European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1091 programme.

European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.29
Task 3

ELRC Workshop Report for Iceland

Author(s): Prof. Gauti Kristmannsson, University of Iceland

Dissemination Level: Public

Version No.: <V3.0>

Date: 2023-01-20

Contents

1 Executive Summary

Language Technology is shaping our multilingual future. It has already been transforming the way we interact with our devices and with each other, the way we shop, work and travel. It reshapes our interaction with service providers, either public or private. Programmes that automatically correct spelling errors and aid sophisticated writing, digital assistants that transform our voices to text messages on mobile phones, bots that answer our calls to the bank or to our social security organisation, systems that automatically translate from a foreign language, and much more, are already empowering our everyday lives, our businesses and our administrations. But can we fully use our own language in our digital interactions? Is our language adequately supported and ready to keep pace with the technological advancements of the AI era?

The third Icelandic European Language Resource Coordination (ELRC) workshop addressed these questions and engaged participants in a fruitful discussion on the status and prospects of Language Technology for the Icelandic language. Developers, integrators and users of Language Technology, both from the private and public sectors shared experiences, requirements and ways for transforming digital interaction in our multilingual Europe with Language Technologies. Finally, there was a discussion how language data, i.e. texts and speech, could fuel development in Artificial Intelligence.

The workshop was organised by Prof. Gauti Kristmannsson and the University of Iceland Translation Centre and took place at the university on the 8th of December 2022.

2 Workshop Agenda

The final workshop programme in English.

09:30 - 10:00	**Registration**
10:00 - 10:15	Welcome and introduction
10:15 - 10:35	**The CEF Automated Translation Platform and services** *Vilmantas Liubinas,* Directorate-General for Translation, European Commission
10:35 - 11:05	**Common European Language Data Space** *Monica PRETTI*, Data Directorate, Directorate-General for Communications Networks, Content and Technology, European Commission
11:05 - 11:30	**Almannarómur presentation** *Jóhanna Vigdís Guðmundsdóttir,* head of Almannarómur
11:30 – 11:50	**Coffee break**
11:50 – 12:30	**The potential of Language Technology and AI – where we are, where we should be heading – LT in Iceland** **Panel:** *Vilhjálmur Þorsteinsson*, CEO of Mideind ehf *Hafsteinn Einarsson*¸Assistant Professor in CS, University of Iceland *Helga Ingimundardóttir*, head of AI at Travelshift
12:30 - 13:10	**LT for Public sector panel discussion** *Kári Örlygsson*, CEO Ministry of Foreign Affairs Translation Centre *Steinþór Steingrímsson*, IT project manager, Árni Magnússon Institute
13:10 - 13:50	**LUNCH and discussions**
13:50 - 14:10	**PRINCIPLE project** *Gauti Kristmannsson,* Professor of Translation Studies, University of Iceland
14:10 - 14:30	**NLTP project** *Bjarni Barkarson*, Researcher, Reykjavik University
14:30 - 14:40	Summary and closing remarks

3 Summary of Content of Sessions

3.1 Welcome and introduction

Prof. Kristmannsson opened the session with introductory words of welcome.

3.2 The CEF AT platform

Vilmantas Liubinas, Directorate-General for Translation, European Commission, gave an enlightening talk entitled „Digital Europe Language Tools Platform“ where he presented a number of use cases of the platform.

3.3 Common European Data Space

Monica Pretti, Data Directorate, Directorate-General for Communications Networks, Content and Technology, European Commission, gave an informative talk on the Language Data Space or LDS, potentials and possibilities, the structure of the project.

3.4 Almannarómur Presentation

Jóhanna Vigdís Guðmundsdóttir, head of the Iceland Centre for Language Technology, gave an overview of the Icelandic LT initiative.

Almannarómur is a non-profit organisation founded by industry partners and institutions to ensure language technology solutions would be built for Icelandic. Its role is to ensure that the public and industry have access to language technology solutions for and in Icelandic.

It is the first language technology programme for Icelandic and it started in 2019, and is led by Almannarómur. It is funded by the government, and is a consortium of institutions, universities and startups building the infrastructure according to Almannarómur´s execution strategy of 2019 – 2023.

The speaker reported on the actions of the project, among which was a visit to Big Tech companies in the USA, where the president of Iceland accompanied the delegation. The project terminates in 2023 and there does not appear to be a successor project in sight.

In the following Q/A session questions regarding the quality of data to be collected from translators was posed, and also, from the translators’ view, the worth of the data they have gathered. A call for a workshop on data was formulated. Upon being asked, Jóhanna Vigdís Guðmundsdóttir, stated that Almannarómur itself does not produce any data, but what data the consortium has gathered has been submitted to CLARIN.

Questions about the future of the government initiative in LT were posed, which could not be answered, since the project terminates later this year. Apparently, the current government budget has allocated some funding, which will mostly pay for maintenance of what has been collected.

3.5 The potential of Language Technology and AI – where we are, where we are heading – LT in Iceland

This was a panel with participants from the industry moderated by Joss Moorkens from the Adapt Centre of Dublin City University, Vilhjálmur Þorsteinsson CEO of Miðeind, an LT start up at the forefront of MT in Iceland, Hafsteinn Einarsson, assistant professor in CS at the University of Iceland and a specialist at deCODE genetics and Helga Ingimundardóttir, head of AI at Travelshift (Guide to Europe).

After an introduction, Hafsteinn Einarsson opened the discussion with a few slides on his previous work on AI and the current development of AI and the rate of progress which he claims is soon to be exponential. Then he introduced the recently opened ChatGPT platform and showed how he had asked the platform about what should be discussed in the panel itself, and the answer appeared competent. He then discussed the possible developments, based on the huge changes on the horizon. The results might be very divisive, with the development of "Compute Rich" companies vs. "Compute Poor" companies. This will then have implications for the competition in the energy sector, since energy will be a limiting factor in the future. In addition, he discussed the possible consequences. The problem in Iceland lies in the need to get more data sets to be used. More content creation is needed as well.

Helga Ingimundardóttir had asked ChatGPT roughly the same question as the previous speaker, and compared Travelshift's templates for travel itineraries with OpenAI: ChatGPT which were pretty comparable, although some improvements were necessary.

Vilhjálmur Þorsteinsson discussed his company's development of MT and the increasing capabilities in MT for Icelandic, noting that some of the recent platforms are doing translation without specifically trained for it. GPT 3 models are already able to summaries very well from Icelandic, but the results are not as good when translating into Icelandic. What this means for smaller languages like Icelandic is based on the quantity of data available, and then he noted how the human input is used to increase the quality. The question is how we can make these models include Icelandic, should there be deals with BigTech or should we take the European angle. This is a major topic in the changes to come.

The outlook, according to Vilhjálmur, is a mix of approaching BigTech, Europe and the Nordic countries in getting Icelandic ahead. Hafsteinn called for a governmental infrastructure for both private and public entities to use, and also increase computer use in primary school teaching. There are also huge opportunities in the public sector, healthcare, for example.

3.6 Language technologies by/for the public sector (Panel session)

The second panel was moderated by Níels Rúnar Gíslason, teacher of translation technology at the University of Iceland and a former employee of the Principle project.

Participants were Kári Örlygsson, director of the Translation Center of the MFA in Iceland, Steinþór Steingrímsson from the Árni Magnússon Institute, who is in charge of large Icelandic corpora at the institute. Þórunn Arnardóttir from the University of Iceland was ill and could not participate.

Kári Örlygsson delineated the use cases of the Translation Center, which has been active since the nineties and which has supplied large bilingual databases for the use of more than one LT project, among them Principle. Their use is primarily they are the end-user of LT. The centre translates EU acts into Icelandic, added to which are texts for use in other areas in government. The centre uses Trados Studio 20121 now, and Multiterm. They have been collaborating most recently with Miðeind and its MT solutions, the so-called Greynir tool, especially when the TMs do not perform well. Translators were sceptical at first, but with time grew to using it more extensively.

Steinþór Steingrímsson noted that public institutions like the Árni Magnússon institute and other public entities are not producing data for translation and other use-cases, but their products could be used as a by-product. The difficulties in gathering data have proven to be a hindrance, probably there would have to be awareness-raising inside the institution and also some legislation might be needed. Government institutions might see the benefits of submitting data when confronted with AI platforms. The question is how to use this technology in the case of public institutions which cannot publish information before it is as good as possible. Qualitative matrixes are not available now. Helga Ingimundardóttir intervened noted that the end-user like the Translation Center of the Ministry for Foreign Affairs should be able to come to an agreement with developers and give them feedback with their data.

Steinþór Steingrímsson pointed out that most parties approached to submit data are very wary of doing it. There is a framework needed. The institute is also collection other kinds of data, for speech recognition for example. They are also working on dictionaries and improving the data sets and corpora under their management. Asked about the status of LT in Iceland with regard to a recent report, he discussed the rapid changes in LT which have been going on and the growing pains the institute has been going through, but ended on - an optimistic note, saying there were now dozens of people working in this area, whereas only few were a decade ago.

3.7 Principle Project

Gauti Kristmannsson, professor of translation studies at the University of Iceland reported on the recent Principle project, which was a two-year project supported by CEF.

Participants were the UoI, Dublin City University, the University of Zagreb, the Norwegian National Library, and the private company Iconic.

The participants in the project were gathered because the four languages in question, Icelandic, Croatian, Irish, and Norwegian were underrepresented in the digitisation of European languages. Therefore, the project focused on two main pillars: firstly, gathering of

bilingual data for the four respective languages, and, secondly, building bespoke translation engines for suppliers of data for them to test and compare with previous MT solutions, such as Google translate and Bing. The bespoke engines performed better than open solutions on the internet, but there is still some way to go for these languages.

3.8 NLTP in Iceland

Bjarni Barkarson from the Reykjavik University gave a talk on the LT platforms already developed within the project, with a stress on national language specifics.

The current development is focused on AI-driven language technology solutions. The project is supported by CEF, and there are eight other institutions and companies taking part in it across four countries: the University of Malta and the Office of the State Advocate in Malta, the Malta Information Technology Agency, Tilde and the Cultural Information System Centre in Latvia, the University of Tartu and the Central State Office for the Development of the Digital Society, the University of Zagreb, Faculty of Humanities and Social Sciences in Croatia.

The focus is very much on MT and e-Translation, with different products being developed, such as an Online Translation Workspace, a Terminology Portal, an Online CAT tool, an Online Translation Memory, a Website Translator, Speech Technology, speech synthesis and recognition. This will be a sustainable open platform.

Finally, Bjarni reported on the status of the platform in Iceland.

The talk was followed by a Q/A session and a lively discussion.

3.9 Take-home message and conclusions

The main conclusion can be that the current situation in LT and MT is a great flux, because of the rapid development of AI and the diverse approaches in the European Economic Area. However, there are also problems of funding in Iceland since the government has reduced its funding for LT projects considerably, in comparison with the last few years.

It remains to be seen what steps will be taken in this regard in Iceland, but there are now many more individuals and companies working in this area than ever before, both in the public and private sectors. The rapid changes happening as the workshop went on (ChatGPT was only nine days old) underlined the need for a small language to find solutions, and the question of government funding remains unanswered yet.

4 Synthesis of Workshop Discussions

The main issues were machine translation and the problem of digitisation for smaller languages like Icelandic. The use of AI was debated, and its influence on the current development assessed.

All participants agreed that multilingual support needs to be increased in public administration and in the private sector, for example about the ongoing globalisation, the massive growth of the information society, and, in Iceland, the rapidly growing numbers of immigrants and tourists.

Iceland has, through the Principle project and NLTP, for example, collected a number of data packages and the process is also ongoing through private firms, such as Miðeind.

The main finding may be that the different smaller projects by both the public and private sectors need to be coordinated better to support a small language. Iceland had an officical Language Technology Programme for Iceland running from 2018-2022, but there is not another comprehensive project to follow it up. Although the programme resulted much greater activity in this field, both in the private and public sector, the question remains whether this will be enough to sustain a small language like Icelandic in LT or whether it will be for BigTech to decide what to do for Icelandic. European solutions and cooperation seem to be the only real alternative.

5 Country Profile: Language data creation, management and sharing

The situation for Iceland has improved considerably in comparison with the 2019 country profile as can be seen in the revised edition of the ELRC White Paper. Greater volumes of data have been collected and the awareness in the field has become much better. Also, the awareness among public institutions and the general public has increased with public actions, such as the gathering of speech data through Almannarómur. The focus hitherto has been on 1) the Icelandic National Corpus 2) the gathering of bilingual data in several projects such as Principle, NLTP, in addition to the work the private company Miðeind is performing in the field of MT and 3) the abovementioned public collecting of speech data by Almannarómur in conjunction with Miðeind. The data gathering is, however, rather sporadic, and very often the greatest bulk of the data is from the The Translation Center of the Ministry of Foreign Affairs. Licensing is a problem, and the relatively small datasets created by individual entities in Iceland. After Almannarómur terminates later this year, there will be no central instance of LT and data gathering in Iceland, so it remains og seen how the development will be.