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

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

[1]Grammatek ehf., Iceland, [2]University of Iceland, [3]The Árni Magnússon Institute for Icelandic Studies, [4]Reykjavik University, [5]Miðeind ehf., Iceland, [6]University of Edinburgh

anna@grammatek.com, thar@hi.is, starkadur.barkarson@arnastofnun.is
jg@ru.is, thorsteinng@ru.is, antoni@hi.is, haukurpj@mideind.is,
hrafn@ru.is, hulda@mideind.is, eirikur@hi.is,
einar.freyr.sigurdsson@arnastofnun.is,
atlisigurgeirsson@gmail.com, vesteinn@mideind.is,
steinthor.steingrimsson@arnastofnun.is, gunnaro@ru.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 LT infrastructure initiative, Language Technology Programme for Icelandic. All deliverables of the programme are published under open licences and are freely accessible for research as well as commercial use. We provide a broad overview of the available repositories and the core publishing guidelines.

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., 2020b) chose CLARIN-IS to be the central hub for all deliverables of the programme.

This paper gives a broad overview of the manifold "buffet" of 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. The main motivation for joining CLARIN was to have a secure and well recognized infrastructure to store all the resources and tools created during the LTPI. But the plan for

Anna Björk Nikulásdóttir, Þórunn Arnardóttir, Starkaður Barkarson, Jón Guðnason, Þorsteinn Daði Gunnarsson, Anton Karl Ingason, Haukur Páll Jónsson, Hrafn Loftsson, Hulda Óladóttir, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, Atli Þór Sigurgeirsson, Vésteinn Snæbjarnarson, Steinþór Steingrísson, Gunnar Thor örnólfsson 2022. Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS. *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 109–125. DOI: https://doi.org/10.3384/9789179294441

the near future is to widen the scope and reach out to researchers of humanities and social sciences. 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.

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
6. Speech Synthesis

Each project area was further divided into work packages with defined goals. In total, 65 work packages were described up front with estimated 1136 man-months over five years to deliver the projects. The work packages have been revised and defined in more detail every year to keep up with developments in the field and to adjust work packages within project areas with the aim of meeting the overall goals of the programme. 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., 2020b).

All deliverables of the programme are published under open licences 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.

In the remainder of this section, we describe each of the core projects, along with evaluation results where applicable.

3.1 Language Resources

A variety of language resources are being compiled or extended within the LTPI. The following list describes the main resources.

- The **Icelandic Gigaword Corpus** (IGC)[3] is a large text corpus containing texts from various sources: news media, parliamentary proceedings, published books, journals, adjudications and more. The first version, published in 2018, contained over 1.2B running words from texts published until the end of 2017 (Steingrímsson et al., 2018), while the latest version contains close to 1.9B words (Barkarson et al., 2022). Within the LTPI, the corpus is being updated yearly with new data sources and updated data from previous ones. Each new edition is annotated using the latest tools. Table 1 shows the development of the corpus, year by year.

[1] https://clarin.is/

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

[3] http://igc.arnastofnun.is/

Version	Words (M)	PoS tagger	Tagset
IGC-2018	1,253	IceStagger (Loftsson and Östling, 2013)	MIM-GOLD 1.0
IGC-2019	1,394	IceStagger	MIM-GOLD 1.0
IGC-2020	1,555	ABLTagger 0.9 (Steingrímsson et al., 2019)	MIM-GOLD 2.0
IGC-2021	1,871	ABLTagger 2.0 (Jónsson et al., 2021)	MIM-GOLD 2.0

Table 1. The table shows the amount of tokens (millions) for the four published versions of the IGC, as well as the PoS taggers and tagsets used.

The corpus is published under two different licences. Approximately half of the corpus uses CC BY 4.0, while the other half is published under the MIM-licence, a custom licence developed for Icelandic text corpora to use in cases where the publishers of the texts cannot accept the terms of CC BY 4.0. Both licences allow use of the data for all research and language modelling.

The first three versions of the corpus were published in two parts, one for each licence. As of version IGC-2021, the corpus is split into eight subcorpora that reflect the different source types: journals, published books, parliamentary speeches, laws, adjudications, social media and two news corpora.

Evaluation sets have been released to evaluate the accuracy of PoS tagging of different text types (Barkarson et al., 2020). This can be used to evaluate the tagging accuracy of different subcorpora. Using ABLTagger 0.9 (see Section 3.2) the accuracy ranges from 94.34% to 97.79%, depending on text type.

- **MIM-GOLD** (Helgadóttir et al., 2014) is a corpus of one million tokens, manually annotated with PoS tags. Within the LTPI, manually checked lemmas have been added to the corpus and the tagset has been revised in order to be able to accommodate for URLs and symbols like emoticons that are common in some modern texts and to make clearer distinctions on how to tag proper nouns, foreign words, abbreviations and more, described in (Barkarson et al., 2021a). A version of this corpus, **MIM-GOLD-NER**, in which named entities (NEs) have been annotated, has also been made available. In MIM-GOLD-NER, about 48,000 Icelandic NEs are tagged with one of eight NE types (Ingólfsdóttir et al., 2020).

- The LT part of the **Database of Icelandic Morphology** (DIM), a multipurpose linguistic resource, has been further developed within the LTPI. The Database of Modern Icelandic Inflection (DMII), which has been in development since 2002, and comprises approx. 300,000 inflectional paradigms (Bjarnadóttir et al., 2019a), is accessible at CLARIN-IS (Bjarnadóttir, 2019), together with valency structures of verbs (Bjarnadóttir, 2021), a list of common abbreviations in Icelandic texts (Bjarnadóttir and Ingimundarson, 2021) and the DMII Core (Bjarnadóttir et al., 2019b), which contains the core vocabulary of contemporary Icelandic. The database of inflectional paradigms has been compressed and encapsulated in a Python package to facilitate quick lookup in programs (Þorsteinsson et al., 2021b).

- **Skiptir** (Rúnarsson, 2020) is a simple command line tool that uses Pyphen to hyphenate text. Along with the tool a new hyphenation dictionary was compiled (Rúnarsson et al., 2020).

- The new **Icelandic Word Web** (Daníelsson et al., 2021) is an LT-focused redesign of a database of semantically related entries. It is stored in a single RDF file accessible directly through CLARIN-IS (Jónsson et al., 2020c).

- Two evaluation sets for word embeddings have been adapted to Icelandic. **IceBATS** (Friðriksdóttir et al., 2021) is an Icelandic adaptation of the Bigger Analogy Test Set (BATS). It contains 98,000 analogy questions that cover inflectional and derivational morphology as well as lexicographic and encyclopedic semantics (Daníelsson et al., 2022). An Icelandic version of **Multi-SimLex** (MSL) has been compiled. MSL is an evaluation protocol and associated dataset for lexical semantics (Daníelsson et al., 2021). The original English-language MSL builds on several older, well-known datasets, most notably SimLex-999, and has been released in more than a dozen languages.

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. Here, we have focused on the tools that are the most helpful for more complex project areas currently within the LTPI, such as spell and grammar checking, in order to maximize the use of time and effort.

- **Tokenizer**: A tokenizer (Þorsteinsson et al., 2021d) has been developed that converts input text to streams of tokens, where each token is a separate word, punctuation sign, number/amount, date, e-mail, URL/URI, etc. It also segments the token stream into sentences, considering various corner cases of abbreviations, dates, etc. to prevent wrong segmentation. It reaches 100% accuracy for sentence detection for texts without (well-documented) edge cases, and 99.7% for token detection. Two modes of tokenization were implemented to serve the widest audience. PoS tagging and machine translation use the *shallow* tokenization, where tokens are separated by white space. Parsing and grammar correction rely on the *deep* tokenization, where the tokens have been annotated with the token type and further information extracted from the token.

- **PoS tagger**: Before the LTPI started, the best performing PoS tagger for Icelandic was ABLTagger 0.9, a BiLSTM model implemented in DyNet, achieving an accuracy of 94.47% when evaluated on the MIM-GOLD corpus with the original tagset (Steingrímsson et al., 2019). During the LTPI, this tagger has been gradually improved. First, it was ported to PyTorch and several parts of it improved, e.g. by adding pre-trained word embeddings (trained on the IGC), resulting in ABLTagger 1.0 on CLARIN-IS, which obtains an accuracy of 95.59% on the revised MIM-GOLD tagset, which all subsequent tagging models use. Second, by incorporating contextualized word embeddings, i.e. ELECTRA-Small trained on the IGC, resulting in ABLTagger 2.0 in CLARIN-IS (Jónsson et al., 2021), the accuracy has increased to 96.95%. Finally, by incorporating larger BERT-like models, e.g. ELECTRA-Base (Clark et al., 2020), the accuracy increases significantly, to 97.71%. This accuracy score refers to a model excluding the tags for non-analysed tokens (*x*) and foreign words (*e*).

- **Lemmatizer**: With resources from Section 3.1, MIM-GOLD and DIM, a RNN lemmatizer accepting the word form as well as the corresponding PoS tag to predict the lemma has been developed (Jónsson and Loftsson, 2021). Latest experiments show an accuracy of 98.9% on known lemmas and 86.6% on unknown lemmas.

- **Named Entity Recognizer**: In parallel to the construction of MIM-GOLD-NER (see Section 3.1), three different machine learning models were evaluated (Ingólfsdóttir et al., 2020). The best performing model was a bidirectional LSTM (BiLSTM) model, obtaining an overall F_1-score of 83.9, for the entities `Person`, `Organization`, and `Location`. In the LTPI, we experimented with fine-tuning BERT-like models using MIM-GOLD-NER, as well as developing a combination method (Guðjónsson et al., 2021). By fine-tuning an ELECTRA-Base model, trained on the IGC, the F_1-score increased dramatically to 91.9. An even higher F_1-score was obtained by fine-tuning a RoBERTa-Base model, trained on the IGC and data from several other sources (Snæbjarnarson et al., 2022), i.e. 92.7. Combining three BERT-like models, using simple voting in CombiTagger (Henrich et al., 2009), further increased the F_1-score to 93.2.

- **Parsers**: Two previously published **parsers** have been updated within the LTPI, a *full parser* and a *shallow parser*. The rule-based full-constituency parser (Þorsteinsson et al., 2021c) relies on a wide-coverage context-free grammar (CFG) and uses a parsing system based on an enhanced Earley parser (Þorsteinsson et al., 2019). The grammar contains over 5,600 nonterminals, 4,600 terminals and 19,000 productions in fully expanded form. It also gives feature agreement constraints for gender, number, case, and person. An enhanced Earley-based parser generates ranked parse trees in shared packed parse forests. The parser is the foundation for the grammar checking module in the spell and grammar checker. The work in the second year led to the tool reaching an F-measure of 81.2.

The shallow parser, IceParser, is useful as a faster, lighter option for basic parsing, where a full parse is not required, for example in information extraction. The parser, which consists of a sequence of finite-state transducers, accepts PoS-tagged input and generates output according to a shallow syntactic annotation scheme (Loftsson and Rögnvaldsson, 2007) . The work on the shallow parser consisted of making it accept tagged text according to the new MIM-GOLD tagset (see Section 3.1) and improving individual components. Evaluation shows that the new version IceParser 1.5.0 on CLARIN-IS (Loftsson et al., 2021) obtains an F-measure of 96.3 for phrases and 83.1 for syntactic functions.

- **Lexicon Acquisition Tool:** ALEXIA (Friðriksdóttir et al., 2021; Friðriksdóttir and Jasonarson, 2021) is used to find neologisms as well as other words that are more frequently used than before. It processes the IGC, but can also be adapted to other data sources. It returns a word list with relevant information, such as frequency per word form.

All the above tools are currently available through CLARIN-IS. By the end of the LTPI, we will also have added pre-trained embeddings, a Universal Dependencies (UD) parser, and BERT-like language models to CLARIN-IS, thus ensuring open access to the most important basic support tools for LT.

A few UD parsing models will be trained using GreynirCorpus (Þorsteinsson et al., 2021e), which is originally a constituency treebank but is being converted to the UD annotation scheme using UDConverter. UDConverter is a tool that has already been used to create two Icelandic UD treebanks by converting constituency treebanks based on the Penn Treebank (Arnardóttir et al., 2020).

Other, more peripheral resources have been added to CLARIN-IS, such as a parsed corpus with a tree search program to search for specific syntactic structures (Þorsteinsson et al., 2021e), and a test suite for different parsing schemas using the parsed corpora.

3.3 Machine Translation

Machine translation (MT) is a substantial part of the LTPI. In its first year, we focused on assessing methods, gathering data and building up infrastructure. Several deliverables were developed as part of this effort and published on CLARIN-IS to share between parties of the consortium. In the second year, the best model methods were further improved and iterated on and collaboration with industry was explored. In addition, organizers of The Sixth Conference on Machine Translation (WMT21) (Barrault et al., 2021) were approached regarding adding Icelandic–English as one of the language pairs in the news translation competition. This was approved and the creation of the datasets used was funded by the LTPI.

Several **resources for MT** between Icelandic and English have been developed and released.

- **ParIce** (Barkarson and Steingrímsson, 2019) is a collection of parallel English–Icelandic corpora suitable for training MT systems. It contains texts from various sources, including the Bible, the European Medicines Agency, open-source software projects, OpenSubtitles, the Nordic Council of Ministers, the European Space Observatory and most substantially European Economy Area regulations (Steingrímsson and Barkarson, 2021). Development and test sets for five different subcorpora (Barkarson et al., 2021b) were labelled and manually reviewed.

- **English-Icelandic glossary** (Steingrímsson et al., 2021) contains over 230K English-Icelandic pairs, single words and multiword units, with probability scores for translations in both directions. The glossary was built using automatic methods for compiling candidate lists, which were then manually checked by human annotators or compared to available manually curated dictionaries and word lists.

- **IPAC** (Símonarson and Snæbjarnarson, 2021) is a parallel corpus extracted from student theses abstracts that cover a wide range of academic topics. This dataset is diverse in its subject matter and contains text in somewhat complex language. It is suitable both for training and as a baseline test set for evaluating general translation performance.

- **Backtranslations** are synthetic parallel corpora created using existing translation systems that have been shown to be greatly beneficial when training neural translation models. Creating the translations requires substantial computational power, so the data has been released publicly (Símonarson et al., 2021a). The data was collected from Wikipedia, legal documents and news articles.

- **Synthetic corpora** with injected proper names were created and released. These contain parallel sentences with names (Símonarson et al., 2020) and entities (Jónsson et al., 2021) substituted and labelled. These are useful for injecting vocabulary and improving performance when translating proper names that should not be translated directly.

Three different MT methods were tried and tested in the first year to understand how more traditional methods and recent advances compared with the available data. The models are available on CLARIN-IS.

(i) **Moses** is a non-neural statistical MT system which has been shown to perform well in low-resource settings with limited computing power.

(ii) **BiLSTM** is a neural MT architecture which was among the best some years ago.

(iii) **Transformers** are state-of-the-art neural MT models for widely used languages with high resources.

All models were compared and evaluated as described in (Jónsson et al., 2020b). The Transformer model showed best results indicating that MT for English-Icelandic is not limited by the amount of available data, i.e. it is possible to make use of the same state-of-the-art methods as for e.g. English-German translations.

In the second year we built on the experience of the first year and more recent developments in NMT. A multilingual language model, **mBART-25** (Tang et al., 2021), was used as a starting point and then fine-tuned for translation between Icelandic and English. This is described more thoroughly in (Símonarson et al., 2021b). The resulting models (Snæbjarnarson et al., 2021) are much improved translation models, and the backtranslation corpus was regenerated using this data. A command line interface has been made available for translation that fetches the necessary models from CLARIN-IS[4]. Finally, the best NMT system for translation between English and Icelandic has been adapted for translation of EEA regulations and has undergone testing at the Translation Center of the Ministry of Foreign Affairs in Iceland. Initial results are good, and the collaboration has been extended for another year.

Besides the datasets collected and models trained, some infrastructure development has taken place to support the translation projects. 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 allow for open discussion about evaluation. The code for the website[5] was packaged and released on CLARIN-IS. **Model serving infrastructure** was implemented for the different methods, and code and configurations to deploy and run translations are distributed on CLARIN-IS (Snæbjarnarson et al., 2020; Jónsson et al., 2020a).

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. Several resources are currently available through CLARIN-IS.

An annotated **general error corpus**, the Icelandic Error Corpus, uses a fine-grained error classification that facilitates performance measurements of the spell and grammar checking software (Arnardóttir et al., 2021). The error corpus consists of three text genres: student essays, online news text and Wikipedia

[4] The code is available in `https://github.com/mideind/GreynirSeq` and the package `greynirseq` is available in PyPI.

[5] See `https://velthyding.is`.

articles. These texts were previously published, without error annotation, as part of the Icelandic Gigaword Corpus. All texts are proofread and errors annotated according to the annotation scheme, which consists of three hierarchical levels: main categories, subcategories, and error codes. The error codes are used when annotating errors, but the main categories and subcategories are used when improving the spell and grammar checker. The corpus is split into a development and test set to enable its usage when developing the spell and grammar checker. The corpus consists of 4,044 texts with a total of 44,268 revisions and 56,794 unique errors. The average number of errors per 1,000 words in the corpus is 45.76, but this count varies depending on text genre.

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 corpora are created using the same methods as used when creating the general error corpus and the same annotation scheme is used in all cases. Texts included in the specialized error corpora are collected particularly for this purpose, so more information on the authors can be obtained with their consent, e.g. their age, native language and name, if they do not wish to be anonymous. The Icelandic L2 Error Corpus is a collection of texts written by second-language learners of Icelandic (Glišić and Ingason, 2021; Ingason et al., 2021c). The corpus consists of 76 texts in which 21,842 errors have been annotated. The authors of the texts are of 16 different nationalities, the most common ones being English and Filipino. The Icelandic Dyslexia Error Corpus is a collection of texts written by native Icelandic speakers with dyslexia (Ingason et al., 2021b). The corpus consists of 26 texts, wherein 5,730 errors have been annotated. The Icelandic Child Language Error Corpus (Ingason et al., 2021a) is the final corpus belonging to the specialized error corpora. It is a collection of texts written by native Icelandic speakers aged 10 to 15 and consists of 119 texts with 7,817 annotated errors. All texts in this corpus are published anonymously.

In addition to these error corpora, various **word lists and language models** were created to further improve the spell and grammar checker. They include aggregated error data from different sources, a database of confusion sets and a trigram language model to help with suggestions for corrections, and are the following:

- A list of Icelandic words that may in some way be considered inappropriate, taboo and/or loaded in use or meaning (Sólmundsdóttir et al., 2021). The list also includes words that are not very inappropriate but can be considered an unfortunate topic for children or questionable depending on context. The words are grouped together in categories depending on either their meaning, form or use.

- A list of common misspellings and their corrections was also created (Arnardóttir and Ingason, 2020a). The word forms originate from the development set of the general error corpus and were annotated as nonwords.

- Yet another word list, created for developing the spell and grammar checker, is a list of automatically prepared word forms containing systematic errors, along with their corrections (Arnardóttir and Ingason, 2020b). The list was prepared using a word list from DIM, which includes different Icelandic words and their inflections. In all cases, one item in the word form is changed, i.e. an accent removed from a letter or added to a letter, or a letter replaced by another letter. These particular errors are common misspellings made in Icelandic text.

- Three datasets related to errors in place names were also prepared. The datasets are in JSON format and encoded in UTF-8. The datasets are isprep4isloc (Þórðarson, 2020c), which contains the correct prepositions for various Icelandic place names, isprep4cc (Þórðarson, 2020b), which contains prepositions for various countries and autonomous territories, and cities_is2en (Þórðarson, 2020a), which maps city names in Icelandic to their English counterparts. By using the last dataset, city names in English can be translated to Icelandic when correcting text.

- A list of systematic inflectional errors was also created. It consists of common erroneous inflectional rules, which have been applied to entries in DIM to produce the error entries. The list is stored as a config file within the checking software.

- In addition to these word lists, a pre-existing trigram language model, Icegrams (Þorsteinsson and Óladóttir, 2020), was re-trained, fine-tuned and expanded to include a wider selection of high-quality texts from the Icelandic Gigaword Corpus. The spell checker was used to correct the trigram data beforehand in a bootstrapping manner, as there were cases where very frequent errors were even more frequent than the correct version in the data. This ensured high-quality trigrams, and that the model did not suggest these errors at the expense of the correct version.

The **spell and grammar checking software** (Þorsteinsson et al., 2021a) is a Python package and command line tool. The version currently available on CLARIN-IS offers token-level correction and some grammar correction. The checker is available on the web[6].

The token-level correction relies on the tokenizer from the support tools (see Section 3.2), along with the aforementioned word lists and language models. The basic tokenizer output, i.e. text split into sentences and tokens, is sent through an error tokenization layer. This layer detects context-independent token-level errors such as duplicated words, single words erroneously written as two or more, and phrases erroneously written as single words.

The sentence-level correction relies on the full-constituency CFG parser from the support tools. One of the first layers provides information on all possible tags and lemmas with the help of DIM and built-in compound analysis. After that, the checker can detect more complex token-level errors, such as capitalization errors and taboo words. Semi-fixed phrases and common erroneous variations are handled with a list of lemmatized forms of all words in the phrase. The trigram model is used to find all possible substitutes for unknown or rare words, ranked by likelihood. All error tags and possible corrections are attached to the corresponding tokens.

The parser chooses the tag that best fits the context and results in a valid syntactic structure. In order to handle known, invalid structures in the grammar checker, special rules were added to the context-free grammar to capture those structures and in some cases map them directly to the correct structure. The syntax tree is also searched for questionable syntactic patterns to detect grammar errors that result in a syntactically valid sentence that is nonsensical in meaning.

For token-level errors, the checker reaches an error detection $F_{0.5}$ measure of 62.32, with typos reaching 92.66. For grammar errors, we currently reach an error detection $F_{0.5}$ measure of 24.64.

The spelling and grammar checker has been integrated into the editorial environment of an international CMS provider used by many Icelandic companies, including large media companies. Collaboration with a media company was used to carry out user tests and improve the user experience. The results show the checker to be a beneficial addition to the workflow.

To get the most usable and complete product for the largest user group by the end of the LTPI, the focus is on incorporating a neural language model, in particular more extensive coverage of 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 on data collection, publication of quality ASR recipes and ultimately providing support for commercial applications depending on ASR. The following **data collections** have been ongoing during the project:

- **Read prompts** have been collected using the **Samrómur** (Mollberg et al., 2020) crowd-sourcing platform. The platform is derived from Mozilla's Common Voice project[7]. The organization of the effort is based on the experience of a previous data collection efforts called **Málrómur** (Guðnason et al., 2012; Steingrímsson et al., 2017) and a platform called *Eyra* (Petursson et al., 2016). The web-based implementation of the platform has enabled easier organization of targeted collection

[6] https://yfirlestur.is/

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

efforts such as competitions aimed at children, teenagers and people who speak Icelandic as a second language. The platform uses a crowd-sourced verification system where users can vote for the correctness of read prompts. An automatic system based on forced-alignment scores (Guðnason et al., 2017) has been used to prioritize this effort. At the time of writing, 4,100 hours have been collected through this system in over 1,000,000 utterances. A part of the corpus has been published on OpenSLR (Mollberg et al., 2020) as well as on CLARIN-IS (Mollberg et al., 2021).

- **Broadcast news** corpus has been collected with the aid of The Icelandic National Broadcasting Service and CreditInfo's news watch service. The post-processing of this corpus has been extensive as the text needs to be aligned to the speech recordings to get a better time-resolution in the recorded segments. Force-alignment tools developed in-house (Guðnason et al., 2017) and the Montreal Forced Aligner (McAuliffe et al., 2017) have been used to create a database suitable for speech recognition training. At the time of writing, 487 hours have been collected through this system.

- **Question Answering** data set was collected using a specialized version of the **Samrómur** platform where the prompts were provided especially as questions. The prompts were obtained by pulling questions from the Icelandic Gigaword Corpus that fit certain criteria. In total, the data consists of 28 hours of recordings, of which 20 hours have been validated and published (Hedström et al., 2021).

- **Recorded lectures** were obtained from nine university lectures. Over 51 hours were transcribed by hand and published on CLARIN (Ragnarsson et al., 2022).

- **Speech dialogue** was obtained from conversations recorded on a specially created platform for two-ways conversations. On the platform a speaker is able to create a virtual chat room and share it with another speaker. The owner of the chat room can record the conversation and submit it. Submissions are then transcribed manually. In total, 21 hours have been collected and transcribed using this platform.

- **Other aligned recordings** have been collected and prepared for ASR before the start of the LTPI project with the collection of 542 hours of parliament speeches (Helgadóttir et al., 2017). Other publicly available sources are being explored. These include open court proceedings and rulings, recorded and transcribed municipality meetings and public parliament's committee meetings.

The utility of developing ASR recipes and applications alongside the data collection efforts is twofold. The obvious one is to create the technology that the data collections are intended to support. It therefore contributes directly to the main aims of the LTPI. The second utility is also very important which is to support and hone the data collection efforts with continuous feedback of quality and efficiency. The ASR recipe developers have had direct say in how the data is collected and curated; and they supported the post-processing of the data with forced time-alignment tools and automatic quality assessments. The **recipes and applications** developed during the project are:

- **ASR Kaldi Recipes** were developed for adult, adolescent, and children voices (Hernández Mena and Guðnason, 2022b; Hernández Mena and Guðnason, 2022a). The recipes were based on the read prompts from the Samrómur and Málrómur data collections (Helgadóttir et al., 2019; Nikulásdóttir et al., 2018b).

- **Web interface for ASR** was set up for both real-time speech streaming and off-line speech file uploads. The repository for setting this up was published with an open-source licence and the service runs online (Ragnarsson, 2021).

- **On-device ASR for smartphones** is in development for the Android operating system, using the Android Speech Recognition Service. The first version is scheduled for release by the end of this year.

- **ASR recipes** for voice control and question answering are being developed with a focus on a few specific tasks. This includes a specialized language model for questions.

- **Specialised ASR acoustic models** adapted to children, adolescents and people who speak of Icelandic as a second language.

- **Punctuation Prediction System** was developed for Icelandic using three different approaches: a BERT-based Transformer, a seq2seq Transformer and a bidirectional RNN.

- **Subword Unit Language Model** was developed for Icelandic which showed an improvement in word error rate.

- **Speaker Diarization Toolkit** for Icelandic using Kaldi (Fong and Guðnason, 2021).

All these **speech data collections** and **speech data recipes and tools** are being prepared for publication on CLARIN-IS.

3.6 Speech Synthesis

The focus of this core project has been on gathering sufficient resources and tools that are critical in developing a state-of-the-art text-to-speech system (TTS). The following resources are currently available on CLARIN-IS:

- **Talrómur** (Sigurgeirsson et al., 2021b) is a corpus containing 213 hours of speech recordings from eight different speakers. The corpus consists of four male voices and four female voices. The 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., 2021a). 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. To take the northern dialect in Icelandic into account, two of the voices were recruited from the north of Iceland and were recorded in a studio at the University of Akureyri. The audio is published in a single-channel 16-bit PCM wave file with a sample rate of 22050 Hz. Recordings were made using the recording platform LOBE (Sigurgeirsson et al., 2020) specifically designed for this purpose.

- **Talrómur 2** (Gunnarsson et al., 2021), is similar to Talrómur in many ways. It includes 80 hours of recordings from 40 different speakers with an even split of female and male voices. The voices were chosen to create four cohorts where each group consists of speakers with similar voice characteristics. Recordings were conducted in the same studio and with the same equipment as Talrómur. Both corpora share the same structure, format and audio specifications.

- Resources for TTS text pre-processing are i) a **text normalization corpus** (Sigurðardóttir, 2021) containing 140,000 sentences in their original form and automatically normalized for TTS (e.g. digits converted to their written-out forms and abbreviations expanded) and 40,000 manually normalized sentences, and ii) a **pronunciation dictionary** (Nikulásdóttir et al., 2018a; Nikulásdóttir et al., 2022) with around 65,000 manually verified entries, covering the four main pronunciation variants in Icelandic.

- Based on the text normalization corpus and the pronunciation dictionary, tools and models for TTS text pre-processing have been developed. For **automatic grapheme-to-phoneme conversion** (g2p) two approaches have been implemented: a rule-based module (Nikulásdóttir et al., 2020a), which is useful in lower resource settings, like on smartphones, and LSTM-based models. There is one model for each of the four pronuncation variants, and one model trained on Icelandic transcriptions of English words (Nikulásdóttir, 2020; Ármannsson, 2021). The **text normalization system** is based on regular expressions (Sigurðardóttir et al., 2021) and handles most cases of text normalization tasks that can be expected in common news and sports texts. Modules for text cleaning, text normalization, adaptation of the spell and grammar checker, phrasing, and a complete g2p module including language detection, syllabification and stress labelling are already available on GitHub and will be published on CLARIN-IS in 2022. These modules build a complete TTS text-preprocessing pipeline, also to be published on CLARIN-IS.

- The evaluation platform **MOSI** (Jónsson et al., 2022) was created and is publicly available. MOSI supports multiple evaluation methods used for TTS systems, including MOS tests and A/B tests.

Training and development of TTS models using the Talrómur and Talrómur 2 corpora is underway. Tacotron2 and FastSpeech2 models have been trained on voices in Talrómur using the ESPNet toolkit, using phoneme inputs. Additionally, a parallel WaveGAN model has been trained on the entire Talrómur dataset. Evaluations are performed using MOSI. The models and evaluation results will be published by the end of the LTPI. The first models are already in use by a smartphone application developed within the LTPI, which will also be published on CLARIN-IS by the end of the programme.

4 Standards and Licencing

One of the core pillars of the LTPI is the publication of data and software under open licences. The guiding licences are CC BY 4.0[8] for data and Apache 2.0[9] for software. In exceptional cases, data have to be published with more restrictive licences, 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 operate 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[10]. Naming, versioning and keyword definitions are coordinated throughout the deliverables. Every software deliverable on CLARIN-IS has a link to the corresponding GitHub repository that is mostly hosted under the account of the developing partner. In general, the deliverables are separated modules, e.g. the tokenizer can be found as a stand-alone project. Other projects combine several modules, like the spell and grammar checker (see section 3.4) and text processing pipeline for TTS (see section 3.6).

Type of Repository	**Number of Repositories**
General text corpora, incl. test/dev	15
Specialized corpora	9
Parallel corpora	8
Lexical resources	12
NLP-tools	12
Machine translation	7
Spell and grammar checking	2
Speech corpora	6
Speech models and related modules	10
ALL REPOSITORIES	81

Table 2. CLARIN repositories from the LTPI. Status as of January 2022. Each project is only counted once but repositories have up to 3 previous versions, also available on CLARIN.

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

[8] http://creativecommons.org/licenses/by/4.0/

[9] https://www.apache.org/licenses/LICENSE-2.0

[10] https://www.go-fair.org/fair-principles/

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.

As an example of products already using core resources, the full parser is the basis of the grammar checker within the LTPI. It is also used to parse questions and form answers for a voice assistant app, and is a module in an automatic term extraction software.

Table 2 shows the number of repositories on CLARIN-IS by January 2022 and how they can be divided into resource categories. We actively reach out to global players in LT to advertise the programme. In particular, we hope that carefully crafted, language-specific resources, like e.g. the TTS recordings and diverse gold and test corpora, will help lower the barrier for including Icelandic in existing global LT products.

It is also worth mentioning that the establishment of CLARIN-IS as the centre for Icelandic LT resources has led to Icelandic LT projects developed outside the programme to be published on CLARIN-IS as well. Thus, the foundation is laid for continuous delivery of LT resources to CLARIN-IS after the LTPI ends.

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

Arnardóttir, Þ. and Ingason, A. K. 2020a. Icelandic Error Corpus Nonwords. CLARIN-IS, http://hdl.handle.net/20.500.12537/63.

Arnardóttir, Þ. and Ingason, A. K. 2020b. nonwords. CLARIN-IS, http://hdl.handle.net/20.500.12537/50.

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., and Steingrímsson, S. 2020. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Arnardóttir, Þ., Xu, X., Guðmundsdóttir, D., Stefánsdóttir, L. B., and Ingason, A. K. 2021. Creating an Error Corpus: Annotation and Applicability. In Monachini, M. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference*, pages 59–63, Virtual Edition.

Ármannsson, B. 2021. Grapheme-to-Phoneme Transcription of English Words in Icelandic Text. Master's thesis, Uppsala University.

Barkarson, S. and Steingrímsson, S. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 140–145, Turku, Finland.

Barkarson, S., Steingrímsson, S., Andrésdóttir, Þ. D., and Hafsteinsdóttir, H. 2020. IGC - Evaluation Set 20.09. CLARIN-IS, http://hdl.handle.net/20.500.12537/51.

Barkarson, S., Andrésdóttir, Þ. D., Hafsteinsdóttir, H., Magnússon, Á. D., Rúnarsson, K., Steingrímsson, S., Jónsson, H. P., Loftsson, H., Sigurðsson, E. F., Rögnvaldsson, E., and Helgadóttir, S. 2021a. MIM-GOLD 21.05. CLARIN-IS, http://hdl.handle.net/20.500.12537/113.

Barkarson, S., Steingrímsson, S., Ingimundarson, F. Á., Hafsteinsdóttir, H., and Magnússon, Á. D. 2021b. ParIce Dev/Test Sets 21.10. CLARIN-IS, http://hdl.handle.net/20.500.12537/146.

Barkarson, S., Steingrímsson, S., and Hafsteinsdóttir, H. 2022. Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)* [to appear], Marseille, France.

Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Bjarnadóttir, K. and Ingimundarson, F. Á. 2021. DMII - Abbreviations 21.10. CLARIN-IS, http://hdl.handle.net/20.500.12537/164.

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. 2019a. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 146–154, Turku, Finland.

Bjarnadóttir, K., Hlynsdóttir, K. I., Þórisson, S., Dagsson, T., and Steingrímsson, S. 2019b. DMII Core. CLARIN-IS, http://hdl.handle.net/20.500.12537/12.

Bjarnadóttir, K. 2019. The Database of Modern Icelandic Inflection (DMII). CLARIN-IS, http://hdl.handle.net/20.500.12537/5.

Bjarnadóttir, K. 2021. DIM Valency Structures 21.10. CLARIN-IS, http://hdl.handle.net/20.500.12537/163.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Daníelsson, H., Friðriksdóttir, S. R., and Steingrímsson, S. 2021. Icelandic Multi-SimLex (21.06). CLARIN-IS, http://hdl.handle.net/20.500.12537/121.

Daníelsson, H., Jónsson, J. H., Árnason, Þ. A., Shaw, A., Sigurðsson, E. F., and Steingrímsson, S. 2021. The Icelandic Word Web: A Language Technology Focused Redesign of a Lexicosemantic Database. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 429–434, Reykjavík, Iceland.

Daníelsson, H., Friðriksdóttir, S. R., Steingrímsson, S., and Sigurðsson, E. F. 2022. IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)* [to appear], Marseille, France.

Fong, J. Y. and Guðnason, J. 2021. RUV-DI Speaker Diarization (2021-10-14). CLARIN-IS, http://hdl.handle.net/20.500.12537/157.

Friðriksdóttir, S. R., Daníelsson, H., and Steingrímsson, S. 2021. IceBATS - The Icelandic Bigger Analogy Test Set (21.06). CLARIN-IS, http://hdl.handle.net/20.500.12537/120.

Friðriksdóttir, S. R. and Jasonarson, A. 2021. ALEXIA: Lexicon Acquisition Tool for Icelandic 3.0. CLARIN-IS, http://hdl.handle.net/20.500.12537/123.

Friðriksdóttir, S. R., Jasonarson, A., Steingrímsson, S., and Sigurðsson, E. F. 2021. ALEXIA: A Lexicon Acquisition Tool. In Monachini, M. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference*, pages 64–67, Virtual Edition.

Glišić, I. and Ingason, A. K. 2021. The Nature of Icelandic as a Second Language: An Insight from the Learner Error Corpus for Icelandic. In Monachini, M. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference*, pages 26–30, Virtual Edition.

Guðjónsson, Á. A., Loftsson, H., and Daðason, J. F. 2021. Icelandic NER API – Ensemble Model. CLARIN-IS, http://hdl.handle.net/20.500.12537/159.

Guðnason, J., Kjartansson, O., Jóhannsson, J., Carstensdóttir, E., Vilhjálmsson, H. H., Loftsson, H., Helgadóttir, S., Jóhannsdóttir, K. M., and Rögnvaldsson, E. 2012. Almannarómur: An Open Icelandic Speech Corpus. In *Spoken Language Technologies for Under-Resourced Languages*.

Guðnason, J., Pétursson, M., Kjaran, R., Klüpfel, S., and Nikulásdóttir, A. B. 2017. Building ASR Corpora Using Eyra. In *INTERSPEECH*, pages 2173–2177.

Gunnarsson, Þ. D., Örnólfsson, G. T., Þórhallsdóttir, R., Sigurgeirsson, A. Þ., and Guðnason, J. 2021. Talrómur 2. CLARIN-IS, http://hdl.handle.net/20.500.12537/165.

Hedström, S., Fong, J. Y., Þórhallsdóttir, R., Mollberg, D. E., Guðmundsson, S. F., Jónsson, Ó. H., Þorsteinsdóttir, S., Magnúsdóttir, E. H., and Gudnason, J. 2021. Samromur Queries 21.12. CLARIN-IS, http://hdl.handle.net/20.500.12537/180.

Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. 2014. Correcting Errors in a New Gold Standard for Tagging Icelandic Text. In *Proceedings of the Ninth Language Resources and Evaluation Conference (LREC)*, pages 2944–2948, Reykjavik, Iceland.

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Guðnason, J. 2017. Building an ASR Corpus Using Althingi's Parliamentary Speeches. In *INTERSPEECH*, pages 2163–2167.

Helgadóttir, I. R., Nikulásdóttir, A. B., Borský, M., Fong, J. Y., Kjaran, R., and Guðnason, J. 2019. The Althingi ASR System. In *INTERSPEECH*, pages 3013–3017.

Henrich, V., Reuter, T., and Loftsson, H. 2009. CombiTagger: A System for Developing Combined Taggers. In *Proceedings of the 22nd International FLAIRS Conference, Special Track: "Applied Natural Language Processing"*, Sanibel Island, Florida, USA.

Hernández Mena, C. D. and Guðnason, J. 2022a. Icelandic Language Models with Pronunciations 22.01. CLARIN-IS, http://hdl.handle.net/20.500.12537/172.

Hernández Mena, C. D. and Guðnason, J. 2022b. Samrómur-Children Demonstration Scripts 22.01. CLARIN-IS, http://hdl.handle.net/20.500.12537/173.

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., and Xu, X. 2021a. The Icelandic Child Language Error Corpus (IceCLEC) version 1.1. CLARIN-IS, http://hdl.handle.net/20.500.12537/133.

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., and Xu, X. 2021b. The Icelandic Dyslexia Error Corpus (IceDEC) version 1.1. CLARIN-IS, http://hdl.handle.net/20.500.12537/132.

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., and Glišić, I. 2021c. The Icelandic L2 Error Corpus (IceL2EC) version 1.2. CLARIN-IS, http://hdl.handle.net/20.500.12537/131.

Ingólfsdóttir, S. L., Guðjónsson, Á. A., and Loftsson, H. 2020. Named Entity Recognition for Icelandic: Annotated Corpus and Models. In Espinosa-Anke, L., Martín-Vide, C., and Spasić, I., editors, *Statistical Language and Speech Processing*, pages 46–57, Cham. Springer International Publishing.

Jónsson, H. P. and Loftsson, H. 2021. ABLTagger (Lemmatizer) – 3.1.0. CLARIN-IS, http://hdl.handle.net/20.500.12537/134.

Jónsson, H. P., Loftsson, H., and Steingrímsson, S. 2020a. MT: Moses-SMT. CLARIN-IS, http://hdl.handle.net/20.500.12537/46.

Jónsson, H. P., Símonarson, H. B., Snæbjarnarson, V., Steingrímsson, S., and Loftsson, H. 2020b. Experimenting with Different Machine Translation Models in Medium-Resource Settings. In Sojka, P., Kopeček, I., Pala, K., and Horák, A., editors, *Text, Speech, and Dialogue*, pages 95–103, Cham. Springer International Publishing.

Jónsson, J. H., Daníelsson, H., Árnason, Þ. A., and Shaw, A. 2020c. The Icelandic Wordweb 21.06. CLARIN-IS, http://hdl.handle.net/20.500.12537/117.

Jónsson, H. P., Snæbjarnarson, V., Símonarson, H. B., and Þorsteinsson, V. 2021. En-Is Synthetic Parallel Named Entity Robustness Corpus. CLARIN-IS, http://hdl.handle.net/20.500.12537/129.

Jónsson, S., Gunnarsson, Þ., Örnólfsson, G., and Sigurgeirsson, A. 2022. MOSI: TTS Evaluation Tool. CLARIN-IS, http://hdl.handle.net/20.500.12537/186.

Jónsson, H. P., Loftsson, H., and Steingrímsson, S. 2021. ABLTagger 2.0. CLARIN-IS, http://hdl.handle.net/20.500.12537/98.

Loftsson, H. and Östling, R. 2013. Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa)*, pages 105–119, Oslo, Norway.

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

Loftsson, H., Rögnvaldsson, E., and Pálsson, G. 2021. IceParser 1.5.0. CLARIN-IS, http://hdl.handle.net/20.500.12537/122.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *INTERSPEECH*, pages 498–502.

Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., and Gudnason, J. 2020. Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 3463–3467, Marseille, France.

Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Guðmundsdóttir, J., Steingrímsson, S., Magnúsdóttir, E. H., Fong, J., Borskỳ, M., and Guðnason, J. 2021. Samromur 21.05. CLARIN-IS, http://hdl.handle.net/20.500.12537/189.

Nikulásdóttir, A. B., Guðnason, J., and Rögnvaldsson, E. 2018a. An Icelandic Pronunciation Dictionary for TTS. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345. IEEE.

Nikulásdóttir, A. B., Helgadóttir, I. R., Pétursson, M., and Guðnason, J. 2018b. Open ASR for Icelandic: Resources and a baseline system. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Myasaki, Japan.

Nikulásdóttir, A. B., Ármannsson, B., and Schnell, D. 2020a. Rule-based G2P for Icelandic. CLARIN-IS, http://hdl.handle.net/20.500.12537/83.

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. 2020b. Language Technology Programme for Icelandic 2019–2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France.

Nikulásdóttir, A. B., Ármannsson, B., and Bryndís, B. 2022. Icelandic Pronunciation Dictionary for Language Technology 22.01. CLARIN-IS, http://hdl.handle.net/20.500.12537/181.

Nikulásdóttir, A. B. 2020. Models for Automatic G2P for Icelandic. CLARIN-IS, http://hdl.handle.net/20.500.12537/84.

Petursson, M., Klüpfel, S., and Gudnason, J. 2016. Eyra – Speech Data Acquisition System for Many Languages. *Procedia Computer Science*, 81:53–60.

Ragnarsson, R. K., Mollberg, D. E., and Magnúsdóttir, E. H. 2022. Kennslurómur 22.01. CLARIN-IS, http://hdl.handle.net/20.500.12537/171.

Ragnarsson, R. K. 2021. Tiro Web Interface for Speech Recognition 1.0. CLARIN-IS, http://hdl.handle.net/20.500.12537/161.

Rúnarsson, K., Jónsson, B., and Gíslason, M. 2020. Icelandic Hyphenation Dictionary. CLARIN-IS, http://hdl.handle.net/20.500.12537/86.

Rúnarsson, K. 2020. Skiptir. CLARIN-IS, http://hdl.handle.net/20.500.12537/87.

Sigurðardóttir, H. S., Nikulásdóttir, A. B., and Guðnason, J. 2021. Creating Data in Icelandic for Text Normalization. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 404–412.

Sigurðardóttir, H. S. 2021. Text Normalization Corpus 21.10. CLARIN-IS, http://hdl.handle.net/20.500.12537/158.

Sigurgeirsson, A. Þ., Örnólfsson, G. T., and Guðnason, J. 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, Marseille, France. European Language Resources association.

Sigurgeirsson, A., Gunnarsson, Þ., Örnólfsson, G. T., Magnúsdóttir, E. H., Þórhallsdóttir, R. K., Jónsson, S., and Guðnason, J. 2021a. Talrómur: A Large Icelandic TTS Corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 440–444, Reykjavík, Iceland.

Sigurgeirsson, A. Þ., Gunnarsson, Þ. D., Örnólfsson, G. T., Þórhallsdóttir, R., Magnúsdóttir, E. H., and Guðnason, J. 2021b. Talrómur. CLARIN-IS, http://hdl.handle.net/20.500.12537/104.

Símonarson, H. B. and Snæbjarnarson, V. 2021. Icelandic Parallel Abstracts Corpus. *CoRR*, abs/2108.05289.

Símonarson, H. B., Snæbjarnarson, V., and Þorsteinsson, V. 2020. En-Is Synthetic Parallel Corpus. CLARIN-IS, http://hdl.handle.net/20.500.12537/70.

Símonarson, H. B., Snæbjarnarson, V., and Þorsteinsson, V. 2021a. En-Is Synthetic Parallel Corpus - 2021. CLARIN-IS, http://hdl.handle.net/20.500.12537/127.

Símonarson, H. B., Snæbjarnarson, V., Ragnarson, P. O., Jónsson, H., and Þorsteinsson, V. 2021b. Miðeind's WMT 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.

Snæbjarnarson, V., Símonarson, H. B., and Þorsteinsson, V. 2020. GreynirT2T Serving - En–Is NMT Inference and Pre-trained Models. CLARIN-IS, http://hdl.handle.net/20.500.12537/72.

Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Jónsson, H. P., Ingólfsdóttir, S. L., and Þorsteinsson, V. 2021. GreynirTranslate - mBART25 NMT Models for Translations between Icelandic and English. CLARIN-IS, http://hdl.handle.net/20.500.12537/125.

Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Ingólfsdóttir, S. L., Jónsson, H. P., Þorsteinsson, V., and Einarsson, H. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. *CoRR*, abs/2201.05601.

Sólmundsdóttir, A., Stefánsdóttir, L. B., and Ingason, A. K. 2021. IceTaboo: a Database of Contextually Inappropriate Words for Icelandic. In Monachini, M. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference*, Virtual Edition.

Steingrímsson, S. and Barkarson, S. 2021. ParIce 21.10. CLARIN-IS, http://hdl.handle.net/20.500.12537/145.

Steingrímsson, S., Guðnason, J., Helgadóttir, S., and Rögnvaldsson, E. 2017. Málrómur: A Manually Verified Corpus of Recorded Icelandic Speech. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 237–240, Gothenburg, Sweden.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Steingrímsson, S., Kárason, Ö., and Loftsson, H. 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)*, pages 1161–1168, Varna, Bulgaria.

Steingrímsson, S., Obrien, L. J., Ingimundarson, F. Á., Magnússon, Á. D., Andrésdóttir, Þ. D., and Eiríksdóttir, I. G. 2021. English-Icelandic/Icelandic-English Glossary 21.09. CLARIN-IS, http://hdl.handle.net/20.500.12537/144.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. 2021. Multilingual Translation from Denoising Pre-Training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Þorsteinsson, V. and Óladóttir, H. 2020. Icegrams (2020-09-30). CLARIN-IS, http://hdl.handle.net/20.500.12537/80.

Þorsteinsson, V., Óladóttir, H., and Loftsson, H. 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)*, pages 1397–1404, Varna, Bulgaria.

Þorsteinsson, V., Óladóttir, H., Arnardóttir, Þ., and Þórðarson, S. 2021a. GreynirCorrect. CLARIN-IS, http://hdl.handle.net/20.500.12537/148.

Þorsteinsson, V., Óladóttir, H., and Þórðarson, S. 2021b. BinPackage. CLARIN-IS, http://hdl.handle.net/20.500.12537/137.

Þorsteinsson, V., Óladóttir, H., and Þórðarson, S. 2021c. GreynirPackage. CLARIN-IS, http://hdl.handle.net/20.500.12537/147.

Þorsteinsson, V., Óladóttir, H., Þórðarson, S., and Ragnarsson, P. O. 2021d. Tokenizer. CLARIN-IS, http://hdl.handle.net/20.500.12537/136.

Þorsteinsson, V., Óladóttir, H., Þórðarson, S., Símonarson, H. B., and Ásgeirsdóttir, K. 2021e. GreynirCorpus. CLARIN-IS, http://hdl.handle.net/20.500.12537/119.

Þórðarson, S. 2020a. cities_is2en (2020-09-28). CLARIN-IS, http://hdl.handle.net/20.500.12537/66.

Þórðarson, S. 2020b. isprep4cc (2020-09-28). CLARIN-IS, http://hdl.handle.net/20.500.12537/59.

Þórðarson, S. 2020c. isprep4isloc (2020-09-28). CLARIN-IS, http://hdl.handle.net/20.500.12537/58.