Chapter 21
Language Report Icelandic

Eiríkur Rögnvaldsson

Abstract In 2019, the Icelandic Government launched a three-year Language Technology Programme for Icelandic (LTPI). Within this programme, a number of language resources and tools have been built from scratch and several pre-existing resources and tools have been enhanced and improved. This programme is now finished and the situation for Icelandic with respect to language technology has improved considerably. In spite of this, Icelandic still remains a low-resourced language compared to most official European languages.

1 The Icelandic Language

Icelandic is a North Germanic language with its roots in Old Norse. It is the only official language of Iceland apart from Icelandic Sign Language. Even though it is only spoken by around 350,000 people in Iceland and by several tens of thousands of Icelanders living abroad, it is not considered endangered according to UNESCO's Language Vitality Scales[1] or EGIDS.[2] The language community is very homogeneous, and dialectal variation is negligible.

Icelandic is a morphologically rich language; nouns, pronouns, adjectives and verbs are inflected for several grammatical features. The language is fusional, such that a single ending usually stands for more than one morphological category. Typologically, Icelandic is an SVO (subject-verb-object) language with a strong V2 rule that requires the verb to appear in the second (or first) position of the sentence. However, because of the rich inflectional system, word order is relatively free.

The Icelandic alphabet is based on the Latin alphabet with a number of additions, especially vowel symbols with an acute accent, *á é í ó ú ý Á É Í Ó Ú Ý*, and the vowel symbols *æ Æ* and *ö Ö* which are also used in a number of other languages. Furthermore, Icelandic employs two more eccentric symbols: *ð Ð* (eth, not to be

Eiríkur Rögnvaldsson
Árni Magnússon Institute for Icelandic Studies, Iceland, eirikur@hi.is

[1] https://ich.unesco.org/doc/src/00120-EN.pdf

[2] https://www.ethnosproject.org/expanded-graded-intergenerational-disruption-scale/

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_21

confused with "d with a stroke", *đ*) which is also used in Faroese, and *þ Þ* (thorn) which is not used in any other language.

Iceland has the highest percentage of internet users in Europe. In 2020, 98% of Icelandic households had internet access.[3] In the same year, 68,344 websites had .is as the top level domain.[4] Icelandic is sufficiently represented on the internet, with a number of media websites and an Icelandic Wikipedia, for instance, but most people also frequently visit news sites in English, access various types of information in English, etc. Even though Icelandic is the main language used on social media, English is also prominent.

2 Technologies and Resources for Icelandic

The Icelandic Government launched the Language Technology Programme for Icelandic (LTPI) in September 2019. The self-owned foundation Almannarómur[5] was entrusted with the role of conducting the programme. Almannarómur, in turn, commissioned the SÍM Consortium,[6] comprising members from academia, NGOs and the private sector, to carry out the research and development work in this project. Researchers, developers and LT users are well represented in the Consortium.

Most of the existing resources and tools for Icelandic are direct or indirect outputs of this programme. Almost all of these resources and tools are stored in the CLARIN-IS repository.[7] They can be downloaded for free, most of them under standard open licences, and used in any kind of application.

The Icelandic Gigaword Corpus (IGC) is a monolingual corpus comprising almost 2.7 billion tokens of different genres. Most of the texts are from 2001-2022. A few parsed corpora exist, most of them having been automatically parsed. Greynir-Corpus contains 10 million sentences from news sources which have been parsed into full constituency trees. The Icelandic Contemporary Corpus is a constituency parsed corpus built by using an Icelandic model of the Berkeley Neural Parser and containing 30 million clauses from the IGC. A number of small specialised corpora have also been developed.

There exist a number of bilingual English-Icelandic corpora. Most of them are domain-specific corpora from ELRC and are not aligned. However, a few general purpose aligned corpora exist, the most important being ParIce with 5.3 million translation units. Much larger bilingual corpora are needed, especially between Icelandic and English but also between Icelandic and other languages such as Polish.

A few audio corpora exist. The most important one is Talrómur which consists of 122,417 short audio clips of eight different speakers reading short sentences, amount-

[3] https://www.statista.com/statistics/185663/internet-usage-at-home-european-countries/

[4] https://www.isnic.is/is/tolur

[5] https://almannaromur.is/en

[6] https://icelandic-lt.gitlab.io

[7] https://repository.clarin.is

ing to 12,780 minutes in total. A large crowdsourcing project, Samrómur, is now ongoing. In May 2022, a total of 2.85 million sentences from 28,000 speakers had been recorded, 247,800 minutes in all. No video corpora have been built for Icelandic.

The Database of Modern Icelandic Inflection (DMII) is supposed to contain the inflectional paradigms of the whole vocabulary of Icelandic. The current version has a vocabulary of about 305,000 lemmas, and 6.2 million inflectional forms. The DMII Core is an extract of DMII and contains the core vocabulary of Modern Icelandic, around 58,000 entries. The monolingual Dictionary of Contemporary Icelandic has 56,000 entries and is constantly being updated. Sound files with recordings of all the headwords in the dictionary are also available.

The company Miðeind, a member of SÍM, has been developing a translation system between English and Icelandic using neural networks. Although still under development, it already gives very promising results. The pilot version is offered as a web-based service.[8] Miðeind is also developing AI models and some of them are already available, such as GreynirTranslate (mBART25 NMT), general domain IS-EN and EN-IS translation models based on a multilingual BART model.

There exist a number of tools for analysing Icelandic text. Among them are two packages that each include various tools. IceNLP is a package which contains a tokeniser, part-of-speech tagger, lemmatiser, and shallow parser. Greynir is a more recent package that can parse text into constituency trees, find lemmas, inflect noun phrases, assign part-of-speech tags and more.

A number of tools for speech processing are currently being developed within the LTPI, among them a new speech recogniser and a speech synthesiser, but these are not yet publicly available although prototypes have been publicly demonstrated.

Embla is the first voice assistant app for the Icelandic language, available both for iOS and Android. It combines a speech recogniser, a speech synthesiser and the Greynir tool which it uses to search for answers to questions that the user poses. Greynir extracts information from Icelandic text which allows natural language querying of that information and facilitates natural language understanding.

In the national AI strategy from April 2021, the importance of developing LT resources and tools for Icelandic is explicitly mentioned.[9] In the policy statement of the new Government that took office in November 2021,[10] it is explicitly stated that the strategic R&D LT programme will be prolonged throughout the current election period, until 2025.

3 Recommendations and Next Steps

Ten years ago, the status of Icelandic LT was rather poor (Rögnvaldsson et al. 2012), but the LTPI has revolutionised the situation (Rögnvaldsson 2022). The forming of

[8] https://velthyding.is

[9] https://www.stjornarradid.is/gogn/rit-og-skyrslur/stakt-rit/2021/04/29/Stefna-Islands-um-gervigreind/

[10] https://www.stjornarradid.is/library/05-Rikisstjorn/Agreement2021.pdf

the SÍM Consortium has led to a very fruitful cooperation among all stakeholders. Researchers who used to work individually on small projects now work together on implementing projects on a much bigger scale. The number of researchers and students involved in LT has multiplied and new startup companies have emerged.

The LTPI has delivered high-quality applications that hopefully contribute to the digital vitality of Icelandic. But even so, Icelandic still lacks a number of important resources now that the LTPI is finished. Among them are spoken language corpora; parallel corpora (Icelandic and other languages than English, such as Polish and the Scandinavian languages); corpora for different purposes (sentiment analysis, question answering, summarisation); annotated multimodal corpora; and term lists. Furthermore, Icelandic lacks tools for sentiment analysis, summarisation, question answering, natural language understanding and generation, dialogue management, disambiguation, text and speech translation, automatic subtitling, advanced speech synthesis (intonation, empathy) and specialised grammar checking.

In order for these resources and tools to be developed, the continuation of the LTPI must be secured. It is also of vital importance that Icelandic is compatible with products of the large international IT companies. A delegation of LT specialists led by the President of Iceland and the Minister of Culture recently visited Amazon, Apple, META and Microsoft in order to convince them to include Icelandic in their products, offering them access to all deliverables of the LTPI. A large-scale European cooperation would also be a welcome assistance in preparing Icelandic for the future.

References

Rögnvaldsson, Eiríkur (2022). *Deliverable D1.19 Report on the Icelandic Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-icelandic.pdf.

Rögnvaldsson, Eiríkur, Kristín M. Jóhannsdóttir, Sigrún Helgadóttir, and Steinþór Steingrímsson (2012). *Íslensk tunga á stafrænni öld – The Icelandic Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/icelandic.