Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS

Anna Björk Nikulásdóttir[1], Þórunn Arnardóttir[2], Jón Guðnason[3], Þorsteinn Daði Gunnarsson[3], Anton Karl Ingason[2], Haukur Páll Jónsson[4], Hrafn Loftsson[3], Hulda Óladóttir[4], Einar Freyr Sigurðsson[5], Atli Þór Sigurgeirsson[6], Vésteinn Snæbjarnarson[4], and Steinþór Steingrímsson[5]

[1]Grammatek ehf., Iceland, [2]University of Iceland, [3]Reykjavik University, [4]Miðeind ehf., Iceland, [5]The Árni Magnússon Institute for Icelandic Studies, [6]University of Edinburgh
anna@grammatek.com, thar@hi.is, jg@ru.is, thorsteinng@ru.is, antoni@hi.is, haukurpj@mideind.is, hrafn@ru.is, hulda@mideind.is, einar.freyr.sigurdsson@arnastofnun.is, atlisigurgeirsson@gmail.com, vesteinn@mideind.is, steinthor.steingrimsson@arnastofnun.is

Abstract

In this paper, we describe how a fairly new CLARIN member is building a broad collection of national language resources for use in language technology (LT). As a CLARIN C-centre, CLARIN-IS is hosting metadata for various text and speech corpora, lexical resources, software packages and models. The providers of the resources are universities, institutions and private companies working on a national (Icelandic) LT infrastructure initiative.

1 Introduction

With the enormous progress in language technology (LT) in the last decades, the use of LT in research and commercial products has greatly increased. LT tools and resources are now not only used by LT specialists but also by researchers and developers from various fields. Beside the improvement in quality and usability, this development is driven by open access to data and software. For such resources to be of broad use, they need to be easily accessible and thoroughly documented. Thus, the large national LT infrastructure initiative *Language Technology Programme for Icelandic (LTPI) 2019–2023* (Nikulásdóttir et al., 2020) chose CLARIN-IS to be the central hub for all deliverables of the programme.

This paper gives a broad overview of the available repositories and the core publishing guidelines.

2 CLARIN-IS

Iceland became a CLARIN ERIC member on February 1, 2020 after having an observer status since November 1, 2018. The Árni Magnússon Institute for Icelandic Studies is the leading partner in the Icelandic national consortium. A Metadata Providing Centre (CLARIN C-centre[1]) has been established at the institute that hosts metadata for Icelandic language resources and distributes them through a Virtual Language Observatory.

As a new member, CLARIN-IS is in the process of establishing a technical Service Providing Centre (CLARIN B-centre), which will maintain language resources among other tasks. For now, we maintain a Gitlab[2], where all relevant GitHub repositories are mirrored, and deliver all resources to the C-centre.

[1]https://clarin.is/
[2]https://gitlab.com/icelandic-lt

3 Language Technology Programme for Icelandic

In October 2019, a consortium of Icelandic universities, companies and institutions (10 in total) started working on the LTPI. The programme aims at making Icelandic viable in future technologies that rely on LT in one way or another. To build foundations for that goal, the LTPI concentrates on developing language resources and infrastructure software, divided into six core project areas: 1) Language Resources, 2) Support Tools, 3) Machine Translation, 4) Spell and Grammar Checking, 5) Automatic Speech Recognition, and 6) Speech Synthesis.

During the preparation work on the LTPI, other European national programmes for LT were reviewed and information from experienced partners collected. Further information on related programmes and the general structure and execution of the LTPI can be found in (Nikulásdóttir et al., 2020).

All deliverables of the programme are published under open licenses and are freely accessible for research as well as commercial use. Therefore, it is of utmost importance to have a stable hosting platform that can ensure access and availability.

3.1 Language Resources

A variety of language resources are being compiled or extended within the LTPI. The **Icelandic Gigaword Corpus** (IGC) (Steingrímsson et al., 2018) is a large text corpus containing over 1.6B tokens. Within the LTPI, the corpus is being updated yearly with new data sources and updated data from previous years. Furthermore, each new edition is annotated using the latest tools. Two versions of a Gold standard corpus, **MIM-GOLD** and **MIM-GOLD-NER**, have been made available, manually annotated with POS tags and named entities, and within the LTPI enriched with manually checked lemmas. Large lexical resources have been redesigned or are being extended and further processed with focus on use in LT: the **Database of Icelandic Morphology** (Bjarnadóttir et al., 2019) and the **Icelandic Word Web** (Daníelsson et al., 2021). Smaller resources, like a hyphenation word list with a hyphenation tool, and a pronunciation dictionary, have also been made available through CLARIN-IS.

3.2 Support Tools

Several NLP tools have been or are currently being developed or improved upon within the LTPI. Each tool is either used as part of a processing pipeline, or as a stand-alone tool. A **tokenizer** has been developed that converts input text to streams of tokens and also segments the token stream into sentences, considering various cases of abbreviations, dates, etc. to prevent wrong segmentation. During the LTPI, a previously published BiLSTM **PoS tagger** for Icelandic (Steingrímsson et al., 2019) has been improved substantially, e.g. by incorporating contextualized word embeddings, resulting in ABLTagger 2.0 in CLARIN-IS with an accuracy of 96.95%. With resources from Section 3.1, a RNN **lemmatizer** accepting the word-form as well as the corresponding PoS tag to predict the lemma is being developed within the LTPI. Latest experiments show an accuracy of 98.9% on known word-forms and 91.7% on unknown word-forms. A **named entity recognizer** based on a fine-tuned ELECTRA-Base model, trained on the IGC, is in development, newest experiments showing an F_1-score of 91.9%, a dramatic improvement from previous models (Ingólfsdóttir et al., 2020). Two previously published **parsers** have been updated within the LTPI, a *full parser* and a *shallow parser*. The rule-based full-constituency parser relies on a wide-coverage context-free grammar and uses a parsing system based on an enhanced Earley parser (Þorsteinsson et al., 2019). The work on the shallow parser (Loftsson and Rögnvaldsson, 2007) consists of making it accept tagged text according to the new MIM-GOLD tagset (see Section 3.1) and improving individual components. A new **lexicon acquisition tool** is used to find neologisms and older words that jump in frequency due to gaining new word senses. All the above tools are currently available through CLARIN-IS and by the end of the LTPI further tools will have been added, thus ensuring open access to the most important basic support tools for LT.

3.3 Machine Translation

Within the machine translation project, the substantial parts are corpus work, translation methods and infrastructure. A collection of parallel English-Icelandic corpora, **ParIce** (Barkarson and Steingrímsson,

2019), contains texts from various sources, most substantially European Economy Area regulations. **Backtranslations** are synthetic parallel corpora created using existing translation systems that have been shown to be greatly beneficial when training neural translation models. ParIce, including development and test sets, and backtranslated corpora have been published on CLARIN-IS. Three different machine translation methods were tried and tested in the first year (Jónsson et al., 2020) to compare traditional methods to recent advances using the available data – the models have been made available on CLARIN-IS. A **Transformer model** showed best results and thus transformers were chosen as the core method for an open Icelandic-English translation system. A **web-based translation interface** was created and set up online to compare the different models, along with translations provided by Google. This served as a way to compare translations between the participating organizations and allows for open discussion about evaluation. Good **model serving infrastructure** is important when sharing translation models based around different methods. The code for the website as well as the code and configurations to deploy and run translations is made available on CLARIN-IS.

3.4 Spell and Grammar Checking

The work in this core project has focused on developing the necessary data and tools for detecting, categorizing and correcting errors for different user groups. The following resources are currently available through CLARIN-IS: An annotated **general error corpus** with a fine-grained error classification that facilitates performance measurements of the spell and grammar checking software (Arnardóttir et al., 2021). Three **specialized error corpora**, each representing a particular user group, have been annotated and published in order to measure the software's performance on errors particular to the respective user groups. The Icelandic L2 Error Corpus is a collection of texts written by second-language learners of Icelandic (Glišić and Ingason, 2021), the Icelandic Dyslexia Error Corpus is a collection of texts written by native Icelandic speakers with dyslexia, and the Icelandic Child Language Error Corpus is a collection of texts written by native Icelandic speakers aged 10 to 15. **Miscellaneous word lists and language models** include aggregated error data from different sources, a confusion sets database and a trigram language model to help with suggestions for corrections. The **spell and grammar checking software** is a Python package and command line tool for checking and correcting spelling and grammar. The version currently available on CLARIN-IS offers token-level correction and some grammar correction. To get the most usable and complete product for the largest user group, the current focus is on grammar errors, error correction in general, and more detailed guidance tailored to different user groups.

3.5 Automatic Speech Recognition

The emphasis of the Automatic Speech Recognition (ASR) project within the LTPI has been data collection, publication of quality ASR recipes and ultimately a support for commercial applications depending on ASR. The data collection effort has many facets. The prompt-based data collection effort was revived through a new crowd-sourcing system based on Mozilla's Common Voice project[3] called **Samrómur** (Mollberg et al., 2020). The Samrómur project continues to collect voice samples from adults but it also reaches out to children, teenagers and people who speak Icelandic as a second language. Transcriptions of broadcast news and media material as well as university lectures are also being produced to create parallel acoustic-text databases for Icelandic ASR. A system to collect prompted questions for Question Answering systems and conversations for spoken dialogue systems have also been set up. All these **speech data collections** are being prepared for publication on CLARIN-IS. Kaldi[4] recipes have been developed for general-purpose speech recognition and for teenage voices. Furthermore, **punctuation models** have been published on CLARIN-IS.

3.6 Speech Synthesis

For speech synthesis (TTS) it is important to have access to a large corpus of high quality recordings. Currently available on CLARIN-IS is the **Talrómur** corpus which includes 213 hours of speech recordings from eight different speakers. The corpus consists of four male voices and four female voices. The

[3] https://commonvoice.mozilla.org/en

[4] https://github.com/cadia-lvl/samromur-asr

voices range in age, from 26 to 71 years old, and speaking style. In total, the corpus is made up of 122,417 single sentence utterances. The reading script was generated to maximize coverage of diphones in the Icelandic language and consists of sentences from multiple different sources (Sigurgeirsson et al., 2020; Sigurgeirsson et al., 2021). The recordings were conducted in 2020 by Reykjavik University and RÚV, the Icelandic National Broadcasting Service, in a professional studio at the headquarters of the latter. Two of the voices were recruited from the north of Iceland and were recorded in a studio at the University of Akureyri. Later this year, **Talrómur 2** will be available on CLARIN-IS with additional two hours of recordings from each of 40 new speakers. For TTS text pre-processing, data, software packages and models for **text normalization** and **automatic grapheme-to-phoneme** (g2p) conversion are already published on CLARIN-IS or will shortly be available.

4 Standards and Licensing

One of the core pillars of the LTPI is the publication of data and software under open licenses. The guiding licenses are CC BY 4.0[5] for data and Apache 2.0[6] for software. In exceptional cases, data have to be published with more restrictive licenses, but all deliverables of the programme will be available for research and commercial use. An important part of ensuring open licensing is the crafting of agreements and consent statements for various data collection efforts.

All teams work by common standards, defined in guidelines for data deliverables, on the one hand, and for software deliverables, on the other. Wherever possible, the guidelines adhere to international standards, e.g. regarding data format, metadata, or coding guidelines. Published data adhere to the FAIR standard[7]. Naming, versioning and keyword definitions are coordinated throughout the deliverables.

Type of Repository	**Number of Repositories**
General text corpora, incl. test/dev	7
Specialized corpora	8
Parallel corpora	4
Lexical resources	7
NLP-tools	7
Machine translation	4
Spell and grammar checking	8
Speech corpora	2
Speech models and related modules	6
ALL REPOSITORIES	47

Table 1: CLARIN repositories from the LT-Programme for Icelandic. Status as of August 2021

5 Usage Scenarios

The aim of the LTPI is that language resources and infrastructure software will be available for research and commercial use. The aimed-at users are LT-specialists and general software developers that need to integrate LT in their products, as well as researchers from various fields.

There are numerous usage scenarios for the "buffet" of the LTPI deliverables. There are several levels of usage possibilities, reaching from low-level development using corpora and basic tools, to the usage of production-ready models or plugins/applications. For speech synthesis, for example, developers can use the speech corpora and necessary language-specific resources, like the pronunciation dictionary, to train and develop their own TTS models and voices. They can use the delivered TTS voices to integrate into their application, or they can use the web reader plugin directly to connect to their website.

[5] http://creativecommons.org/licenses/by/4.0/
[6] https://www.apache.org/licenses/LICENSE-2.0
[7] https://www.go-fair.org/fair-principles/

Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019-2023. The programme, which is managed and coordinated by Almannarómur (https://almannaromur.is/), is funded by the Icelandic Ministry of Education, Science and Culture.

References

Þórunn Arnardóttir, Xindan Xu, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, and Anton Karl Ingason. 2021. Creating an Error Corpus: Annotation and Applicability. In *Proceedings of CLARIN*.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and filtering ParIce: An English-Icelandic parallel corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland.

Hjalti Daníelsson, Jón Hilmar Jónsson, Þórður Arnar Árnason, Alec Shaw, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2021. The Icelandic Word Web: A language technology focused redesign of a lexicosemantic database. In *Proceedings of NODALIDA 2021*, pages 429–434, Reykjavík.

Isidora Glišić and Anton Karl Ingason. 2021. The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic. In *Proceedings of CLARIN*.

Svanhvít L. Ingólfsdóttir, Ásmundur A. Guðjónsson, and Hrafn Loftsson. 2020. Named Entity Recognition for Icelandic: Annotated Corpus and Models. In Luis Espinosa-Anke, Carlos Martín-Vide, and Irena Spasić, editors, *Statistical Language and Speech Processing*, pages 46–57, Cham. Springer International Publishing.

Haukur Páll Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Hrafn Loftsson. 2020. Experimenting with Different Machine Translation Models in Medium-Resource Settings. In Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák, editors, *Text, Speech, and Dialogue*, pages 95–103, Cham. Springer International Publishing.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 128–135.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jón Guðnason. 2020. Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3463–3467, Marseille, France.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language Technology Programme for Icelandic 2019–2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France.

Atli Sigurgeirsson, Gunnar Örnólfsson, and Jón Guðnason. 2020. Manual Speech Synthesis Data Acquisition – From Script Design to Recording Speech. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 316–320.

Atli Sigurgeirsson, Þorsteinn Gunnarsson, Gunnar Örnólfsson, Eydís Huld Magnúsdóttir, Ragnheiður Kr. Þórhallsdóttir, Stefán Jónsson, and Jón Guðnason. 2021. Talrómur: A large Icelandic TTS corpus. In *Proceedings of NODALIDA 2021*, pages 440–444, Reykjavík.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Burgaria.

Vilhjálmur Þorsteinsson, Hulda Óladóttir, and Hrafn Loftsson. 2019. A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System. In *Proceedings of the International Conference on Recent Advances in Natural Language Proceedings (RANLP 2019)*, pages 1397–1404, Varna, Bulgaria.