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

Deliverable D3.2.4
Task 3

ELRC Workshop Report for Finland

Author(s):	University of Helsinki, FIN-CLARIN
Dissemination Level:	Public
Version No.:	V 1.0
Date:	2022-03-02

Contents

1 Executive Summary

The third ELRC workshop in Finland took place as an online event via Zoom on 15th December 2020 at 9.30-12.40 local time. The event was organized by the Department of Digital Humanities, University of Helsinki, a member of the ELRC consortium. The local organization committee consisted of Krister Lindén, Mietta Lennes, Tommi Jauhiainen and Erik Axelson. Additionally, ELRC provided advice and support with the practical arrangements.

Being a follow-up of previous national ELRC workshops and related events, the third ELRC workshop was built around the general theme of Language-Centric Artificial Intelligence.

Several of the workshop sessions referred to the ongoing Donate Speech campaign (Lahjoita puhetta, https://www.kielipankki.fi/donate-speech/), started as a joint project by Vake Oy (currently Ilmastorahasto Oy), the Finnish Broadcasting Company YLE and the University of Helsinki. The campaign is collecting Finnish speech from any Finnish speakers who would like to participate via mobile and web applications specifically designed for this purpose. The collected speech samples will be stored and distributed via the Language Bank of Finland under a special license that allows the material to be used for AI development and language research in academia as well as in the private sector.

The event was interpreted live into Finnish and English by interpreters from Interactio and to Finnish Sign Language by interpreters from the University of Jyväskylä.

In the same online meeting room as the ELRC workshop an ELG workshop was organized in collaboration with the European Language Grid (ELG).

2 Workshop Agenda

09:30 – 09:40	**Welcome and introduction** ***Krister Lindén***, *University of Helsinki / FIN-CLARIN*
09:40 – 10:00	**The potential of Language Technology and AI – where we are, where we should be heading** ***Jörg Tiedemann***, *University of Helsinki*
10:00 – 10:30	**Language Technologies for the Languages of Finland – Panel session** ***Filip Ginter***, *University of Turku* (Moderator) ***Sebastian Andersson***, *Lingsoft* ***Jörg Tiedemann***, *University of Helsinki* ***Sampo Pyysalo***, *University of Turku* ***Pasi Tapanainen***, *Etuma* ***Kaarina Hyvönen***, *Kielikone*
10:30 – 10:45	Coffee Break
10:45 – 11:15	**The CEF AT Platform** ***Vilmantas Liubinas,*** European Commission
11:15 – 11:45	**Language technologies by/for the public sector – Panel session** ***Jouko Salonen***, *Finnish Immigration Service* (Moderator) ***Osma Suominen***, *National Library of Finland* ***Ville Viitasaari***, *Kela* ***Kaisamari Kuhmonen***, *Prime Minister's Office*
11:45 – 12:15	**Language data creation, management and sharing: existing practices and challenges – Panel session** ***Aleksi Rossi***, *YLE* (Moderator) ***Krister Lindén***, *University of Helsinki / FIN-CLARIN* ***Mikko Kurimo***, *Aalto University* ***Tommi Kurki***, *University of Turku*
12:15 – 12:30	**The EU Council Presidency Translator – Finnish presidency success story and what's beyond** ***Pekka Myllylä***, *Managing Director at Tilde Eesti OÜ*
12:30 – 12:40	**Conclusions** ***Krister Lindén***, *University of Helsinki / FIN-CLARIN*
12:40 – 14:00	Break
14:00 – 16:30	***European Language Grid (ELG): Introduction and overview. 4th Regional European Language Grid (ELG) Workshop in Finland***

3 Summary of Content of Sessions

3.1 Welcome and introduction

Krister Lindén welcomed the participants, many of whom had also participated in some of the previous ELRC events in Finland. Lindén introduced the goals of the ongoing Donate Speech campaign, which was closely related to the workshop theme, LT and Artificial Intelligence. After summarizing the current goals of ELRC, he introduced the program of the workshop.

During the session, general information about the workshop and about the online polls was provided via the chat on Zoom.

3.2 The potential of Language Technology and AI – where we are, where we should be heading

Jörg Tiedemann talked about the relationship between human intelligence and language and about the important role that Language Technology plays in AI development – LT is already all around us in various applications. Traditional LT approaches have been mostly modular, theory-driven and built for a single purpose, often also for a single language, and such solutions have been found to be difficult to embed in more complex applications and real-life situations. Current trends in LT emphasize the dynamic, multimodal and interactive nature of language and aim to process raw data, possibly in several modalities, for training big language models that can be used for various specific tasks. The goal is to map language input to meaning representations that can then be used for solving real-life problems.

Tiedemann gave an example of his own research area, Machine Translation, where traditional rule-based translation systems have been replaced by neural machine translation systems. Neural models are data hungry, which can be a serious disadvantage for stakeholders without data access. Fortunately, the number of public data sets is increasing and there are policies for sharing data, models and tools. Tiedemann talked about the projects Fiskmö and OPUS MT, the collaboration with ELRC and the European Language Grid, and the Finnish Center for Artificial Intelligence (FCAI) that is a community of experts with the slogan "Real AI for real people in the real world". He explained why the AI system of the future requires both language and interactive communication in order to achieve understandability, trust and self-awareness.

During the session, a comment was made by one of the members of the audience via the chat, stating that at the beginning of year 2017, the number of foreign-language (not Finnish or Swedish) speakers in Finland was 354.000, and the number of their languages was around 500. According to the predictions of the City of Helsinki City Research and Statistics Unit, the expected percentage of the so-called foreign-language speakers in 2035 will be 26% in Helsinki, and 25% in the whole region of 4 municipalities in the metropolitan area in Finland. Another participant noted via chat that there are also sign languages in addition to spoken and written languages. These are very good points to bear in mind, since it can be difficult to fully participate in Finnish society without knowing any Finnish or Swedish. The non-Finnish speaking and sign language communities are a relevant target group for LT.

3.3 Language Technologies for the Languages of Finland (Panel session)

After an initial round of introductions, the panelists discussed the current status of Language Technologies in Finland. The panelists were in accord about Finland doing quite well in the LT field. This is not surprising since Finland has strong traditions in computational linguistics and software engineering. When asked about what might be "the next killer app" in Language Technology, the panelists mentioned virtual assistants and simultaneous interpretation, or some sort of "invisible" machine translation that would work behind most web services. As an example of the future improvement possibilities for MT, Amazon was mentioned as having made a bold move when opening their online store in Sweden recently, without making sure that all the translations in their services were 100 % correct. In addition to MT, search facilities could still be improved a lot for Finnish. The "killer app" could also be a more general type of helper with functionalities that might not necessarily be built around any specific language technology.

The panel was nearly unanimous on the most important bottleneck for LT, namely that LT development requires large quantities of data that are currently not easy to obtain. To alleviate the data issue, some sort of intervention from the state might be in order. On the other hand, ensuring a constant influx of data would be useful, since existing collections of data can become outdated very quickly. It is necessary to invest sufficient human resources as well in addition to access to data.

Open source tools have sometimes been considered as a potential threat since big companies can exploit the same open resources for building their own systems, with which it is difficult for smaller players to compete. According to the panelists, however, these threats had actually not realized. In practice, open source can offer huge opportunities for small companies, too.

3.4 The CEF AT platform

Vilmantas Liubinas, a computational linguist from the Directorate-General for Translation (DGT), discussed the limitations of human beings in understanding languages. According to the statistics he presented, about half of the population in Europe are either monolingual or speak only one foreign language. In other words, machine translation could offer many citizens straightforward assistance in different situations.

In 2006, Google started offering automated translations for three different languages. Today, the list of translated languages is impressive. The eTranslation system of the European Commission offers 29 different languages, Finnish included. The advantage of using a European service is in the security of data and intellectual property. Legal restrictions and confidentiality can be taken into account when translating official documents.

Liubinas also presented statistics regarding the quality of machine translations for different languages. The context affects the quality of translation, so choosing the right domain for the translation engine is important as it improves the quality of the output. Liubinas demonstrated several features that eTranslation service currently offers, such as the support for TMX format.

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

According to Ville Viitasaari from the IT innovation unit at Kela, language technology is used extensively in the public sector in Finland. The especial challenge at Kela is the privacy concerns about their own material, which can be used for only a few specific purposes. Kaisamari Kuhmonen, who works at translation services at the Finnish Prime Minister's Office, talked about the use of machine translation at governmental institutions and about their cooperation with Tilde at the European level. Machine translation has become a useful tool in translation workflows both to and from Finnish. Data confidentiality issues at the ministry level are similar to those in Kela. Jouko Salonen discussed the quality of machine translation in cases where the translations are used to decide the fate of the applicants at the Finnish Immigration Service (Migri). He also pointed out issues related to data sharing and to the use of data to improve Migri's own services. Osma Suominen talked about the materials stored in the the National Library of Finland. He described problems related to detecting new terminology and concepts as well as the challenges of using general language models such as the Finnish BERT.

The participants then discussed the use of LT powered products by private citizens at public institutions. Kela utilizes text-based chatbots and has also experimented with speech-based bots. Many government web pages have been generated with the help of machine translation.

The session had to be cut short, but the discussion was continued via the chat during the following session. Osma Suominen added some notes about the National Library services via the chat. Finto AI was mentioned by him as an example of publicly available AI services in Finnish libraries. When an academic thesis is uploaded to one of the publication archives, the system will provide the user with suggested topics that the thesis is about. A test service is provided at https://ai.finto.fi and an API is also available.

3.6 Language data creation, management and sharing: existing practices and challenges (Panel session)

The session started with initial introductions led by the moderator, Aleksi Rossi. The first round of comments dealt with the challenges encountered in data collection and distribution. Tommi Kurki, who has a background in sociolinguistics, presented examples of how speech data was collected in the past and how the first steps in digitization of speech data collection were taken in 2013. More challenges were presented by Mikko Kurimo, who told us about the difficulties in data distribution due to copyright law. The distribution of data across national borders is crucial to the success of research when working with international partners, and sometimes it demands a great deal of stamina and extra paperwork to achieve this goal. Krister Lindén described some of the challenges in applying LT in the research of a small language with only a limited number of resources.

Rossi asked the panelists to name some issues that one should become aware of when talking about LT and speech data collection. Kurimo emphasized the importance of understanding the richness of language. If we want to have an AI that understands us, it cannot be achieved without a comprehensive collection of data that includes different styles and groups of speakers. Kurki pointed out that the speech of an individual tends to vary from one interactional situation to the next, and to get the AI to learn this, the system must be trained with prosodic data as well. Lindén added that many non-technologically orientated researchers seem to believe that speech recognition would already be fully functional and complete. They tend to underestimate the efforts that are still required in order to make this technology work.

In the third round, the panelists discussed the changes that have taken place in the operational environment. What has changed or should change in the future? Each panelist brought out issues regarding legislation. The new Copyright Directive will affect data mining and it will probably bring about some new use cases. Another factor is the principle of open data, which encourages organizations to find safe ways for providing access to confidential or legally restricted data. Data management proficiency cannot be sufficiently highlighted. It is a basic requirement in order to work efficiently with data sets.

3.7 The EU Council Presidency Translator – Finnish presidency success story and what's beyond

Pekka Myllylä presented Tilde and described the machine translation systems developed by the company. Tilde has also developed the machine translation system that was used during the period of EU Council Presidency by Finland.

During EU Presidency, public administrations tend to face similar challenges: huge volumes of documents, requirements for confidentiality, safety and fast transfer of information and communication, a high amount of collaboration between different communities and multilingual groups, etc. In the EU Council Presidency Translator, all 24 official EU languages were available via the generic translation motors, whereas tailored translation modules were built for the language pairs English-Finnish and Swedish-Finnish.

The number of words translated via the EU Council Presidency Translator did not increase very much until the latter half of the six-month presidency. However, after the presidency, the translator has been used quite regularly, and the translation volumes in 2020 have stayed at the level of approximately 3 million words per month, summer months excluded. Myllylä also pointed out that during the past few years, the attitudes of professional translators towards machine translation have changed to the positive, which has probably contributed to the general increase of MT volumes.

As a second example, Myllylä presented the EU Council Presidency Translator built for Germany. The translator included some more advanced features, such as MT motors that were pre-trained with domain-specific data for various purposes, tools for translating web pages, integration of terminologies and translation tools for professional translators and a simplified online translation tool for public sector employees.

The Prime Minister's Office has been the most active user of the MT system, but a lot of other entities in Finnish public administration, including municipalities, are already using the tool in their daily work.

3.8 Conclusions

In his summary of the issues raised during the workshop, Krister Lindén stated that the general attitude toward Language Technology is quite positive in Finland. Although Finnish may not be one of the languages facing digital extinction, children may learn to rely on English in cases where Finnish language support is not readily available, e.g., in games and entertainment, whereas they might use Finnish for school and for communicating with their parents. LT research and industry can help in supporting the native language by increasing the amount and quality of multilingual services in various areas of life.

In order to provide language-centered AI solutions, large quantities of training data are required. Lindén invited all the participants to join the Donate Speech campaign, the project where Finnish speakers can contribute to the development of AI with better support for spoken Finnish.

3.9 Afternoon sessions: European Language Grid (ELG): Introduction and overview

The 4th Regional European Language Grid (ELG) Workshop in Finland took place in the afternoon, after the lunch break. The event was organized in collaboration with ELG and the agenda consisted of an overview of ELG by Katrin Marheinecke, an online demo by Nils Feldhus, presentations of Finnish pilot projects where ELG has already been used, a presentation of the expectations of Finnish Language Technology providers regarding ELG by Marko Turpeinen, and a short summary and discussion. The agenda is available on the workshop website, https://www.kielipankki.fi/elg-workshop-2020/.

In her introductory talk, Katrin Marheinecke described the variety of LT tools that are already around and pointed out that, although Europe is strong in research and innovations, the European LT landscape is very fragmented and not very successful in scaling innovations and capturing the market. This is the main motivation behind the European Language Grid (ELG) platform project that aims to enable the European LT community to upload services and data sets and to make it easy to use and to connect with the resources. The ELG includes facilities for different user groups and use cases, e.g., for browsing the catalogue, for downloading data, for calling services from the command line via APIs, and for uploading resources. It is also possible to make resources available for a fee. The number of the available resources is constantly growing. However, the details of the business model of ELG were still open.

Nils Feldhus provided an overview and demonstration of the current facilities and features of the ELG platform. The live ELG platform can be accessed at https://live.european-language-grid.eu. The catalogue provides a faceted search for finding various tools, corpora and other resources, and after logging in, users can try out many tools via the web interface.

Three Finnish pilot projects for ELG were then presented as examples. Filip Ginter talked about the Paraphrase data set for deep language modelling, Sebastian Andersson described the LSDISCO project that involves the various LT tools developed by Lingsoft, and Jörg Tiedemann discussed the OPUS-MT project involving publicly available machine translation tools. Lastly, Marko Turpeinen talked about the Donate Speech campaign in Finland. This example potentially provide some ideas for similar projects in other countries where public-private partnerships might be useful.

In the final discussion of the ELG workshop, it was concluded that sharing LT models and other data would be very useful especially for smaller countries and smaller companies with limited resources. It was foreseen that LT tools and technologies may soon become common knowledge. In the future, it will be even more important to provide tools that can be readily applied via accessible interfaces and flexibly integrated as part of larger and more complicated workflows.

4 Synthesis of Workshop Discussions

Finland is traditionally strong in Language Technologies. The attitudes towards MT have become more positive in Finland during the past few years. MT systems are developed actively and used more and more widely. Especially in the public sector, the EU Council Presidency Translator has been a success story.

For developing LT models and tools with good coverage for various domains, it is essential to have access to large quantities of training data. However, there are still legal, technical and practical issues to solve in order to put more incentive on sharing data and to make data sharing and reuse more convenient for various stakeholders.

The current trends in LT tools include multimodality and interactivity. User interfaces must be able to process dialogue and conversation. Speech data is recorded in audio and video format instead of text, and sign languages will also need attention. Many stakeholders are looking forward to speech interfaces that could support Finnish sufficiently well. One of the topics highlighted during the workshop day was the Donate Speech campaign that aims to collect Finnish speech data that can be legally used for academic research as well as for AI development in the private sector. The general concept, the speech recording interface and the experiences obtained during the campaign will be useful in other countries where similar data are required.

A total of 32 people participated in the afternoon session organized by ELG. All participants except for three of the speakers had already been present in the ELRC workshop in the morning. This indicates that the interest groups of the two workshops were quite similar and it is probably a good idea to try and combine the efforts of ELRC and ELG in future events as well.

Regarding future online events, it is vital to have a sufficient number of available people who can actively support the online conference. The person(s) monitoring the chat cannot actively take notes on the talks and panel discussions. Of course, the online event can be recorded for future reference, but if there are no reliable automatic captioning services available for the language (as is the case for Finnish), it will take a lot of time to review the entire recording afterwards for note-taking. According to the accessibility requirements in Finland, if the videos are published afterwards for the audience and kept online for a longer time, they must be equipped with captions in the original language, which takes a lot of additional resources.

For the chat, it is a good idea to prepare a list of links in advance, so that they can be pasted in the chat at suitable moments. When the speakers mention online services, the chat moderator should be ready to look them up and to post the link in the chat as soon as possible. This can increase the feeling of involvement by the participants.

The polls on Zoom only work for participants who join the meeting via the desktop app. Instead of the poll feature in Zoom, we tried asking for feedback via Presemo, which can be used on any web browser. The participants were notified and reminded about the poll questions via the chat and by the chair of the event. The poll requires a moderator who can take care of making the relevant questions visible at specific moments during the event. We received quite many answers to the first questions, but the response rate decreased towards the end of the event, although nearly all of the participants stayed online until the end of the workshop.

5 Country Profile: Language data creation, management and sharing

Only two people filled in the ELRC Country Survey for Finland after the event. On the basis of the presentations, the discussions by the panels and the points raised via the chat, we can at least conclude that data management skills are becoming more and more important for researchers as well as for companies. Legal issues, including copyrights and the constantly evolving and country-specific practices with regard to personal data processing, are still problematic for data sharing on many levels.