Language Technology Programme for Icelandic 2019-2023

Abstract

In this paper, we describe a new national language technology programme for Icelandic. The programme, which spans a period of five years, aims at making Icelandic usable in communication and interactions in the digital world, by developing accessible, open-source language resources and software. The research and development work within the programme is carried out by a consortium of universities, institutions, and private companies, with a strong emphasis on cooperation between academia and industries. Five core projects will be the main content of the programme: language resources, speech recognition, speech synthesis, machine translation, and spell and grammar checking. We also describe other national language technology programmes and give an overview over the history of language technology in Iceland.

Introduction

During the last decade, we have witnessed enormous advances in language technology (LT). Applications that allow users to interact with technology via spoken or written natural language are emerging in all areas, and access to language resources and open-source software libraries enables faster development for new domains and languages.

However, LT is highly language dependent and it takes considerable resources to develop LT for new languages. The recent LT development has focused on languages that have both a large number of speakers and huge amounts of digitized language resources, like English, German, Spanish, Japanese, etc. Other languages, that have few speakers and/or lack digitized language resources, run the risk of being left behind.

Icelandic is an example of a language with almost a negligible number of speakers, in terms of market size, since only about 350,000 people speak Icelandic as their native language. Icelandic is therefore seldom on the list of supported languages in LT software and applications.

The Icelandic Government decided in 2017 to fund a five-year programme for Icelandic LT, based on a report written by a group of LT experts BIBREF0. After more than two years of preparation, a consortium consisting of universities, institutions, associations, and private companies started the work on the programme on the 1st of October 2019. The goal of the programme is to ensure that Icelandic can be made available in LT applications, and thus will be usable in all areas of communication. Furthermore, that access to information and other language-based communication and interaction in Icelandic will be accessible to all, e.g. via speech synthesis or speech-to-text systems.

The focus of the programme will be on the development of text and speech-based language resources, on the development of core natural language processing (NLP) tools like tokenisers, taggers and parsers, and finally, to publish open-source software in the areas of speech recognition, speech synthesis, machine translation, and spell and grammar checking. All deliverables of the programme will be published under open licenses, to encourage use of resources and software in commercial products.

While the government-funded programme for the development of resources and infrastructure software builds the backbone of the Icelandic LT programme, another branch is a competitive fund for research and development. This Strategic Research and Development Programme for Language Technology is managed by the Icelandic Centre for Research, Rannís, which publishes calls for applications on a regular basis.

The third pillar of the programme is the revival of the joint Master's programme in LT at Reykjavik University (RU) and the University of Iceland (UI). The goal is further to increase the number of PhD students and to build strong knowledge centres for sustainable LT development in Iceland.

The budget estimation for the programme, including the competitive fund, education plan and infrastructure costs, is around 14 million euros. Additionally, around 3.6 million euros is expected to be the contribution of the industry through the competitive fund.

This paper is structured as follows: In Section SECREF2 we discuss national LT programmes that have been run in other European countries and helped developing the Icelandic project plan. Section SECREF3 gives an overview over the 20 years of LT development in Iceland. Section SECREF4 shows the organisation of the new programme, and in Section SECREF5 we describe the core projects that have been defined for it. Finally, a conclusion is presented in Section SECREF6.

Other European LT Programmes

In recent years, there has been much international discussion on how the future of languages depends on them being usable in the digital world. This concern has led to a number of national LT programmes. We studied three of these national programmes: the STEVIN programme in the Netherlands which ran between 2004 and 2011, the Plan for the Advancement of Language Technology in Spain, and, in particular, the Estonian LT programmes that have been running since 2006.

Other European LT Programmes ::: The Netherlands

The STEVIN programme was launched in 2004 to strengthen the position of Dutch in LT by building essential resources for the language. Its objectives were to raise awareness of LT in order to stimulate demand for LT products, to promote strategic research in the field and develop essential resources, and to organise the management, maintenance and distribution of language resources that have been developed BIBREF1. The programme was based on cooperation between government, academia and industry, both in Flanders and the Netherlands. It encompassed a range of projects from basic resources to applications for language users, and attention was paid to distribution, dissemination and valorisation of project results by means of the HLT Agency, which also had a role in clearing intellectual property rights (IPRs) and issuing licence agreements BIBREF2.

The general targets of the STEVIN programme were reached to a large extent. According to a report on the results of the programme BIBREF3, it resulted in a network with strong ties between academia and industry, beneficial for future utilisation of the STEVIN results. The evaluators of the programme qualified it as successful, but had recommendations for a future programme, if initiated. They suggested more interaction with other similar (inter)national R&D programmes, asserted that the complexity of IPR issues had been seriously underestimated and called for a better clarification of the role of open-source. The total cost of the STEVIN programme was over 10 million euros, of which well over 80% was spent on R&D projects.

Other European LT Programmes ::: Spain

The Spanish LT programme Plan for Advancement of Language Technology started in 2016, and is scheduled to finish in 2020. Its aims are to develop infrastructure for LT in Spain, specifically for Spanish and the co-official languages, Basque, Catalan, Galician and Aranese. Furthermore, to promote the LT industry by boosting knowledge transfer between research and industry actors, and to improve the quality and capacity of public services by employing NLP and machine translation (MT) technology. Government should be the leading participant in LT with high-profile projects in healthcare, as well as in the judicial and educational systems, and in tourism BIBREF4.

The plan was to facilitate the development of tools and linguistic resources. Examples of tools are named entity recognisers, word-sense disambiguation, tools for computing semantic similarity and text classification, automatic summarisation and MT. Examples of linguistic resources to be developed in the programme are parallel corpora, lists of proper nouns, terminology lists and dictionaries.

The estimated total cost of the programme was 90 million euros. As the programme had just recently started when the Icelandic programme was being planned, we did not have any information on what went well and what could have been done better.

Other European LT Programmes ::: Estonia

Regarding LT, the Estonian situation is, in many ways, similar to that of Iceland: It has too few users for companies to see opportunities in embarking on development of (costly) LT, but on the other hand society is technologically advanced – people use, or want to be able to use, LT software. In Estonia, the general public wants Estonian to maintain its status, and like Icelandic, the language has a complex inflection system and very active word generation. The problems faced by Estonia are therefore not unlike those that Iceland faces.

In Estonia, three consecutive national programmes have been launched. The third national programme, Estonian Language Technology 2018–2027, is currently under way. While the Estonian Ministry of Education and Research has been responsible for the programmes, the universities in Tallinn and Tartu, together with the Institute of the Estonian Language, led the implementation.

The National Programme for Estonian Language Technology was launched in 2006. The first phase ran from 2006 to 2010. All results of this first phase, language resources and software prototypes, were released as public domain. All such resources and tools are preserved long term and available from the Center of Estonian Language Resources. 33 projects were funded, which included the creation of reusable language resources and development of essential linguistic software, as well as bringing the relevant infrastructure up to date BIBREF5. The programme managed to significantly improve upon existing Estonian language resources, both in size, annotation and standardisation. In creating software, most noticeable results were in speech technology. Reporting on the results of the programme BIBREF5 stress that the first phase of the programme created favourable conditions for LT development in Estonia. According to an evaluation of the success of the programme, at least 84% of the projects had satisfactory results. The total budged for this first phase was 3.4 million euros.

The second phase of the programme ran from 2011 to 2017 with a total budget of approx. 5.5 million euros. It focused on the implementation and integration of existing resources and software prototypes in public services. Project proposals were called for, funding several types of actions in an open competition. The main drawback of this method is that it does not fully cover the objectives, and LT support for Estonian is thus not systematically developed. Researchers were also often mostly interested in results using prototypes rather than stable applications. As most of the projects were instigated at public institutes, relation to IT business was weak. Furthermore, the programme does not deal explicitly with LT education. On the other hand, the state of LT in Estonia soon become relatively good compared to languages with similar number of speakers BIBREF6.

History of Icelandic LT

The history of Icelandic LT is usually considered to have begun around the turn of the century, even though a couple of LT resources and products were developed in the years leading up to that. Following the report of an expert group appointed by the Minister of Education, Science and Culture BIBREF7, the Icelandic Government launched a special LT Programme in the year 2000, with the aim of supporting institutions and companies to create basic resources for Icelandic LT work. This initiative resulted in a few projects which laid the ground for future work in the field. The most important of these were a 25 million token, balanced, tagged corpus, a full-form database of Icelandic inflections, a training model for PoS taggers, an improved speech synthesiser, and an isolated word speech recogniser BIBREF8.

After the LT Programme ended in 2004, researchers from three institutions, UI, RU, and the Árni Magnússon Institute for Icelandic Studies (AMI), joined forces in a consortium called the Icelandic Centre for Language Technology (ICLT), in order to follow up on the tasks of the Programme. In the following years, these researchers developed a few more tools and resources with support from The Icelandic Research Fund, notably a rule-based tagger, a shallow parser, a lemmatiser, and a historical treebank BIBREF9.

In 2011–2012, researchers from the ICLT also participated in two speech technology projects initiated by others: A new speech synthesiser for Icelandic which was developed by the Polish company Ivona, now a subsidiary of Amazon, for the Icelandic Association for the Visually Impaired, and a speech recogniser for Icelandic developed by Google BIBREF9.

Iceland was an active participant in the META-NORD project, a subproject of META-NET, from 2011 to 2013. Within that project, a number of language resources for Icelandic were collected, enhanced, and made available, both through META-SHARE and through a local website, málföng.is (málföng being a neologism for `language resources'). Among the main deliveries of META-NET were the Language White Papers BIBREF10. The paper on Icelandic BIBREF11 highlighted the alarming status of Icelandic LT. Icelandic was among four languages that received the lowest score, “support is weak or non-existent” in all four areas that were evaluated.

The White Paper received considerable attention in Icelandic media and its results were discussed in the Icelandic Parliament. In 2014, the Parliament unanimously accepted a resolution where the Minister of Education, Science and Culture was given mandate to appoint an expert group which should come up with a long-term LT plan for Icelandic. The group delivered its report to the Minister in December 2014. The result was that a small LT Fund was established in 2015.

During the last years, a strong centre for speech technology has been established at RU, where development in speech recognition and synthesis has been ongoing since 2011. Acoustic data for speech recognition was collected and curated at RU BIBREF12, BIBREF13, BIBREF14 and a baseline speech recognition system for Icelandic was developed BIBREF15. Specialised speech recognisers have also been developed at RU for the National University Hospital and Althingi BIBREF16, BIBREF17, BIBREF18. A work on a baseline speech synthesis system for Icelandic has also been carried out at RU BIBREF19, BIBREF20.

The AMI has built a 1.3-billion-word corpus, the Icelandic Gigaword Corpus (IGC) BIBREF21, partially funded by the Icelandic Infrastructure Fund. Further, a private company, Miðeind Ltd., has developed a context-free parser BIBREF22 partially funded by the LT Fund.

In October 2016, the Minister of Education, Science and Culture appointed a special LT steering group, consisting of representatives from the Ministry, from academia, and from the Confederation of Icelandic Enterprise (CIE). The steering group commissioned three LT experts to work out a detailed five-year Project Plan for Icelandic LT. The experts delivered their proposals, Language Technology for Icelandic 2018–2022 – Project Plan BIBREF0 to the Minister in June 2017.

Organisation of the Icelandic LT Programme 2019–2023

The Icelandic Government decided soon after the publication of the report Language Technology for Icelandic 2018–2022 – Project Plan to use the report as a base for a five-year government funded LT programme for Icelandic. The self-owned foundation Almannarómur, founded in 2014 to support the development of Icelandic LT, was to be prepared to take over a role as a Centre of Icelandic LT and to elaborate on how the programme could be organised and executed to meet the goals defined in the report.

The Icelandic Ministry of Education, Science and Culture signed an agreement with Almannarómur in August 2018, giving Almannarómur officially the function of organising the execution of the LT programme for Icelandic. Following a European Tender published in March 2019, Almannarómur decided to make an agreement with a consortium of universities, institutions, associations, and private companies (nine in total) in Iceland (listed in Table TABREF6) to perform the research and development part of the programme. This Consortium for Icelandic LT (Samstarf um íslenska máltækni – SÍM) is a joint effort of LT experts in Iceland from academia and industry. SÍM is not a legal entity but builds the cooperation on a consortium agreement signed by all members. During the preparation of the project, an expert panel of three experienced researchers from Denmark, the Netherlands, and Estonia was established to oversee the project planning and to evaluate deliverables at predefined milestones during the project.

SÍM has created teams across the member organisations, each taking charge of a core project and/or defined subtasks. This way the best use of resources is ensured, since the team building is not restricted to one organisation per project. One project manager coordinates the work and handles communication and reporting to Almannarómur and the expert panel.

Besides the role of the executive of the research and development programme itself, Almannarómur will conduct communication between the executing parties and the local industry, as well as foreign companies and institutions. Together with the executing parties, Almannarómur will also host conferences and events to promote the programme and bring together interested parties.

Core Projects

In this section, we describe the five core projects that have been defined in the Icelandic LT programme.

Core Projects ::: Language Resources

As mentioned above, a number of language resources have been made available at the repository málföng. Most of these are now also available at the CLARIN-IS website and will be integrated into the CLARIN Virtual Language Observatory. Below we give a brief and non-exhaustive overview of language resources for Icelandic which will be developed in the programme.

Tagged corpora. The IGC BIBREF21 contains 1.3 billion running words, tagged and lemmatised. It is much bigger than previous tagged corpora, most notably the Icelandic Frequency Dictionary (IFD; Pind et al., 1991), which was the first morphologically tagged corpus of Icelandic texts, containing half a million words tokens from various texts, and the Tagged Icelandic Corpus (MÍM; Helgadóttir et al,. 2012), a balanced corpus of texts from the first decade of the 21st century, containing around 25 million tokens. A gold standard tagged corpus was created from a subset of MÍM BIBREF23. Some revisions of the morphosyntactic tagset used for tagging Icelandic texts will be done in the programme, and the gold standard updated accordingly.

We will update the IGC with new data from more sources and continue collecting data from rights holders who have given their permission for using their material. A new version will be released each year during the five-year programme.

Treebanks. The largest of the syntactically parsed treebanks that exist is the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al., 2011; Rögnvaldsson et al., 2011, 2012), which contains one million words from the 12th to the 21st century. The scheme used for the syntactic annotation is based on the Penn Parsed Corpora of Historical English BIBREF24, BIBREF25. On the other hand, no Universal

Dependencies (UD)-treebanks are available for Icelandic. Within the programme, a UD-treebank will by built, based on IcePaHC, and extended with new material.

Morphological database. The Database of Icelandic Morphology (DIM; Bjarnadóttir et al., 2019) contains inflectional paradigms of about 287,000 lemmas. A part of the database, DMII-Core, only includes data in a prescriptive context and is suited for language learners, creating teaching material and other prescriptive uses. It consists of the inflection of approx. 50,000 words. We will extend it by reviewing ambiguous inflection forms. We will define format for data publication as the core will be available for use by a third party. For the sake of simplifying the process of adding material to the database and its maintenance, we will take advantage of the lexicon acquisition tool described in Section SECREF16 and adapt it for DIM.

Hyphenation tool. Hyphenation from one language to another often seems rather idiosyncratic but within one and the same language, such as Icelandic, such rules are often reasonably clear. A list of more than 200,000 Icelandic words with permissible hyphenations is available in the language resources repository. It will be expanded based on words from the DIM. A new hyphenation tool, trained on the extended list, will be built in the programme. The tool makes a suggestion for correct hyphenation possibilities of words that are not found on the hyphenation list.

Icelandic wordnet. The Icelandic wordnet BIBREF26, which contains 200,000 phrasemes of various kinds and about 100,000 compounds, is not a traditional dictionary as it analyses internal connections semantically and syntactically within Icelandic vocabulary. We will define a more appropriate data format and convert the wordnet data to that format. In addition, we will work on improving the wordnet itself by filling in gaps in various categories.

Core Projects ::: NLP Tools

A wide variety of NLP tools are to be developed or improved upon within the programme. It is of vital importance to develop quality NLP tools, as many tools often form a pipeline that analyses data and delivers the results to tools used by end users, and, in the pipeline, errors can accumulate and perpetuate.

When the programme started, there were a few available tools for Icelandic. IceNLP BIBREF27 is a suite of NLP tools containing modules for tokenisation, PoS-tagging, lemmatising, parsing and named entity recognition. Greynir BIBREF22 is a full parser which also includes a tokeniser and recognises some types of named entities. Nefnir BIBREF28 is a lemmatiser which uses suffix substitution rules, derived from the Database of Icelandic Morphology BIBREF29, giving results that outperform IceNLP. ABLTagger BIBREF30 is a PoS tagger that outperforms other taggers that have been trained for tagging Icelandic texts.

Some of these tools give good results, but can be improved upon. For other tasks, new tools need to be built. As part of the release process care will be taken to ensure all resulting software are up to high quality standards, and well documented to facilitate use by third parties. Where applicable, RESTful APIs will also be set up to further promote the usage of the products.

Tokeniser. A basic step in NLP is to segment text into units, normally sentences and tokens. Since any errors made at this stage will cascade through the process, it is important that the tokeniser is as accurate as possible. A tokeniser for Icelandic needs to be able to correctly recognises abbreviations, time units, dates, etc. It must also be adjustable and able to run using different settings, since its output must be adaptable to different projects and different uses.

Previously, two tokenisers have been built for Icelandic, one is a part of IceNLP and the other a part of Greynir. As Greynir is still in active development, it will be used as a base for the LT project's development. In order to be able to test the tokenisers' accuracy, a test set that takes different tokeniser settings into account will be developed.

PoS tagger. Precise PoS-tagging is important in many LT projects because information on word class or morphological features is often needed in later stages of an NLP pipeline. Improved tagging accuracy, thus often results in an improvement in the overall quality of LT software.

A number of PoS-taggers have been developed for Icelandic, with the best results achieved by a recent bidirectional LSTM tagging model BIBREF30. While developing PoS taggers for Icelandic further using state-of-the-art methods, we will also study and try to estimate how much accuracy can theoretically be reached in tagging a variety of Icelandic text styles, using the tag set chosen for the LT programme (see Section SECREF7).

Lemmatiser. A new lemmatiser for Icelandic, Nefnir, has recently been published BIBREF28. It has been shown to be quite accurate, although a standardised test set is not available to compare it to other lemmatisers, like Lemmald BIBREF31. Its main weakness is in lemmatising unknown words, which is a hard problem for inflected languages. We will study if its accuracy can be improved in that regard.

Parser. Three parsers have previously been developed for Icelandic. IceNLP includes a shallow parser based on a cascade of finite-state transducers BIBREF32. Greynir, on the other hand, fully parses sentences according to a hand-crafted context-free grammar. A parsing pipeline for Icelandic based on the IcePaHC corpus and the Berkeley-parser has also been released BIBREF33. No Universal Dependencies (UD) parser is available for Icelandic and no UD treebank, but in a project that started in 2019, independent of the LT programme, IcePaHC BIBREF34 will be converted to a UD treebank.

The IceNLP and Greynir parsers will be evaluated and either one of them or both developed further. We will also adapt a UD-parser to Icelandic UD-grammar.

Named entity recogniser. Some work has been carried out on named entity recognition for Icelandic. IceNLP contains a rule-based module that has achieved 71-79% accuracy and a recent tool based on a bidirectional LSTM BIBREF35 obtained an F1 score of 81.3%. There is also a named entity recogniser for proper names in Greynir, but its accuracy has not yet been evaluated. Within the programme, different training methods will be experimented with and evaluated, and the most promising tools evaluated further.

Semantic analysis. A variety of different tasks involve semantic analysis, including word-sense disambiguation (WSD), anaphora resolution, identifying co-references, analysing semantic similarity between compound verbs and phrases, and more.

We will work on these four aspects of semantic analysis listed above. In recent years, not much work has been carried out in this field for Icelandic. This part of the LT programme will thus start with researching the current state-of-the-art and defining realistic goals.

Lexicon acquisition tool. When constructing and maintaining lexical databases, such as DIM, the Icelandic wordnet or other related resources, it is vital to be able to systematically add neologies and words that are missing from the datasets, especially those commonly used in the language. Within the LT programme a flexible lexicon acquisition tool will be developed. It will be able to identify and collect unknown words and word forms, together with statistics, through structured lexical acquisition from the Icelandic Gigaword Corpus, which is constantly being updated, and other data sources in the same format.

Core Projects ::: Automatic Speech Recognition (ASR)

The main aim of the automatic speech recognition (ASR) project is to gather all necessary language and software resources to implement and build standard speech recognition systems for Icelandic. The project should enable developers to either research, develop or implement ASR without having to gather language resources. To achieve this goal, the project is divided into data gathering, recipe development, and software implementation and research.

Data gathering. The data gathering part of the project encompasses a wide variety of speech and transcript resources. A continuation of the Málrómur project BIBREF14 has already been implemented using Mozilla Common Voice. Here the aim is to double the size of the existing data set, get a more even distribution of speakers across geographic locations and age groups, and gather data from second language speakers. Additionally, radio and television transcripts are being gathered on a large scale and prepared for publication for ASR development. Conversations, queries and lectures will also be transcribed and published, and large open historical data sets will be aligned and prepared for publication.

Recipe development. ASR recipes for Icelandic will be developed further using more language resources BIBREF15 and specific application areas such as conversations, question answering and voice commands will be given a special attention. ASR systems that focus on teenagers, children and second language speakers are also within the scope of the project. These recipes are then used to create resources for smart-phone and web-based integration of ASR for Icelandic.

Software implementation and research. The research areas are chosen so to enhance the language resource development for Icelandic. A punctuation system for Icelandic will be analysed and implemented. Compound words are common in Icelandic and the language also has a relatively rich inflection structure so it is important to address those features for language modeling. Pronunciation analysis, speaker diarization and speech analysis will also be addressed especially for Icelandic, and acoustic modelling for children and teenagers receive attention in the project.

Core Projects ::: Speech Synthesis (TTS)

. The text-to-speech project will produce language resources that enable voice building for Icelandic.

Unit selection. Eight voices for unit-selection TTS will be recorded, with the aim of attaining diversity in age and dialect, with an equal number of male and female voices. The reason why unit-selection is chosen is to increase the likelihood that the project will produce useful and viable voices that can be used in addition to the two unit-selection voices that already exist for Icelandic.

Statistical parametric speech synthesis. Forty voices for statistical parametric speech synthesis (SPSS) will be recorded during the project. The plan is to publish open-source unit-selection and SPSS recipes with all necessary language resources so that programmers and researchers can continue to develop voices for Icelandic.

Suitable TTS voices for web-reading and smartphones will be developed within an open-source paradigm. This will allow the industry to use the voices developed within the project.

Research. The targeted research part of the project will facilitate the recipe development and software implementation. Quality assessment systems will be set up, text normalization for Icelandic will be developed fully, and intonation analysis for Icelandic will be implemented and applied to TTS.

Core Projects ::: Spell and Grammar Checking

The Spell and Grammar Checking project will develop and make freely available, under open-source licensing, important data sets and tools for further establishment of automated text correction systems for Icelandic. The project makes extensive use of other resources that have been developed independently, or will be developed within the larger framework of the current LT Programme for Icelandic, including the Database of Icelandic Morphology BIBREF29, the Greynir system BIBREF22, and the Icelandic Gigaword corpus BIBREF21. On the one hand, the project focuses on developing error corpora for Icelandic, and on the other, it focuses on creating a set of correction tools. Challenges associated with richly inflected languages continue to be a matter of central interest in this project, like previous work on Icelandic spelling correction BIBREF36.

Text correction data. The data construction aspect of the project will develop three error corpora that can be used for quantitative analysis of errors in written Icelandic text. The error corpora will also serve as a foundation for training data-driven training correction systems. One corpus will focus on the written language of Icelandic speakers who are not known to have unusual language properties. Another corpus will focus on speakers who are in the process of learning Icelandic as a second language, and a third one will include data from dyslexic speakers.

Software development. The software development tasks of the spell and grammar checking project will build a working open source correction system whose development is informed by the analysis of the data sets created within the project. The spell and grammar checker will be based on the foundation for processing Icelandic text provided by the Greynir system.

Core Projects ::: Machine Translation

The purpose of the MT project is to build open-source systems capable of translating between Icelandic and English, in both directions, is$\rightarrow $en and en$\rightarrow $is. The goal is that the translation quality will be good enough to be useful for translators in specific domains. A part of the MT project is indeed to define in which translation domain most value can be gained with the systems.

Very limited work on MT for Icelandic has been carried out since the turn of the century. A prototype of an open-source is$\rightarrow $en rule-based MT system has been developed using the Apertium platform BIBREF37, but this system is not currently in public use.

The AMI has recently compiled an English-Icelandic parallel corpus, ParIce, the first parallel corpus built for the purposes of MT research and development for Icelandic BIBREF38. The primary goal of the compilation of ParIce was to build a corpus large enough and of good enough quality for training useful MT systems. ParIce currently consists of 39 million Icelandic words in 3.5 million segment pairs. The largest parts of ParIce consists of film and TV subtitles from the Opus corpus BIBREF39, and texts from the European Medicines Agency document portal, included in the Tilde MODEL corpus BIBREF40.

Google Translate supports translations between Icelandic and various languages and is currently used widely by Icelanders and foreigners for obtaining understandable translations of given texts (the task of assimilation). The problem with Google's system is, however, that neither the source code nor the training data is publicly available. Moreover, the system is a general translation engine, but not developed specifically for translating texts in a particular domain.

Our MT project in the new LT programme consists of the following sub-parts:

Parallel data. Icelandic's rich morphology and relatively free word order is likely to demand large amount of training data in order to develop MT systems that produce adequate and fluent translations. The ParIce corpus currently consists of only 3.5 million sentence pairs which is rather small in relation to parallel corpora in general. The goal of this phase is to create an aligned and filtered parallel corpus of translated documents from the European Economic Area (EEA) domain (e.g. regulations and directives). As of

2017, around 7,000 documents were available in Icelandic with corresponding documents in English. The aim is to pair all accessible documents in the course of the project.

Back-translation. In order to augment the training data, back-translated texts will be used. Monolingual Icelandic texts will be selected and translated to English with one of the baseline system (see below). By doing so, more training data can be obtained for the en$\rightarrow $is direction. An important part of using back-translated texts during training is filtering out translations that may otherwise lead to poor quality of the augmented part.

Baseline system. In this part, three baseline MT systems will be developed. First, a statistical phrase-based MT system based on Moses BIBREF41, second, a bidirectional LSTM model using the neural translation system OpenNMT BIBREF42, and, third, a system based on an attention-based neural network BIBREF43 using Tensor2Tensor. All the three systems will be trained on ParIce, and the additional data from tasks 1 and 2 above. Eventually, the goal is to choose the best performing MT-system for further development of MT for Icelandic.

MT interface. An API and a web user interface for the three baseline systems, mentioned in item 3 above, will be developed to give interested parties access to the systems under development, and to establish a testing environment in which members of the public can submit their own text. Thus, results from the three systems can be compared directly, as well as to the translations produced by Google Translate. Moreover, in this part, a crowd-sourcing mechanism will be developed, i.e. a functionality to allow users to submit improved translations back to the system for inclusion in the training corpus.

Pre- and postprocessing. Preprocessing in MT is the task of changing the training corpus/source text in some manner for the purpose of making the translation task easier or mark particular words/phrases that should not be translated. Postprocessing is then the task of restoring the generated target language to its normal form. An example of pre- and postprocessing in our project is the handling of named entities (NEs). NEs are found and matched within source and target sentence pairs in the training corpus, and replaced by placeholders with information about case and singular/plural number. NE-to-placeholder substitution is implemented in the input and placeholder-to-NE substitution in the output pipelines of the translation system.

Conclusion

We have described a five-year, national LT programme for Icelandic. The goal is to make Icelandic useable in communication and interactions in the digital world. Further, to establish graduate and post-graduate education in LT in Iceland to enable the building of strong knowledge centres in LT in the country.

After studying somewhat similar national programmes in other European countries, we have defined the most important factors that in our opinion will help lead to the success of the programme: First, we have defined core projects that comprise the most important language resources and software tools necessary for various LT applications. Second, all deliverables will be published under as open licenses as possible and all resources and software will be easily accessible. The deliverables will be packaged and published for use in commercial applications, where applicable. Third, from the beginning of the programme, we encourage innovation projects from academia and industry through a competitive R&D fund, and fourth, constant communication with users and industry through conferences, events and direct interaction will be maintained, with the aim of putting deliverables to use in products as soon as possible. The cooperation between academia and industry is also reflected in the consortium of universities, institutions, associations, and private companies, performing the R&D work for all core projects.

The described plan is tied in with 20 years of LT history in Iceland, and despite the steep path to getting where we are, we have every reason to be optimistic about the future of Icelandic LT.