Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages', SALT MIL workshop at the 6th International Conference on Language Resources and Evaluation, LREC 2008*. Marrakech, Morocco.
- Rögnvaldsson, E. and Helgadóttir, S. 2008. Morphological Tagging of Old Norse Texts and Its Use in Studying Syntactic Variation and Change. In *Proceedings of the 2nd Workshop on Language Technology for Cultural Heritage Data, 6th International Conference on Language Resources and Evaluation, LREC 2008*. Marrakech, Morocco.

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. Our LT Masters program will start for the second time in Fall 2008. The program will hopefully result in more LT research projects being carried out in the field of LT in Iceland.

Furthermore, ICLT will try to continue attracting foreign speakers to its seminar series.

Finally, in August 2008 ICLT decided to host IceTAL 2010, the 7th International Conference on Natural Language Processing, in Reykjavik¹.

8 Summary

During the past year ICLT has expanded its activities and achieved a level of recognition in the local LT community. By inviting foreign speakers to its seminar series, organising an LT conference and an NGSLT course, and last but not least, publishing articles in international journals and conferences, ICLT is gradually becoming known in the global LT community as well.

Further information is available from the ICLT web page at <http://www.tungutaekni.is>.

¹The last TAL conference was held in Gothenburg in August 2008 (<http://www.cse.chalmers.se/gotal/>).