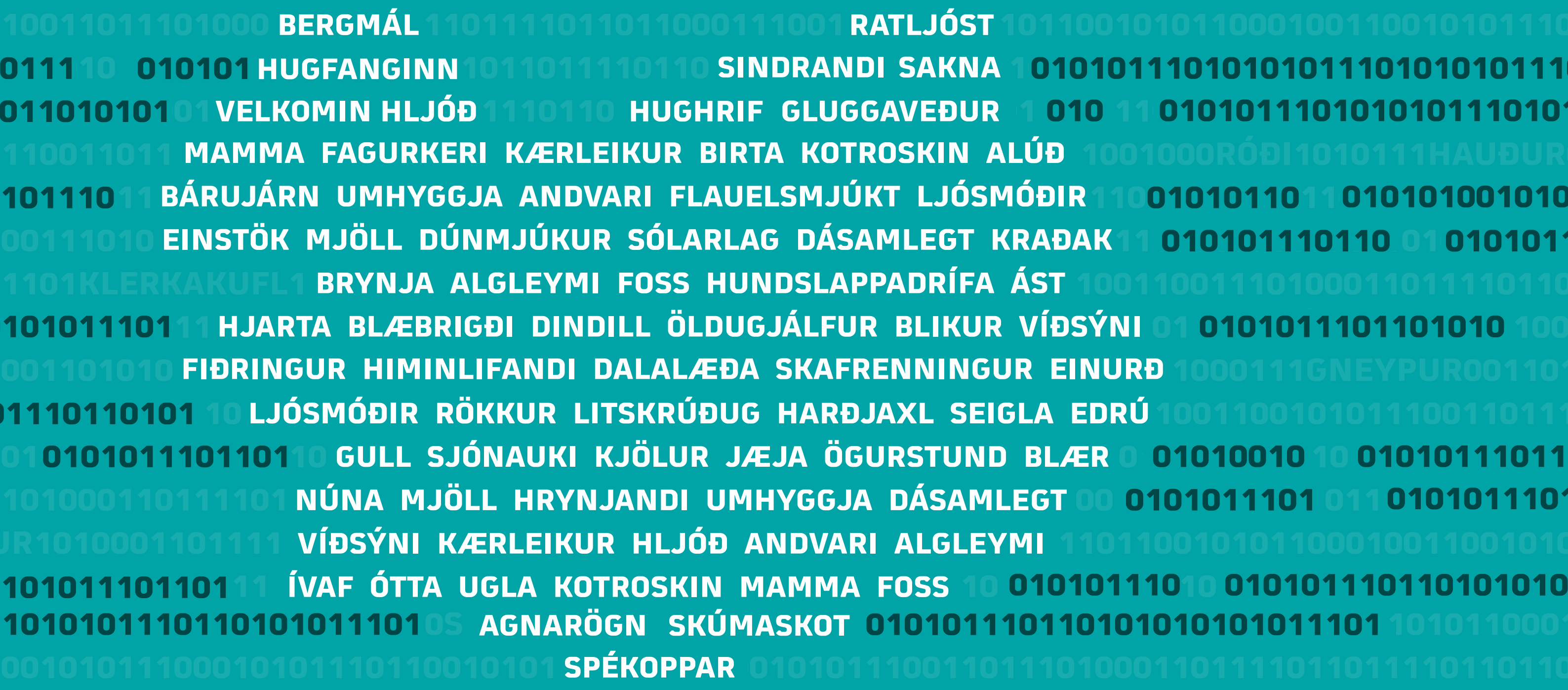

BERGMÁL RATLJÓST
HUGFANGINN SINDRANDI SAKNA
VELKOMIN HLJÓÐ HUGHRIF GLUGGAVEÐUR
MAMMA FAGURKERI KÆRLEIKUR BIRTA KOTROSKIN ALÚÐ
BÁRUJÁRN UMHYGGJA ANDVARI FLAUELSMJÚKT LJÓSMÓÐIR
EINSTÖK MJÖLL DÚNMJÚKUR SÓLARLAG DÁSAMLEGT KRAÐAK
BRYNJA ALGLEYMI FOSS HUNDSLAPPADRÍFA ÁST
HJARTA BLÆBRIGÐI DINDILL ÖLDUGJÁLFUR BLIKUR VÍÐSÝNI
FIÐRINGUR HIMINLIFANDI DALALÆÐA SKAFRENNINGUR EINURÐ
LJÓSMÓÐIR RÖKKUR LITSKRÚÐUG HARÐJAXL SEIGLA EDRÚ
GULL SJÓNAUKI KJÖLUR JÆJA ÖGURSTUND BLÆR
NÚNA MJÖLL HRYNJANDI UMHYGGJA DÁSAMLEGT
VÍÐSÝNI KÆRLEIKUR HLJÓÐ ANDVARI ALGLEYMI
ÍVAF ÓTTA UGLA KOTROSKIN MAMMA FOSS
AGNARÖGN SKÚMASKOT
SPÉKOPPAR

LANGUAGE TECHNOLOGY FOR ICELANDIC 2018-2022

PROJECT PLAN

ANNA BJÖRK NIKULÁSDÓTTIR
JÓN GUÐNASON
STEINÞÓR STEINGRÍMSSON

INDEX

1. PROGRAMME OBJECTIVES

2. CORE PROJECTS

AUTHORS OF THE REPORT

Anna Björk Nikulásdóttir completed her MA in Language Technology from Heidelberg University in 2007. During her studies she worked at the Fraunhofer Institute, in Darmstadt, and the European Media Laboratory, in Heidelberg. She participated in the project, "Viable Language Technology Beyond English – Icelandic as a Test Case" and worked on her doctorate in Language Technology at the University of Iceland between 2009 and 2012. Nikulásdóttir's masters and doctoral projects revolved around the automatic extraction of semantic relations in dictionaries and texts. She later worked as a software developer and project manager at VICO Research & Consulting, in Stuttgart, a company that specialises in analysing social data. She is currently employed at the Centre for Analysis and Design of Intelligent Agents, at Reykjavík University.

Jón Guðnason is an Associate Professor in Electrical Engineering at Reykjavík University, the Director of the Centre for Analysis and Design of Intelligent Agents, at the same university, and an expert in speech signal processing. In 2007, he completed his Signal Processing doctorate from the Imperial College, London. His thesis discussed how to recognise a voice by extracting characteristic features from a voice source. Guðnason completed his master's degree at the University of Iceland in 2000; his thesis discussed non-linear system identification of speech with recurrent neural networks. Between 2006 and 2008 he worked as a speech recognition expert at SpinVox, in the UK, and in 2008 and 2009 he was a Resident Scholar at Columbia University, New York. In 2009 he started working at Reykjavík University, where he leads projects in addition to lecturing in Signal Science and Machine Learning. He has set up a research group in language technology and voice signal processing.

Steinþór Steingrímsson holds a degree in Computer Science and Icelandic from the University of Iceland. He completed his MSc in Speech and Language Processing from Edinburgh University in 2005. Since 2011 he has worked on language-technology projects in Iceland, first on the META-SHARE project, and later on developing corpora at the Árni Magnússon Institute for Icelandic Studies, Tagged Icelandic Corpus, Icelandic Gigaword Corpus, The Malromur Corpus, and other lexicographic resources. He previously worked as a software developer for two banks, as a computer games programmer, a teacher and a news reporter. He is now the Information Technology Program Manager at the Árni Magnússon Institute for Icelandic Studies.

In planning this project, in order to gather information and advice on language technology and/or language technology programmes, the authors consulted the following experts:

Hynek Hermansky, Julian S. Smith Endowed Professor in Electrical Engineering and Director of the Johns Hopkins University Center for Language and Speech Processing, USA;

Kadri Vider, Research Fellow at the Institute of Computer Science at the University of Tartu, Estonia; Heiki-Jaan Kaalep, Senior Research Fellow at the Institute of Computer Science at the University of Tartu;

Mark Fišel, Chair of Language Technology and Associate Professor of Natural Language Processing at the Faculty of Mathematics and Computer Science at Tartu University:

Tanel Alumäe, Senior Researcher at the Department of Software Science, Tallinn University of Technology;

Einar Meister, Senior Researcher at the Department of Software Science, Tallinn University of Technology;

Meelis Mihkla, Head of Department and Senior Researcher at the Institute of the Estonian Language; Tõnis Nurk, Head of Department and Senior Researcher at the Institute of the Estonian Language; Martin Eessalu, Chief Expert at the Estonian Ministry of Education and Research;

Andero Adamson, Linguistics Expert at the Estonian Ministry of Education and Research;

Etienne Roth, Language Technology Expert at EPC Consulting & Software GmbH, Germany;

Markus Foti, Project Manager at MT@EC/eTranslation, Directorate-General for Translation (DGT); Andreas Eisele, Expert at MT@EC/eTranslation, DGT;

Michael Jeelinghaus, Expert at MT@EC/eTranslation, DGT;

Szymon Kocek, Expert at MT@EC/eTranslations, DGT;

Oddur Kjartansson, Program Manager in Language Technology and Data Collection, Google, London, UK;

Alexander Gutkin, Expert in Speech Synthesis, Google, London;

Hanna Silen, Speech Signal Specialist at Google, London;

Bente Maegaard, Senior Executive Consultant at the Centre for Language Technology at Copenhagen University and the national coordinator for CLARIN in Denmark;

Koenraad De Smedt, Professor in Language Technology at the University of Bergen and the national coordinator for CLARIN in Norway;

Jens Edlund, Lecturer in Speech Technology at KTH, Stockholm, Sweden;

Eiríkur Rögnvaldsson, Professor of Icelandic at the University of Iceland;

Anton Karl Ingason, Lecturer in Icelandic Linguistics at the University of Iceland;

Sigrún Helgadóttir, Specialist in Language Technology at the Árni Magnússon Institute for Icelandic Studies;

Kristín Bjarnadóttir, Research Reader at the Árni Magnússon Institute for Icelandic Studies;

Jón Hilmar Jónsson, Research Reader at the Árni Magnússon Institute for Icelandic Studies;

Jón Friðrik Daðason, Computer Scientist;

Vilhjálmur Þorsteinsson, Programmer.

We would like to thank all of them for their invaluable help and advice.

PREFACE

In the autumn of 2016, the Minister for Education, Illugi Gunnarsson, put together a steering group to oversee the mapping of language technology, and to define a strategy for Icelandic. The group was entrusted with the role of carrying out a status analysis of Icelandic language resources and of making an exact finance and five-year language technology programme. The announcement for the formation of the group said that "in coming years, the increasing effect of computers on our daily lives will demand action from the government to ensure that using Icelandic will be an option in all communications using computers and telecommunications technology. Knowledgeable opinion states that, should no action be taken, the Icelandic language is in grave danger. If the language can be used for communication on all types of smart devices, this will also present great opportunities for Icelandic society."

This report is the conclusion of a workgroup, which the steering group assembled to evaluate the status of Icelandic language technology and to create a five-year project plan. The workgroup suggests that four open core-solutions are created:

Speech Recogniser;

Speech Synthesiser;

Machine Translation System; and

Spell and Grammar Checker.

These will be developed and adapted to Icelandic to the point where they are fully usable, and used by the public, companies, governmental bodies, and institutions in Iceland. The principal prerequisite in building language technology tools is to have language resources and support tools in place, and the report describes the measures required to achieve this. In addition to the project schedule, suggestions are made as to how it will be carried out, with a description of how similar schedules have been executed in other countries.

Anna Björk Nikulásdóttir
Jón Guðnason
Steinþór Steingrímsson

SUMMARY

The future of computers will be intertwined with language technology. New artificial intelligence technology enables the use of vast collections of informational, textual and other linguistic data, in ways that previously were impossible. As a result, language will increasingly be used to communicate with devices, and devices will increasingly be used for working with language. These changes offer immense possibilities – and new demands of users. Automatic dialogue systems and question-answering systems can increase efficiency and improve the services of companies and institutions and machine translation systems can increase the productivity of translators and make more material accessible in any language. High-quality speech synthesisers can make far more books accessible in audio form than would otherwise be possible in a small market. Software allowing people to read and write, who were otherwise unable to read or write due to disability or illness, will transform their quality of life. It is vital for small language communities to employ this technology: it is not merely of an advantage to the language but can be crucial for its survival.

We are at a crossroads. We must choose whether to develop the infrastructure needed, so that the Icelandic language will be used in a transformed and technological world, or whether simply to wait and see what happens. If Icelandic does not embrace this new technology, devices will only use and be used in other languages. The availability of language technology for Icelandic will be accidental and haphazard, opportunities will be lost and speakers of Icelandic who are not adept at other languages will be left behind.

Developing technology can, of course, be expensive, but the cost of lost opportunities and retaining practices that are rapidly becoming obsolete can also be dear. The real choice lies between accepting losing out on the one hand, with lower quality of life and the inherent cost of not using the best available technology and investing in the technological development that increases the competitiveness of the economy, society and the language on the other hand.

To ensure that Icelandic will be an option in the technological world, we must make sure that the public, companies and institutions can use language technology and implement language technology solutions without being hampered by complicated and expensive infrastructure development.

In this context, this report emphasises the following key factors.

1 Infrastructure development: Basic tools and lexicographical and linguistic data are known as language resources. If these are underdeveloped, or do not exist, it is impossible to develop language technology. These data can be large and structured collections of text, audio recordings or glossaries which have been adapted for use in language technology. It is possible to build on previous work on data in Iceland, but in order to make full use of the technology's possibilities, it is imperative to invest more in development.

The basic tools in language technology are, for example, open-source speech recognisers and synthesisers that deal with everyday language and can be adapted for specific use. These may be tools for analysing speech and pronunciation, or the support tools necessary for the end-user of language technology. They may also be general machine translation systems that can be adapted as fit. It is vital that the tools will be open and accessible to all, so that anyone wanting to develop language technology solutions for Icelandic or software that incorporates such systems, can use these resources without having to carry out time-consuming basic research and development.

2 Language technology innovation: it is important to support, and to ensure the participation of, companies that practise language technology innovation and/or can use language technology tools to improve their services or production. These companies will create solutions to society's need for language technology – the infrastructure set out in this schedule will enable them to realise those solutions. This must be encouraged through the means of an incentive programme, as well as good interaction and co-operation between participants in each stage of the project.

3 Collaboration and clustering: International collaboration is of the utmost importance, in order to guarantee that the Icelandic language will be available in the devices and computers of the future. We must use great determination in order to ensure that large international corporations offer Icelandic in their systems, same as other languages. This can be achieved by regular communication with these companies, as well as with universities and institutes abroad, that work towards the same goals for other languages. It is important to collaborate on language technology development for less-resourced languages. In the initial stages, leading enthusiasts and other interested parties in Iceland will form a cluster, thereby creating an opportunity for local collaborative ventures and participation in international projects, that are useful for Icelandic language technology.

Good organisation and joint effort can make Icelandic a part of the digital world of the future.

Strong development of Icelandic language technology will enable us to integrate Icelandic into technology and services, so that using the language becomes a real option in all user interfaces and information processing.

EXCERPT

Language technology comprises everything that enables software to deal with language. It is straightforward for computers to carry out complex calculations and work with large amounts of data in closed systems, in which the rules are clear. Natural language, however, cannot be fully captured with rules. Although we have an innate ability to understand and speak a language, we do not fully understand how it is processed. Living languages are subject to change: new words are created, words change meaning, and there are infinite possibilities for forming sentences and connecting ideas and concepts. It is, therefore, challenging to make computers work with language in the same way people do, to understand speech sounds, words and sentences and to connect to the general understanding of the language and the world around us.

Since the middle of the 20th century, multiple different methods for speech and language processing have been developed. For a long time, language technology was mostly an academic subject with limited working solutions, but for the past few years a revolution has occurred: an unknown abundance of data, more (and cheaper) processing power, and novel ways of utilizing powerful algorithms, although they may be based on decades old foundations. It has also transpired that the prerequisite for the successful development of language technology is to have access to large amounts of data. Development hardware and algorithms has also enabled the processing of the necessary amount of data.

At this point, language technology equipment has become immersive, and it has become clear that languages that are not a part of this development will be under threat. People in technologically advanced communities become used to talking to devices and having their queries answered; they get used to searching for information, not only by using search words, but also with the help of software that extracts information from large amounts of different data, connects the information and draws conclusions. It is also possible to dictate a text for a computer to transcribe, and for a computer to read texts aloud; to be able to understand foreign-language speech and text with the aid of machine translations, and so on.

Language technology does not only have a great impact on people's daily lives and communications. We live in the information age, where data and information are among the most valuable assets that companies and institutions can have; those that utilise data in an intelligent way, increase their competitiveness. Participation in this development is of the utmost importance to sustain in a global market; the need to respond to this development is reflected in the fact that scarcely any modern company or institution does not provide information or services on the internet - something which did not seem overly important in earlier times.

Language technology comprises many specialised fields, that deal with specialised areas. Different subfields demand different and often diverse specialised knowledge: computer science, linguistics, engineering, mathematics, philosophy and statistics are some of the areas that can be applied. Traditional language technology education tends to revolve around interweaving computer science and linguistics. It is necessary to have a core of language technology educated experts, but many other specialists can participate in creating a strong knowledge industry in Iceland. Although specialised language technology solutions need to be developed for Icelandic, these may very well be adapted for other languages and, therefore, for a larger market.

The Icelandic language technology programme 2018-2022 aims to ensure that Icelandic can be used for communication with machines, with other people through machine intermediaries and in processing written and spoken information. What needs to be done in order to achieve this has been reported on before, first in a 1999 report by a workgroup on language technology. The current report discusses what needs to be done in the time allowed to achieve it, how the work and project should be organised, and the importance of interaction between local participants and international enterprises and collaborators.

A short excerpt of the report's content follows; references are made to further discussion where applicable.

PRIORITY PROJECTS FOR THE LANGUAGE TECHNOLOGY PROGRAMME

The priority projects in the programme are those that form the necessary foundation for further development in the different areas of language technology for Icelandic. We divide these into speech recognition, speech synthesising, machine translation, spell and grammar checking, and language resources.

Speech Recognition (Chapter 2.1) is about converting speech into written text. It is the prerequisite for us being able to communicate with computers and machines in a way that is natural to most of us – by speech. The potential for voice-controlled communication is particularly important in circumstances where you cannot use your hands – or where it is too distracting to do so – for example while driving, or for people who are less able to write or have trouble using a keyboard or touch-screen.

Speech recognition can be used to make a written record of a long, continuous recital, speeches or dictations; to follow dialogue, or even participate in it; or to take voice orders for further analysis. Speech recognition software is a "keyboard" for the voice. In common with other fields of language technology – and with artificial intelligence in general – specialised software is often a more realistic solution than software that is supposed to work in all circumstances. A number of things can influence the development of a speech recognition system: it matters who is speaking, in what circumstances and what the content of the speech is.

- **Who is speaking?** Voice and pronunciation are individual. Certain groups have, however, more in common with one another than with others: female voices are more similar to one another than they are to male voices, people of a similar age from the same area often use similar pronunciation, etc. It is necessary to have audio recordings that are typical of the group that the technology is meant to recognise. For example, a traditional speech recogniser does not work well for children because their voices are too different from adult voices.

- **What are the circumstances?** A number of conditions can have an impact on speech recognition: background noise (traffic, nature sounds, chatter or other arbitrary sounds, such as a bell in parliamentary chambers); more people talking close to the recorder; how direct the speech is (not too hesitant and repetitive); and the quality of the recording. For example, in designing speech recognition for automobile navigation systems, vehicle noise and the background chatter of passengers must be considered.

- **What is the content?** Simple voice control can mean that a system is programmed to differentiate between certain single words. Communications are, however, normally much more complicated and involve diverse vocabulary and sentence structures. The system can be restricted to certain topics, for example medicine, or be completely open, but it must be adapted to the topic it is likely to be used for.

Objective: To create a general Icelandic speech recognition system that is accessible through a web service. All procedures and data will also become accessible as a basis for developing specialised speech recognisers; to make Icelandic speech recognisers available in smart devices; and to develop speech recognition as part of voice control, question answering and dialogue systems. That work will be carried out on developing speech recognisers for children and young adults.

What is needed: A vast quantity of varied data; software that contains thoroughly tested algorithms for speech recognition and the potential to adapt it to individual systems; knowledge of pronunciation, dialects and language structures; and the tools to use that knowledge.

Speech Synthesis (Chapter 2.2) converts written text into speech. Two main areas of speech synthesis software are dictation and (verbal) communication. Speech synthesisers are used to read text, for example from websites or books. People who for some reason are unable to read, or struggle with reading, rely on speech synthesis in their daily lives. Communication systems, in which a speech recogniser hears what the user is saying, need speech synthesisers to be able to answer in a human-like voice. The answers should sound normal and in harmony with the topic being discussed. Speech synthesisers have different requirements, depending on the domain they are designed to work in.

- **Listener's objective.** A listener whose objective is to read certain material, for example a course book, as quickly as possible needs a speech synthesiser which reads quickly and clearly. In normal communications or when reading a novel, however, the intonation and emphasis is far more important than speed.

- **The role of the speech synthesiser.** Speech synthesisers that read books or newspaper articles are one-sided in that they do not need to react to anything the user might say. These speech synthesisers should be designed to read long texts in a way that meets the listener's demands; synthesisers that react to what users say are designed to read shorter questions, answers and messages.

- **Listeners' preferences.** People do not necessarily agree which voices are most agreeable. Speech synthesisers that speak in different male and female voices, and that may be adjusted to things such as strength and stress, should be available.

Objective: To create an Icelandic speech synthesiser, accessible through a web portal. To create an open environment for the development of various speech synthesisers that speak Icelandic.

What is needed: Specialised recordings to develop the speech synthesisers; a quality dictionary on pronunciation that contains a sufficient number of words and different pronunciations and dialects; a system to prepare text for reading by a speech synthesiser (text standardisation), and knowledge and tools for speech and intonation; software to develop speech synthesisers, using tested methods.

Machine translation (Chapter 2.3) is the automatic translation from one language to another. This has already become useful for many language pairings, to help people understand the content of text that is written in a language they are not familiar with, and to expedite the work of translators in languages they speak fluently. As of yet, however, no translation software is capable of delivering translations that are satisfactory for more than a few domains – the text must be reviewed if the translation is to be accurate.

- **Which two languages form the language pair for translation?** Each machine translation system is normally restricted to one language pair. For translating between languages for which insufficient data exist, some systems use another language as a pivot language, and thus project the translation into a number of others.

- **What is being translated?** A machine translation system that is designed to be able to translate anything, struggles with ambiguity to a much greater extent than a system which is tailor-made for particular domains. Translation systems must recognise the types of text they are meant to translate; a system that has been designed using administrative texts is not likely to be a very good translator of sports news.

- **What is the purpose of the translation?** Machine translation systems are unlikely to make translators obsolete in the near future by delivering perfectly translated texts. They can, however, be an invaluable support tool and can save companies and institutions a great deal of money and time: without them translators would have to carry out much more work from scratch, and as systems may be able to learn from translators' corrections, they can get better over time. Another important use of machine translation is gisting, giving the system's users a general idea about the content of a given text, when there is no need for a perfect translation.

Objective: To create an open-source machine translation system which translates between English and Icelandic. An important goal should be to create a useful system for particular domains, to enable translators to work more quickly.

What is needed: Large, parallel corpora of Icelandic and English texts, along with open-source software to develop machine translation systems, using common and well-known methods.

Spell and grammar checking (Chapter 2.4) assists in correcting and writing text. There may be various types of errors in any written text: typographical, spelling, grammatical or incorrect word usage. Software for automatic error correction is an important aid, for the general public as well as for companies and institutions, when writing text. This technology is also imperative in developing other kinds of language technology software and is necessary for OCR-read texts to become fully useful. The usefulness and importance of automatic spell and grammar checking, and correction increases in correlation with the level of skill of the writer as well as on the time available and the level of quality required.

Points of focus for developing spell and grammar checking software.

- **There are many different types of errors.** It is necessary to define what the spell and grammar checker is supposed to correct so that users – people or other software – can rely on the results.

- **Text origin.** Spell and grammar checking software may need to analyse OCR-read text, text from other software, or text written using word processors and must be specifically adapted to each type. It must also take into consideration the variation in people's writing skills. Traditional spell and grammar checking are developed for those who are skilled in their language, have received training and find it relatively easy to write. Other groups, including dyslexics, people who have a different first language, and children who are learning to read and write, need a different type of support.

- **The purpose of correction.** As a part of word processing software, spell and grammar checking offers the user a chance to make a correction when it finds an error. More comprehensive support also shows why it is a mistake, for example, by referring to a grammatical rule. As a part of other language technology software, spell and grammar checking must, however, be integrated into the system and decide which correction to use each time.

Objective: To develop a general spell and grammar checker which can find and correct the most common errors found in normal Icelandic text; to gain knowledge of the nature of errors made by different groups of people; to develop methods to adapt the system to different needs, for example in relation to training and teaching; to make sure that people are able to use the spell and grammar checking software, regardless of their operating system or word processing software; to ensure that other language technology software can use the spell and grammar checker and adapt it.

What is needed: Corpora of misspellings and grammar errors; dependable support tools for grammatical and semantic analysis.

Language resources (Chapter 2.5). All language technology is built on language resources: texts and/or audio recordings. These are necessary in the analysis of language, identification of new and emerging words and senses, and in finding rules and patterns. From language resources it is possible to "teach" the computers what is important for the software being developed. It is, however, becoming increasingly common for software to be made to find rules and patterns by itself and to learn to analyse the language largely without ready-made rules. These methods, which often give much better results than manual methods, demand a great deal of data, which must sometimes be prepared in a particular way. Different software may also demand different data. The project is focused on establishing important corpora of spoken and written language. It is also necessary to prepare data, such as vocabulary, pronunciations, and semantics, for use in language technology. Chapters 2.5.1.8 – 2.5.1.18 give a detailed description of this.

Although many specialised solutions exist for language technology, there is certain basic software which is used at all stages. These are normally hidden tools which analyse core units in texts, from analysing what does or does not constitute a word, to analysing a text's complex grammatical and semantic context. These support tools are not fully rounded software solutions in themselves but are a necessary component of language technology software and data processing (Chapter 2.5.3). Good support tools, which are easy to use and give reliable results, are vital for quality language technology solutions.

A necessary amount of relevant data, together with reliable support tools, is the solid foundation for all further development. These are not only extremely important to the quality of complex software, but also make for speedier development.

Objective: To collect data and extract corpora of spoken and written language from it. The focus will be on further work with large corpora of written and spoken language, but also specialised corpora that are defined for individual core projects; to continue working on important resources, such as a pronunciation dictionary, a morphological dictionary and a wordnet; and to continue to develop, and add to, the support tools we already have.

What is needed: Data must be collected from various sources – from the internet or from institutions and companies – and necessary licences be acquired for its use; to a considerable extent, speech data must be created by, for example, recording people's conversations; experts must prepare all data for usage in language technology; lexical data needs to be formatted for use in software development; some support tools need to be created from scratch; licences must be acquired for those that are already in place and need further work; and test data are necessary for all support tools.

Chapter 3 deals with licensing and accessibility. All data and tools will be released with as open a licence as possible to enable the infrastructure to be used as widely as possible. Standards, access and maintenance of all infrastructure and other projects must be considered.

Once the development of basic infrastructure is well established, we must start work on developing various language technology tools currently unavailable for Icelandic (see Chapter 4): information extraction, sentiment analysis, information retrieval, question answering systems, dialogue systems and multimedia analysis. These projects should be scheduled as soon as possible.

Innovation must take the central role in Icelandic language technology. The infrastructure that will be developed within the language technology programme should enable companies to develop language technology solutions and to use Icelandic language technology without having to take on comprehensive and specialised basic development. It is important to create incentives for individual innovators, as well as for companies, to nourish innovation in the field.

This report outlines project plans for a language technology infrastructure. Each core project is, however, accompanied by a description of the possible use of the particular infrastructure for technical conversion, and chapter 5 contains examples of more comprehensive language technology software, such as teaching aids, automatic call centre software, semantic search, and so on. Through innovation, the main focus will initially be on making Icelandic available in diverse language technology software. A strong language technology industry in Iceland offers innumerable opportunities. Many languages of the world need language technology software; the transfer of solutions and services could create a sizeable market for language technology from Iceland. Diverse international collaboration is important: Iceland could even become a leading force in the development of language technology for smaller language communities.

LANGUAGE TECHNOLOGY PROGRAMME COORDINATION

For the aims of the language technology programme to be attained – to bring Icelandic and Icelandic language technology into general use in computers and devices – all organisations must strive for this from the start. It is important to define clearly the role of institutions, universities, the State and the industry, and to plan the work with the collaboration of all interested parties (Chapter 6).

It is proposed that the non-profit organisation Almannarómur will be a centre for the Language technology programme. Its main objective will be to ensure that the scheduled projects are carried out by the experts, institutions and companies that are entrusted with executing them; to co-ordinate the projects with one another and with the industry; and to ensure good communication between the project parties, the local industry, and foreign companies and institutions, so that the infrastructure and technology being developed will be put to use.

THE PROGRAMME IN A NUTSHELL

This report details all the projects that we consider essential to establishing the infrastructure for Icelandic language technology. We describe the human resources and the professional knowledge which will be required

Core teams

T1: Speech recognition

speech recognition, computer science, deep neural networks

T2: Speech synthesis

speech synthesis, semantics, computer science

T3: Machine translation

natural language processing, machine translation, translation, computer science, deep neural networks

T4: Spell and grammar checking

language technology, grammar, computer science, deep neural networks

T5: Data architecture

collection and set-up of speech data, grammar, computer science

T6: Language technology

language technology, grammar, computer science

Other teams

T7: Recordings and set-up of audio data

computer science, graduate students in related subjects, readers/voice providers for speech synthesis

T8: Smartphone

programming smartphone operating systems (Android, iOS, WindowsPhone)

T9: Web portals and interface

computer science/programming

T10: Licencing

data licences, law

in each project, along with an estimate of the time and work needed. The workgroup has not estimated other items of expenditure, such as the cost of technical equipment, network administration, hosting or cloud services, but it will be important to include these items when more exact cost estimates are made.

Core teams have been defined for each of the five projects that will be carried out. The teams possess the necessary knowledge to work on basic developments in the relevant field. Considerable experience has already been gained at the Árni Magnússon Institute for Icelandic Studies, Reykjavík University and the University of Iceland. Other teams will work across the core projects, on tasks that do not necessarily demand knowledge or experience in language technology.

It is recommended that each core team is a strong group that will work on the project in question for its entire lifetime. In this way, experience and knowledge will accumulate and capable people will be employed for longer. Individual projects will, however, require the temporary participation of additional staff and university students. The extent and diversity of projects relating to data, support tools and general language technology mean that a few teams will need to be established. Suggestions on the construction of teams and analysis of the skills they must possess can be seen in the image on page 24.

The table on page 25 shows an overview of the core projects: the teams that will be needed to participate in their execution and a summary of estimated man months are included. A detailed description of the individual parts of each project can be found in the sections referred to in the table. Software and databases will need maintenance and further development once the individual parts are completed. This work is not included in this report.

Project/ team	Speech recogniser H.1-H.16, p. 21-28	Speech synthesiser T.1-T.13, p. 33-39	Machine translation V.1-V.5, p. 46-49	Spell and grammar checker M.1-M.14, p. 57-62	Data G.1-G.9, p. 66-72	Support tools I.1-I.8, p. 75-80	Total man months
Team 1	125						**125**
Team 2		117					**117**
Team 3			114				**114**
Team 4				143			**143**
Team 5		2	60	18.5	97.5	3	**181**
Team 6	12	18		42	38	94	**204**
Team 7	80	30			4		**114**
Team 8	24	18		12			**54**
Team 9	12	33	18	6			**69**
Team 10	6	6		1	2		**15**
Total man months	**259**	**224**	**192**	**222.5**	**141.5**	**97**	**1136**

1 PROGRAMME OBJECTIVES

1. PROGRAMME OBJECTIVES

We search for information, order goods and services, sign up for events, and read, listen and watch news and entertainment on computers and smartphones. Communication between man and devices is rapidly becoming more interactive and the media more prolific. Language plays a large role in the way people experience devices, particularly in more common language areas where voice control, dialogue systems and speech synthesis enable people to seek information and send spoken messages.

The objective of the language technology programme for Icelandic is to ensure that Icelandic can be used to communicate in the technological world. We are in the midst of a technological revolution that is based on the analysis of vast amounts of data, big data, and on using artificial intelligence to build models based on these data. The models, the data, and the analysing tools form the core of intelligent machines that are able to use the diversity inherent in the data. In the case of language technology, this means using artificial intelligence on a large amount of language data to enable intelligent machines to work with language in the same way as people do: to be able to complete a wide range of tasks, including writing out the spoken word, creating speech from texts, analysing dialogue and making decisions based on it, processing information from text, and identifying spelling and grammatical errors.

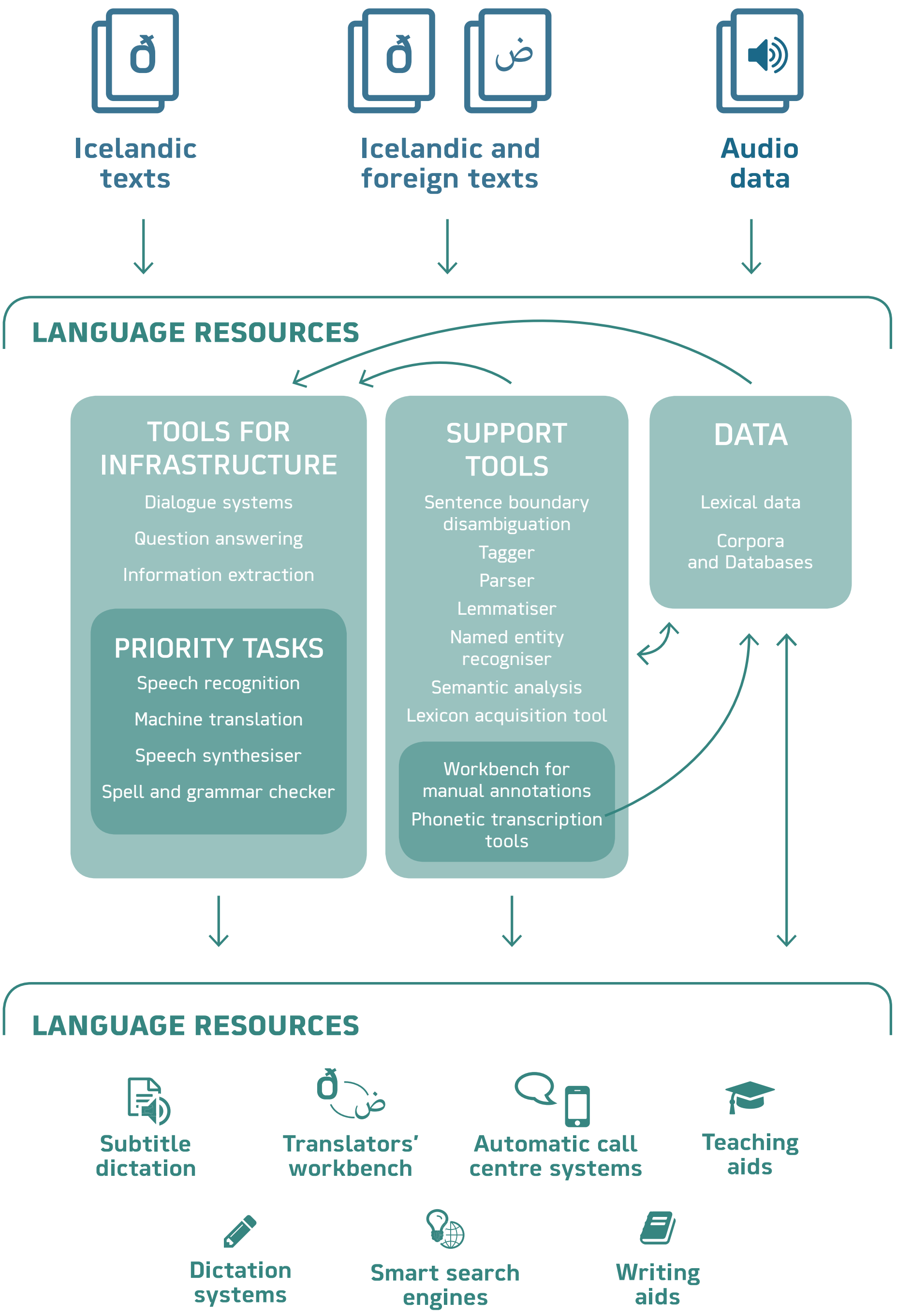
The principal focus of the programme will, therefore, be on building the required infrastructure to develop language technology/language resources: the data and support tools that are needed to develop software solutions that use language technology, and the basic tools such as spell and grammar checkers, machine translation systems, speech synthesisers and speech recognition systems. If we take an example of the data needed to build a speech recogniser, it consists of a large text corpus, a pronunciation dictionary, and audio recordings with accompanying text. The support tools required for building the data may be specialised recording software and workbenches for transcription, software that generates pronunciation for unknown words, or programs that train acoustic and language models, and are used to create models for the development of speech recognition technology. The speech recogniser itself, is then a tool built on this technology, that is possible to integrate into other software.

Another objective of the programme is to encourage the utilization and development of language technology in Iceland. The development of infrastructure will lower the threshold that software developers must surmount to introduce language technology into their software solutions.

This introduction is, however, by no means given and it is proposed that solutions using language technology should receive special support. Innovation in the field of language technology in Iceland must be encouraged by supporting start-up companies, as well as technological development in larger enterprises. It is also necessary to communicate and collaborate in advance with international parties and large language technology companies.

For this to succeed, it is essential to establish a knowledge industry around language technology in Iceland. The knowledge we already have must be nurtured by organising collaboration and by co-ordinating the initiative of those who work in the field, but also by enabling more to join in the development. The collaboration of universities, research facilities and the economy must be co-ordinated through a programme centre, in charge of assessing and choosing project proposals and evaluating the results. Such centre should also be able to initiate collaboration with international companies and universities.

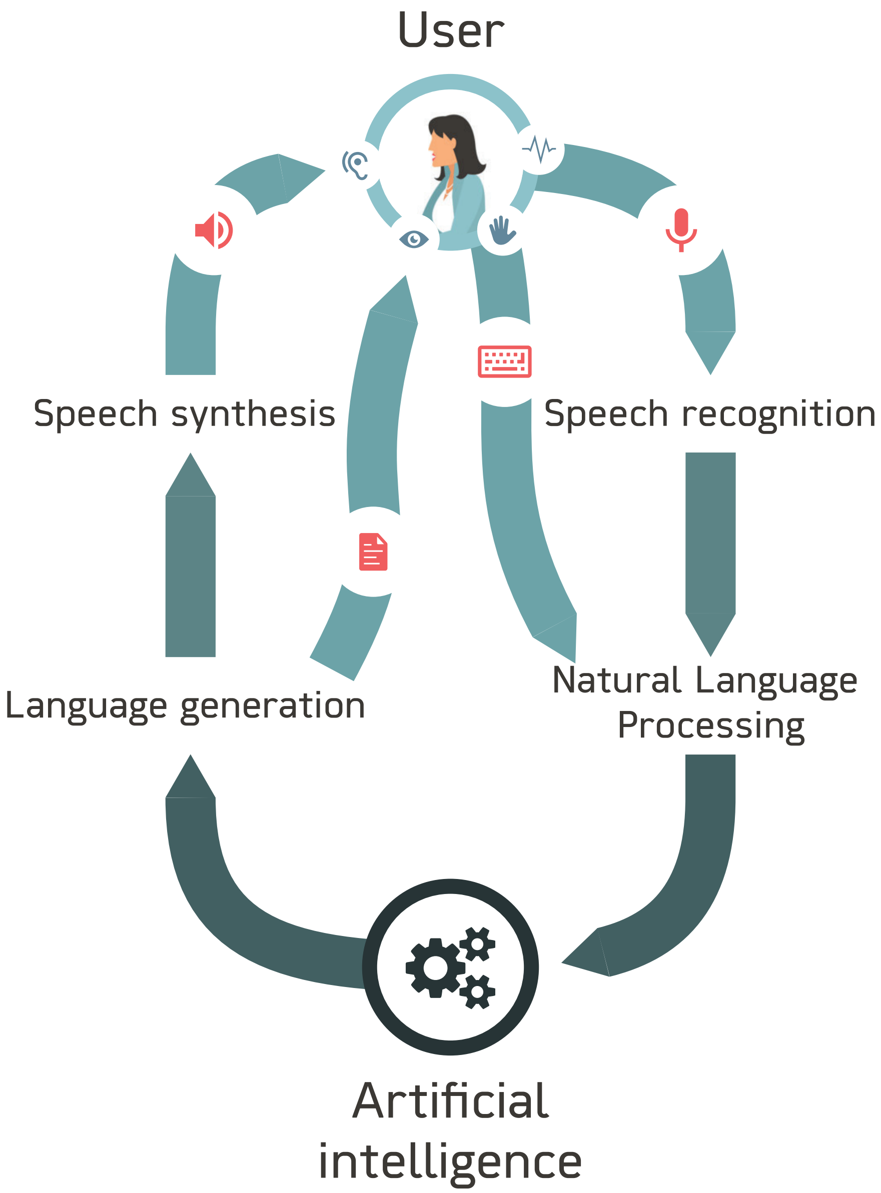
"THE LANGUAGE TECHNOLOGY ECOSYSTEM"



1.1 IMMERSIVE TECHNOLOGY

In Iceland, much groundwork has been done in the field of language technology, and a few tools that are used by particular groups have been built. A small number of speech synthesisers have also been developed, mainly for the use of blind and visually impaired people; spellcheckers have existed since the 1980s, and a speech recogniser has been in development for the past few years. The quality of these tools, as well as the possibilities for their usage has, however, been limited. In many other countries, specialised or composite systems have spread widely: automatic telephone answering and information providers that use speech synthesisers and speech recognisers; systems which enable people to search in large quantities of audio recordings of, for example, TV and radio programmes; speech recognisers that assist in transcribing medical records or parliamentary speeches; speech recognition systems that are used for authentication; speech synthesis software that enables people, who have lost the ability to speak through neurodegenerative diseases, to speak using only their eyes; machine translation systems that aid translators and increase their productivity; and, in areas affected by natural disasters, language technology is used in rescue operations. In the future, intelligent machines with multilingual skills could save lives. The list goes on – language technology can be used in most areas of society.

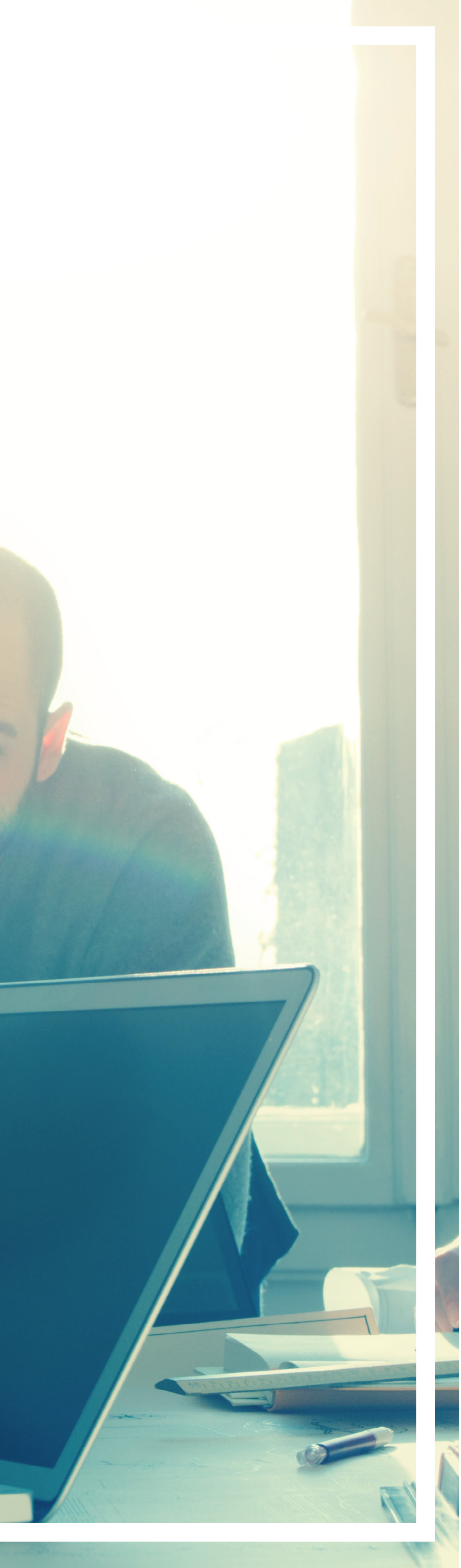
The world's largest technology companies now all put a great deal of effort into developing virtual assistants: Apple Siri, Google Assistant, Microsoft Cortana, Samsung Bixby and Amazon Alexa are new types of software that users communicate with verbally. These virtual assistants can search for information on the internet, aid users in buying goods or services, operate domestic appliances and do everything else a computer is capable of doing. The key factor in their usefulness is their ability to use language. In future years, voice-controlled interfaces, such as in virtual assistants, will increasingly be used in devices. In cars, for example, they increase safety: by operating the navigational system, radio and telephone by voice, drivers can focus on the traffic. This also adds convenience by, for example, having a voice-controlled TV, lighting and other electrical equipment. Language technology is also being increasingly used to improve the quality of life of those who live with disability or illness. Language technology brings different language communities closer together and may help in overcoming problems associated with diverse language environments.



If Icelandic is to be a part of this development and is to be used in communicating with machines to help us express ourselves in Icelandic, we must act now.

Tools, such as those listed above, can be made accessible to everyone who speaks Icelandic. This would be of great advantage for society, since language technology can have a considerable economic impact. Tools, which can be used to increase productivity and improve services in the health and administrative sectors, have already been mentioned. Speech technology and machine translation create enormous marketing opportunities in the telecommunications and entertainment sectors. Software for computer-aided language learning, distance-learning environments and programs for the detection of plagiarism, can increase the quality of education and access to it. Language technology can be used to analyse social media discussions, which may then be used to develop products or services, or to understand social issues better.

All of this can have a direct and obvious impact, but the indirect influences would be even greater. Computer technology has increased productivity in most lines of work, and better access to information technology and increased possibilities for the use of it will, similarly, have a considerable impact on the language community's well-being, although this may be hard to evaluate in full.



2 CORE PROJECTS

2. CORE PROJECTS

The development of the language technology infrastructure projects is organised into five cores: four of which aim to develop language resources and other infrastructure for speech recognition, speech synthesis, machine translation and spell and grammar checking, while the fifth revolves around the development of general language resources, the creation of general corpora and lexical data, and the development of support tools. This chapter discusses each core project in detail and proposes a schedule of work within the language technology programme.

2.1 SPEECH RECOGNITION

Speech recognition for Icelandic will be developed to enable people who design and develop voice-based user interfaces for devices, the web, and information providers, or who process information from spoken language, to add Icelandic easily. An open environment will be established for the development of speech recognition systems, and recipes for common usage will be made open and accessible.

Speech recognition systems convert speech to written text. Although people do this instinctively, a great deal of technology and resourcefulness is required in order to achieve this on a computer.

Considerable advances in technology in recent years have made the development and use of speech recognition a manageable project, given that the language resources and the necessary knowledge is in place. Despite this, speech recognition is a very diverse task because of variation in spoken language. Speech recognition can be used to make a written record of a long, continuous recital, speeches or dictation, to follow dialogue and even participate in it, or to use spoken instructions for further processing.

Speech recognition is, therefore, a kind of "keyboard" for the voice that enables people to communicate with one another, and with computers and systems.

Utility value and usage

Speech recognition is the prerequisite for humans to be able to communicate with computers and devices in a way that is natural for most people, i.e. by speaking. Voice-controlled communication is particularly important in situations where it is difficult to use your hands - or where it is too distracting to do so - for example while driving, or for people who are less able to write or who have trouble using a keyboard or touch-screen.

A number of things can influence the development of a speech recognition system: it matters who is speaking, in what circumstances and what the content of the speech is.

- **Who is speaking?** Voice and pronunciation are individual. Certain groups have, however, more in common with one another than with others: female voices are more similar to one another than they are to male voices, people of a similar age from the same language area often use similar pronunciation, etc. It is necessary to have audio recordings that are typical of the group that the technology is meant to recognise. For example, a traditional speech recogniser does not work well for children because their voices are too different from adult voices.

- **What are the circumstances?** A number of conditions can have an impact on speech recognition: background noise (traffic, nature sounds, chatter or other arbitrary sounds, such as a bell in parliamentary chambers); more people talking close to the recorder; how direct the speech is (not too hesitant and repetitive); and the quality of the recording. For example, in designing speech recognition for automobile navigation systems, vehicle noise and the background chatter of passengers must be considered.

- **What is the content?** Simple voice control can mean that a system is programmed to differentiate between certain individual words. Communications are, however, normally much more complicated and involve diverse vocabulary and sentence structures. The system can be restricted to certain topics, for example medicine, or be completely open, but it must be adapted to the content for which it is likely to be used.

2.1.1 UNDERLYING TECHNOLOGY IN SPEECH RECOGNITION (STATE OF THE ART)

Speech recognisers are trained using recordings of many different voices. It is common to include a few hundred voices in the so-called training set of speech recognition systems. This number of voices is necessary for the creation of a general acoustic model for the speech recognition system, and to enable it to process new voices. For the speech recognition system to be able to identify as many adult voices as possible the voices need to be male and female, and from different age groups. If the speech recognition system is intended to handle different dialects, a number of dialect examples need to be included. In addition, since children's voices are so different from the voices of adults, a special training set is needed to create a speech recognition system for them.

A traditional speech recognition system is trained by pairing texts with audio recordings. The continuous audio stream of the recordings is divided into 25-30 millisecond periods that are frequency analysed. This analysis is employed to select certain features that are represented as a numeric vector for each interval, using standard pattern-recognition methods. The objective is to differentiate between features, to normalise the feature vectors over different voices, and to exaggerate the differences between phones. Once the acoustic signal has been segmented and analysed in this way, it is connected to the phonetic transcription of the text that comes with the recording. The entire training set, transcriptions and recordings, is then used to create an acoustic model, which can comprise individual phones or triphone models. Currently, the best results are achieved with methods that use recurrent neural networks with long-short term memory units to train the acoustic models.

If the speech recognition system is required to write text, a comprehensive language model must be created with the aid of a pronunciation dictionary that uses the same phonetic transcription system as the acoustic model, and a large corpus. The dictionary and texts are chosen in relation to the speech recognition system's intended use: for example, simple instructions or queries have a more limited vocabulary and a simpler sentence structure than open communication systems.

Finally, the acoustic model and the language model come together in a finite-state machine called a hidden Markov-model. This model computes the probability of certain feature vectors applying to particular phones and calculates it against the prior probability of certain di- or triphones, words and a particular word order. This large model (or acoustic model with a small language model, plus a larger language model), uses a ready-made speech recognition system to process speech in the same way as it did during the training process. The audio stream is sometimes pre-processed to minimise background noise and define the speech signal, and the audio signal is segmented and analysed. The Markov-model is used to find the likeliest text match, word or sentence, but it can also deliver a lattice of likely results if the speech recognition is part of a larger system that adds information later in the process.

2.1.2 KALDI AND OTHER OPEN-SOURCE SOFTWARE

In recent years, immense advances have been achieved in the development of speech recognition systems. This is largely related to the progress of (deep) neural networks, but also to the availability of open-source software that contains all tried and tested algorithms and methods for speech recognition.

Kaldi is now the most common software environment for the development of speech recognition systems. Released under the very liberal Apache 2.0 licence, Kaldi has been thoroughly tested and documented, and its algorithms are flexible and offer myriad settings. While experts in universities and companies have undertaken the execution and maintenance of the latest and best algorithms that are being examined for speech recognition, a larger group worldwide has assembled a greater selection of recipes for the different usage, languages and methods offered by speech recognition. This has significantly reduced the complexity of the development of speech recognition systems and has made it a much more accessible technology.

This programme recommends the use of Kaldi software to create recipes for Icelandic speech recognition. Different methods and settings that Kaldi offers need to be explored, and data needs to be collected and Icelandic language resources prepared. Other open-source software, such as HTK, from Cambridge University, and Sphinx, from Carnegie Mellon University, could also be useful for this programme and should be considered. The development should be as diverse as possible: if there is a simple and cheap way of making comparable recipes for Icelandic in other software environments, it should be taken.

2.1.3 SPEECH RECOGNITION FOR ICELANDIC AND IN ICELAND

In 2012, Google made an Icelandic speech recognition system accessible: Google can be searched in Icelandic and the voice interface for Android smart devices (excluding Google Assistant) can be used in Icelandic. The Icelandic speech recognition system is also accessible by connecting individual software to Google's web service. The Málrómur corpus, which contains around 150 hours of audio recordings of Icelandic voices, was created for the speech recognition system in collaboration with Reykjavík University and Google. It is stored in Iceland and is open, but the speech recognition system itself, and all technology related to it, is the property of Google. It is important for Iceland to create its own open-source speech recognition technology that is adaptable to different needs.

In the past two or three years, Reykjavík University has built a basic knowledge of speech recognition technology. A speech recognition system for air-traffic control was developed in partnership with Tern Systems ehf., based on the limited English that is used in communication between aircraft captains and air-traffic controllers. Reykjavík University, The National University Hospital of Iceland, and the newly established Læknarómur, are collaborating on the development of a speech recognition system for radiologists. Furthermore, a project for the Parliament Office is underway, on the automatic transcription of the speeches of Members of Parliament, thereby increasing the quality and speed of their release; and an open environment for Icelandic speech recognition is being developed, with a plan to release methods and instructions to Kaldi to make it easier for technicians to start developing specialised speech recognition systems for Icelandic.

This work is all based on language resources that often depend on the subject that the speech recognition system deals with. The language resources that are used for speech recognition for radiologists are, therefore, a collection of written and dictated medical reports, where the vocabulary is highly specialised. These extremely sensitive data cannot be used for general development. In other instances, for example the Parliament Project, the language resources are more accessible and can be used for further speech recognition work. A 550-hour database of Parliamentary speeches was recently created; it is estimated that the database will increase to ten times that size. The Málrómur data, together with that of around 40 hours from the Hjal-project, is available at www.málföng.is.

Other language resources necessary for the development of speech recognisers are pronunciation descriptions and text corpora. It is, therefore, important to develop those resources in parallel with the collection of speech data.

2.1.4 NORMAL ERROR FREQUENCY

It is easy to assess the quality of a speech recognition system if its output text can be compared with the original. Comparison creates the opportunity to assess three types of errors: words that are missing, have been inserted or have been replaced. The ratio of these errors to the total word count, word-error-rate (WER), is most frequently used to evaluate speech recognition results. The WER not only depends on the quality of the speech recognition system, but also on the subject. For example, it is possible for the WER to be less than 1% if the vocabulary is limited and/or the interlocutors are few, whereas if the system is to be interrogated, it needs considerable vocabulary and to be adapted to many different voices. The best speech recognition systems in existence can achieve a WER of 5-7%, but 10-15% is more common.

2.1.5 INFRASTRUCTURE FOR SPEECH RECOGNITION

In the language technology programme for Icelandic 2018-2022, there are 16 speech recognition projects aimed at collecting and organising data, developing core infrastructure for speech recognition, and preparing the support tools that are necessary for technical solutions containing speech recognition.

At the outset, the data collection must be constant, extensive and take the diverse area of usage into account. The focus must be on data collection, since this is the basis for all development in speech recognition; without data, nothing much can happen. The work is divided into five parts: recording data with Eyra, transcribing TV and radio material, recording and transcribing dialogue and queries, aligning large corpora, and licensing.

Once data collection is underway, various types of speech recognition systems can be developed and designed. A general system will enable the creation of a web portal that can be inserted into the majority of smartphone operating systems. It will also enable the creation of recipes for speech recognition for dialogue, and queries and experiments with speech recognition for children and teenagers. Speech recognition systems for Parliamentary speeches and debates, recitals, and entertainment material on TV and radio will also become a reality.

Developing speech recognition systems demands many additional projects that improve and simplify their execution and use. In continuous speech it is not obvious where, for example, to punctuate a speech recognition system's output, but without punctuation the text becomes difficult to read. Better word- and sentence-analysis can also be useful for creating language models for Icelandic, which has a more diverse inflection system and more active word formation than, for example, English or Spanish. Knowledge of general articulation, distribution of amplitude and length of sounds in Icelandic, as well as differences in dialect, can be used to make more accurate speech recognition systems and to enrich their output. In dialogue systems it needs to be possible to identify a speaker and discern when one voice takes over from another. Special attention must be paid to the development of acoustic models for the voices of children and teenagers: this work is not as advanced as it is for adults. Designing recipes for software solutions other than Kaldi will strengthen Icelandic speech recognition and therefore recipes for HTK and Sphinx will also be developed.

2.1.5.1 RECORDING DATA WITH EYRA

It is proposed to continue recording text which has been pre-defined and is in the same format as that which was collected in collaboration with Google in 2012. Reykjavík University has maintained the collaboration and developed software, Eyra, to simplify recording. A compilation of sentences, for adults and teenagers, began during the first year of the project and will continue for three years. The plan is to record 200,000 expressions each for adults and teenagers – around 150 hours of recordings for each group. Parents' informed consent must be given for teenagers, but both databases will be published with an open CC BY 4.0-licence. In addition, 100,000 expressions from people who use Icelandic as a second language will be collected. In the project's second half, the software will be adapted to freer speech, in which images will be described and recordings repeated. All recordings must be transcribed, which is more time-intensive than the actual reading. The advantage of these recordings is, however, that the speech is more free-flowing and the participation of children who are unable to read, again with parents' informed consent, will become possible.

H.1 Data recording using Eyra

Work packages:

- Audio recordings from teenagers (200,000 utterances)
- Audio recordings from adults, dialects considered (200,000 utterances)
- Audio recordings of adults who use Icelandic as a second language (100,000 utterances)
- Audio recordings and transcribing of children's voices (image descriptions)
- Audio recordings and transcribing of adult voices (image descriptions)

Human resources:

- Data collection: 30 months
- Programmer: 2 months

Total: 32 man months

2.1.5.2 TRANSCRIBING TV AND RADIO MATERIAL

More diverse data than that set out above, is needed to facilitate speech recognition in other areas. In the project's first year, transcription of radio and TV talk shows will start. This work is expected to take three years; in the second and third years, entertainment material and news programmes will be added. This will be undertaken in collaboration with radio and TV stations, and care will be taken to release the data collection with as open a licence as possible. Some support programming must take place, but it is assumed that programs, such as SoundScriber, can be adapted. In the latter half of the project, recordings of meetings, dialogue, question answering systems and specialised oral description projects will be transcribed.

H.2 Transcribing TV and radio material

Work packages:

- Transcribed discussion programmes from radio/TV (250 hours)
- Radio and TV news programmes (transcribed or aligned)
- Transcribed entertainment material (50 hours)

Human resources:

- Transcriber: 24 months
- Programmer: 2 months

Total: 26 man months

2.1.5.3 TRANSCRIBING DIALOGUE AND QUERIES

Spoken language is very diverse and speech in dialogue and queries can be very different from speeches, readings, and TV and radio material. This section is aimed at recording and transcribing meetings and phone queries. It will be necessary to find people who are willing to have their meetings recorded and released, but this could be in collaboration with companies and/or institutions where work meetings or committee meetings are already recorded. The companies could be given the opportunity to edit content before it is released. It should also be possible to release open Parliament committee meetings in the same format.

It is also necessary to set up a system to record people who phone to ask for information and service. Companies and institutions with large call centres could possibly participate in this project. Banks, airlines, utility companies and other companies, along with various government institutions operate expensive human services to provide customers with information that could be done automatically.

H.3 Transcribing dialogue and queries

Work packages:

- Recording and transcribing meetings
- Recording and transcribing call queries

Human resources:

- Data collection: 6 months
- Transcriber: 6 months
- Programmer: 2 months

Total: 14 man months

2.1.5.4 TRANSCRIBING LECTURES

In this segment, lectures – course, conference or open lectures – will be recorded, transcribed and made accessible, and it will be necessary to find lecturers who agree to do that.

H.4 Transcribing lectures

Project segments:

- Recording and transcribing lectures

Human resources:

- Data collection: 3 months
- Transcriber: 3 months
- Programmer: 2 months

Total: 8 man months

2.1.5.5 ALIGNING LARGE SPEECH CORPORA

Large collections of written text with parallel recordings already exist, but in their present form are not necessarily ideal for building a speech recognition system. Examples of this are Parliamentary speeches, accessible court data, and data from the Icelandic Audio Library. Considerable experience has been built up on preparing such collections for speech recognition and the results of the collaboration between Reykjavík University and Parliament will be used for this project. A considerable amount of data should be available from the above-mentioned sources and they will be useful in building speech recognition systems for similar environments.

H.5 Aligning large speech corpora

Work packages:

- Aligning Parliamentary speeches (approx. 5,000 hours)
- Aligning recordings from the District Courts of Iceland
- Aligning Audio Library data and storing intonation profiles

Human resources:

- Data programmer: 18 months

2.1.5.6 LICENSING

It is vital to license the data that is created in the project. It is intended to release most of the data under a CC BY 4.0, or comparable, licence and, since this is a legal rather than a technical task, it will demand a discrete type of work. The following issues should be kept in mind, although more might arise during the project period.

Recording childrens' and teenagers' speech: the main issue with collecting this data is to acquire informed consent from parents. A project description needs to be prepared for the Ethical Review Committee, and a process created whereby parents can easily agree, or refuse, to let their child participate.

Reading and text from the Audio Library: it is obviously not intended to re-release Audio Library material under open-source licences, but it should be possible to label and rearrange the data so that it is usable for speech recognition and other language technology.

Transcribing news material: negotiations must take place with the media on the transcription of discussion and news programmes, and entertainment material.

H.6 Licensing speech recognition data

Work packages:

- Designing a process for the recording of children and teenagers (under 18)
- Work on licensing issues for Audio Library data
- Work on licensing for data from news media (RÚV, 365 media, etc.)

Human resources:

- Expert in data licensing: 6 months

2.1.5.7 DEVELOPING A GENERAL SPEECH RECOGNITION SYSTEM

A general speech recognition system, using Kaldi software, will be developed based on existing language resources: Málrómur speech corpus, Tagged Icelandic Corpus, and a pronunciation dictionary. A method, which will be created for training new speech recognition systems on more data, will be useful for other speech recognition projects of the programme. In addition, it will become possible to begin work on other developments, including web portals and smartphone input, which are based on this speech recognition system. This will be possible even if the accuracy of the speech recognition system is not yet sufficient.

H.7 Developing a general speech recognition system

Work packages:

- Recipe based on Málrómur data

Human resources:

- Speech recognition expert: 12 months

2.1.5.8 WEB SERVICE FOR SPEECH RECOGNITION

With the creation of a general speech recognition system it will be possible to establish a web service in which speech is analysed and the result given in text. The service can accept whole audio files or continuous speech input. These web portals can be established with a version of the Icelandic speech recognition system that is based on existing data. Initially, the speech recognition system's quality may not be particularly good, but it will be possible to update it when more data is available and thus the web portals' ability to recognise speech will increase. The portals serve an extremely important role: they will make it easier to test the speech recognition system that is being developed and to monitor usage and results.

H.8 Web portals for speech recognition

Work packages:

- Web service for speech recognition of audio files
- Web service for speech recognition of audio streaming

Human resources:

- Web programmer: 12 months

2.1.5.9 SPEECH RECOGNITION AS INPUT FOR SMART-PHONES

The three main smartphone operating systems, Android, iOS and Windows Phone, offer third parties the opportunity to build keyboards for their systems. This is an easy way to insert Icelandic speech recognition into Android smartphones, because the keyboard has a speech recognition button that could be connected to an Icelandic speech recognition service. Similar methods might be employed for the other operating systems but will probably need to be worked on individually.

H.9 Speech recognition as input for smartphones

Work packages:

- Icelandic keyboard with a speech recognition button for Android operating system
- Icelandic keyboard with a speech recognition button for iOS operating system
- Icelandic keyboard with a speech recognition button for Windows Phone operating system

Human resources:

- Web programmer: 12 months

2.1.5.10 VOICE CONTROL, QUERIES AND DIALOGUE

To facilitate technological transfer, speech recognition recipes will be prepared for voice control, queries and dialogue. Starting the development of these procedures will be possible before the necessary data is collected, but it is assumed that the data will increase the quality of the systems and make them functionable. It is possible to narrow down a general acoustic model with isolated recordings. Most of the work will be to prepare an appropriate language model which controls the flow that this usage offers.

H.10 Voice control, question answering and dialogue

Work packages:

- Recipe for voice control
- Recipe for queries
- Recipe for dialogue

Human resources:

- Speech recognition expert: 24 months

2.1.5.11 SPECIALISED SPEECH RECOGNITION

The project's particular focus is on preparing systems for the speech recognition of children's and teenagers' voices for two principal reasons: children and teenagers have the most influence on changes in the use of technology, and their voices can be so different from those of adults that they need separate speech recognition systems.

H.11 Specialised speech recognition

Work packages:

- Speech recognition for teenagers
- Speech recognition for children
- Speech recognition for people who speak Icelandic as a second language

Human resources:

- Speech recognition expert: 32 months

2.1.5.12 PUNCTUATION AND SENTENCE BOUNDARY DETECTION

It is important for a speech recognition system to deliver legible text. When acceptable word precision has been achieved, it is important to be able to break it up with punctuation, but it is not always straightforward to decide where to insert it. Punctuation models based on text corpora achieve reasonable results but taking the speech into account as well helps improving the models. This is done by listening for longer silences, stresses and style of speech, which can be achieved by prosodic analysis.

It is particularly important to recognise full stops in the analysis of long, continuous speech. The speech recognition system must be able to sense where to stop, resolve and complete word lattices. The size of the word lattice increases with longer speech and as the number of possible sentences increases. If the length of the speech being analysed is not limited, the speech recognition system's memory can overload and create too many possible sentences, which need analysing. It is, therefore, important to break down long, continuous speech into segments through full-stop recognition. Acoustic analysis for punctuation can also be used for sentence boundary detection. Long silences, stresses and intonation indicate that the speech recognition can be stopped and restarted for the next segment.

H.12 Punctuation and sentence boundary detection

Work packages:

- Punctuation, sentence boundary detection and prosodic analysis

Human resources:

- Language processing expert: 12 months

2.1.5.13 LANGUAGE MODEL FOR WORD FORMS AND COMPOUND WORDS

Standard language modelling for speech recognition is primarily based on English; it rarely includes morphologically rich languages, such as Icelandic, with all their inflections, compound words, prefixes and suffixes. Some work has, however, been carried out on, for example, Czech and other Slavonic languages. It should, therefore, be investigated as to whether it is possible to create a more complex language model that is based on this work, so that speech recognition will become more accurate.

Compound words are much more common in Icelandic than in English. As a result, it can be difficult to accommodate all possible words into the list that is used for speech recognition. Because it is difficult to identify words that are not on the list, this makes the speech recognition process harder. The advantage of compound words, and words with prefixes and suffixes is, however, that it is possible to create a model of them using the pronunciation of shorter words and morphemes. Language technology research has been carried out on languages such as Dutch and German, which both have these qualities.

H.13 Language model for word forms and compound words

Work packages:

- Language model for word forms and compound words

Human resources:

- NLP expert: 12 months

2.1.5.14 DIALECTS, PHONETIC ANALYSIS AND SPEAKER DIARIZATION

An increased understanding of dialects, stress and intonation, and the statistical distribution of speech sounds is particularly useful for speech recognition. It makes it easier to achieve greater accuracy; sentence boundary detection and punctuation become easier; and voices are better identified. The output of the speech recognition system could, therefore, become more than just the text – it could identify the type of speech, who is talking and in which dialect.

Speaker diarization offers the analysis a chance to react to fast communication where two or more voices take turns in talking. Good intonation analysis, knowledge of statistical distribution of speech sounds, and identification of who is talking can not only give a lower word-error-rate in the speech recognition but can also tag the dialogue for further analysis and usage of artificial intelligence.

H.14 Dialects, phonetic analysis and speaker diarization

Work packages:

- Dialects and phonetic analysis and diarization

Human Resources:

- Acoustic analysis expert: 12 months

2.1.5.15 ACOUSTIC MODELS FOR CHILDREN'S AND TEENAGERS' VOICES

Since the voices of children and teenagers are different from adult voices, the project assumes the collection of data, and the development of systems, for recognising their voices. The results are, however, not guaranteed to be as good because the systems are not as evolved as those for adults. This might be because much less data has been collected for children and teenagers. It is assumed that a separate acoustic model for children and teenagers will be developed with the methods used for foreign languages.

H.15 Acoustic models for children's and teenagers' voices

Work packages:

- Adapting acoustic models to children's and teenagers' voices

Human resources:

- Speech recognition expert: 12 months

2.1.5.16 SPEECH RECOGNITION METHODS IN OTHER SYSTEMS

This programme assumes all speech recognition development is carried out using the Kaldi software. We expect that once the main recipes for that software have been released, they will be able to be easily transferred to other open-source speech recognition software, which is also used internationally. In this regard, the focus is mainly on the Hidden Markov Model Tool Kit (HTK), from Cambridge University, and Sphinx, from Carnegie University. The advantage of the creation of these recipes is that it makes it easier for people who use them to add Icelandic to their research and development.

H.16 Speech recognition methods in other systems

Work packages:

- Speech recognition methods in other systems

Human resources:

- Speech recognition expert: 3 months

2.1.6 TECHNOLOGICAL TRANSFER

Once the infrastructure is in place, data has been collected, recipes released, and support tools and services are ready, it will be possible to transfer knowledge and technology to business solutions and to integrate speech recognition in existing systems and programs. The following examples show the usefulness of the project's infrastructure in software that is developed for general use.

2.1.6.1 SPEECH RECOGNITION FOR TV AND RADIO

It is extremely important to be able to utilise language technology with TV and radio material, whether through traditional media or on the web. Speech recognition is one of the key factors since it enables speech to be automatically transcribed. There are many uses; an obvious example is the subtitling of material in real time. This enables people to follow broadcasts without sound, which not only helps deaf people, but also viewers who for some reason are unable to listen to the material. Speech recognition for TV and radio will also make it possible to search material for keywords and analyse it based on use of terms and style.

2.1.6.2 SPEECH RECOGNITION FOR PARLIAMENT AND COURTS OF LAW

Public bodies have a great need to transcribe spoken material. Every speech given in Parliament is transcribed and published. In the courts, witness examinations and reports are also transcribed if cases are appealed. Speech recognition simplifies this work and offers new opportunities for searching and analysing information in discussions and court cases.

2.1.6.3 QUESTION ANSWERING SYSTEM FOR BANKS AND INFORMATION PROVIDERS

Speech recognition systems will increase the quality of service in companies and institutions which need to provide their clients with diverse information. A question answering system enables individuals to phone for information without having to wait for a person to pick up the phone. Their query is analysed, and information given via a speech synthesiser. The possibility of developing these solutions increases in tandem with infrastructure development, but a good speech recognition system, with a language model built from the language resources system's domain, is the key to this process.

2.2. SPEECH SYNTHESIS

Speech synthesis for Icelandic will be developed so that it will be possible to produce multiple different voices. An environment, and language resources, will be created and released to enable synthetic voices to be built quickly and in a simple manner. It will also enable the integration of speech synthesis into software, where e.g. automatic recital or voice answering is needed.

Speech synthesis converts written text into speech. When it comes to spoken language, speech synthesis is a key component in language technology: it enables computers and communications systems to deliver information vocally. Speech synthesisers need to be designed and built to accommodate a large, and diverse, range of subjects. For example, a speech synthesiser that synthesises speech from long texts requires different facilities from one that provides short replies in a question answering or dialogue system. It is, therefore, self-evident that speech synthesis must offer a variety of interpretations, since much information that is not present in text is communicated by voice. It needs to be possible to program a speech synthesiser to a predetermined style of speech, to keep cadence and strength steady, and to add other information, such as tempo and stress, in real time.

Development of speech synthesisers has a long history: inventors in the past attempted to imitate the human voice through different mechanical approaches and electrical circuits, but real progress was not made until the advent of computer technology. The "voice" of physicist Stephen Hawking is a famous example of a synthetic voice that was based on first-generation speech synthesisers. The technology has greatly improved since then; the biggest progress was made around the turn of the century with so-called unit selection, which is built on a large list of diphone units that are used to form the acoustic signal. Diphone units are based on one voice, but in creating the acoustic signal the best unit is chosen for the sequence, which is then stretched and adjusted to create the best possible speech.

Speech synthesisers based on unit selection can give excellent results, but the focus has more recently shifted to parametric systems, which define parameters that describe a voice to create the most normal speech signal. For some time, these speech synthesisers were distinctly less proficient than those based on unit selection, but the advances of deep neural networks balanced the scales and it is now unclear as to which method is better. The new parametric speech synthesisers have the advantage of being able to create smoother and more diverse voices using the same quantity of recordings and language resources, whereas with unit selection it is it is not possible to change voice identification and is more difficult to adjust cadence and intonation. Parametric voices are, however, still at the experimental stage.

2.2.1 QUALITY OF SPEECH SYNTHESISERS

Building a good speech synthesiser presents many challenges, some of which are significantly different from those faced by other language technology projects. One of the principal challenges is to overcome the so-called "uncanny valley" of quality and clarity. Normally people see nothing wrong with the quality of speech synthesisers, which sound robotic enough to avoid confusion with human voices, but when the quality of speech synthesisers improves, and the artificial voices become more similar to human voices, mistakes and mispronunciations may sound strange. The main challenge in developing speech synthesisers is, therefore, to overcome this barrier.

An environment to test the quality of synthetic voices needs to be established. Quality evaluation is always subjective to some extent, since the usefulness of synthesisers is based on how easy they are on the ears and how comprehensible they are. In addition, not everyone uses speech synthesis in the same way: some people use it to scan texts quickly and want to be able to listen to text read at high speed, while others prefer a pleasant voice in dialogue systems, for the reading of announcements, newspaper articles, course books, or even literature.

We propose three different methods of evaluating synthetic voices, based on their intended use.

1. General users listen to a text read by a speech synthesiser and press a button each time they find the reading unnatural or uncomfortable. This enables the identification of whether certain elements bother a vast array of users and taking measures towards correcting them. It is also possible to compare how many observations are made on each of the different voices.

2. General users listen to a long text read by a speech synthesiser. The reading is stopped intermittently, and users are asked to answer questions from the text. Should they be unable to answer, they need to listen again until they can answer all the questions correctly. The listening time is noted. The best speech synthesiser is the one that people take the shortest time to listen to and answer everything correctly.

3. Mean opinion score is used to compare different speech synthesisers. This simple method evaluates how close the synthetic voice is to normal speech. Listeners are asked to grade sentences from 0-5, where 0 is incomprehensible and 5 sounds like a human voice.

The advantages of these three methods is that they primarily assess whether the users like the speech synthesisers. They also make it easy to compare different synthesisers, and to compare the quality of the reading against that by actors or other human voices.

2.2.2. UNDERLYING TECHNOLOGY IN SPEECH SYNTHESIS (STATE-OF-THE-ART)

For the most spoken languages in the world the available language resources, and the accumulated software and knowledge, make it easy to build a new voice for software solutions that demand a speech synthesiser.

The creation of speech synthesisers may be separated into language processing and speech generation. The speech synthesiser's input is text, which needs to be prepared to enable it to form a speech signal. The main steps in this are tokenising, text normalisation, phonetic analysis, stress analysis and intonation. The tokeniser delimits all word units (tokens) in the text and identifies units such as abbreviations and numbers for the text normaliser to expand them into their corresponding word forms. For the phonetic transcription a pronunciation dictionary is used, but for words that do not exist in the dictionary, an automatic phonetic transcription has to be performed. Finally, stress and intonation are analysed to create the correct stress and intonation for the speech signal. The language processing output, which is also the speech generation input, is a sequence of phonetic units, marked with stress and intonation.

Speech generation can be built on unit selection or on a parametric system. Unit selection systems are built on a large database of voice signals, so that each diphone in the language appears at least once. The units are concatenated, according to the phone sequence that is being generated and, for this process, optimisation algorithms are used to take account of predetermined objectives, and of intonation and stress marking. Parametric systems generate speech with an acoustic model and a vocoder rather than with recordings. The parameters of the vocoder, which decide which speech signal is the outcome, are controlled by the acoustic model, which finds the optimal parameters based on the phone sequence, and stress and intonation markings.

While unit selection systems often overcome much of the "uncanny valley", the last steps can be difficult, principally because unit selection systems tend to be quite cumbersome. The best way to increase the quality of unit selection systems is to enlarge the database to make it more likely to cover all the diversity needed by speech and text. Parametric systems, on the other hand, are more manageable because it is easier to model diversity of texts. This property, however, makes it harder to find the best way to generate speech. Development is ongoing and is promising.

There has been some success recently in using deep neural networks. Google, for example, has released a speech synthesiser, WaveNet, which is a type of a parametric speech synthesiser, but instead of using a statistical acoustic model and a vocoder, a neural network is used to generate speech directly from the phone sequence. This technology is new, and still very cumbersome, but the quality which can be attained seems to exceed previous methods.

2.2.3 OPEN-SOURCE SOFTWARE FOR SPEECH SYNTHESIS

Festival has long been the most common software for designing and building speech synthesisers. Maintained and supported by Carnegie Mellon University and the University of Edinburgh, many operators have relied on the tools that have been the result of this collaboration. The main advantages of developing speech synthesisers based on *Festival*, are that the software is well designed and implemented, has been thoroughly tested through much use, and has a community with a great deal of experience and knowledge. The problem with *Festival*, however, is that it is based on *Scheme*, a programming language that is now rarely used. The cost of teaching people to use *Festival* is, therefore, greater than if another software were used.

Further open-source software solutions for building speech synthesisers have recently been developed. *Idlak*, which is a variant of the speech recognition software *Kaldi* (the word has been reversed for speech synthesis), is based on a combination of C++ and shell scripts. *MaryTTS*, which is based on *Java*, has been developed by the German Research Centre for Artificial Intelligence, DFKI, and Saarland University, and is widely used for languages other than English. *Merlin* software, developed by the University of Edinburgh, is based on the AI framework, *Theano*, and on the *Python* programming language. *Merlin* is designed so that experts can easily create and share their own recipes, in a similar way to that which *Kaldi* does for speech recognition.

Evidently, are clearly many different software solutions available for the development and design of an Icelandic speech synthesiser. A decision has to be made as to which is the most suitable for an open-source speech synthesiser that can be accessible to anyone who wishes to create and use speech synthesis technology for Icelandic.

2.2.4 SPEECH SYNTHESIS IN ICELAND

Speech synthesis has existed in Iceland for about 30 years. Towards the end of the 1980s, *Sturla*, a voice based on formant synthesis, was jointly developed by the University of Iceland, The Organisation of Disabled in Iceland, and KTH, the Royal Institute of Technology in Stockholm. Around the turn of the century, *Snorri*, which is based on the same language resources as *Sturla* with additional voice recordings, was developed. The questionable quality of *Sturla's* voice prompted the development of a speech synthesiser that would be based solely on unit selection. More language resources were collected and the international company, Nuance, was employed to build a speech synthesiser, named Ragga. An accompanying web interface was also designed, and the voice was therefore also called *Vefþulan* (Web Announcer). The voice quality was still deemed to be unsatisfactory, but further development proved difficult – ownership was unclear, and the necessary expertise was not available in the country. In 2010, the Icelandic Association of the Visually Impaired employed the Polish company Ivona to build two voices, *Karl* and *Dóra*, which were also based on unit selection. They are thought to be a considerable improvement on previous voices and are still maintained by the association. There has, however, been some criticism of *Karl* and *Dóra*, primarily of their pronunciation of foreign words, but also because their speeding up is near impossible to achieve.

The history of speech synthesis in Iceland clearly shows the importance of creating language technology knowledge and skills in Iceland, rather than employing foreign companies to develop language technology solutions for Icelandic (although it can be argued that this is better than doing nothing at all). We need local organisations that own the technology and can provide expert knowledge when further development is needed. Software is a living technology: systems, computers and software are updated regularly, and new methods of communication and computation appear frequently. We need the knowledge and skills to adapt existing technology so that expensive investments will not become obsolete.

2.2.5 INFRASTRUCTURE FOR SPEECH SYNTHESIS

The main objectives of the speech synthesis development in the programme is to create a diverse environment for the design and development of the technology, with the aim of making its use as widespread as possible. Unlike previous language technology programmes, the focus will be on diverse methods for speech synthesis and on how local entities can have their own voices created to their needs. The emphasis will be on collecting language resources for unit selection and parametric systems, as well as on defining and guaranteeing the ownership and management of the resources.

Towards the end of the project, systems will be created to easily enable experts to adapt standard or specialised voices to their own software. This presents the potential for creating a diverse selection of speech synthesisers for Icelandic – if one voice does not work in a particular context, another one might[1]. News media and information providers could, for example, each rely on an individual set of voices when building speech synthesisers.

The prerequisite for ensuring that there is variety in synthetic voices is the recording of many types of voices. Two types of voice recording collections need to be established: one for unit selection systems and one for parametric systems. For the unit selection system, one synthetic voice can be created from each reader and, although the plan is to have a variety in age and dialect, since each participant needs to read for more than 20 hours the number of participants will never be large. Many more voices must, however, be recorded for the parametric systems, although each recording need not be longer than one or two hours. The voices must be similar enough to enable a mix of the recordings into one synthetic voice. In both cases, care must be taken to include an equal number of male and female voices.

The primary focus on the development of speech synthesisers will be on creating recipes for unit selection and parametric speech synthesisers, and on preparing web portals on which the public and experts can test standard voices that will be released in the project.

[1] Speech synthesis projects for Icelandic have all produced one or two voices. Because individual needs and tastes vary, there has never been widespread satisfaction with the results. New methodology will resolve this problem.

To create infrastructure and to build speech synthesisers, it will be necessary to measure spoken language, handle text and develop various support tools. It will also be necessary to carry out work on automatic phonetic transcription of word forms, phone sequence generation, text normalisation, and prosodic and intonation analysis. In addition, parametric speech synthesis will need particular attention because it relies on a vocoder, which must be adapted to Icelandic.

2.2.5.1 RECORDING UNIT SELECTION DATA

For a unit selection system, one synthetic voice can be created from each reader. The plan is to record eight voices and attain a diversity in age and dialect, with an equal number of male and female voices. Ideally, to achieve a good distribution of diphones, which are used in the unit selection systems, each participant should read for at least 20 hours. A script must be prepared, participants chosen, and data managed. A principal product of this part of the project will be a detailed recipe and process that can be emulated once it is completed. It will be possible, therefore, to continue to record voices and create speech synthesisers after the project schedule has been completed.

T.1 Recording unit selection data

Work packages:

- Preparing a script for reading
- Recording voices
- Data management

Human resources:

- Language technology expert: 3 months
- Data expert: 6 months
- Participants: 6 months
- Programmer: 3 months

Total: 18 man months

NB: the recordings must be made in a near-anechoic studio, similar to a broadcasting studio.

2.2.5.2 DATA COLLECTION FOR MULTI-SPEAKER TRAINING

For parametric speech synthesis systems, the plan is to record voices from 40 participants who will each read for up to two hours; 20 female and 20 male voices are required. The readers' voices need to be similar, so that they blend well to create a parametric voice. The participants must read clearly and carefully.

T.2 Data for multi-speaker training

Work packages:

- Preparing scripts for reading
- Recording voices
- Data management

Human resources:

- NLP expert: 3 months
- Data expert: 6 months
- Participants: 6 months
- Programmer: 3 months

Total: 18 man months

2.2.5.3 USE OF EXTERNAL DATA IN SPEECH SYNTHESIS

Recordings of the Audio Library, Parliament and broadcast media can be used to create data for speech synthesis. All licensing matters must be taken care of and data converted to a format in which recordings and transcriptions match. These recordings can not only be used to create speech synthesisers, but also for analysing intonation and prosody, and for creating acoustic profiles for different areas of usage.

T.3 Use of external data in speech synthesis

Work packages:

- Work on licensing matters
- Architecture of new data
- Usage of older speech synthesis data

Human resources:

- Data licensing expert: 6 months
- Data expert: 6 months

Total: 12 man months

2.2.5.4 WEB PORTALS FOR SPEECH SYNTHESISERS

Two web portals will be established, where text can be converted into speech. The first will be for short texts typed into a web window, and for which audio may be played back directly or sent as an audio file. The second will be for longer texts where synthesis might take more time since intonation and prosody will be created based on the circumstances and the context of the text.

T.4 Web portals for speech synthesisers

Work packages:

- Web portal for short texts
- Web portal for long texts

Human resources:

- Web programmer: 6 months
- Programmer: 9 months

Total: 15 months

2.2.5.5. SPEECH SYNTHESISERS AS OUTPUT FOR SMARTPHONES

Most operating systems offer an application programming interface for speech synthesisers. A speech synthesiser with a small footprint will be created for Icelandic, so that it can be installed in smartphones with Android, iOS and Windows Phone operating systems.

T.5 Speech synthesisers as output for smartphones

Work packages:

- Speech synthesis for Android operating system
- Speech synthesis for iOS operating system
- Speech synthesis for Windows Phone operating system

Human resources:

- Programmer: 18 months

2.2.5.6 WEB READER

An interface will be installed to enable the addition of an Icelandic speech synthesiser to websites. This will enable web designers to offer site visitors the option to listen to text being read. Since this is an important access issue for people who are unable to read websites, it is important to enable web designers to add this service to their sites.

T.6 Web reader

Work packages:

- Designing and implementing a centralised web reader
- Installing the web reader for the web server
- Installing a web reader plug-in for browsers

Human resources:

- Programmer: 18 months

2.2.5.7 RECIPES FOR VOICES

When texts have been normalised, phone sequences generated, pronunciation descriptions defined and voice recording completed, it will be possible to create a recipe for building unit selection speech synthesisers. A decision needs to be made on which speech synthesis software to use: at the time of writing this report, some of the best options were Festival/Festvox, Merlin and MaryTTS. The recipe will be released on the internet, together with appropriate language resources, to enable anyone to reproduce the voices offered on the recordings, and to continue developing and improving them with, for example, more precise pronunciation descriptions, better intonation models or updated methods offered by the speech synthesis software.

Voices based on data from more than one participant can be built with parametric speech synthesisers. The parametric acoustic model is either trained using statistical methods or with neural networks. A recipe for such a system will be released for the software that is the most promising at project time. This recipe can be used to continue developing voices with better methods, increased language resources and more recordings.

T.7 Recipes for voices

Verkþættir:

- Releasing a recipe for unit selection voices
- Releasing a recipe for parametric voices

Human resources:

- Speech synthesis expert: 18 months

2.2.5.8 EVALUATING QUALITY OF SPEECH SYNTHESIS

While evaluating the quality of speech synthesis (mentioned in chapter 2.2.1) is important, it is also difficult. Users may be critical of quality and user testing must be established for comparing and assessing the results.

T.8 Evaluating quality of speech synthesis

Work packages:

- Installation of software to evaluate the quality of speech synthesis
- Organising user testing

Human resources:

- Programmer: 9 months
- Speech synthesis expert: 9 months
- Data expert: 6 months

Total: 24 man months

2.2.5.9 TEXT PRE-PROCESSING, NORMALISATION AND STRESS ANALYSIS

In speech synthesis, pre-processing revolves around converting written text into a sequence of phonetic symbols. The first steps are tokenisation and text normalisation (see 2.5.3.3). Text normalisation converts all non-word units, such as abbreviations and numbers, into words. For example, *"Þjóðhátíðardagur Íslands er 17. júní"* is converted to *"Þjóðhátíðardagur Íslands er sautjándi júní"*.

Before the transcription of a text, it is good practice to decide which part of the word is stressed. In most cases this is fairly straightforward, but stress rules are more complicated for compound words. Programming these rules could be very beneficial for the quality of Icelandic speech synthesis.

T.9 Text pre-processing, normalisation and stress analysis

Work packages:

- Text normalisation for Icelandic
- Stress analysis

Human resources:

- NLP expert: 18 months

2.2.5.10 AUTOMATIC PHONETIC TRANSCRIPTION

A speech synthesiser must be able to convert text into phonetic units. It must be able to transcribe all possible word forms and to adapt the transcription of single words into continuous speech. Good and well-defined pronunciation description is necessary for the training of software, which can automatically transcribe unknown word forms. Once individual words have been transcribed, the equipment needs to examine the last and first phonetic units of adjoining words and use assimilation rules where applicable.

When developing automatic phonetic transcription, close monitoring of the error rate will be necessary. If the pronunciation dictionary is found to have gaps or faults that increase the error rate in a system based on its data, improvements will be necessary.

If a speech synthesiser is designed to use different dialects, the pronunciation dictionary must contain enough examples of all the main variations in each dialect to enable an automatic transcription tool to learn the rules.

In speech synthesisers, the transcription tool would be used to transcribe unknown words and word forms phonetically, but it must also be possible to use it on its own, for example when building speech synthesisers that have a limited and specialised vocabulary. In those instances, it is possible to incorporate a pronunciation dictionary, but also to include pronunciation that the program has created based on the rules it has learnt. The builders of the speech synthesiser can review the latter to ensure that it is correct.

The tasks of this project are on the one hand to release a tool for automatic transcription (grapheme-to-phoneme) with documented results, and on the other hand to build a web interface on top of it to facilitate the use of the tool in isolation.

T.10 Automatic phonetic transcription

Work packages:

- Training and testing transcription tools with pronunciation description
- Phonetic transcription environment
- Web interface

Human resources:

- NLP expert: 6 months
- Transcriber: 9 months
- Programmer: 3 months

Total: 18 man months

2.2.5.11 PROSODY AND INTONATION ANALYSIS

It is necessary to identify the speaking style of the reader, as well as the adjustment to the voice projection that the speech synthesiser demands. Prosody must vary and must reveal whether the speech synthesiser is giving a lecture or a speech, reading advertisements, participating in conversation, or reading aloud from a book. It must also be able to change intonation and stress, based on the subject and the style of the text. If this is for a unit selection system, a mixture of text analysis and an analysis of the phonetic units available in the speech synthesiser must, therefore, be used. In speech synthesisers with parametric acoustic models, the analysis must be able to influence the parameters to create the correct stress and intonation.

T.11 Prosody and intonation analysis

Work packages:

- Building an intonation analyser

Human resources:

- Signal processing expert: 12 months

2.2.5.12 PATTERNS AND SENTENCES

Patterns and Sentences is a list of rare letter patterns in Icelandic and of sentences that contain words with those patterns. The sentences were taken from novels written around the year 2000. The purpose of the list was to ensure that when all the recordings for the *Hjal* project were completed, they would contain an example of all possible letter patterns in Icelandic. The database contains 1,433 sentences, which are accessible with a CC BY 3.0-licence.

In building modern speech synthesisers, many more sentences are needed than exist in this database. Therefore, it is important to define ways of building a larger reading list containing all possible Icelandic patterns. Based on the number of sentences being recorded, this reading list would be used to create the optimal lists for recording.

T.12 Patterns and sentences

Work packages:

- Defining methods to create reading lists
- Creating reading lists

Human resources:

- NLP expert: 2 months

2.2.5.13 PARAMETRIC SPEECH SYNTHESIS FOR ICELANDIC

Parametric synthesis is a new technology which, at the time of writing this report, seems to be achieving good results. It is assumed that this technology will achieve the best voice quality. The aim of this work package is to adapt parametric synthesis to Icelandic. The main focus will initially be on creating a statistical parametric acoustic model, which will be used to control a vocoder that generates speech. A statistical parametric model for Icelandic has not previously been built and it is important to allow enough time for this part of the project. General vocoders can be used to create speech, but they have generally been developed for English. As a result, it is important to adapt the vocoders to Icelandic.

The latest technology uses deep neural networks instead of a statistical acoustic model, but it is also possible to use neural networks in the steps preceding and following the acoustic model, for example for automatic phonetic transcription and vocoding. The adaption of this technology to Icelandic is expected to take some time.

T.13 Parametric synthesis for Icelandic

Work packages:

- Statistical parametric acoustic model
- Vocoder for Icelandic
- Parametric acoustic model based on a neural network

Human resources:

- Speech synthesis expert: 21 months
- Signal processing expert: 12 months

Total: 33 man months

2.2.6 TECHNOLOGY TRANSFER

Direct use of speech synthesis primarily revolves around making written text accessible by reading aloud. This technology is important for blind and visually impaired people, but can also be useful for people who, for some reason, can listen but are unable to read text. Speech synthesis is also particularly useful as part of larger language technology solutions, in which speech recognition is intertwined with text analysis. Examples of projects can be found in Chapter 5.2.

2.2.6.1 WEB READER

Speech synthesisers are important to make written language accessible to those who are unable to read the material but can listen to it. It is popular to add a reading button to websites to enable users to listen to content. This technology can be developed in various ways and presents many opportunities for innovation.

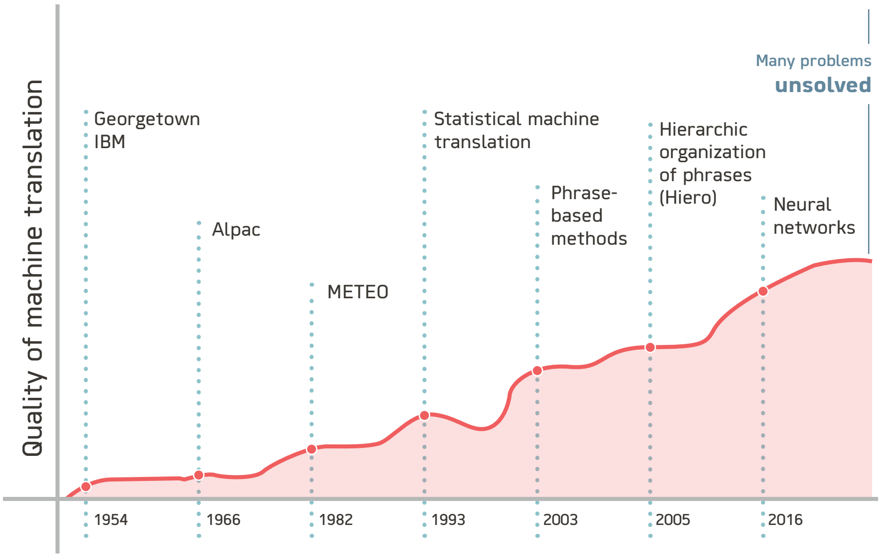
2.2.6.2 READER FOR E-BOOKS

A well-built speech synthesiser enables e-books and screen text to be read aloud automatically. People can use a speech synthesiser to enjoy literature or to listen to course material. In all cases, intonation and stress must be accurate and a wide selection of voices be available.

2.3 MACHINE TRANSLATION

An open-source machine translation system capable of translating between English and Icelandic will be built. Translation quality should be high enough to be useful for translators in certain domains, making it possible for them to work faster than they would if they had to translate everything from scratch.

The use of computers to translate text was first envisaged in the late 1940s. The initial ideas were to use similar methods as were used in cryptography in the Second World War and/or few basic rules of the languages being translated. Within a few years, research began in a number of universities, mainly in the US. The first systems were rule-based and despite great optimism in the beginning, now, nearly 70 years on, we still have a long way to go if we want to build a machine translation system able to match a good human translator.



Main stepping stones in machine translation since the start of the computer age

Machine translation (MT) has recently reached the point where it is useful to those who want to understand text in a language with which they are unfamiliar, and to expedite the work of translators in languages in which they are expert. Machine translation is, however, not yet perfect. All technological approaches to the problem have some limitations.

When modern MT-systems translate text for publication – in any language pairing – the text must be reviewed and proofread. With this being said, in many cases the technology has reached such a high standard that the use of MT makes it possible for translation professionals to work much more efficiently. MT-systems are used by translators, worldwide, to increase productivity and quality.

The world translation industry has a turnover of around US$40 billion a year; a large proportion of which is in Europe. There is a large and growing need for technology able to help translators in their work and increase their productivity. The European Union (EU) has invested considerable efforts into developing MT-systems for all its official languages: MT-systems are used to help translate, for example, administrative documents, laws, regulations, contracts and other official documents. Recent experience shows that the technology can be useful in translating up to 50% of official EU documents and that it can increase productivity by up to 20-30%, although this varies considerably between languages.

2.3.1 STATE-OF-THE-ART

The first MT-systems were based on rules built for pairs of languages, but statistical methods based on a large amount of data have dominated for the last 20 years or so. Traditional statistical machine translation systems use predictive models, which employ parallel corpora in two or more languages, to teach a computer how to translate: the text is usually written in one language and translated into the other, and the parallel sentences are then aligned. The use of parallel corpora enables the most probable translations to be achieved, but for the results to be useful the texts must be quite large: if a particular word or phrase is not found, the system is unable to translate it.

There are a number of variations of MT-systems that are based on statistical models: they use individual words, sentences or parts of sentences, syntax analysis, and other methods. The advantage of this is that the systems can work with many different language pairs without having to be specially adapted – the adaptation is mostly based on the data. To correct systematic errors and improve translations, there are some rules that run on input or output text, but the disadvantage of this is that it may be difficult and costly to assemble large enough parallel bodies of text, particularly for less-common languages.

In the last few years, MT-systems based on deep neural networks have shown great promise. They are reputed to be able to achieve better results than traditional, statistical methods. The inspiration behind the technology comes from the neural networks in the brain. These methods are devised to identify patterns or hidden structures in the input data. These patterns can be projections between word forms, words or phrases in one language to another, they can be the grammar or exceptions to the grammar of the languages in question or some other structures in the language. By decrypting these patterns, the system "learns" the languages and that what is necessary to convert text from one language to the other. In addition to parallel corpora, the system may be trained to use monolingual data to improve its understanding of how sentences are formed in the target language, and even to augment the parallel data. All the "knowledge" it gains from analysing the data can then be used in the translation process.

The quality of machine translation depends on a number of factors, including the characteristics of a particular language, and the quantity and quality of the training data. The fewer unknown words, the better. In the case of morphologically rich languages, it is all the more important for the MT-system to be trained on large training sets, as it thus will be less likely to come across rare word forms that it has not previously encountered.

There are indications that systems employing deep neural networks will be considerably more effective than older methods when it comes to morphologically rich languages. A recent paper written by experts from MT@EC (the European Commission's machine translation service, now CEF eTranslation - developed by the Directorate-General for Translation), states that by using neural networks it was possible to build an MT-system for Hungarian that was useful for, and is being used by, translators. The system that the institution had previously built for Hungarian, which was based on statistical methods, did not give sufficiently good results. Usefulness for translators is probably the most important element in evaluating the quality of MT-systems for professional purposes – if they are no use to translators, they have little practical value.

The first results of deep neural network MT-systems are promising: they are fast, and they are more likely to produce text that resembles natural languages than statistical MT-systems. Conversely, this could create problems because the more natural the translated text appears to be, the harder it can be for translators to identify errors; although there will probably be fewer than in older systems, they will always exist. Since this could have a negative impact on the efficiency of translators who rely on the systems, it must be examined and measured. Translators are, therefore, the most important collaborators in the development of MT-systems. Not only do they create the data that is used to train the systems, but they can also provide necessary feedback during the process.

2.3.2 OPEN-SOURCE SOFTWARE FOR MACHINE TRANSLATION

2.3.2.1 MOSES

Moses is a statistical MT- system that is licensed under an open-source LGPL-licence (see Chapter 3.1.1). It was initially released in 2005 and has been actively developed since. *Moses* is used for research purposes and live systems for business alike. Google and Microsoft have for example both used it in their systems.

Moses, and other statistical MT-systems, are trained by using vast amounts of parallel data to learn how to translate sentences or phrases. The system also utilizes monolingual corpora, normally a much larger dataset than the parallel one, to learn the structure of the target language.

In the training process, *Moses* identifies examples of the same words and phrases appearing in parallel sentences or parts of sentences. The system deduces from these examples. It is also possible to use grammatical information, such as part-of-speech, to make the system more accurate.

We can say that the system is twofold. On the one hand, there are training tools that are used to create translation tables and models to use in the translation process, on the other hand, there is the MT-system that uses the models to translate text from one language to another.

During this process, it is possible to tune the system in different ways and to attach different weights to the resulting translation models. It is also possible to choose between different translation algorithms and various types of language models. The most practical choices depend on the language in question and it can take many attempts to arrive at the most accurate results.

2.3.2.2 NEMATUS AND OPENNMT

Nematus and *OpenNMT* are neural networks MT-systems, which are under very active development. *OpenNMT* is licensed under an MIT-licence by Harvard University and SYSTRAN. *Nematus*, a co-operative project of scientists at universities in USA, Germany and Scotland, is licensed under a BSD-licence. Both licences are open and both systems were introduced in peer-reviewed papers in early 2017. In May 2017, Facebook released its MT-system, *fairseq*, under an open-source licence and in a peer-reviewed paper.

Since the authors of this report were not in a position to get acquainted with that system in time [the Icelandic version of the report was published in early June 2017], it is not discussed here.

The reason for multiple systems of the same kind to appear simultaneously is the vastly increased interest in the use of deep neural networks to solve language-technology problems. This interest is due to the fact that in the recent few years, research employing these methods have achieved very good results. Not much attention was paid to, and minimal research done on, machine translation with deep neural networks until 2014. Only two years later the situation had changed completely, and a great number of researchers in the field are now turning their focus to neural networks.

The basic concept of neural networks was explained at the beginning of this chapter. The two systems, *Nematus* and *OpenNMT*, are based on the same technology and are similar in many aspects. There is no reason to delve into the differences between the two in this report, but it must be investigated as to which of them is better suited to building an MT-system for Icelandic.

2.3.3 MACHINE TRANSLATION FOR ICELANDIC

Two rule-based MT-systems have been developed for Icelandic. *Tungutorg*, by Stefán Briem, translates between Icelandic and English, from Icelandic to Danish, and from Esperanto to Icelandic. It is a closed-source system for which the source code has not been released. *Apertium-is-en*, a prototype based on the *Apertium* translation system, was developed by students at Reykjavík University in 2009-2010. Neither of these systems is in public use and the technology they are based on is unlikely to give better results than more recent methods, such as neural machine translations (NMT).

Furthermore, *Google Translate* also translates between Icelandic and other languages. The accuracy is low as it does not differentiate between texts from different domains and, as a result, is prone to making errors in translating ambiguous words, phrases and concepts. It is a closed-source software that only Google can develop or adapt to special needs.

International consultancy on machine translation for Icelandic

The authors of this report consulted with international experts in the field of machine translation on how best to develop a policy for Icelandic machine translation. Meetings were held with experts from MT@EC, and from Estonia, who have worked on the development of machine translation in the Estonian language technology programme.

MT@EC's role is to develop machine translation for public institutions in the EU, in Norway and in Iceland. It is responsible for all 24 official EU languages, as well as Norwegian and Icelandic. Because of a lack of Icelandic training data, MT@EC has not, however, been able to provide for Icelandic, and now works with 24 languages in 78 language pairs. The department's service to administration bodies will continue to be free of charge until 2020, after which it is planned for a larger project to take over, eTranslation, that will not be confined to translating official documents.

The first version of MT@EC's translation system was released in 2013; it was tested and advised by selected translators. MT@EC used statistical models and the open-source statistical MT-system Moses. The group started to experiment with neural network systems in the latter part of 2016.

MT@EC's MT-systems are custom-built to enable it to fulfil its primary objective, to translate administrative documents. The system, which can be used through a web interface or web services, is targeted at translating entire documents, rather than individual sentences. The normal procedure is to download a document to be translated and for the translation to be emailed when it is ready. The group plans to run a dual system: one will focus on speed rather than quality, the other on quality rather than speed.

There are great variations in the quality of machine translation systems, depending on the languages in which they are working. MT@EC has experimented with Icelandic, but without any functional results because its training set has been small: only around 400,000 parallel sentences or phrases. The experts at MT@EC's consider that it may take 25-35 million parallel sentences or phrases to form a solid basis for training good systems. Its largest, English-Spanish, contains about 50 million parallel sentences or phrases.

The complex morphology of Icelandic and the number of unknown, mainly compound, words, is likely to create difficulties for traditional, statistical MT-systems. While this applies to other languages with similar structures, MT@EC's experiments with neural networks have proven successful in the case of Hungarian, which used to be in the same position as Icelandic. Further experiments with deep neural network systems are now being carried out on other languages that have hitherto not enjoyed great results.

The greatest advantage of neural networks is their ability to understand complex grammar.

MT@EC uses open-source software: Moses, for statistical MT-systems, and Nematus and OpenNMT for neural networks. Their experts say that if machine translations are to achieve decent accuracy for Icelandic, it will probably be necessary to establish a very large training database. For the first experiments, however, they expect that around two million parallel sentences or phrases will be adequate to establish a base-line for minimum quality and accuracy. They also say that experiments and research will be needed to define how much data are needed to achieve useful results.

MT@EC have experimented with using part-of-speech taggers and sentence parsing in machine translation. This appears to increase the quality of the translations a little, but is very time consuming and, therefore, not common practice. Their MT-systems pre-process data: normalise text, such as punctuation and numbers; ensure that capital and lower-case letters are used in training and translating; and fix other minor details. For certain languages, particular items are more likely than others to be translated incorrectly. The post-processing of output text, individually adapted to each language, can remedy this and help translators by drawing their attention to possible errors. The value of this phase should not be underestimated: pre- and post-processing can improve translation quality greatly, not least when translating numbers or other elements that follow clear, systematic rules, but vary between languages.

It is vital to work with as many Icelandic translators as possible to build parallel corpora, and to provide important feedback, throughout the development process for an MT-system for Icelandic, a language spoken by relatively few people. MT@EC suggests that we should try to get as many translators as possible to lodge their translations into a common system in return for being allowed free access to the system.

Work on machine translation in the Estonian language-technology plan, has been ongoing since 2015. The principal objective is for the quality of machine translation to become adequate for it to be useful in translations for specific domains, and for the translators to complete translations more quickly than if they had translated from scratch. The bulk of the work has been in adapting methods which have proven useful with other languages, and in experimenting with translation companies in building equipment to translate texts from particular domains. The first stages have involved a great deal of work in collecting usable data and establishing parallel source material. The problems that have been emphasised in language technology research for Estonian have included the handling of compound words; productive word formation is commonplace in the Estonian language, just like in Icelandic. To solve this issue, they have tried dividing the words into morphemes for training and in the pre-processing of text. Better results can be achieved if this is done correctly. It has been difficult for the MT-systems to form a normal word order in the target language, particularly when the target language is Estonian (and not English) in translations between English and Estonian. The best results, as might be expected, come from short sentences. As of now, the results already seem good enough to be useful, although this has not been thoroughly evaluated. It is currently being researched to see whether, as the first findings indicate, neural MT-systems give better results. Word order and grammar are mostly correct, but there are some omissions and/or additions to the original text in roughly a quarter of the translated texts. Errors of this type are always difficult to analyse, but research must be carried out on how much they affect translators' post-processing. Since it is likely that Icelandic will confront many of the same problems, it may be beneficial for researchers working on Icelandic MT to follow the development in Estonia.

Emphases in the development of machine translation for Icelandic

The focus must be on developing practical machine translation for Icelandic. The simplest way to evaluate its potential use is to monitor whether translators use it. It is not expected that MT-systems will be useful in all domains from the outset, but it must be defined as to where most value can be gained. In developing a useful MT-system for Icelandic, the first steps will involve collecting training data, deciding on what requirements to fulfil, and which technology to employ.

It is MT@EC's experience that the quality of translations varies greatly between languages. Icelandic's complex morphology and productive word formation is likely to demand relatively large amounts of data. Looking at results for languages facing similar obstacles, including Slavonic and Finno-Ugric languages, and German, there is some indication that neural networks, rather than statistical MT-systems, are likely to give more accurate results for Icelandic. Emphasis should therefore be on the development of neural network systems, but for comparison, and in order to ensure that the right route has been chosen, it would also be useful to experiment with traditional statistical methods.

2.3.4 QUALITY EVALUATION

There are many ways of evaluating the quality of machine translations. The approaches taken depend on what in particular is being evaluated. Automatic processes can be used to assess whether an MT-system's performance has improved or deteriorated after changes in the software, but other methods are more useful in evaluating if the machine is useful, and whether it expedites the translators' work in finalising text for publication.

We will focus on four methods of evaluating the quality of translation machines. They are either used by MT@EC, or were used in the SUMAT-project, which focussed on developing software for translating subtitles. Two of these methods were used at both MT@EC and in the SUMAT-project.

1. MT@EC monitors how large a percentage of translators in each language pairing (for example English > German or Romanian > English) looks at machine translation or takes it into account while working. This has shown that there is considerable difference in usage between language pairs, which is likely to stem from the fact that, depending on the languages, machine translations are not equally good. In the pairs with the best machine translations, translators are more likely to take them into account. On average, machine translations were looked at in around 50% of cases. Since translators at the Translation Centre have continuous access to machine translations, over time it can be used to see whether MT-systems are improving and for which languages they are most useful.

2. TER (Translation error rate) measures how many corrections must be made on a machine translation to achieve the desired result. The percentage of words that must be changed is calculated. By examining how much post-processing is required for each language pair, the systems' accuracy can be compared. This method is used both by MT@ EC and in the SUMAT-project.

3. The BLEU-score is commonly used in machine-translation research and development. It is used at MT@EC and in the SUMAT-project and is suitable in evaluating whether changes made on a translation machine are likely to give better results. It is not always useful to evaluate the quality of a finished translation, nor to compare systems that are based on different underlying processes.

4. In the SUMAT-project, translators' work was timed to find the usefulness of machine translation systems for each language pairing. The translators worked both with and without an MT-system. The results varied with each pair, but on average, based on subtitles per minute, the translators' productivity increased by around 35.5%.

In developing an MT-system for Icelandic, it is essential to use as many methods as possible to evaluate the quality. The choice of methods is governed by the stage of development that is being tested. TER should be used at all stages when it is possible to get translators to evaluate the quality of the machine translation; and when evaluating the quality of MT-systems, measuring time spent and thus productivity is probably the best way to judge usefulness.

2.3.5 DEVELOPING INFRASTRUCTURE FOR MACHINE TRANSLATION

The main objective of machine translation in the Icelandic language technology programme is to build useful MT-systems that will increase productivity of translators. With that in mind, focussing on the language pair English/Icelandic would yield the biggest benefits. To achieve this objective, we need to establish adequate data collections, build a general, open-source MT-system and corresponding support system, and install work processes to adapt the system to specified fields.

2.3.5.1 COLLECTING DATA AND BUILDING CORPORA

There is no open parallel corpora available for Icelandic, but they are key to the development of MT-systems. It is important to start as soon as possible on creating a parallel corpus. We should aim for 25-30 million sentence pairs during the period of the programme.

Two projects must be undertaken on the construction of open parallel corpora for machine translation. The first revolves around using automatic processes to create parallel text from material that is accessible on the web. We should investigate how texts from Wikipedia and OpenSubtitles have been employed elsewhere, as well as material from CommonCrawl, which contains a large amount of accessible texts from the web. The methods that have given the best results elsewhere should also be used here.

2. CORE PROJECTS

V.1 Parallel Corpora built from material available on the web

Work packages:

- Parallel Wikipedia corpus
- Parallel OpenSubtitles corpus
- Parallel CommonCrawl corpus

Human resources:

- Computer scientist: 24 months
- Linguist: 6 months

Total: 30 man months

NB: the material should be constructed automatically to enable it to be enlarged easily as accessible web material increases.

In the latter project, a parallel corpus will be constructed from documents translated by the Translation Centre of the Ministry of Foreign Affairs for the EEC Treaty. Around 7,000 documents are available in Icelandic, corresponding documents in English and, in many cases, also in other European languages. Documents and sentences from the documents will be paired to create a parallel source from this data.

V.2 Parallel source material

Work packages:

- Pairing Icelandic and English documents
- Pairing sentences

Human resources:

- Computer scientist: 12 months
- Linguist: 18 months

Total: 30 man months

NB: this material should be constructed automatically, but the data should, at least partly, be reviewed to ensure the creation of a parallel source material that can be relied on to be correct.

In this work, whether in regard to the EEC files or other data, alignment software such as hunalign will be used to pair sentences from two languages. For the parallel material to be totally accurate, the pairing must be reviewed and corrected where the software has made errors. Since the goal is to ensure that the data is trustworthy, it must be evaluated how many texts will be required to fulfil those conditions. To achieve the desired results, it will also be necessary to examine the acceptable error rate for a corpus that has not been checked manually for errors.

There are other possible sources for parallel corpora. TV stations and others who publish films for the Icelandic market produce a lot of subtitles, and book publishers have foreign books translated. This material is, however, mostly copyright protected and would either need an amendment to the law, or substantial negotiations to obtain a licence to use it for MT-systems. It is not, therefore, assumed that the use of this material will be possible in the core language technology projects.

When MT-systems are tailored towards a particular domain, a parallel corpus with texts in that domain is important. Using translations of EEC regulations could be a first attempt at this, but it is also necessary to investigate whether, and then in which domains, building parallel corpora for Icelandic MT is feasible.

ELRC is an EU project that is aimed at collecting parallel corpora and other useful data for MT-systems. It has obligations towards 30 European languages, including Icelandic. We need to explore the possibility of collaborating with ELRC in building parallel corpora for Icelandic.

A large, monolingual corpus can be very useful for increasing accuracy of MT systems. The Icelandic Gigaword Corpus can be used for these purposes (it is described further in 2.5.1.3).

Bilingual dictionaries and glossaries are useful in helping to fill gaps in parallel texts, as well as for choosing the right translations in specific domains. Part of the content of the Icelandic Term Bank of the Árni Magnússon Institute for Icelandic Studies is accessible under a CC BY-SA-licence, but other bilingual sources are not. It is important to investigate whether other bilingual lexicons can be made available for language technology projects. No English-Icelandic dictionary has been compiled by any public bodies, which might pose some limitations in this area. The Translation Centre of the Ministry for Foreign Affairs' terminology is accessible online. It contains translations of concepts that have been used in documents translated by the

Translation Centre, but not general English vocabulary. It has not been licensed in a way to render its contents available for developing language-technology software.

2.3.5.2 INFRASTRUCTURE

MT-systems, preferably both statistical and neural, must be experimented with. For initial training, we need training material such as the translation memories from the Translation Centre of the Ministry for Foreign Affairs (see 2.5.1.19 Translation memory of Translation Centre), to train the software and then add data from parallel corpora as they become available. The Translation Centre's translation memories contain around 1.2 million paired sentences. A pilot project should determine which open-source software should be used and the development continue on the basis of the pilot project. In the pilot project experiments should be carried out on Moses, OpenNMT and Nematus, or other common tools in general use.

V.3 Baseline in machine translation

Work packages:

- Installing selected open-source MT-systems and adapting available data to each one
- Examining the benefits of each system
- Analysing the need for pre-processing and post-processing the texts
- Choosing the best MT-system and estimating the needed size of parallel corpora for it to achieve the desired results

Human resources:

- Computer scientists: 18 months
- Translator: 6 months

Total: 24 man months

Once certain milestones have been reached, the MT-system and all open data will be released. The first milestone is simply for the system to deliver a sentence in Icelandic when it is fed with an English sentence, regardless of quality. The system used will be the one that was chosen in V.3.

V.4 Open-source MT-system for Icelandic

Work packages:

- Developing the MT-system, establishing pre-processing and post-processing rules, and other items that are needed to minimise the TER (see 2.3.4 Quality evaluation).
- Installing instructions and the programs necessary to facilitate the installation of specialised MT-systems

Human resources:

- Computer scientist: 72 months
- Translator: 18 months

Total: 90 man months

NB: this should be a three-year project with milestones that must be reported on every six months.

2.3.5.3 USAGE SUPPORT

An Application Protocol Interface (API) must be installed to give interested parties access to systems under development, and to establish a testing environment in which members of the public can submit their own text, get results from different MT-systems and choose the result they like best. This will be carried out in tandem with the development of the MT-system to give as many people as possible access to it during the development process and to receive instant feedback. By doing this, it will, for example, easily compare the results of the language technology project to closed-source tools, such as Google Translate.

V.5 Translation machine interface

Work packages:

- Implement API
- Building testing environment
- API and testing environment put to use

Human resources:

- Computer scientist: 18 months

Total: 18 man months

2.3.6 TECHNOLOGY TRANSFER

The accuracy and quality of MT-systems largely depends on the success in translating content correctly from one language to another, and on how comprehensible the translation is in the target language. When a word or phrase in one language can have a different meaning in another, it is possible to increase accuracy and to decrease ambiguity by limiting translation to a specific domain: in this context, the machine is more likely to choose the correct translation in the target language.

2.3.6.1 SUBTITLE TRANSLATION

Systems have been developed to expedite the work of TV and movie translators. The SUMAT-project, an EU research project in 2011-2013, was designed to create a system for subtitles to translate between European languages. The system was developed for 11 language pairs and, by using it, the translators' productivity increased in all but two cases. The average performance of translators using the system increased by around 40%.

If a system that increased translators' productivity existed for subtitles between English and Icelandic, translations could become cheaper and better. To create a powerful translation system, however, a great deal of data is needed. It would be ideal if Icelandic companies needing subtitling for screen material were to collaborate on such a project.

2.3.6.2 TRANSLATION OF OFFICIAL DOCUMENTS

At the Translation Centre of the Ministry for Foreign Affairs, a great deal of work goes into translating official papers. MT@EC has developed a system to assist in translating documents between most official EU languages. If the project's experts were to have access to Icelandic data, they would also try to develop machine translation for Icelandic, at no cost to Icelanders. Once the MT@EC project is complete, we would be able to draw on their experience of working with Icelandic machine translation. The system would have to be accessible for Icelandic, and its development continued for the benefit of the Translation Centre and others who work on administrative translations.

2.4 SPELL AND GRAMMAR CHECKING

Icelandic spell and grammar checking will be developed to detect and correct writing errors. By recognising grammatical and semantic context it will be able to handle many more errors than is currently possible. Open-source spell and grammar checking can be connected to word processing, smart devices and other language technology (LT) software.

Software that corrects errors is not only important to help people write, but also as a part of other LT systems. There are a few programs available for Icelandic that correct errors in individual words if the erroneous word cannot be found in a dictionary. It is necessary to develop open-source software that can analyse grammatical and semantic context, to be able to recognise context-based errors. To enable this, data on common errors needs to be collected and objectives defined for each step in the development of the correction software.

In many languages, LT software for finding and correcting errors in text has become standard. The errors can be of many types: typographical, misspelling, grammatical, and lexical. This document refers to them all as writing errors, and to software that detects and corrects them as spell and grammar checking. Software can also help with writing in other ways, for example by reviewing style and register. The introduction of Icelandic spell and grammar checking software enables the development of such tools.

People sometimes make mistakes when they write; through, for example, carelessness, insufficient practice or dyslexia. The reading of texts that are full of errors can prove difficult, and their meaning may also become unclear. In the modern digital environment, texts and information are often found through search engines, which need to be able to identify the appropriate information even when the query or text they are searching for contains errors. Other LT software, for example speech synthesisers that read text, similarly rely on texts being close to error-free. Spell and grammar checking is, therefore, a necessary basic software in any kind of automatic or manual text processing.

The automatic spell and grammar checking for English and other languages, embedded in Microsoft Word, is a well-known example of text correction software. This kind of writing support in text processing software, e-mail editors, etc., is the type of text correction software usage with which the majority of people are familiar. The usefulness of such support can be defined from different perspectives: the point of view of the writer, the circumstances, or the value of quality texts in different domains.

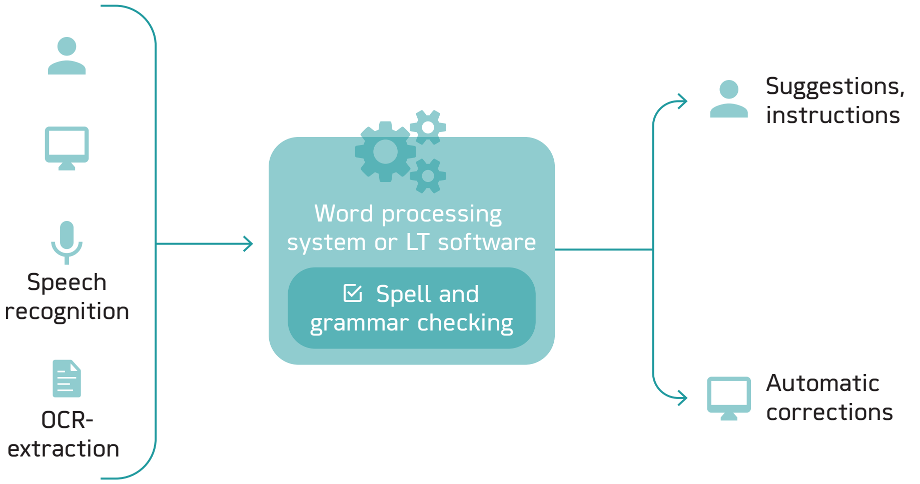
- **Usefulness based on the skill of a writer:** the quality of people's writing not only depends on education and practise, but also on age, on whether the person is writing in his/her native language, and on conditions such as dyslexia.
- **Usefulness based on a writer's circumstances:** there is often little time to review texts before they need to be handed in or published. Typographical errors and other lapses can occur, even when the writer is skilled.
- **Usefulness based on quality:** some texts need to be of a higher quality than others. Published material, books, newspapers, etc., should normally consist of high-quality content since it is the text that is the product. Where large volumes of text are created, it is important that it is written according to a certain standard, in order to facilitate finding what is being looked for and to avoid unnecessary interpretation problems caused by errors. This can apply to health institutions, administrative bodies and the judicial system. Texts are also influential in creating a company's image through, for example, written communications with customers, terms and contracts, and information and advertising.

The importance of automatic spell and grammar checking depends on a combination of the above: the lower the level of skill, the less the time available, and the higher the demand for quality, the more important such writing aids become.

Spell and grammar checking also plays an important role in other LT software. This can be programs for which language checking is the core, such as in writing enhancement software or in teaching programs, but also software such as speech recognition, speech synthesis, machine translation or search engines. Almost all the more complex LT software relies on automatic spell and grammar checking.

The third field of application for spell and grammar checking is the correction of texts extracted using Optical Character Recognition (OCR). A great deal of work has been done in Iceland on OCR-read text from printed material from the second half of the 19th century and onwards (see http://timarit.is/). To be able to use these texts in a digital environment, they need to be checked for writing errors that are unavoidable when they are converted into digital form.

Spell and grammar checking may need to be adapted to each user group and application. For example, errors made by children are likely to be different from those made by adults, and errors made by speech recognisers and OCR systems will be different from human errors..


Speech recognition
OCR-extraction
Word processing system or LT software
Spell and grammar checking
Suggestions, instructions
Automatic corrections

Spell and grammar checking in various contexts

2.4.1 WHAT ARE WRITING ERRORS?

Errors in a written text are words or sentences that do not conform to the spelling or grammar rules/linguistic use of the language in question. In a living language, these rules are not always clear and, as time passes, some misuse becomes so widespread that it becomes the norm. An example could be the infamous “dative sickness” in Icelandic (for example, mér langar (I want, dative) instead of mig langar (I want, accusative)) that is still not considered a valid (written) language but is increasingly widespread and accepted in the Icelandic language community. In the development of spell and grammar checking it needs, therefore, to be kept in mind that it is not always easy to tell “right” from “wrong”.

In creating a spell and grammar checking system, it is useful to divide errors into categories based on the type of analysis that is needed to detect them. First, the errors are divided into non-word errors and real-word errors. The first category contains errors that create non-existent word forms in

Icelandic (isolated word errors and, in many cases, hyphenation/spacing errors); real-word errors, on the other hand, are where the "incorrect" words do exist in Icelandic but are used in the wrong context. While the following list is not exhaustive, it gives an idea of the various points that need to be considered in creating spell and grammar checking software.

- **Isolated word errors** – spelling errors that can be detected without looking at other words. This is possible if the resulting word form is invalid in Icelandic, for example pisla instead of pistla (opinion columns), but not if the wrong word form exists as another word, for example neita (deny) instead of neyta (eat).
- **Grammatical errors** – errors where it is necessary to make a grammatical analysis of words – or sentences or phrases –to find them. Common errors relate to, for example, declension and verb moods/tenses: Hann er langbestur leikmaður í heimi instead of Hann er langbesti leikmaður í heimi (He is by far the best player in the world); Guðni segir í samtali við Vísi að undirbúningur málþingsins hófst í byrjun árs, instead of Guðni segir í samtali við Vísi að undirbúningur málþingsins hafi hafist í byrjun árs (in a conversation with Vísir, Guðni says that the preparation for the symposium started a year ago).
- **Context-sensitive errors** – to find these errors, it is necessary to analyse the semantic and/or grammatical context: Sem var við líði á þeim tíma instead of sem var við lýði á þeim tíma (that was in fashion at the time); græða sem mest á sem skemmtum tíma instead of græða sem mest á sem skemmstum tíma (make as much money in as short a time as possible).
- **Hyphenation/space errors** – errors where a space is missing or is in the wrong place: Sá hvorugur þeirrahjólreiðamanninn (Neither of them saw thecyclist); sækjast eftir háskóla námi (pursue university studies); réttis em innihalda kjöt instead of rétti sem innihalda kjöt (dishes that contain meat).
- **A word missing or added** – these errors are often connected to set phrases: Bíða með fram vormánuði instead of bíða með fram á vormánuði (wait until spring to do it); þar biðu þeir bara á þangað til instead of þar biðu þeir bara þangað til (there, they just waited until).
- **Discourse/dialogue errors** – detecting these errors demands an analysis of the text. To spot the error in the second sentence, the first one needs to be interpreted: Katrín Tanja hafði á endanum betur eftir hörkukeppni en það munaði aðeins þremur sekúndum á stelpunum. Katrín kláraði á 6:53 mínútum en Sara á 6:53 mínútum (Katrín Tanja

won after a very exciting match, the time difference was only 3 seconds. Katrín finished in 6:53 minutes and Sara in 6:53).

- **Incorrect use of terms** – meaning and nuances of meaning, but also tradition, decide that words fit semantically with other words and that words are normally used in a certain context: Ýsan er útrunnin (The haddock is expired) instead of ýsan er gömul/úldin/óæt (the haddock is old/rotten/inedible). Here, a wrong word, not necessarily similar to a word that could be used, is used in a certain context and the possible correction is not one definitive word.
- **Punctuation errors** – a punctuation mark is missing, is superfluous, or in the wrong place: Síðan kom annar þáttur til greina en það var snákurinn sem sagði þeim að borða ávöxtinn tel ég að ráða megi bót á öllum þessum … (There was also another factor the snake that told them to eat the fruit I think this could all be sorted out …)

Writing errors happen because of carelessness, by mistake, or because the writer doesn't know any better or is unsure. It is often difficult to tell which: does the person who writes noskra instead of norskra ((adj. Norwegian)) not know better or are they simply in a hurry? Even if it is not necessary to be able to detect the reason for each and every error, a knowledge of the way errors are created is, however, useful when developing spell and grammar checking software.

2.4.2 STATE-OF-THE-ART AND LEADING METHODS

Automatic spell and grammar checking is done in two stages: first, errors are detected and, second, corrections are suggested or made.

2.4.2.1 SPELL AND GRAMMAR CHECKING

Automatic spell and grammar checking methods have been developed at least since about 1960. For a long time, the main focus was on detecting isolated word errors by comparing words in a text with valid n-grams (a sequence of letters that can stand together in a language) or by using word lists. Today, word lists are more likely to be used for traditional error checks, but n-grams are more commonly used for correcting OCR-read texts. The basic method in isolated word error search is, therefore, simple: if a word is not found in the software's word list, it is marked as an error. This is, however, only the first step in the error check. The word list of a spell checker can never contain all possible words and word forms of any language. In order not to confuse the user with too many false positives (marking a correct word as an error), unknown words need to be analysed further. In languages such as Icelandic, German and Estonian, which contain a large number of compound words and in which new compound words are constantly being created, valid unknown words are most likely to be either compound words or proper names. Spell and grammar checking for these languages must therefore use some sort of analysis of unknown words to assess the probability of them being valid compound words or proper names. It is also possible to use parsing to assess whether the word is a valid form of a known word that is missing from the word list.

To assess whether a word found in the word list is an error, the text needs further analysis. How this is done and how detailed the analysis is depends on the type of error. One relatively simple way is to define confusion sets. These are pairs or lists of words that are commonly confused because, for example, they are pronounced in the same way. In Icelandic these are words such as leiti – leyti; neita – neyta; sína – sýna; list – lyst, etc. Each time the program detects one of these words, it looks up the rules of the confusion set in question and compares them to the context in which the word is used. Example: Verkinu var að miklu leiti lokið (The work was largely finished). The word leiti is found in a confusion set, with examples of context. The rule is that in the context að miklu …, leyti should always follow: {leiti, leyti: að litlu/miklu/öllu leyti; á næsta leiti}. The rules can be in various forms, for example with detailed context, collocates, part of speech patterns or syntactic patterns.

It is difficult to find errors relating to words from the word list without a grammatical analysis of the text. Words are tagged with the appropriate grammatical information (for example, for Icelandic nouns: noun-gender-number-case-article-proper noun), and sentences analysed based on certain grammatical factors. The results of the tagging are compared to the program's collection of rules, and errors are marked where they are found. For example, the Swedish spell and grammar checker GRANSKA defines rules such as: the free-standing article must have the same gender and number as its noun. If the program detects inconsistency in the tagging of article and noun, it marks it as an error and the user is notified.

2.4.2.2 SPELL AND GRAMMAR CORRECTION

Once an error has been detected, a likely correction must be found. In a rule-based system, the corrections or instructions can be part of the rules.

The most common way of correcting isolated word errors is based on research from the 1960s. Dameraus (1964) used the following definition of edit distance: one character is superfluous (insertion), one character is missing (deletion), one character is replaced with another (substitution), or two characters in a row switch places (transposition). These definitions have subsequently been known as the Damerau-Levenshtein distance. In English, the fact that around 80% of misspelled words contain only one error – an insertion, a deletion or transposition ¬– has formed the foundation for corrections using the Damerau-Levenshtein distance. Whether systems are based on rules or on statistical methods, most use calculations of character distance to find probable corrections.

Although most incorrect words are close to the correct one, researchers have found that the difference between the incorrect and correct version of a word can be by as much as seven letters as, for example, in German.

The most popular statistical method for correcting spelling errors is based on the noisy channel model: the concept is that the word is spelled wrongly because of a “noise” in the communication channel. The spell and grammar checker need to calculate the probability of a certain word having been distorted in such a way. First, the theory of short character distance (Damerau-Levenshtein) is used to find the likeliest correct words, then Bayesian inference is used to assess which of them is the most likely correction. For example, if the misspelling firir (for) is detected, the likeliest proximate words are fyrir (substitution i - y) and firðir (insertion ð). The calculation takes into consideration how likely it is that the wrong word is in fact a misspelling for the word in question, and how likely it is that this word belongs in this place in the text, regardless of the type of error.

When spell and grammar checking is used to assist writing, it is enough to show the user likely corrections in the order of probability, but when correcting a text in a software system, the spell and grammar checker must decide on the most probable correction and correct automatically.

As a result of the vast amount of text now available on the internet, as well as the advance of new methods and computer hardware to process it, experiments have been made in using internet texts to train spell and grammar checkers that need neither manual rules nor word lists.

2.4.2.3 DUDEN KORREKTOR

The authors of this report met with the creators of two rule-based spell and grammar checkers for German and Estonian. EPC (http://www.epc.de/) holds the rights for the Duden Korrektor spell and grammar checker for German, which has been in development since 2001. This software uses the comprehensive databases of Duden, the leading publisher of dictionaries and books on grammar in Germany. The rules are developed on a powerful parser, designed by another company. Duden Korrektor can be used with MS Office programs, but more extensive software, Duden Proof Factory, which offers hyphenation, language register testing, dialect analysis, different levels of correction, etc, in addition to the spelling and grammar corrections of Duden Korrektor, has also been developed. In EPC's experience, to avoid problems with updates, it is important to adhere to Microsoft standards when software is intended to connect to MS Office. Duden Korrektor is in constant development; daily changes are tested against a special corpus to evaluate their effect on the overall quality of the software.

2.4.3 SPELL CHECKING FOR ICELANDIC

Although a few spell-checking tools have been developed for Icelandic, none of them are open-source; therefore, they are not available for further development by third parties. The most commonly used are Púki Writing Error Protection, Skrambi, and an Icelandic version of the Hunspell-spell checker. These have primarily been designed to detect and correct isolated word errors, but they also perform some analysis of compound words, and Skrambi has a few built-in confusion sets.

Púki Writing Error Protection has been in development since 1984, and the first version was released in 1987. It was originally tailored to Word and Windows and has been updated and improved alongside new versions of MS Office and the Windows operating system. In 2014, it was released for OSX operating system, where it works with the MS Office-package and all programs that support system-wide corrections. Much like the built-in spell and grammar checkers, it is connected to programs in the MS Office-package and thus uses many MS Office options. It is possible to add words to Púki's dictionary that is used to check the validity of a word, find synonyms, use automatic correction, ignore error messages, and activate a hyphenation program. It also comes with a program to look up word inflection forms.

Púki only detects isolated word errors and does not analyse context. Unknown compound words are not marked as errors if the program evaluates them in line with Icelandic word-formation rules. It is built and sold by a private company, Friðrik Skúlason ehf, and its correction methods are not open. The software description, however, states that by identifying inflection forms it can assess whether a word is spelled wrongly or is foreign. Púki works with word lists, morphological analysis, and methods to analyse compound words, and is most likely completely rule-based.

The Skrambi spell checker was developed as part of, and pursuant to, Jón Friðrik Daðason's Master's thesis, Post-Correction of Icelandic OCR Text (Leiðréttingar á ljóslesnum texta, 2012), which describes Skrambi's methods. The program, which can be used for short texts through a web interface, is based on the noisy channel theory (see 2.4.2.2) and contains a few confusion sets. The correct word from a set is chosen according to context, with the aid of classifiers. Skrambi's word list consists of all word formations in The Database of Modern Icelandic Inflection (DMII), as well as a list of indeclinable words. It also contains specific word lists for the analysis of compound words: it checks whether an unknown word can be a valid compound word before it marks it as an error.

Hunspell is open-source spell checker software that is used in a number of browsers and programs, including LibreOffice, Photoshop and InDesign. It is not tailored to any particular language: it simply needs word and affix lists of the language that needs correcting. Such lists exist for Icelandic that can be obtained from the internet to use with programs that use Hunspell.

When testing Icelandic spell and grammar checkers, it soon becomes clear that the huge number of word forms in Icelandic makes it extremely difficult to detect isolated errors out of context. With the exception a few handpicked words in Skrambi, none of the programs detect context dependent errors. Typographical and other isolated errors are often impossible to identify without context, since the incorrectly spelled word becomes a word form of another word. Example: Þeir vilja bara græða á sem skemmtum tíma (They only want to make money in as short a time as possible). The incorrect word, "skemmtum" (entertain) should be "skemmstum" (shortest). None of the programs would detect the error in "skemmtum" because it is a valid Icelandic word. The three programs were tested on several sentences from the web, and on text fragments from a collection of error-marked sentences, in which each sentence contained at least one error. The collection was assembled in a Norwegian project: Feilkorpus for å testa stavekontrollar for grønlandsk, islandsk, lulesamisk og nordsamisk, in 2013. The results are shown the chart below. A total of 151 errors were manually labelled, both isolated word and context-sensitive errors. A review was made of how many of the errors the programs detected, and how many correct words were presumed to be errors (the text contained foreign names that would normally be marked as errors, i.e. not all the words marked as errors were valid Icelandic words).

	Errors detected	Correct words marked as errors	Precision	Recall	F-value
Púki	96	11	90%	63.6%	0.745
Skrambi	71	16	81.60%	47%	0.597
Hunspell	88	41	68.20%	58.30%	0.629

The test result of spell checking for Icelandic

Although the reviewed spell checkers certainly useful, they may offer false security to a certain extent. While users do not expect them to detect mistakes in grammar, or other context-sensitive errors, the programs also overlooked many simple typographical errors.

Chapter 2.4.1 listed nine categories of errors in writing. This is not an exhaustive list, but it gives some idea of the project scope. The existing programs deal almost solely with one of the categories, isolated word errors, independent of context. Skrambi also checks the context of a few predefined words.

Therefore, it is imperative to develop a foundation for spell and grammar checking that can analyse text grammatically and semantically, thereby detecting more error categories. If an open-source software of this kind is released, it may be further developed and used in general-user software, teaching programs and other LT software.

2.4.4 QUALITY EVALUATION

The quality of spell and grammar checking can be evaluated in a number of ways. The basic standard should be precision, recall and F-value, based on a standardised test corpus. Precision needs to be measured for error detection and error correction. The precision of error correction reveals how many of the words/phrases that the program marks are true errors, and also whether the program gives the accurate correction. In testing, the values should not only be measured against a compendium of errors, but also against the errors that the program is expected to be able to find and correct. For comparison, an independent expert, such as a proof-reader, may be employed to correct the test corpus and to calculate precision, recall and F-values based on the previously labelled errors.

User testing is a further form of quality evaluation: users evaluate whether the program is a help or even a hindrance: does it, for example, mark too many correct words as errors? Finally, evaluation must be done on the impact an integrated spell and grammar checker in a larger LT-software system has on the overall quality of that system.

2.4.5 DEVELOPING INFRASTRUCTURE FOR ICELANDIC SPELL AND GRAMMAR CHECKING

The development of infrastructure for spell and grammar checking is twofold.

- **Collecting and analysing data** to map common writing errors, and to train and test spell and grammar checking software.
- **Developing spell and grammar checking methods** that are able to correct predefined error types to a certain level.

2.4.5.1 ERROR CORPORA

There are many types of writing errors (see 2.4.1). Since different error often demand different means of detection and correction, the mapping of errors is important in order to be able to define the objectives of spell and grammar checking. Two things must be kept in mind: how common a particular type of error is, and how easy it is to detect and automatically. Considerable work needs to be invested in detecting and correcting the most common errors while, for the time being, ignoring less-common errors which are harder to correct.

Another purpose of data collection and error analysis is to prepare good-quality test data that measure the quality of spell and grammar checking.

The one error corpus that has been created for Icelandic contains around 167,000 words – -drawn from high-school students' final essays, and from news and blogs – and is part of another project. Its licensing, and access to it, is unclear. It is necessary to create a larger and more detailed error corpus that is tailored to the development and testing of Icelandic spell and grammar checking: a collection of around 500,000 words must be assembled. This would be incorporated into the Icelandic Gigaword Corpus (see 2.5.1.3) but should be highlighted since it is likely to contain a number of writing errors. The bulk of the work in creating an error corpus is in annotating writing errors. This will be done with a tool for manual annotations (see 2.5.3.1), but at the beginning it will be necessary to define error types and the way they will be labelled. This must be supervised by, preferably, more than one highly experienced linguist since various differences of opinion may arise. While the annotations can be done by postgraduate Icelandic students, it is important to co-ordinate and review this work regularly. From experience gained during the creation of the above-mentioned error corpus, it could take up to 1,000 hours to review 500,000 words.

In order to train a spell and grammar checking system using neural networks, a choice of two corpora must exist: either a very large database of relatively error-free high-quality text, or a huge general corpus that, by containing many billions of words, is large enough for the checker to learn the correct and incorrect versions of particular words and phrases. In this project we only describe the development of annotated error corpora.

M.1. General error corpus

A large corpus in which writing errors are annotated using a certain system.

Work packages:

- Collecting data and preparation for annotation
- Defining types of errors and their annotations
- Installing annotation labels in an annotation tool
- Annotation

Human resources:

- Data expert: 1.5 months
- Senior linguist: 2 months
- Linguist or postgraduate linguistic student: 6 months

Total: 9.5 man months

For spell and grammar checking to be adapted to different needs, texts must be collected from different groups: from children and teenagers, from dyslexics, and from those who use Icelandic as a second language. Licences must be obtained for the handling of data from minors. Corpora must be created from all the collected files, and errors detected and annotated according to a defined system. The order in which the specialised information is collected is unimportant. After preparation work, it should be evaluated as to which data is most easily accessible and the data collection prioritised accordingly.

M.2 Specialised error corpora

Corpora that contain text from particular groups. Writing errors annotated according to M.1.

Work packages:

- Error corpus containing texts from children and teenagers
- Error corpus containing texts from dyslexics
- Error corpus containing texts from people who use Icelandic as a second language

Human resources:

- Data expert: 3 months
- Linguist or postgraduate linguistic student: 6 months
- Licensing: 1 month

Total: 10 man months

Once work on the error databases is well established, statistical work must be carried out on the annotations to map the error types. Since it is simple to repeat calculations during annotation, it is feasible to install the process before the annotation is complete. In addition, high-quality test corpus from each error corpus must be created and used for each stage in the spell and grammar checking development. Certain parts of the test corpora should be closed: testing and development should be separate when testing a new version of the spell and grammar checker software.

M.3 Statistical processing and test corpora

Work packages:

- Error statistics
- Test corpora

Human resources:

- LT or statistics expert: 2 months

2.4.5.2 WORD LISTS AND LANGUAGE MODEL

Spell and grammar checker must contain a word list with valid Icelandic word forms, which should be created from the Icelandic Gigaword Corpus and lexical data, principally the Database of Modern Icelandic Inflection (DMII) (see 2.5.1.8). Spell and grammar checking may also need other word lists, for example to analyse compound words. When the general error corpus is ready, common misspellings can be collected into a type of error dictionary (for example, "eitthver" is always a misspelling of "einhver"). To enable the use of statistical methods in corrections, language models must be built with the aid of The Icelandic Gigaword Corpus.

M.4 Word lists and Language Models

Work packages:

- Assembling a word list with the aid of lexical data (DMII), the error corpus and The Icelandic Gigaword Corpus
- Training language models on The Icelandic Gigaword Corpus

Human resources:

- LT expert: 3 months

2.4.5.3 METHODOLOGY AND SOFTWARE ARCHITECTURE

Unlike the other core projects, no open-source software for the development of spell and grammar checking exists and, therefore, a great deal of work will go into developing the basic software.

The basic methods for the building of a spell and grammar checking system must be chosen; it is probable that a combination of rules and statistical methods will give the best results. A suitable environment must also be chosen (programming languages, etc), design decisions made, connections to support tools established, and issues important for the technology transfer identified. It is essential that an experienced software expert is involved in this work.

M.5 Methodology and software architecture

Choosing methods and choosing and installing a development environment

Work packages:

- Defining the main methods that will be used in spell and grammar checking
- Choosing the development environment, designing software, connecting to existing support tools, defining missing support tools; taking items related to technology transfer into account as much as possible during this stage

Human resources:

- Software and LT experts: 2 months

2.4.5.4 THE DEVELOPMENT OF SPELL AND GRAMMAR CHECKING FOR ICELANDIC

A systematic development, which aims at correcting the most common errors, can begin once the error corpus has been analysed. The development of an isolated word error checker can, however, start as soon as the development environment has been installed (M.5). The first version of the spell and grammar checking system must detect and correct non-word errors. Unknown words, for example proper nouns, compound words, or foreign words, must be analysed before they are tagged as errors. When an error is detected, the most likely correction must be found. Once this has been completed, the first version of the spell and grammar checking system should be released.

M.6 Spell checking for isolated word errors

Work packages:

- A simple spell checker for non-word errors that tags unknown words and offers probable corrections
- Further analysis of unknown words: compound words, proper nouns, etc
- Installing a web interface and open-source access

Human resources:

- An LT expert, an LT expert with the appropriate programming knowledge, or a specialised programmer: 12 months

When the results of the error analysis are clear (M.2), the spell and grammar checking system's next objectives must be defined. Particular error types must be chosen for the next version of the program to correct. When the first work package is complete, quality will be assessed using the test corpus, and development will continue: particular and measurable objectives will be defined, software extended and, finally, quality evaluated. In evaluating each iteration, care must be taken to ensure that each extension is an improvement and does not have a negative effect on previous iterations.

Possible objectives for each iteration should be defined and prioritised only after the error corpus analysis is complete. The following are examples of possible goals:

- **The program** uses confusion sets to recognise common errors that are easily corrected (leiti – leyti; list – lyst etc), simple context-sensitive corrections
- **The program** understands particular prepositions and errors in relation to them, for example leita að/*af, víst/fyrst að
- **The program** analyses particular grammatical rules/detects grammatical inconsistency between words that are in close proximity: Telur að mergæxli séu *algengir (Thinks myeloma is common); mörgum vantar skriffæri (many people need writing appliances)

In the latter stages of this development, the program should be able to detect errors that need more context: Guðni segir í samtali við Vísi að undirbúningur málþingsins *hófst í byrjun árs (In conversation with Vísir, Guðni says the symposium's preparation started at the beginning of the year).

It is important to realise that, while the spell and grammar checker is developed in stages over a long period, it can be useful from its first iteration. The software cannot, however, correct all errors, or badly constructed or illogical sentences, such as, for example: Eigum við í forræðishyggjunni sé að gefa okkur að þær sem mæta ekki séu að mæta út af trassaskap? Ég gleymdi næstum að segja að atvinulausir og lágmarkslauninn eru svo há (Shall we in our prescriptivism be to assume that those who do not show up are showing up because of negligence? I almost forgot to say that unemployed and the minimum wage is so high). In such cases, a good analytical program should be able to detect errors, even when it is unable to suggest corrections. The objective should be to make significant improvement in the recall of error detection by current spell checkers, without this being at the expense of precision. At the end of each iteration, the open-source access should be updated, together with documentation and instruction.

M.7 Systematic development of spell and grammar checking

Work packages:

- Developing general spell and grammar checking software for Icelandic in an iterative manner. Each iteration will take around six months, focusing on improving the software based on prioritised issues defined with respect to error analysis (M.2).

Human resources:

- LT expert, LT expert with appropriate programming knowledge, or a specialised programmer: 84 months

2.4.5.5 ACCESS AND ADAPTATION

M.8 Spell and grammar checking in smart devices

It is proposed that the speech recognition project should collaborate on the development of keyboards for smart devices with the aim of enabling keyboards to contain a spell and grammar checking function, together with a speech recognition button, and to be able to use autocompletion to predict which word the user is writing or will write next.

Work packages:

- Icelandic keyboard with spell and grammar checking and autocompletion for Android-operating system
- Icelandic keyboard with spell and grammar checking and autocompletion for iOS-operating system
- Icelandic keyboard with spell and grammar checking and autocompletion for Windows Phone-operating system

Human resources:

- Programmer: 12 months
- LT expert: 4 months

Total: 16 man months

M.9 Spell and grammar checking in word-processing systems

The spell and grammar checker must be connected to popular word-processing software and operating systems, most of which have standardised methods of connecting to spell and grammar checking. It is important to adhere to those standards to enable all updates to be implemented without problems.

Work packages:

- Connecting to MS Office
- Connecting to Mac OS
- Connecting to software, including InDesign, browsers and other systems as requested

Human resources:

- Programmer: 12 months

To use the spell and grammar checker as a part of LT software, it will be necessary to adjust programming interfaces and output. The spell and grammar checker must be able to receive information from other software, process it and give explicit results, such as an automatic correction, or deliver information to the next part of the software. Since each software package has its own requirements, the primary role of the spell and grammar checking team is to prepare the system so that connections and changes are as easy as possible. This may involve special projects that will be in the domain of technology transfer, but which will need the participation of the spell and grammar checking team.

M.10 Adaptation to LT software

Work packages:

- Adapting and implementing a version of the spell and grammar checker that can be used in other LT software

Human resources:

- LT expert, programmer: 6 months

M.11 Error models for Optical Character Recognition (OCR)

Errors that occur in OCR-read texts are different from typographical errors. Typographical, and other writing errors are of course also a part of OCR-read text, but the errors as a result of OCR require a different error model from those in typed text.

Work packages:

- Designing an error model for OCR

Human resources:

- LT expert: 6 months

2.4.5.6 SUPPORT TOOLS FOR SPELL AND GRAMMAR CHECKING

The development of a spell and grammar checking system must be carried out in close co-operation with the development of support tools. The methodology, and corresponding tagger and parser all must be defined. The parser is an elemental tool in spell and grammar checking development: either a parser such as IceParser will be developed further, or a new parser based on dependency grammar will be developed. Close collaboration is necessary with teams working on support tools, such as tokeniser, tagger, hyphenation tool, and named-entity recogniser.

M.12 Parser and other support tools for spell and grammar checking

Work packages:

- Continuing development and adaptation of a parser, or the development of a new parser, in line with the chosen methodology for language checking
- Collaborating with other support-tool teams on language checking's requirements

Human resources:

- LT expert: 12 months

For the first part of the project on the development of error search in context-sensitive writing, it is likely that confusion sets will be used. A list of often-confused words must be established (such as leiti – leyti) and, for example, a Winnow-classifier trained to analyse the context of each word form[2]. Spell and grammar checking, and writing-enhancement software are developed in an iteration process and each iteration creates a demand for more complex semantic analysis. Examples of semantic analysis projects that are required for spell and grammar checking and writing enhancement include, phrase analysis, synonym analysis, content analysis, context analysis, and vocabulary analysis.

M.13 Semantic analysis for spell and grammar checking

Work packages:

- Preparing tools and data for correction with the aid of confusion sets, for example with a Winnow-classifier
- Developing and adjusting other necessary semantic analysis tools for spell and grammar checking and writing enhancement

Human resources:

- LT expert: 24 months

2.4.5.7 SPELL AND GRAMMAR CHECKING WITH NEURAL NETWORKS

M.14 Spell and grammar checking with deep neural networks

The development of neural network language checking should run concurrently with the development of a traditional spell and grammar checking system, for at least the first two years. After that, it should be evaluated if, and how, its development should be continued.

Work packages:

- Developing neural network-based spell and grammar checking

Human resources:

- Deep neural network experts: 24 months

[2] Two projects that have revolved around such analysis for Icelandic are: Anton Karl Ingason et al, 2009 and Jón Friðrik Daðason, 2012. See bibliography for further details.

2.4.6 TECHNOLOGY TRANSFER

From its initial version, the spell and grammar checker will be open-source and accessible for specialised development. The knowledge learned during the project will enable the development of specialised spell and grammar checking systems and other related software. In each case, it will be necessary to evaluate whether the development of specialised spell and grammar checkers has a business value (specialised writing enhancement for individual companies or disciplines) or should be open and accessible to everyone (specialised writing enhancement for dyslexics). It is assumed that after the second year of the LT project, spell and grammar checking should be sufficiently developed for technology transfer to take place.

Technology transfer occurs in two different ways: one, as specialised spell and grammar checking system for particular user groups and, two, as a connection to LT software.

2.4.6.1 SPECIALISED SPELL AND GRAMMAR CHECKING

The development of general spell and grammar checking aims at assisting the average user – an adult who has a reasonable grasp of written Icelandic. The project will also communicate knowledge of particular writing problems that are faced by certain groups: schoolchildren and teenagers, dyslexics and people who use Icelandic as a second language. The basic general spell and grammar checking software will be adaptable for these groups by emphasising other types of errors, and by showing rules and instructions.

2.4.6.2 ICELANDIC LANGUAGE LEARNING SOFTWARE

The technology transfer that takes place in the core project is limited to connecting language checking to word-processing systems. A good spell and grammar checking tool also opens-up possibilities for the development of software that relies on such analysis. Programs that teach spelling and grammar must be able to detect errors and instruct students; programs that help people to learn Icelandic may become better and more flexible if the spell and grammar checking system can analyse the student's writing and offer advice.

2.4.6.3 WRITING ENHANCEMENT SOFTWARE

Software that includes writing enhancement also needs a spell and grammar checking core. Writing enhancement can analyse more than general writing errors. It can review text and check, for example, whether there is much repetition – and suggest replacement words or phrases; whether the style is appropriate, for example, if it is formal enough; whether the use of terms is appropriate according to any applicable standards, etc. In large institutions and companies, it can be important to be consistent in word usage and writing style. In such cases, a specialised writing-enhancement tool can save time and ensure that vital information is not lost. People who work on texts for publishing, in publishing houses, advertising agencies or news sources, for example, are another target group for writing-enhancement software that can help to ensure the quality of text.

2.4.6.4 SPELL AND GRAMMAR CHECKING IN LT SOFTWARE

In the core project, a spell and grammar checking module will be created, which can be adapted and connected to more complex LT software. While spell and grammar checking might not be the core purpose of such a larger system, it can considerably improve its function. This includes search engines that can detect errors in search queries, and improved search in OCR-read texts. Further, spell and grammar checking can, for example, analyse and correct speech recognisers' output and correct text for input to speech synthesisers and machine translation systems.

2.5 LANGUAGE RESOURCES

In this section we discuss language resources, that is, data and support tools for language technology.

Data for language technology is divided into text collections and corpora, lexical data, and speech data. These data are necessary for developing and testing LT software, and often lexical data need to be connected directly to software.

2.5.1 TEXT RESOURCES

Text resources that are useful in language technology can be text corpora, parallel texts, language descriptions, or lexical and dictionary data. Language descriptions and lexical data are data like dictionaries and semantic databases, terminology dictionaries, morphological databases, pronunciation descriptions, and other data that expound word meaning or use. Corpora are collections of texts stored digitally in a standardised format. For the texts to be as useful as possible in language research and in building language models, they are analysed in a number of ways.

Parallel texts are normally in two languages, one the translation of the other, and sentences or paragraphs with the same meaning side by side. A collection of parallel texts is the prerequisite for the development of machine translation systems.

A corpus is said to be annotated when each word form has been analysed and tagged to show the word class and grammatical attributes, such as case, number, gender of nominal words and person, number and tense of verbs. In addition, each word form has a lemma, for example the singular nominative for nouns, and the infinitive of verbs.

Corpora in a tree bank have been syntactically parsed and the sentence parts tagged to show syntactic structure.

Each text in a corpus has metadata that expounds the text, its origin, what type of text it is, who the author is, and other information that might be useful.

Text corpora are important for all LT projects that are based on statistical methods. They make it possible to build the language models that are necessary to compute probabilities, but such computations are the foundation of statistical LT methods. As a general rule, the larger the corpora, the better the language models that are built on them. It is, however, also important to bear other factors in mind, including data quality, the type of texts in the corpus and how they have been processed.

2.5.1.1 AVAILABLE TEXT CORPORA

There are three Icelandic corpora accessible with well-defined licences: Icelandic Frequency Dictionary, The Saga Corpus and Tagged Icelandic Corpus (MÍM).

Icelandic Frequency Dictionary

The Icelandic Frequency Dictionary (IFD), published in 1991, presents the results of extensive research on modern Icelandic that targeted the frequency of words and grammatical features of various kinds of text.

A special text collection, which contains fragments of 100 texts that were published between 1980 and 1989, each of around 5,000 words, was created for the making of the book; as a result, the corpus now contains approx. 500,000 words. The texts were tagged and lemmatised and have been made openly accessible for search in that form. Most can be downloaded for use in language research and LT projects, although this does not apply to all the texts in the Icelandic Frequency Dictionary because translated texts are not in this category. The texts are accessible through a special licence.

The Saga Corpus

This corpus holds the electronic texts of the Old Icelandic Family Sagas, Sturlunga Saga, Sagas of the Kings of Norway (Heimskringla) and the Book of Settlement (Landnámabók). They have been converted into modern Icelandic spelling. Several inflectional endings were also changed to modern Icelandic form. The texts can be searched and downloaded for use in language research and LT projects. They are tagged and lemmatised. The Saga Corpus is licensed under a CC BY 3.0-licence.

Tagged Icelandic Corpus

Work on the Tagged Icelandic Corpus, which contains text from 2000 to 2010, was completed in 2012. It was planned to hold around 25 million words from various texts, giving as clear a picture as possible of contemporary written Icelandic in this period. The corpus has approx. 25 million words, from a variety of sources, stored electronically in a standardised format. The words have been grammatically analysed and each text is accompanied by metadata and bibliographical information about its source. The corpus is intended, and is available, for use in various language research and LT projects. It is accessible under a special licence, the MÍM-licence. The right holders to the corpus have all agreed to its use under those terms.

2.5.1.2 MIM-GOLD

MIM-GOLD contains one million words. Texts, which were tagged automatically and manually corrected, were sampled from Tagged Icelandic Corpus (MÍM); the MÍM-licence is, therefore, valid. The final version of the MIM-GOLD corpus is intended to be used as the gold standard for the training of statistical taggers.

2.5.1.3 THE ICELANDIC GIGAWORD CORPUS

In 2015, work began on the creation of a gigantic new corpus, supported by Rannís' (The Icelandic Centre for Research) Infrastructure Fund. The objective is for it to contain texts from various sources, with a total of at least one billion running words. Agreements have been reached with large rights holders, for example Iceland's largest newspaper and magazine publisher, and the bulk of the collection will come from the media. It will also contain official texts, Parliamentary speeches, and more.

The texts in the corpus are tagged and lemmatised. They will be searchable through custom-built search engines for linguistic research and downloadable for LT use. Since not all rights holders were prepared to agree on a completely open-source licence for their data, the texts will be available under two types of licence: a special licence, similar to the MÍM-licence, and an open-source CC BY 4.0-licence.

This corpus is intended to be constantly updated. Once the first version has been released, data collection needs to continue from the rights holders who have given their permission. This has a number of advantages: it enlarges the corpus, which will be useful for all LT projects that will use it; it facilitates contemporary language research; and an ever-increasing corpus would be immensely important for lexical acquisition, for example in the DMII-project (see 2.5.1.8).

Funding from the Infrastructure Fund originally made it possible to begin this project and to finish around 75% of the work needed to release the first version. To complete the first version, an estimated eight additional man months are needed, and to install and adapt the software for the corpus's use, an additional four months. To maintain The Icelandic Gigaword Corpus after its release, an approximately half-time-equivalent position will be needed, but this person could also be employed on other LT projects, particularly data projects.

G.1 The Icelandic Gigaword Corpus

Work packages:

- Finishing the collection and processing of licensed texts
- Installing research software for corpora
- Installing an n-gram viewer for all time-stamped data
- Projecting data to a standardised format for distribution

Human resources:

- Data programmer: 6 months
- Expert in linguistic data collections and corpora: 6 months to finalise the first version and then 6 months each year to maintain it

Total: 36 man months

2.5.1.4 A LARGE HISTORICAL CORPUS

It must be examined whether it is possible to create a large historical corpus that uses all the accessible material from bækur.is, timarit.is, and from older manuscripts that have been transcribed. This would enable the creation of a language model for Icelandic from as much text data as is available.

2.5.1.5 CORPUS FOR NAMED-ENTITY RECOGNISER

There is, as yet, no corpus in Icelandic that has been tailored to the training of a named-entity recogniser (see 2.5.3.7) and research on accuracy, but it is planned to create one by revise the MIM-GOLD corpus and adding the appropriate annotation.

G.2 Corpus for named-entity recogniser

Work packages:

- Building a program to annotate all units in the MIM-GOLD corpus that could be named-entity units
- Manually reviewing and correcting annotated units

Human resources:

- Programmer: 2 weeks
- Linguist: 4 months
- Linguist/student (annotating): 3 months

Total: 7.5 man months

2.5.1.6 LEIPZIG CORPORA COLLECTION

The Leipzig Corpora Collection contains an untagged Icelandic corpus, based on web texts from the internet archives of The National and University Library of Iceland. Leipzig Corpora Collection is a project of the University of Leipzig, where the corpora are created and stored.

2.5.1.7 THE ICELANDIC PARSED HISTORICAL CORPUS (ICEPAHC)

The Icelandic Parsed Historical Corpus is a collection of syntactically analysed texts from the 12th-to-21st century. The treebank contains around one million words from more than 60 texts, which span that 800-year period. The initial analysis was done automatically and then reviewed manually. The tree bank shows syntactical tags, lemmas and cases of nominal words.

The treebank was developed with a twofold purpose in mind: it is intended for LT use, but precise syntactic information is necessary in the building of LT tools, such as spelling and grammar checkers, automatic translators, etc. It is also intended for linguistic research, particularly on syntax and syntactical changes.

The treebank is open-source and free to use without restrictions and licences.

In recent years, the most common format of treebank data has been that used for Universal Dependencies treebanks. They have a compatible form, which is intended to fulfil the requirements of any language and enable the shared use of tools between languages that could otherwise be difficult or even impossible. As of 1 March 2017, 70 UD-treebanks in 50 languages were accessible through the CLARIN language resources infrastructure. Moving the Icelandic treebank to this format would considerably increase its usability and people who work with language resources would be more likely to do research on Icelandic, which would in turn support the development of Icelandic language technology.

G.3 The Icelandic Parsed Historical Corpus

Work packages:

- Transforming the Icelandic treebank to the Universal Dependencies (UD) format

Human resources:

- Treebank expert: 8 months

2.5.1.8 THE DATABASE OF MODERN ICELANDIC INFLECTION

Icelandic's complex inflection system has many special cases and exceptions. It is common to use the same ending for many categories of inflection and, in some cases, the same words have more than one possible inflection form in the same category. As a result, it is impossible to construct universal rules to work with inflection forms in LT software. The Database of Modern Icelandic Inflection (DMII) is, therefore, hugely important and it is vital to enable its use as widely as possible. The data must also be prepared to ensure its correct use.

It is essential that the DMII is licensed in such a way that it enables and encourages developers to use the data for training and developing LT tools, without setting any restrictions on the software, including deployment of the data. At the same time care has to be taken that the data from DMII is not published in a way that conflicts with the legal obligations of the Árni Magnússon Institute for Icelandic Studies, where the DMII was developed. This means e.g. that the published data should not provide contradictory information.

Therefore, it is necessary to start improving the DMII database and administration system, with the objective of enabling access to the published data through access-controlled programming interfaces, and of distributing inflection forms in a special version-controlled package. Since the package is only intended for LT work, it will be distributed in XML-format under another name, for example Icelandic Language Technology Inflection Database (ILTID). Access to the DMII programming interface will be granted for conditional use, but ILTID will be accessible to download under an open-source licence.

It is imperative to license the data in the two databases under open-source licences. To facilitate this, the data must be treated in the correct way, but a few problems have to be solved first.

It will take considerable work to prepare the data to enable it to be used for its intended function, and to facilitate work on the DMII files:

- DMII will be connected to a text corpus, for example The Icelandic Gigaword Corpus (see 2.5.1.5).
- The DMII data structure will be reviewed in light of these requirements. The potential publishing needs have to be explored, as well as the possible use of the inflection forms (according to style, semantic category, and less-well-known word forms or writing forms from, for example, spoken language, etc). This is the foundation for any output that the DMII programming interface must be able to offer.
- Words with more than one possible inflection form in one or more inflection categories must be reviewed and the use of each inflection form defined.
- A powerful editorial mode for DMII, which does not currently exist, must be created to facilitate the maintenance of, and additions to, the database.
- A powerful lexicon-acquisition tool must be installed to simplify decisions on maintenance and additions to the database.

G.4 The Database of Modern Icelandic Inflection for LT

Work packages:

- Adapting a lexicon acquisition tool to DMII
- Revising the data model
- Reviewing ambiguous inflection forms
- Creating an editorial user interface
- Building API
- Defining XML-format for LT data and projecting the data to the format

Human resources:

- Programmer: per segment: 3+1+1+4+2+1 months
- Linguist: per segment: 2+3+12+1+1+1 months

Total: 32 man months

2.5.1.9 HYPHENATION

A list of 203,964 hyphenated Icelandic words exists. The word list is based on lemmas from the Icelandic-Danish Dictionary of Sigfús Blöndal; the list was created at the Icelandic Language Institute in the middle of the 1980s.

This word list needs to be extended with data from DMII. A list of the hyphenation of all words from DMII would be extremely useful in spell and grammar checking tools, as well as for hyphenation tools in desktop publishing software. This can be done by training a hyphenation tool with the existing data and running it on a word list from DMII. A linguist would then need to review the program's suggestions.

G.5 Hyphenation

Work packages:

- Training a hyphenation tool
- Enlarging the hyphenation word list with words from DMII
- Reviewing and correcting new hyphenated words

Human resources:

- Programmer: 1 month
- Linguist: 5 months

Total: 6 man months

2.5.1.10 THE PRONUNCIATION DICTIONARY

The Pronunciation Dictionary for Icelandic was created as a part of the Hjal project in 2003. It includes 50,000 to 60,000 phonetically transcribed word forms and is accessible under a CC BY 3.0-licence. The words were transcribed using the SAMPA phonetic alphabet, and the transcriptions have also been converted to the International Phonetic Alphabet (IPA). The words in The Pronunciation Dictionary are often transcribed in more than one dialect. Information on the dialects is not currently available, but for the dictionary to be useful in speech synthesis, it must be on record. For the creation of speech synthesisers with different dialects, care must be taken that examples cover all dialects, and that there are examples of every variation that makes one dialect distinctive from another.

Information on word classes is also important, particularly in words with two different pronunciations. Examples of this are most often connected to proper names and to –ll- within words: Gellur (verb), gellur (noun); Valla (proper name), valla (noun), etc. There is, therefore, a real need to enlarge and review The Pronunciation Dictionary. The objective should be to cover all exceptions and to include enough examples of pronunciation to enable a phonetic transcription tool to be trained to transcribe the main dialects in Icelandic almost perfectly.

G.6 The Pronunciation Dictionary

Work packages:

- Building a tool to review The Pronunciation Dictionary. The tool will ensure that only the correct phonetic symbols are used and, using an older speech synthesiser, will offer a phonetically transcribed string to be played back
- –Tagging pronunciation forms with dialect and word classes
- Reviewing and improving coverage of the dictionary

Human resources:

- Linguist: per segment: 0+4+8 months
- Programmer: per segment: 3+1+2 months

Total: 18 man months

2.5.1.11 ICELANDIC WORDNET

The Icelandic wordnet describes the semantic and syntactic relations of Icelandic words and phrases. It is based on more than 200,000 phrases of various kinds, and around 100,000 compounds. Since the data comes from continuous texts, it is evidence of semantic and syntactic relations.

The data are useful in many ways. It is possible to search the wordnet for information in a similar way to that in thesauri and conceptual dictionaries, but its possibilities for use in language technology is underdeveloped. It has been used to improve the results of translation systems, but the semantic information it offers could be used in improving search engine results, information retrieval and content analysis, and tagged phrases could be used for context-sensitive correction in writing-enhancement systems. The problem with it, however, is that the data are accessible only for searching.

To enable access to the wordnet, it will be necessary to define an appropriate data format and convert the data to that format. This will require some research to ensure that the correct format is chosen. During the conversion process it would be useful to review, and report on, whether some word categories are sparse, and to work on filling any gaps. A large corpus, and the lexicon-acquisition tools mentioned in the chapter on DMII, will be useful to do that. Finally, metadata must be created, and the entire package licensed under an open licence, for example, CC BY-SA 4.0. Since new words and phrases frequently come into use and meaning sometimes shifts and changes, the database has to be constantly maintained. Once the first version of the data package has been released, it is likely that a quarter-time position would be sufficient to do that successfully.

G.7 Icelandic wordnet

Work packages:

- Choosing or defining data format
- Reviewing data and transferring it to the chosen format
- Reviewing gaps
- Adjusting a lexicon-acquisition tool to wordnet
- Filling the gaps
- Creating metadata, preparing data for release

Human resources:

- Programmer: per segment: 2+1+1+3+1+1 months
- Linguist: per segment: 2+0+2+1+6+1 months

Total: 21 man months

NB: the review and preparation of the data must be done in consultation with the author, Jón Hilmar Jónsson. He has expressed an interest in the data being managed in the way described, and access being open.

2.5.1.12 MERKOR

To distinguish it from more traditional wordnets, MerkOr has been called a Semantic database for Icelandic. Created between 2010 and 2012, it has two outcomes: first, algorithms and methods of automatically processing semantic information from text and, second, the semantic database itself, which contains about 110,000 words. It is intended for use in language technology and the database and API are accessible at https://github.com/bnika/MerkOrCore. The database contains semantic relations based on syntactic patterns, and relations and word clusters based on statistical methods. By looking at these, it is often possible to arrive at a clearer overall view of the semantic environment of words than is offered by traditional dictionaries and wordnets. It is also possible to find words that are typical for a particular subject (sports, politics, etc). This type of information is useful in deciding the topic of a text or the semantic relatedness between texts.

The MerkOr database, is ready-to-use, but for it to become truly useful it must be updated by extracting semantic relations from far more data than was accessible at the time it was built. The software also needs reviewing and updating. A connection to the Icelandic wordnet would open up the possibility of an even more powerful semantic database.

G.8 Updating MerkOr

Work packages:

- Updating software
- Running software on a large corpus, for example, the Icelandic Gigaword Corpus
- Releasing the software and semantic database

Human resources:

- LT expert: 6 months

2.5.1.13 ICEWORDNET

IceWordNet, which is the Icelandic version of Princeton Core WordNet, is accessible under a CC BY 3.0-licence. It contains nearly 5,000 Icelandic translations of words from the Princeton core list, together with Icelandic synonyms of the words. Princeton WordNet is used in search engines and software for information retrieval and a large, Icelandic version could be useful. It is probable, however, that the Icelandic wordnet (2.5.1.11), which is already comprehensive, could prove to be adequate.

2.5.1.14 DICTIONARY OF MODERN ICELANDIC

The Dictionary of modern Icelandic, created at the Árni Magnússon Institute for Icelandic Studies, uses the same basic word list as ISLEX, approx. 50,000 entries. Licensing for the dictionary has not yet been concluded.

2.5.1.15 WRITTEN LANGUAGE ARCHIVE

The Written Language Archive contains 2.6 million examples of word usage, dating from 1540 to the end of the 20th century. The spelling in the corpus is the same as in the original texts, but words are linked to their standardised forms with modern spelling where needed. This data can be used to develop software for search and information retrieval from older texts.

2.5.1.16 THE ICELANDIC TERMINOLOGY COLLECTION

The Icelandic Terminology Collection (Íðorðabankinn) comprises collections of terminology glossaries that have been created by committees, in particular by people in scholastic or specialised fields. While the collections remain the property of the respective committees, more than half of the committees have given their permission for the collections to be distributed under a CC BY-SA 3.0-licence. The vocabulary can be useful in machine translation and text classification.

2.5.1.17 ISLEX

ISLEX is a multilingual dictionary. It has modern Icelandic as its source language and the Nordic languages – Icelandic-Swedish, -Danish, -Norwegian (bokmål and nynorsk), -Faroese and -Finnish – as its target languages. The dictionary, which contains about 50,000 entries for each language pair, can be useful for machine translation. Since each country's institutions manage the licensing for its language, licensing is not simple, but the data has been distributed under a CC BY-NC-ND 3.0-licence.

2.5.1.18 TERMINOLOGY COLLECTION FROM THE TRANSLATION CENTRE OF THE MINISTRY FOR FOREIGN AFFAIRS

The Translation Centre of the Ministry for Foreign Affairs has been in operation since 1990, with the remit of translating legal documents and regulations that fall under the EEA agreement. Since the work includes a great deal of terminology the Translation Centre has built up a terminology dictionary that contains around 80,000 headwords. It can be useful in content analysis and machine translation, but the data is not accessible: it has not been released, neither under an open-source nor a restrictive licence.

2.5.1.19 THE TRANSLATION CENTRE'S TRANSLATION MEMORY

The translation work at the Translation Centre of the Ministry for Foreign Affairs has resulted in the creation of a large translation memory, which is used by translators to speed-up their work and to ensure consistency. The memory, which contains around 1.2 million pairs of sentences, is of limited use because it is not licensed under an open-source licence. The universities in Iceland and the Árni Magnússon Institute for Icelandic Studies are, however, in discussions on the use of the data for machine translation research.

2.5.1.20 OPEN SUBTITLES

The collection of subtitles on the web page, Open Subtitles, is fully accessible. Its 1.4 million pairs of sentences in English and Icelandic can be found on the website of the OPUS-project, which holds open-source parallel-text collections of European languages, that can be used to help train machine translation systems.

2.5.2 AUDIO RESOURCES

2.5.2.1 THE MÁLRÓMUR CORPUS

The Málrómur corpus is an open-source database of Icelandic speech recordings that was compiled in collaboration with Google between 2011 and 2012; 563 people participated in the project and 127,000 speech samples – 152 hours – were recorded. When the recordings were reviewed, the files in which the text did not correspond with the audio were identified. More than 108,000 of the files – approximately 135 hours of recordings – are correct. The Málrómur corpus is licensed under a CC BY 4.0-licence.

2.5.2.2 GOOGLE'S AUDIO RECORDING FOR SPEECH SYNTHESIS

Reykjavík University collected the voices of 20 participants for a speech-synthesiser project in collaboration with the Google project, Unison. The ChitChat tool was used to collect the data and around 250 sentences were recorded for each participant (45-60 minutes each of recorded material). The objective of the collection was to create a parametric speech synthesis system, in which the synthetic voice did not necessarily replicate that of any of the readers. The database will be released under an open-source CC BY 4.0-licence.

2.5.2.3 ISLEX RECORDINGS

Each headword in the ISLEX dictionary is accompanied by an audio file of its pronunciation. Roughly 49,000 words, and more than 700 phrases, were recorded. The recordings were released in full acoustic quality under a CC BY-NC-ND 3.0-licence.

2.5.2.4 PARLIAMENTARY DEBATES

These audio resources include recordings of Parliamentary debates from the winter of 2004-2005 - a little less than 21 hours in total. The recordings have been precisely transcribed, and time-stamped. They are accompanied by text files, together with information – such as age and gender – on the speakers. The data is accessible under a CC BY 3.0-licence.

2.5.2.5 HJAL

The Hjal-recordings, which were carried out during the development of an isolated word recogniser in 2003, were made by telephone from 883 speakers. Each speaker produced around 47 recorded files – from a single word to a full sentence – resulting in a total of more than 40,000 audio files. The total length of the files, which were recorded at 8 kHz, is around 52 hours although, since the recordings contain some silences, actual speech is probably around 40 hours. The Hjal-recordings are accessible under a CC BY 3.0-licence.

2.5.2.6 ÍSTAL, DATABASE OF SPOKEN ICELANDIC

ÍsTal, a database of spoken Icelandic compiled between 1999 and 2002, contains around 20 recorded hours of spontaneous and natural conversations, which have been analysed and tagged. The transcribed conversations are annotated to indicate speakers, insertions, interruptions, hesitations, etc. The data have, however, not been made accessible, for a number of reasons: the participants' permission was not sought at the time, and the recordings contain some delicate material that would need to be deleted. If ÍsTal were to be made accessible, considerable work would need to go into reviewing the material and acquiring the necessary permissions from the participants.

G.9 ÍsTal

Work packages:

- Contacting participants and acquiring written permissions
- Reviewing data and preparing it for release

Human resources:

- Linguist: per segment: 2+4 months

2.5.2.7 OTHER SMALLER DATABASES

There are a few smaller speech databases, including the Jensson-corpus, Þór-corpus and RÚV-corpus at www.málföng.is. Since these smaller collections contain only a few hours of data, and licensing work has not been concluded, they are not covered in this document.

2.5.3 SUPPORT TOOLS

Support tools either assist in the processing corpora and other annotated data from raw data or perform basic language analysis and processing. They normally deal with a single task, for example recognising word classes and inflection forms, and often form an analysis pipeline that delivers results for more complicated software. Since any errors in the pipeline are perpetuated in the latter stages of processing, the quality of support tools is vitally important and has a significant influence on the quality of the final LT solutions.

There are two software packages that include support tools for Icelandic: IceNLP and Greynir. IceNLP contains modules for the analysis of Icelandic text: a tokeniser; a part-of-speech tagger; a lemmatiser; a shallow parser; and a named-entity recogniser. Since its release, modules from IceNLP have probably been used in most LT projects in Iceland. It is written in Java and JFlex and the code is accessible on the software development platform, GitHub, under GNU LGPL v3-licence (see 3.1.1); https://github.com/hrafnl/icenlp. The package in binary form can be obtained at Sourceforge, under IceNLPCore (https://sourceforge.net/projects/icenlp/).

Greynir, a recently developed natural language parser and analyser, analyses Icelandic sentences and offers various alternatives in text analysis (https://greynir.is/about). It collects text from news media, analyses the sentences and retrieves information on people, related texts, etc. Greynir's infrastructure is described in the sub-chapters on tokenisers and parsers.

Greynir is accessible on GitHub: https://github.com/vthorsteinsson/greynir. It is necessary to download the DMII with the appropriate licence (see 2.5.1.8), to run it independently. Primarily written in Python, it is licensed under a GNU GPL v3-licence.

To what extent certain parts of IceNLP and Greynir can be used for building open-source support tools within the LT project has to be evaluated and based on the quality of individual units, and the requirements of other LT projects.

The following support tools must be developed, or developed further, during the project.

2.5.3.1 ANNOTATION TOOL

Tools for manually annotating text are important in a number of areas of work with data. Texts often need manual annotation to some, or all, extent when they are being pre-processed for LT tools. Texts can be annotated with grammatical and semantic information; individual aspects, such as personal names, place names or dates; writing errors, etc. Annotation, manual or automatic, is carried out to enable other software to read information from standardised labels.

The objective in the manual annotation of data is usually to create enough data to train software so that the annotation process eventually becomes automatic. It can also be important in testing the quality of software intended to perform the same annotations automatically.

Annotations of proper names, an example:

> GGuðni Th. Jóhannesson **[PERSONAL NAME]**, the President of Iceland **[LOCATION]**, lightly rebuked Vladimir Putin **[PERSONAL NAME]**, the President of Russia **[LOCATION]**, at the Arctic Forum **[EVENT]** conference in Arkhangelsk **[LOCATION]** in Russia **[LOCATION]** yesterday.

If a tool is able to annotate texts with the appropriate information, it speeds-up the process if annotation is run automatically and then manually corrected. It is important to be able easily to convert the output of the automatic annotation tool to the format in which the tool for manual annotation works. Similarly, it is important that the manual annotation output is easy to use as input for the LT tools and, finally, it is essential that the annotation tool is easy to use, robust and efficient. General computer knowledge should be adequate for manual text annotation.

To date, no particular tool has been used for manual annotations in creating Icelandic corpora. Some existing open-source tools are usable, but due consideration must go into choosing which is the most appropriate, with regard to the aforementioned factors. It would be best to establish an all-round environment in which texts could be annotated with all the information needed by different LT tools. The definition of new annotation label sets for new projects must be possible, including relationships between annotated units. Initially this could call for more work, but in the long run it will simplify development.

- GATE https://gate.ac.uk/family/developer.html
- brat http://brat.nlplab.org/index.html
- Flat https://github.com/proycon/flat. A web environment based on FoLiA annotation-form, which in turn is based on XML: http://proycon.github.io/folia/

I.1 Choosing and adjusting a manual tagging tool

Work packages:

- Defining the requirements, choosing open-source software
- Implementing, installing and adjusting, instructions for annotators

Human resources:

- Data expert and/or LT expert, programmer: 3 months

2.5.3.2 LEXICON ACQUISITION TOOL

When constructing and maintaining databases, such as DMII and the Icelandic wordnet, there must be an overview of gaps in the database, information given in examples, etc. Through structured lexical acquisition from large corpora, in which unknown word forms – together with statistics on known and unknown words – are collected, the editors of the databases can find gaps in their data, monitor when new words gain a foothold, and see in what context the words are used. A lexicon acquisition tool can also be useful in composing dictionaries and in research on modern language.

I.2 Lexicon acquisition tool

Work packages:

- Building a flexible lexicon acquisition tool that can use The Icelandic Gigaword Corpus and other sources that are in the same format

Human resources:

- Programmer: 4 months

2.5.3.3 SENTENCE DETECTOR, TOKENISER

A basic step in all language processing is to segment text into units, normally sentences and tokens. Since any errors made at this stage will continue through the process, it is important for this to be as precise as possible.

The main challenge faced by a sentence detector is to decide whether a full-stop means the end of a sentence and whether a capital letter denotes the start of a sentence. In its simplest form, a tokeniser divides text into symbols (tokens)tokens that are separated by spaces. In most cases, however, this is not enough: various signs, symbols and numbers can form a token without being delimited by a space. A simple example are punctuation marks, which often need to be separated from the preceding word.

In this respect, Icelandic faces challenges similar to those of other languages, but a unique Icelandic tokeniser is needed: one that recognises Icelandic abbreviations, time units, dates, etc. It must also be adjustable, since its output must be adaptable to different projects. For example, the symbols # and @ have a particular meaning on Twitter and a flexible tokeniser could see them as part of a token ("#keynotespeech" vs "#", "keynotespeech").

Tokenisers in IceNLP and Greynir:

IceNLP's sentence detector divides texts into sentences according to so-called Segmentation Rules eXchange (SRX). The general rule is that full-stops, question marks and exclamation marks denote the end of a sentence. SRX-rules deal with exceptions from that rule. Most of the rules contain abbreviations, which dictate that a full-stop in an abbreviation should not be treated in the same way as a full-stop at the end of a sentence. There are also rules for dates, time and other ordinal numbers. However, the sentence detector does not always identify the end of a sentence correctly: In spite of flooring the car at ca. [end of sentence] 100 km an hour …

The tokeniser in IceNLP contains the sentence detector and further divides the sentences into tokens. It is possible to choose from different settings, for example whether the abbreviations are separated, and whether or not the tokeniser should divide in a strict manner; for example: delta§(4) or delta § (4). Despite these settings, tokenisation is not predictable enough – it is not possible to rely on the results given by the chosen setting.

Greynir is primarily intended as an integral LT tool, installed as a web interface. There is no module or a programming interface within Greynir that offers text tokenisation only, without language analysis. The sentence detector processes the results for a tokeniser that, for example, marks tokens as punctuation marks and, possibly, the beginning or end of a sentence, if applicable. This differs from the traditional way of first dividing text into sentences, and then letting the tokeniser process each sentence. Greynir delivers results on sentence division that are similar to those of IceNLP: by interpreting too many full-stops as sentence endings it puts sentence boundaries into the middle of sentences.

Greynir's tokeniser does not offer different settings. It separates strings strictly into different tokens on punctuation marks or symbols, although it does recognise full-stops in ordinal numbers – "4. september" – and in large numbers – "40.000" – and does not separate them. While the output is quite predictable, it is also inflexible.

The tokeniser in IceNLP is a ready-made tool that has already been used in Icelandic LT projects. Greynir's tokeniser does not exist as an independent tool or interface. The holder of the rights to Greynir has granted permission for the development of an independent module based on its tokeniser, licenced under the Apache 2.0-licence that corresponds to the LT project.

Both tokenisers need further development. At least one very good tokeniser for Icelandic text that offers flexible settings, is necessary for Icelandic LT. To focus the development, a test set that considers different input and output requirements must be prepared.

I.3 Tokeniser

Work packages:

- Preparing an independent tokeniser that can be run using different settings. Isolating and developing the tokeniser from Greynir could be the best solution
- Preparing a test set for tokenisers that takes different settings into account

Human resources:

- LT expert (programmer): 6 months

2.5.3.4 MÁLFRÆÐILEGUR MARKARI (E. PART-OF-SPEECH TAGGER)

The first part-of-speech tagger for Icelandic was built by Stefán Briem in preparation for the Icelandic Frequency Dictionary, which was published in 1991. While the tagger was never released, it was used to tag the text in the book before the tags were reviewed manually. Between 2001 and 2003, a considerable number of taggers were trained on the tagged texts from the Icelandic Frequency Dictionary. TnT-tagger gave the best results. CombiTagger, a package that used five to six different taggers was created, and the tag with the most votes was chosen. In this way, up to 93.41% accuracy was achieved. IceTagger is part of the IceNLP-package; during tests in 2009 it recorded an average accuracy of 91.59%. The best results reported up to date are 93.84% accuracy with IceStagger.

Little research has been carried out into the tagging of Icelandic texts since the building of IceStagger in 2012.

A project must be undertaken to research how much accuracy may be achieved with the latest technology, for example neural networks such as LSTMs. It would also be useful to explore how much of what is not detected by an automatic tagger is actually impossible to detect because of, for example, ambiguity in the text.

Precise tagging of text is important in most LT projects because it is the computer's main aid in understanding a language's underlying systems. In a large number of cases, improved tagging accuracy can result in an improvement in the overall quality of LT software. It is important to keep improving tagging methods for Icelandic, and to analyse how far it is possible to progress, both theoretically (how similar the results would be if two capable linguists were to tag the same text) and technically.

I.4 Part-of-speech tagger

Work packages:

- Researching how much precision can be reached in tagging Icelandic texts, using the existing tag sets
- Experimenting with the latest versions of taggers with the aim of reaching more than 94% accuracy

Human resources:

- Linguist, LT expert: 18 months

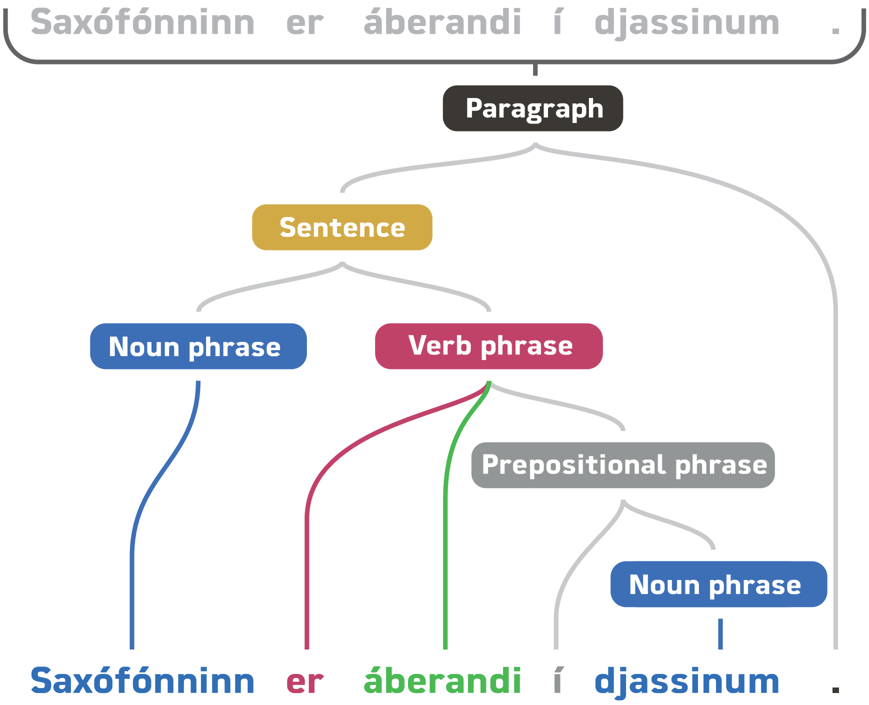
2.5.3.5 PARSER

A sentence analyser, or parser, uses tagging output to analyse sentence structure based on the predefined syntax. Greynir and IceNLP can both carry out sentence analysis, but direct comparison is impossible since they use different methods that deliver different results. There no "gold standard" to test the quality of parsing, either.

IceParser in IceNLP analyses tagged texts (from IceTagger) through phrase structure rules. Sentences are not fully analysed, but are shallow-parsed or chunked, i.e. each part of a sentence is analysed independently without an exhaustive analysis of the sentence as a whole. IceParser is able to identify syntactical roles: subject, object, etc, as well as the corresponding predicate.

In Hrafn Loftsson's and Eiríkur Rögnvaldsson's 2007 testing of the IceParser, the shallow parsing reached an F-value of 96.7% when the input tagging was from a manual tagged test, but 91.9% if IceTagger was used. Parsing with syntactical roles reached 84.3% F-value on manually tagged input, but 75.3% with input from IceTagger.

The Greynir parser fully parses sentences according to the context-free grammar that was written for the software. Greynir generates all possible parses for a sentence and chooses the most probable result. For example, Greynir finds 167 tree analyses in Saxófónninn er áberandi í djassinum (The saxophone is prominent in jazz music). There are thousands of options for more complex sentences. The image below shows how the most probable analysis is shown on Greynir's web interface. Since there is no available testing data for Greynir, it is impossible to evaluate its quality.


Saxófónninn er áberandi í djassinum .
Paragraph
Sentence
Noun phrase
Verb phrase
Prepositional phrase
Noun phrase
Saxófónninn er áberandi í djassinum .

IceParser and Greynir are two very different parsers for Icelandic. Shallow parsing, like the one provided by IceParser, is often sufficient, but full parsing, like the one Greynir offers, is required for other projects. IceParser would need to be quality-tested in its current form because the most recent testing was in 2007. Greynir is relatively new and its parsing has not been quality-tested. To enable the use of its parser as an individual tool, the rights holder must grant permission to its release under an open-source licence, such as Apache 2.0.

Other parsing techniques may possibly also be suitable for individual projects. The tagging and parsing following Universal Dependencies (UD) (http://universaldependencies.org/) has, for example, been gaining momentum. UD-treebanks exist for many languages, including English, Danish, Norwegian, Swedish, Finnish, Estonian and German. One of UD's objectives is to facilitate the sharing of LT tools, in which tagging, and sentence analysis are standardised between languages.

I.5 Parsers

Work packages:

- Testing the IceNLP and Greynir parsers, and evaluation for further development
- Adapting a Universal Dependencies-parser to Icelandic UD-grammar
- Collaborating with other teams on requirements for automatic parsing

Human resources:

- LT expert: 18 months

2.5.3.6 LEMMATISER

Using information such as part-of-speech tagging, inflectional information, and context, a lemmatiser searches for the dictionary form of words in a text. Two programs have been built to lemmatise Icelandic text. The first gave good results when it was tested during development, but when it was used in the Tagged Icelandic Corpus, it was prone to errors; as a result, work began on another tool that used different methods. While this work is promising, it is nowhere near completion. The new system, which uses DMII to find known inflection forms, uses rules to identify unknown words and to differentiate between ambiguous word forms. It is licensed under an Apache-licence. However, quite a lot of work remains to be done and, to enable comparison with the previous version, it must be tested for accuracy.

I.6 Lemmatiser

Work packages:

- Finishing work on the lemmatiser that is in development
- Testing its quality and comparing it to the older lemmatiser

Human resources:

- LT expert: 6 months

2.5.3.7 NAMED ENTITY RECOGNISER

The role of a named entity recogniser is to detect and categorise proper names, such as those of companies and institutions, personal names and place names, as well as numerical units, including time, quantity, price and ratio. Some work has been carried out on named entity recognition for Icelandic. IceNLP has a module that has achieved 71-79% accuracy. There is also a named entity-recogniser for proper names in Greynir, but its accuracy has not yet been tested. The most effective systems in other languages have reached 93-94% accuracy. Once the corpus for named entity recognition is ready (see 2.5.1.5), different methods must be tested, including neural networks that have produced good results in recent years.

I.7 Named Entity Recogniser

Work packages:

- Experimenting with different training methods
- Choosing a method and releasing a named entity recogniser

Human resources:

- LT expert: 6 months

2.5.3.8 SEMANTIC ANALYSIS

Semantic analysis includes many different aspects, for example: connecting, and disambiguating if necessary, words in a text to a particular meaning; anaphora resolution; identifying co-references; assessing the semantic relations between words or similarity in the meanings of words or text; and detecting semantic roles (agent, patient, etc.). It is also possible to use semantic analysis to extract semantic information from a large number of texts to create or add to semantic databases, such as wordnets.

Semantic analysis tasks are defined in 2.4.5.6, spell and grammar checking. Other urgent semantic analysis projects pertain to disambiguation, analysis of phrases, and the resolution of anaphora and co-references. The image on page 138 shows examples: Léttari hluti (lighter things) must be clarified, here it means the fun side of something but not an actual thing that is not heavy; the verb phrase unnið fyrir því (work for it) must be analysed so that verb unnið (work, win) is not interpreted as að vinna fótboltaleik (win in a football match); anaphora resolution is needed where hann (he) or honum (him) refers to Gylfi Þór Sigurðsson and he is referred to as Gylfi, Gylfi Þór, Gylfi Sigurðsson or Gylfi Þór Sigurðsson; úrvalsdeildinni (Premier League) and deildinni (league) are co-references to ensku úrvalsdeildinni (English Premier League) and lið sem berst við falldrauginn (team that is fighting relegation) means Swansea City FC.

> „**Gylfi Sigurðsson** á skilið sæti í liði ársins."
>
> Þetta er fyrsta setningin í myndbandi sem Facebook-síðan Dream Team gerði um **Gylfa Þór Sigurðsson**, leikmann **Swansea** í **ensku úrvalsdeildinni**. Dream Team er vinsæll hluti enska götublaðsins The Sun þar sem farið er aðeins yfir léttari hluti fótboltans.
>
> Í myndbandinu er bent réttilega á það að **Gylfi** er búinn að spila frábærlega fyrir **Swansea** og er stoðsendingahæstur í **úrvalseildinni**. Kevin de Bruyne hjá Manchester City er reyndar búinn að ná **honum**.
>
> Tekið er fram að **Gylfi Þór** er búinn að skora fleiri mörk í **deildinni** en Philippe Coutinho hjá Liverpool og Pedro hjá Chelsea. Þá hefur **hann** komið að fleiri mörkum en Dele Alli hjá Tottenham.
>
> Ólíkt öðrum sem verða vafalítið í liði ársins er **Gylfi Þór** að spila fyrir **lið sem berst við falldrauginn** en samt sem áður er **hann** búinn að skora eða leggja upp ríflega helming allra marka **liðsins**.
>
> „**Gylfi Þór Sigurðsson** á skilið sæti í liði ársins því **hann** hefur unnið fyrir því"
>
> (http://www.visir.is/g/2017170409191/-gylfi-sigurdsson-a-skilid-saeti-i-lidi-arsins-)

Anaphora resolution and co-references: the same colour indicates the same referent.

I.8 Semantic analysis: disambiguation, phrases, anaphora resolution and co-references

Work packages:

- Disambiguation of individual words
- Analysing compound verbs and phrases
- Developing methods for anaphora resolutions
- Developing methods to analyse co-references

Human resources:

- LT expert: 36 months

3 LICENSING AND ACCESS TO LANGUAGE RESOURCES

3. LICENSING AND ACCESS TO LANGUAGE RESOURCES

To ensure that the tools and data created within the LT program are accessible, certain parameters need to be set for each project. The objective is for all the projects to be usable for further development, whether research or business.

3.1 LICENCE

If software or other data is released without a special licence, it is not free and open-source in the sense that others are able to adapt, use, distribute or sell it. It is always, therefore, important to consider carefully what licence is applicable. For copyright and data protection purposes, it is suggested that all software created within this project is licensed under an open-source licence, such as Apache 2.0, and that all data have the most open-source licences possible.

3.1.1 COMMON SOFTWARE LICENCES

Apache License 2.0

Apache Licence 2.0 is an open-source licence that allows unlimited use of software and source code. That includes business use and the re-releasing of the code with individual alterations. The original or altered code needs to be distributed or re-released under an Apache 2.0-licence. It is not necessary, however, to release the code of any software that is developed using code or software libraries that are released under Apache 2.0.

GNU General Public Licence (GNU GPL v3)

GNU GPL v3 focuses on the "four freedoms": 1) Freedom to use the software for any purpose; 2) freedom to adjust it to one's own needs; 3) freedom to share it; and 4) freedom to share any alterations.

There are certain obligations attached to using software that has a GPLv3 licence. Since all software that uses code released under GPLv3 needs to be open-source, it is not always feasible for companies to use code under this licence.

GNU Lesser General Public Licence (GNU LGPL v3)

The principal difference between GNU GPL and GNU LGPL is that LGPL allows code to be utilised in software without demanding all code to be open source.

3.1.2 DATA LICENCES

As a rule, all data in the Language Technology Programme must have defined licences. Licences for data that is created, or developed, within the programme should be open-source and follow international standards. The objective of the LT programme is to maximise the use of the data. It is envisaged that all third-party data that is used for LT projects within the programme will be licensed under open-source licences. Where this is not possible, data will be released with as few restrictions as possible. The following points need to be considered in obtaining licences: can the data be re-released and, if so, on what terms; can it be altered; and can it be used in for-profit projects? The majority of licences demand that the origin or author of the data is included.

Creative Commons (CC) data licences are well-known. The particular advantages in using them are that it is easy for users to find out what is, and is not, allowed and under what terms. Custom-built licences, particularly those that are detailed and complicated, can be a deterrent. The most open-source type of CC licences allow anyone to obtain data and to use it in any way, provided the origin is specified; to copy and resell it; to release an altered version; or to use it for any other purpose, including research.

Some CC-licences limit their use in some way: they cannot be used for profitable purposes, they prohibit any alterations to the data, or they demand that all resulting products from the data are distributed under the same licence as the original. The ideal is to aim to use CC-licences for data, or comparable licences.

3.1.3 STANDARDS

Since different standards have been defined for different types of data – for example, dictionary and recorded data each have specific standards – it needs to be decided at the start of each project, which standards should be adhered to. It is important for the standards to be open-source and well-known in the specific field, particularly in neighbouring and European countries which might be interested in collaborating with Iceland. Software needs to adhere to the standards of a particular environment, and in choosing third-party software, care needs to be taken that it is actively maintained and widely used. All software that the project delivers as open-source software, must be thoroughly tested and documented.

3.2 ACCESS AND TRANSFER

In the preparation and planning of projects, to facilitate transfer of the technology it is important to establish communication with possible users at an early stage. APIs, programming languages and operating systems need to be taken into consideration. The software developers should strive for it to be used as widely as possible, and to recognise what needs altering/adjusting for different environments.

All projects should be delivered in a well-documented state to a central storage facility for Icelandic language technology. The facility needs to be actively managed to enable anyone who wants to use the data and/or software to have easy access to the centre at any time, for example through a web interface. Suggestions on the arrangement of storage, maintenance and access are discussed in 6.3.

4 OTHER LANGUAGE RESEARCH PROJECTS

4. OTHER LANGUAGE RESEARCH PROJECTS

The previously defined core projects form the basis for many more complex LT tools. The foundation for all "smart" communication is the analysis of speech and text (queries, comments, longer texts), and the processing of information and relations, possibly with the help of artificial intelligence, utilising text, databases, knowledge bases or other information. Information extraction; sentiment analysis, dialogue and multimedia analysis; and information retrieval, are some of the components that make the development of powerful and smart communication- and information-systems possible.

A short description of a few tools that currently stand outside the core projects is set out below. The development of other devices, such as text summarisation and document classification, should be undertaken as soon as the human resources and necessary basic tools are available.

4.1 INFORMATION EXTRACTION

Information extraction uses particular patterns to process facts from text and uses those facts for further processing and/or storage in a knowledge database. The first step in information extraction is to use a named entity recogniser to locate proper names and other units (see 2.5.3.7). Various methods can then be used to analyse the connections and relations. For example, from the news text, Guðni Th. Jóhannesson, the President of Iceland, lightly rebuked Vladimir Putin, the President of Russia, at the Arctic Forum Conference in Arkhangelsk, Russia, yesterday, the following facts can be identified:

Guðni Th. Jóhannesson is the President of Iceland

Vladimir Putin is the President of Russia

Arctic Forum is a conference

Arctic Forum was held in Arkhangelsk

Arkhangelsk is in Russia

Guðni Th. Jóhannesson and Vladimir Putin attended the Arctic Forum

It can also be important to date events. By linking yesterday to the news story's publishing date, the Arctic Forum Conference can be dated, and it can be confirmed that, at that time, Jóhannesson and Putin were presidents of their respective countries.

Greynir already contains the foundation for a named entity recogniser and information extraction. From part of a sentence such as, Ellen Calmon, Chairman of the Organisation of the Disabled, Greynir can identify Ellen Calmon's occupation and store it in the database. At the time of writing, information that has been extracted about people does not seem to be analysed further but is been stored in its entirety; if the occupation were worded differently (Chairman of ÖBÍ, Chairman of the association) new roles are created in the database. It is still unclear whether, or in what way, independent development of the named entity recogniser and information extraction might be possible in Greynir.

4.2 SENTIMENT ANALYSIS

Sentiment analysis detects whether a certain expression or text is negative or positive towards, for example, a company or a product. Sentences can be assessed as positive ("Great restaurant, the best food I´ve eaten!"); negative ("I will never shop here again, terrible service"); or neutral ("I went out for dinner last night"). To categorise such sentences automatically, it is possible either to use special sentiment word lists, in which never again, terrible, would be marked as negative, and great, best as positive; or sentiment annotated texts in which a system can learn how sentences and texts are positive/negative/neutral. An important element in sentiment analysis is the analysis of "emojis", since subjective texts often contain such symbols to large extent.

Popular opinion now has far more effect than advertising on potential customers; as a result, it is hugely important for companies to analyse online discussions. Automation will make this much more efficient and will facilitate greater coverage.

4.3 INFORMATION RETRIEVAL

The objective of information retrieval is to locate documents or web pages that are likely to contain answers to a particular query. The language technology aspect of information retrieval includes analysing text, and even audio data, but it can also include analysis of images and movies. This is an extensive discipline that revolves around, for example, the organising and analysis of data, query analysis, and ways of finding the most appropriate response as quickly as possible. The most obvious example of information retrieval in day-to-day use is search engines such as Google, that give a list of search results with the most relevant at the top.[41]

Development in this area is moving farther away from general search and towards more specialised solutions. The focus is on analysing and connecting particular information from a large amount of different forms of data, which is often part of more-complex business software. Because companies store vast amounts of data in various formats and languages, specialised information retrieval systems can help to expedite processes and increase the data's value.

4.4. QUESTION ANSWERING

Information retrieval delivers data that is likely to contain answers to user queries, but the files still need to be read. There are, however, question-answering systems (QA), that provide a simple answer to a question instead of only returning related data. Such systems need to analyse a question before searching for possible answers. The most powerful of these systems consult knowledge and ontology databases direct and use information retrieval methods, in which answers are based on the most promising results. This applies, for example, to IBM's Watson system that famously won a large quiz in 2011.

[1] Google does not, however, rely solely on traditional information retrieval methods in choosing and listing search results.

Google's search engine has integrated information retrieval and question answering. If you ask in English, Who is the president of Finland? the answer, Sauli Niinistö, is returned immediately, together with an image. The search engine understands the question and delivers the answer and references the web pages where the answer can be found. If, however, you were to ask the same question in Icelandic, Hver er forseti Finnlands? the top result would be a link to the Finland Wikipedia page. The search engine searches only for the individual words in the search box.

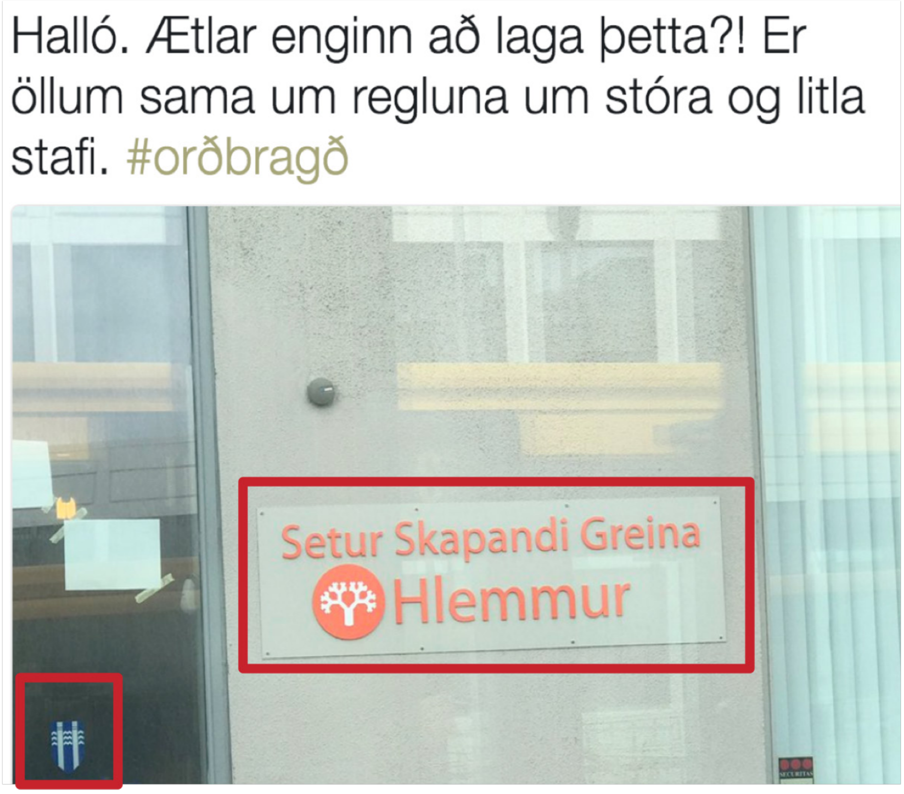
4.5 SPOKEN DIALOGUE SYSTEMS

Spoken dialogue systems enable spoken interaction between people and computers. A common example of this is telephone-answering systems, through which computer systems can handle simple tasks and queries. Spoken dialogue systems can also be found elsewhere, including in cars. The simplest, and most common, way is for the system to be designed so that the computer leads a conversation on a specific topic; open systems are, however, more complex in content and the control of the conversation. The latest development in this field are the virtual assistants Alexa (Amazon), Google Assistant, Cortana (Microsoft) and Siri (Apple), that locate information and carry out tasks, such as making phone calls, ordering products or services, choosing music, etc. These systems are also increasingly being connected to "smart homes", where lights, curtains, radiators and other household equipment is computer controlled.

Traditional spoken dialogue systems are constructed as shown in the image on page 31. The artificial intelligence is the dialogue manager: once speech recognition has been enacted, the language processing software delivers the results to the dialogue manager (dialogue analysis and control), which then delivers the information to the language generator and, finally, a speech synthesiser reads what the system wants to convey. There are a few different designs for a dialogue manager; the design of the language processing and language generation modules has to adapt to the chosen method: graph-based, frames, artificial intelligence, or statistical methods. The design of dialogue systems is closely aligned to current development in the use of deep neural networks.

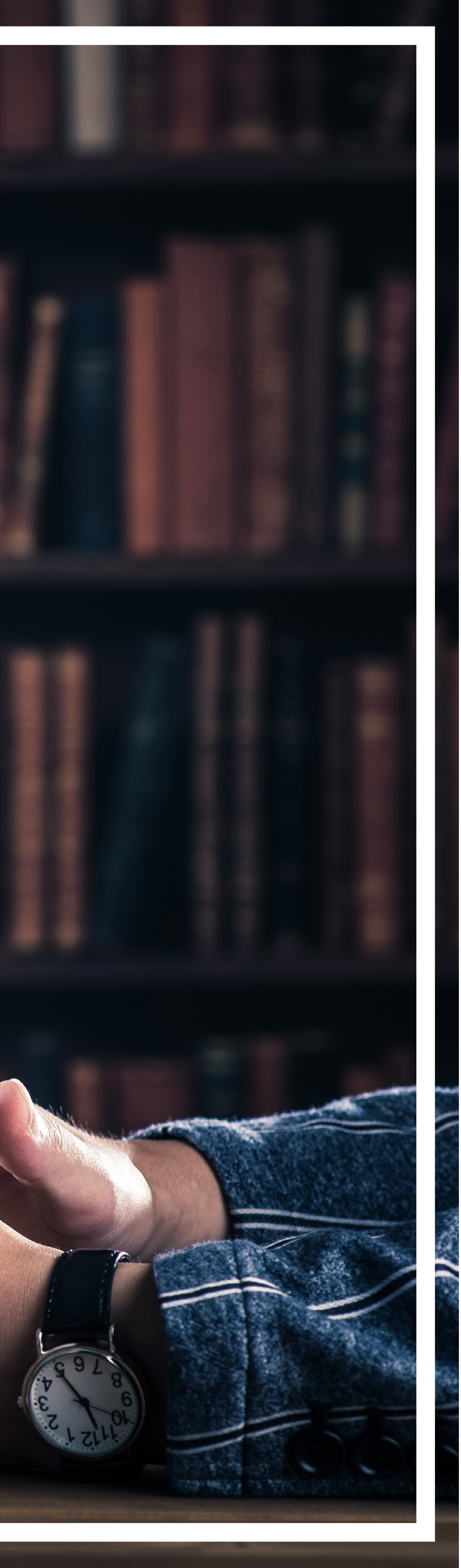
4.6 MULTIMEDIA-CONTENT ANALYSIS – AUDIO AND VISUAL

In multimedia content analysis, the content of movies, images, audio (speech) and text is analysed. It is, for example, possible to analyse speech from TV programmes, to categorise and connect it to material with similar content, to find related material from news and social media, to analyse logos and text in images, to connect to other discussions and, in general, to connect the content of different media together, even regardless of language. Since an increasing amount of material is issued in movie/video format, it is important to find ways to process content from it in the same way as from any other source. An example of an image, accompanied by a "tweet", is shown below. It is necessary to use image analysis for the text and the symbols within the image to understand the tweet.



Halló. Ætlar enginn að laga þetta?! Er öllum sama um regluna um stóra og litla stafi. #orðbragð

Hello. Is no one going to fix this? Does nobody care about the rule for capital or lowercase letters. #languageuse



5 INNOVATION IN LANGUAGE TECHNOLOGY

5. INNOVATION IN LANGUAGE TECHNOLOGY

The objective of the LT Programme for Icelandic 2018-2022 is to enable people to use Icelandic in communicating with software and information systems. It is, therefore, not sufficient to build the necessary infrastructure; it is also necessary to encourage the industry to be innovative in using language technology. Infrastructure can be used to solve the need for creating room for Icelandic in the digital world. A good and extensive LT infrastructure would create numerous business opportunities, and the potential for exciting solutions would multiply. Since it is important for Icelandic that these opportunities are exploited, it is suggested that technological development and LT innovation should be supported in the wake of the construction of the program's infrastructure.

It is also suggested that an incentive system for LT innovation should be established, with specific technology development subsidies to companies and institutions that develop LT solutions. The incentive system would be based on subsidies that meet investment in LT innovation, in a similar way that the Technology Development Fund subsidises general innovation. Since companies or institutions that embark on this type of development would be investing in the project, the resulting software would be their property and they make the decision on whether or not it is open-source. The organisation of this part of the LT programme is discussed and examples of possible projects are given. In addition, the knowledge-transfer that language technology normally brings is described, together with a brief account of opportunities for knowledge and service export.

5.1 TECHNOLOGICAL DEVELOPMENT IN LANGUAGE TECHNOLOGY

The LT programme's overall objective is to promote Icelandic language technology with the Icelandic public in as many areas as possible. It is, therefore, important to create and nurture an innovative environment to enable companies and institutions to start creating solutions and providing LT-based services. The focus will be on building a network of contacts between the companies and institutions that are involved in LT development and business, and those that are building and maintaining the infrastructure.

To encourage participation in this work, a competition fund should be established around LT technological development, and a group of professionals appointed to evaluate applications to the fund and to allocate grants. The members of the group must have sufficient knowledge of language technology and related subjects to evaluate the applications reliably, and conflicts of interest must be avoided. Since the programme has a limited timeframe, it is recommended that grants should be awarded two to four times a year, that no deadline be set for applications, and that applications are assessed at the time of reception.

To ensure that there are many promising applications to the fund, we will work to connect companies and institutions that plan to pursue technological development in language technology with those who will work on creating the infrastructure (see Chapter 6 for the organisation and management of the work.)

5.2 EXAMPLES OF TECHNOLOGICAL DEVELOPMENT PROJECTS

In the infrastructure projects described in Chapter 2, a few technological transfer projects were mentioned that were a logical continuation of the programme's infrastructure. More examples are set out below. They do not necessarily need to be connected to individual parts of the infrastructure but may use language resources and tools from more than one area of language technology. The objective in listing these ideas is to show some of the opportunities that would occur through the development of good LT infrastructure, although it comes with some provisos. First, this type of list quickly becomes obsolete: technology is constantly changing and some of these ideas are likely to have been replaced in a few years' time. Second, the project assumes that the initiative in working on these, or similar, projects will come from innovative companies in the technology industry. Thus the users of language technology products will influence their development through their interaction with those companies. The economy's initiative would support the speedy development of technology in the LT programme, as well as reacting to changes. Therefore, it is important to bear these provisos in mind.

5.2.1 AUTOMATIC READING INSTRUCTION FOR CHILDREN

Speech recognition, and its accompanying technology, offers many opportunities as an automatic learning aid. It is, for example, possible to create a training program for children to read to a computer that evaluates the quality of their reading. This could be done directly through a speech recogniser, but more precise results could be achieved through alignment if the text were known beforehand; if the child were to receive points by reading a sentence, the process of learning to read could be turned into a game.

5.2.2 COMPUTER-ASSISTED LANGUAGE LEARNING

It is possible to use language technology to help in teaching Icelandic to non-native speakers. Glossaries could be created to help people increase their vocabulary and learn word forms such as nominal word cases, and moods and voices of verbs. Reading and pronunciation may be taught in a similar way to that which is suggested in the project, Automatic reading instruction for children.

5.2.3 AUTOMATIC TELEPHONE ANSWERING

Companies and institutions can use speech synthesisers and recognition to automate call centres, and to provide information to customers and clients. Simple queries are answered right away, leaving staff free to deal with more complex requests. This not only reduces waiting time, but also increases efficiency. To attain this objective, however, infrastructure such as speech recognition and speech synthesis would need to be fully operational.

5.2.4 VOICE-CONTROLLED DEVICES AND WEBSITES

Many devices are produced with an application programming interface (API), that enables voice-control. Once the necessary infrastructure and technological development support was in place, it would be easy to support these devices with Icelandic speech recognition and speech synthesisers. Similarly, it would be possible to develop voice-control for websites, to enable material to be retrieved and listened to instead of being read on the site. It would, for example, be possible to ask for the news from web media and listen to it being read, making news consumption hands-free.

5.2.5 SEMANTIC ANALYSIS, SEMANTIC SEARCH AND INFORMATION SYSTEMS

Companies and institutions store vast amounts of valuable information in various forms, in organised databases and all types of unstructured files. This data is difficult, and sometimes impossible, to locate: the user must know where to look for it and exactly how to phrase their queries. Even if a query gives a result, the search will almost certainly bypass related data where similar material is worded differently. By using automatic semantic analysis to analyse a query, the data may be searched for meaning, rather than solely by the search string. Semantic search systems can find related information either in unstructured text or in a database, and they also contain some sort of a knowledge and/or ontological network in which important knowledge is organised. This enables semantic search systems not only to find isolated information, but also to analyse it further, to connect it and to offer valuable insight into data, often as part of a business-intelligence system. In addition to being obviously more efficient for larger companies, smart information systems can improve service and decision-making in, for example, the healthcare and justice systems; since less time is spent on collecting the information, the process is speeded up.

5.2.6 EYE-CONTROLLED WRITING FOR ICELANDIC

In spring 2017, Microsoft released GazeSpeak, a program that enables people who suffer from ALS (Amyotrophic Lateral Sclerosis) to "speak" with their eyes. Some ALS sufferers are unable to move anything but their eyes. The software enables them to choose words from lists by looking up, down or to the sides, to pick the first letter in a word or a word category, and then to move their eyes to the word they wish to say. A camera that is watching their eyes enables the computer to find the correct word quickly, using the word lists and language models. Computer vision watching eye-movement gives a language technology module the necessary information to predict the correct word. Tests reveal that users can "speak" up to 15 words a minute using this technology. There would probably not be many people in Iceland who would use such a system, but the need is there for the few who do.

5.3 KNOWLEDGE TRANSFER

The building of infrastructure and LT development for Icelandic has advantages for the language, for communications between people, for businesses, and for access to public services. The programme would create knowledge and skills in Iceland that would be useful in language technology, but as international experience has shown, specialised education and LT knowledge can also be very beneficial, through knowledge transfer, in other fields. LT experts can practise statistical modelling, signal processing or data analysis, they can train and apply deep neural networks, design intelligent systems and analyse dynamic models, to name only a few benefits. At the same time, LT projects often demand diverse expert knowledge. People with different backgrounds and knowledge are, therefore, needed to establish a strong LT industry. It is often possible to use the same technology being used in other fields of language technology, such as in bioscience, financial engineering and engineering management. Therefore, it is clear that the additional job opportunities in language technology would have a positive effect on other areas: Iceland's economy would benefit.

5.4 LANGUAGE TECHNOLOGY AS AN EXPORT COMMODITY

There is still a great demand, of various kinds, for LT knowledge in the world, although much has already been gained and technological advancement in the more common languages has been considerable. Work must be done in implementing many LT solutions, even in the larger language areas; the technology is not available to everybody; and many unexplored opportunities exist. In addition, LT solutions have yet to be developed for many language areas that have increasing technological needs. The pool of expertise that would be created by this LT programme would undoubtedly not be confined to Iceland or Icelandic: the programme's experts would be able to make their mark internationally.

Direct export of products and services based on language and related technology, would be an obvious way to use the Icelandic expert knowledge that will be created through the LT programme. An innovation company could carry out trials in the Icelandic market before competing in international markets, as is common with other Icelandic innovation companies. They would be able to use the open-source and free infrastructure for Icelandic to test their technical solutions before developing language resources for their target markets.

Participation in international co-operative projects is also an excellent means of exploiting the expert knowledge that would be created. EU or the Nordic Council projects, for example, could provide interesting opportunities for exporting Icelandic LT knowledge. In addition, international collaboration on LT development for under-resourced languages[5] would present the perfect opportunity for Icelandic experts to find new ways of improving Icelandic language technology by comparing development in Iceland to other areas, and by supporting language communities that are less able to develop language technology infrastructure.

[5] The conference Spoken Language Technologies for Under-resourced Languages, which is held every two years, revolves around making language technology accessible for as many world languages/languages of the world as possible.

6 PROGRAMME ORGANISATION

6. PROGRAMME ORGANISATION

The collaboration of institutions, universities, and industry must be well defined. It is suggested that a group of enthusiasts and interested parties should be brought together to collaborate and communicate at each stage of the LT programme. The group should be based at the LT programme centre, preferably located at Almannarómur, the privately-owned foundation, which should be reorganised and strengthened to enable it to undertake this assignment. An advisory panel should be established, with the remit of organising the work on infrastructure and supervising the teams that will work on the relevant projects. It is assumed that a large number of the projects will be carried out in institutions that have previously worked on language technology, and that have the expert knowledge in place, but the possibility of involving more organisations should be investigated where possible. A complementary fund, which has annual funding, should also be established to enable companies and institutions to put the technological solutions from the LT programme into use quickly.

These recommendations are based on: the group's analysis of similar overseas programmes, interviews with LT experts in Iceland and Estonia, where a programme that has been running since 2011 will be completed this year, interviews and correspondence with LT experts in the Germany and UK private sectors, and with machine translation experts at MT@EC. Representatives from CLARIN have also been consulted about Iceland's possible participation in that project and what the development of Icelandic language technology could gain from it. These interviews and correspondence are described at the end of this chapter, together with descriptions of LT programmes in other countries.

6.1 OVERVIEW

It is suggested that a centre for the organisation of the Language Technology programme for Icelandic 2018-2022 should be established, with the role of implementing its objectives in a service agreement with the Ministry of Education, Science and Culture.

The Almannarómur foundation was established in 2013 to work on the development of LT solutions for Icelandic, and more than 20 companies, institutions and organisations are founding members. It is recommended that Almannarómur should become the centre for the LT programme, with the objective of ensuring that the programme's projects are carried out by the experts, institutions and companies that are most qualified to do so. It would also co-ordinate projects and, where possible, would involve economic and other interested parties in them. The centre should also emphasise the importance of the co-operation of all participants in Iceland, and work in collaboration with international companies and institutions to ensure that the infrastructure and technology created from this programme will be put to use. Almannarómur would be responsible for creating an advisory panel that organises and reviews projects annually and provides professional supervision.

6.1.1 EXECUTION OF CORE PROJECTS

The infrastructure projects that are discussed in Chapter 2 of this report would undergo further cost analysis, and targets and objectives would be defined in more detail. Almannarómur would promote the projects and the advisory panel, appointed by Almannarómur, would select the applicant deemed most suitable for each one. Almannarómur may also approach specific parties for participation in certain projects, should the panel believe this is beneficial. An overall project manager would be appointed to work with Almannarómur and to monitor progress. It is recommended that a special core team, which would develop certain infrastructures during the LT programme's time period, be formed for each core project. This would help to accumulate knowledge and experience. A central team should supervise the development of each infrastructure project (speech recognition, speech synthesis, machine translation and spell and grammar checking), but the diversity and extent of language resource projects would necessitate more teams in that category. A suggestion of possible teams is made in "The LT Programme in a Nutshell" on page 22 of this report.

When looking for parties that will apply to implement particular projects, the focus should be on Iceland's leading institutions in LT development and the collection of language resources: the Árni Magnússon Institute for Icelandic Studies, Reykjavík University, and the University of Iceland, which should be consulted from the outset. Almannarómur should also endeavour to bring in other participants, such as The University of Akureyri and the Innovation Centre of Iceland, which are also sources of knowledge that could be useful in collecting language resources and developing language technology. Many of the projects would also demand the co-operation of companies and institutions that are in charge of data or activities, which could be used to collect and prepare language resources and test new technology.

These include, the Audio Library, RÚV public broadcasting company, District Courts of Iceland, 365 Media, and Creditinfo. Almannarómur could initiate contact and accelerate their participation by providing the basis for collaboration, for example expert knowledge in data licensing and technological knowledge.

6.1.2 EXECUTION OF APPLIED TECHNOLOGICAL DEVELOPMENT PROJECTS

It is recommended that innovation and LT start-up companies should be supported through a competitive development fund. It is currently not possible to estimate how large the fund should be, but it is likely that an average of 50-200 million ISK will be needed annually – less in the beginning, but more in the later stages of the programme. A 50% contribution would be requested, but the projects would be evaluated from a business point of view, similar to the Technology Development Fund, as well as from the view of bringing language technology into use for Icelandic. This way, the industry will participate in creating a fruitful environment for the development of language technology in Iceland.

To encourage companies and innovators to apply to the fund, Almannarómur should run regular promotions for the infrastructure projects that are in development, and bring together people from universities, institutions, and business, with the aim of using the infrastructure being developed to implement practical LT projects. In this way, innovators could, e.g., apply for a technology development grant with the aid of, or in collaboration with, the people developing basic software and technology. Almannarómur would be responsible for establishing an important network of contacts to foster this type of collaboration in order to ensure that the business community is aware of the potential of the developing infrastructure offers.

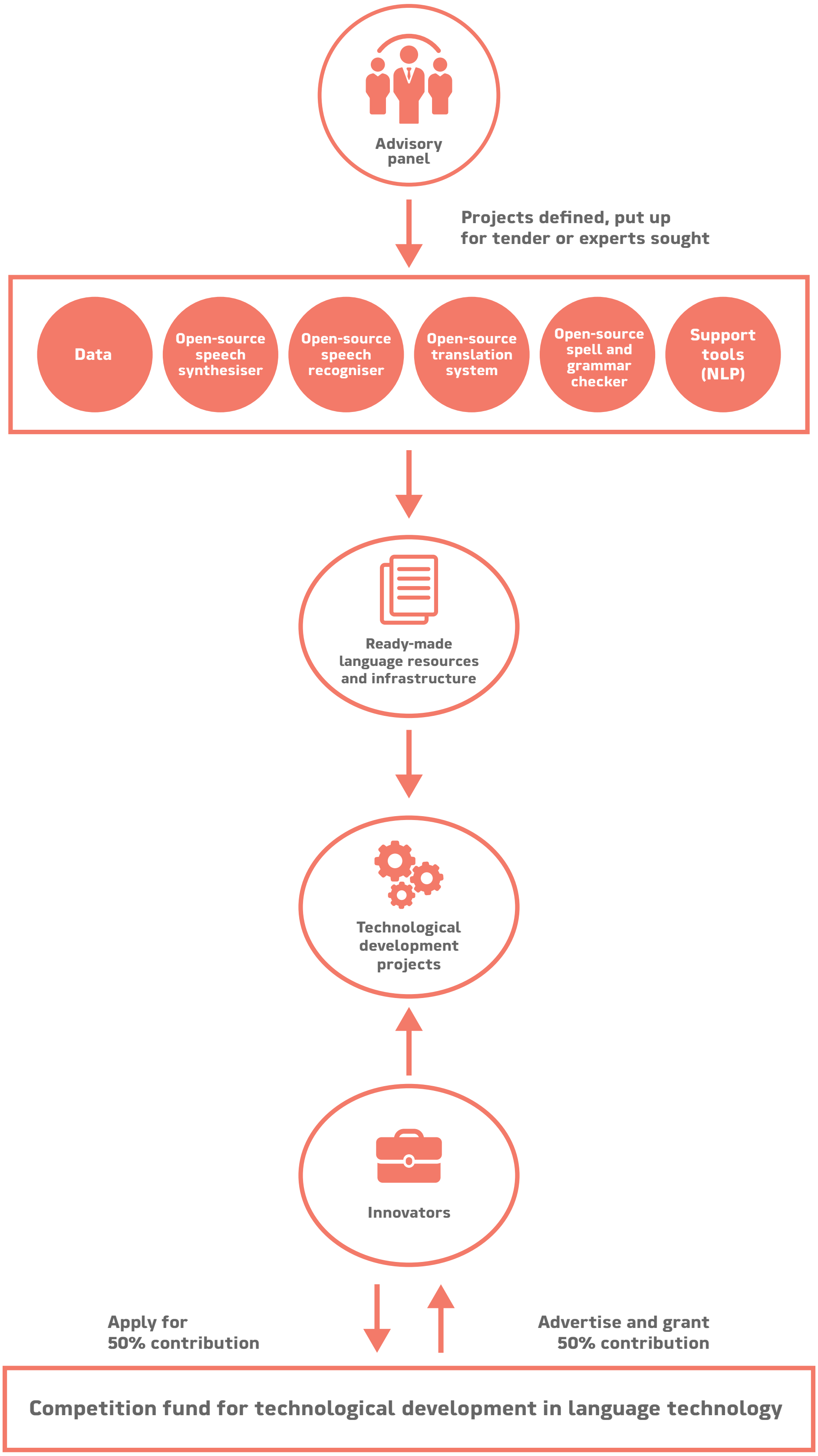
6.1.3 INTERNATIONAL COLLABORATION

It is extremely important for the LT programme to communicate and collaborate with international projects: it is vital that open-source language resources being created in the programme are made accessible to overseas designers and developers. This is particularly pertinent in the case of large organisations, such as Google, Apple, Microsoft, Amazon and Samsung, that prepare their technology with the use of virtual assistants – Google Assistant, Siri, Cortana, Alexa and Bixby. Since the potential for international software solutions is large and diverse, it is important to take advantage of all available opportunities to advance Icelandic language technology.

It is not yet clear how we can install Icelandic into international companies' technologies, but it would appear that the most feasible option would be to make Icelandic language resources and other tools open and accessible enough for them to be incorporated easily. It would also be a good idea to use the same standards of data and tools as are being used by these companies.

Participation in international projects is a further important element in making Icelandic a part of the digital world of the future. Participation in the co-operative European CLARIN project on language resources would help in creating a better foundation for promoting our resources and tools. In addition, this would ensure the preservation of our data and would give us access to expert knowledge that does not exist in Iceland. Opportunities to participate in this type of projects should be taken. In addition, the knowledge, experience and technology gained from these projects could prove to be useful for other under-resourced languages.

6. PROGRAMME ORGANISATION


Advisory panel
Projects defined, put up for tender or experts sought
Data
Open-source speech synthesiser
Open-source speech recogniser
Open-source translation system
Open-source spell and grammar checker
Support tools (NLP)
Ready-made language resources and infrastructure
Technological development projects
Innovators
Apply for 50% contribution
Advertise and grant 50% contribution
Competition fund for technological development in language technology

Implementation of infrastructure projects is confirmed and tendered Technological development develops from innovators' projects.

6.2 THE CENTRE FOR THE LT PROGRAMME

In recommending Almannarómur as the centre for the LT programme 2018-2022, and its adaption to fulfil the role, the following tasks would be scheduled:

- **Ensuring active cooperation and communication** between those who work in language technology in Iceland and those interested in participating.
- **Prioritising and organising work on the programme's infrastructure projects,** with the assistance of an advisory panel.
- **Supervising ongoing projects** and organising the promotion of project.
- **Providing assistance and advice** to organisations wishing to carry out technological development projects, based on language technology.
- **Promoting the benefits offered by language technology** to companies and institutions which could use the technology in their operations.
- **Establishing collaboration with international companies** developing LT solutions, on including Icelandic in their solutions.
- **Pursuing opportunities for multinational development** co-operation in language technology and exploring the possibilities of Icelandic participation in the projects.
- **Promoting the LT programme** and its results internationally.

It is clear that the centre would need a full-time employee and an office. The project's budget should, therefore, include grants to Almannarómur.

6.2.1 ADVISORY PANELS

For infrastructure and innovation projects, advisory panels, which include experts and other representatives of the programme's participants, would need to be appointed. The councils advise on allocating grants and supervising project distribution. To ensure that the project is not delayed, it would be important that they meet frequently and make decisions quickly.

6.2.2 LOCAL CLUSTER

In addition to the LT programme's central organising body, a wider co-operative forum must be established for all those who work in language technology: an Icelandic LT cluster, which would be the platform for development collaboration, promotional work and brainstorming. In co-operation with the universities, the cluster could also be involved in LT education.

6.2.3 ICELANDIC FOR ALL DEVICES

The software of large enterprises, including Google, Apple and Microsoft, is widely used in Iceland, but since the market is so small, these companies largely do not see any advantage in providing specific Icelandic LT solutions. It is, therefore, vital to stress the importance of good communications with overseas software companies and to strive to provide the necessary resources for the adoption of Icelandic as widely as possible. The LT programme centre must inform the people who will work on the preparation and development of the resources, that the data has to be universally useful.

6.2.4 PARTICIPATION IN INTERNATIONAL PROJECTS

The LT programme centre would need to familiarise itself with, and follow, international LT projects which could benefit Icelandic and should endeavour to participate in as many of them as possible. The centre should also promote the programme and its results internationally and follow ongoing developments so that all decision-making in and around the programme is up to date and pertinent.

6.3 MAINTENANCE AND PRESERVATION OF LANGUAGE RESOURCES

All data and software created within the LT programme should be distributed with open licences to allow for universal usage. It has to be ensured that registration, maintenance, preservation and access to the products is permanently available. The necessary infrastructure must, therefore, be maintained within institutions that are likely to survive any future economic turbulence. The Nordic countries and 15 European countries have all joined CLARIN (Common Language Resources and Technology Infrastructure), which has the primary objective for all digital language resources – Europe-wide data on language, language corpora and tools, and other linguistic resources – to be accessible through a single sign-on online environment for research in the humanities and social sciences, and for technological development. It is recommended that Iceland applies for participation in CLARIN and makes use of the infrastructure development which has already taken place.

6.3.1 CLARIN

Since CLARIN was established in 2012, it has been engaged in building an infrastructure to support the construction, registration, maintenance, preservation, usage, and sharing of linguistic data and tools for research in the humanities and social sciences. To facilitate search, and to ensure that the correct data is located for a specific search, it focuses on the precise and detailed registration of metadata. This speeds-up development and enables the location of extensive reusable data for research and development.

CLARIN uses this infrastructure to grant easy and permanent access to digital language resources (text, audio, visual) that can be used by scientists in any discipline. A network of CLARIN centres, combined with a single sign-on access for the entire scientific community in the participating countries, offers advanced tools for the examination, analysis and processing of data, wherever it is located. Software and data from different sources are standardised and accessible to all CLARIN centres.

The CLARIN centres are run by consortia that benefit from the collaboration. There is usually one consortium in each country, the members of which are generally universities, libraries, research facilities, and other institutions that work with linguistic data. A CLARIN centre in Iceland would be established at one of the member institutions.

It is possible to store, and grant access to, data in CLARIN so that it can be of the most use: it recommends that licences for data and software are as open as possible, although it is possible to use more restrictive licences if necessary.

By participating in CLARIN, Iceland's language-technology specialists would gain access to a wide range of tools and data resources to use in research and technological development. CLARIN also offers the permanent storage of new data and tools, which are all given a unique – ISLRN – number. This is not only useful in referencing databases being used in research or development, but the numbers also ensure that the same data is used when experiments are to be repeated, or if new and improved methods are developed. Participation in CLARIN would thus offer significant benefits to Icelandic language technology.

LT data and tools must be securely stored and permanently accessible, regardless of any potential changes in the technological environment, and must also be easily located by as many parties as possible, in Iceland and overseas. Since the number of language technology experts in Iceland is rather small, the knowledge communication that is emphasised by CLARIN would be extremely important for Icelandic.

While participation in CLARIN would provide great support for Iceland's LT programme it should not be dependent on the programme, which will probably run for five years, but should continue to ensure that the programme's results, as well as other data that are not necessarily related to language technology, would continue to be accessible. CLARIN has a wider use than simply for language technology: its network could, for example, be useful in the Ministry of Education, Science and Culture's current plans to preserve cultural heritage in digital form, and to ensure more effective use of the results of the work, in and outside Iceland.

The cost of participating in CLARIN is in the form of participation fees, and in running the CLARIN centre. Iceland's participation fee would be based on the cost for smaller countries and is based on a percentage of the EU countries' GNP. The fee, which for 2017 would have been €13,028, increases by 2% a year so that the amount is known for years to come.

In most countries, CLARIN centres are run in one of the participating institutions, such as universities or research institutions. See https://www.clarin.eu/content/participating-consortia.

It is probable that at least one full-time equivalent employee, whose time would be equally divided between technology management, and administration and the provision of information, would need to be added to the institution that would house Iceland's CLARIN centre. It is likely, however, that during the first 18 months or so, while the project is being launched, two employees would be needed.

In addition to employment costs, and the cost of the employees' facilities, operating costs for computer equipment and travel expenses for one employee to the CLARIN annual meeting would need to be taken into consideration. The annual cost can, therefore, be roughly estimated as:

Participation fees:	1,500,000 ISK
Computer operating costs:	around 500,000 ISK
Travel expenses:	around 500,000 ISK

In addition, to employing personnel, the institution that would house the CLARIN-centre would need sufficient funds to cover its expenses.

Total cost:	2.5 million ISK + wages

6.4 OTHER 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. As a result, LT programmes have been launched for many languages. The workgroup acquainted itself particularly with the Dutch/Flemish programme in The Netherlands, STEVIN (Essential Speech and Language Technology Resources), which ran between 2004 and 2011, a Spanish LT programme that is ongoing, and two programmes for Estonian.

The Netherlands

The STEVIN programme was launched to strengthen the position of Dutch and Flemish in language technology. Its objective was to increase general knowledge of language technology, particularly in the industry, to organise research and develop language resources to fill gaps in the LT infrastructure, and to organise the management, maintenance and distribution of language resources. There were research and development projects, demos, and education projects in the programme, including the collection of speech data from children, older people and people who use Dutch as a second language; projects related to semantic analysis, parallel texts and collections of text; special writing enhancement for dyslexic children; a search engine for law courts, with speech recognition for trials; and many others. A programme board (the STEVIN Board), which comprised representatives from participating institutions, and LT experts, was established. The board supervised the programme and made decision on grants. Nederlandse Taalunie (The Dutch Language Institution) was in charge of co-ordination and financial management, but a special programme committee, which planned the programme, ensured that its policies were enforced, advertised for, and advised the board on, grant applications. In addition, the grant applications were reviewed by an international advisory panel. A special LT programme office, which was run by two institutions, was responsible for project management.

A large conference was organised to increase awareness of language technology in the industry: LT experts and organisations were invited to promote LT solutions in a number of areas. The emphasis was on the conference being as diverse as possible, and on introducing the possibility of using LT for the media, the education system, health-care institutions, transport, tourism, administration, telecommunications and the financial sector.

Spain

The Spanish LT programme Plan de Impulso de las Tecnologías del Lenguaje is ongoing. Started in 2016, and scheduled to finish in 2020, it is designed to ensure the competitiveness of Spain and South-America, and to avoid the digital extinction of other official Spanish languages, including Catalan, Galician, Basque and Occitan. By introducing machine translation and other language technology, the objective is to expand language resources; to improve quality and access to them; to focus on technology transfer from research to the economy; and to improve the quality and efficiency of public institutions. By emphasising the visibility of language technology to companies as well as to students, the Spanish programme is not only facilitating the transfer of technology to the economy, but also encouraging talented people to specialise in language technology and related fields.

Government should be the leading participant in language technology: high-profile projects are being developed in the healthcare, judicial, and educational systems, and in tourism. The estimated total cost of the programme is €90 million, the equivalent of around 10 billion ISK – around 2 billion ISK for each language.

Estonia

The workgroup met with the representatives of the Estonian LT programmes, which have been running since 2011 and conclude this year, who shared their experience and answered questions on their work. There have been two LT programmes in Estonia: the first was the National Programme for Estonian Language Technology 2006-2010, and the second was the National Programme for Estonian Language Technology 2011-2017. The Estonian Ministry of Education and Research has been responsible for both programmes, but the universities in Tallinn and Tartu, together with the Institute of the Estonian Language (Eesti Keele Instituut), were leading their implementation. The Estonian language and environment is, in many ways, similar to that of Iceland: It has too few users for companies to see the advantage of embarking on costly development of language technology, but society is technologically advanced – people use, or want to be able to use, LT software. The people want Estonian to maintain its dominance and the language, like Icelandic, has a complex inflection system and very active word generation. Therefore, the problems faced by Estonia are similar to those that Iceland is facing.

6. PROGRAMME ORGANISATION

In the Estonian programmes, projects were carried out in speech recognition, speech synthesis, lexica and corpora, spell checking and machine translation. The infrastructure development has been quite successful and software, based on the infrastructure, is already up and running. This includes the automatic reading of audio books and subtitled TV material; a web interface that enables the submitting of audio files and receiving text from a speech recogniser by email; and an extensive collection of open-source dictionaries. An Estonian spell checker is included in MS Office, and machine translation has provided good results. The translation system can be tested through a web interface, and its results compared with Google Translate, although the objective of the machine translations is primarily to build specialised translation systems.

The core projects of the Estonian LT programme are proceeding well and are already partly being used in general software. An important lesson is, however, to be learned from Estonia: work on connecting language technology to the economy came too late and was ineffective. It was assumed that there would be little interest, and regulations on direct support to companies were also a hindrance. Nevertheless, there are at least two significant LT companies in Estonia: Tilde-Eestim and Filosoft. In the opinion of the programme's representatives, it might well be advisable to set up a specific organisation, for example a separate association with its own office and id-number, to manage the implementation of the programme rather than leaving it to an employee of a separate institution. Tartu University houses Estonia's CLARIN centre, which they said was important in managing language resources.

EPILOGUE

0101011101 arðmiði geisp griðungur 1010111 Skrúður 0101110

Róði gauskur 0101011101 glundroði 0101

101 arðmiði geisp griðungur 1010111 Skrúður

This report discusses the opportunities that lie in the development of language technology for Icelandic. It will be necessary to start the organised construction of LT infrastructure, to create knowledge clusters and to create a strong innovation environment for LT solutions. In light of the revolution that is now taking place in artificial intelligence and language technology, it is important that the LT programme for Icelandic should be implemented as quickly as possible, to enable us to participate in the development and use of our own language, Icelandic, in the technology of the future.

BIBLIOGRAPHY

ABOUT ICELANDIC IN THE DIGITAL AGE

- Eiríkur Rögnvaldsson, Kristín M. Jóhannsdóttir, Sigrún Helgadóttir & Steinþór Steingrímsson. 2012. Íslensk tunga á stafrænni öld [Icelandic in the Digital Age] Meta-Net White Paper Series. Springer

- Eiríkur Rögnvaldsson, Haraldur Bernharðsson, Sigrún Helgadóttir, Björgvin Ívar Guðbrandsson, Jóna Pálsdóttir & Sigurbjörg Jóhannesdóttir. 2012. Íslenska í tölvuheiminum [Icelandic in the Computer World]. Ministry of Education and Culture.

- Íslenska til alls [Icelandic for Everything]. Tillögur Íslenskrar málnefndar að íslenskri málstefnu [Icelandic Language Committee suggestions for a language policy for Icelandic]. 2008. Ministry of Education.

- Tungutækni [Language Engineering]. Skýrsla starfshóps [Workgroup Report]. 1999. Ministry of Education.

OTHER LT PROGRAMMES AND LT PROGRAMME REPORTS

- Language Technologies. 2013. LT2013: Status and potential of the European language technology markets. LT-Innovate.

- Liin, K., Muischnek, K., Müürisep, K., & Vider, K. 2012. Eesti keel digiajastul – The Estonian Language in the Digital Age. Meta-Net White Paper Series. Springer.

- National programme for Estonian Language Technology 2006-2010. 2007. Estonian Ministry of Education and Research.

- National programme for Estonian Language Technology 2011-2017. 2011. Estonian Ministry of Education and Research.

- Plan for the Advancement of Language Technology (The Spanish LT Program). 2015. Agenda Digital para España. Madrid, Spain.

- Rafel Rivera Pastor et al. 2017. Language equality in the digital age – Towards a Human Language Project. Scientific Foresight Unit (STOA).

- Spyns, Peter & D'Halleweyn, Elisabeth. 2012. The STEVIN Programme: Result of 5 years cross-border HLT for Dutch Policy Preparation. In Peter Spyns and Jan Odijk (Hrsg.): Essential Speech and Language Technology for Dutch. Theory and Applications of Natural Language Processing. pg. 21-39.

SCIENTIFIC SOURCES

- Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior & Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. ArXiv. org:1609.03499.

- Anton Karl Ingason, Skúli B Jóhannsson, Eiríkur Rögnvaldsson, Hrafn Loftsson & Sigrún Helgadóttir. 2009. Context-Sensitive Spelling Correction and Rich Morphology. Proceedings of the 17th Nordic Conference of Computational Linguistics, NODALIDA. pg. 231-234.

- Carlberger, J., R. Domeij, V. Kann & O. Knutsson. 2004. The Development and Performance of a Grammar Checker for Swedish: A Language Engineering Perspective.

- DePalma, Donald A., Vijayalaxmi Hegde, Hélène Pielmeier, Robert G. Stewart & Stephen Henderson. 2016. The Language Services Market: 2016. Common Sense Advisory, Boston, USA.

- Edlund, Jens, C. Tånnander & J. Gustafson. 2015. Audience response system-based assessment for analysis-by-synthesis, Proceedings of ICPhS.

- Etchegoyhen, Thierry, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard Van Loenhout, Arantza Del Pozo, Mirjam Sepesy Maucec, Anja Turner & Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

- Helfrich, Antje & Bradley Music. 2000. Design and Evaluation of Grammar Checkers in Multiple Languages. Proceedings of the 18th conference on Computational linguistics, Vol. 2, bls. 1036-1040.

- Hrafn Loftsson, Ida Kramarczyk, Sigrún Helgadóttir & Eiríkur Rögnvaldsson. 2009. „Improving the PoS Accuracy of Icelandic Text.“ In: Jokinen, Kristiina and Eckhard Bick (eds.): Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009, pp. 103- 110. NEALT Proceeding Series 4. Northern European Association for Language Technology (NEALT), Tartu University Library.

- Ingibjörg Elsa Björnsdóttir. 2016. Vélþýðingar á íslensku og Apertium-þýðingarkerfið

- [Machine Translation in Icelandic and the Apertium Translation System]. Orð og tunga 18, Reykjavik.

- Jón Friðrik Daðason. 2012. Post-Correction of Icelandic OCR Text. University of Iceland, MA-thesis.

- Jurafsky, Dan & James H. Martin: Speech and Language Processing. Draft of third version, 2017. https://web.stanford.edu/~jurafsky/slp3/ Retrieved 25.04.2017.

- Klein, G., Y. Kim, Y. Deng, J. Senellart & A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. ArXiv e-prints.

- Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin & Evan Herbst. 2007. Moses: Open source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic.

- Kukich, Karen. 1992. Techniques for automatically correcting words in text. ACM 24(4):377-439.

- Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry & Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In Proceedings of the Demonstrations at the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain.

- Strömbergsson, Sofia, C. Tånnander & J. Edlund. 2014. Ranking severity of speech errors by their phonological impact in context. Interspeech, 1568-1572.

- Taylor, P., Black, A. & Caley, R. 1998. The architecture of the Festival Speech Synthesis System. Proc. 3rd ESCA Workshop on Speech Synthesis, bls. 147-151, Jenolan Caves, Australia.

- Tihanyi, László, Csaba Oravecz. 2017. First Experiments and Results in English-Hungarian Neural Machine Translation. XIII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Hungary.

- Tånnander, Christina. 2012. An audience response system-based approach to speech synthesis evaluation. In The Fourth Swedish Language Technology Conference (SLTC 2012), pg. 74-75. Lund, Sweden.

- Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages. In Proceedings of the RANLP 2005, pg. 590-596.
- Whitelaw, C., B. Hutchinson, G. Chung et al. 2009. Using the Web for Language Independent Spellchecking and Autocorrection. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pg. 890-899.
- Wu, O. Watts & S. King. 2016. Merlin: An open source neural network speech synthesis system, in 9th ISCA Speech Synthesis Workshop (SSW9), Sunnyvale, CA, USA.
- Zhang, X., H. Kulkarni & M. Morris. 2017. Smartphone-Based Gaze Gesture Communication for People with Motor Disabilities. CHI 2017.

WEBLINKS

- http://malfong.is/
- http://malid.is/
- http://bin.arnastofnun.is/
- https://greynir.is/
- http://nlp.cs.ru.is/icenlp/
- http://puki.is/
- http://skrambi.arnastofnun.is/
- http://www.epc.de/
- http://hunspell.github.io/
- https://www.clarin.eu