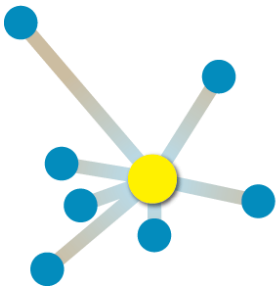
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.



European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.15
Task 3

ELRC Workshop Report for Estonia

Author(s):	Tilde Eesti
Dissemination Level:	public
Version No.:	V3.0
Date:	2021-10-25

Contents

1 Executive Summary

The Estonian ELRC Workshop took place in Tallinn, on the 30th of September 2021, at the office of the European Commission Representation in Estonia. The event was also broadcasted via internet.

The workshop was co-organized by the ELRC Representatives in Estonia – Tilde Eesti and the Office of the European Commission Representation in Estonia, and local NAPs. In general, the participants appreciated the workshop very much, finding a well-organised arrangement, technically perfect and interesting presentations and discussions.

An event-dedicated web part in Estonian was set up at the ELRC website before the event, at https://lr-coordination.eu/estonia3rd. The page was populated with the agenda and the online registration form.

Invitations to the participants were sent out a couple of weeks before the event. The invitation was followed by a personal e-mail to the persons to whom invitations were sent.

The event attracted more than 200 participants. All feedback forms that were handed in showed a positive experience.

You can find the full recording of the event on ELRC website: https://lr-coordination.eu/estonia3rd

2 Workshop Agenda

The final workshop program in English

09:45	**Opening remarks** by the Acting Head of Estonian Language Department, DGT, Ms Leila Anupõld
10:00	"Has the machine translation really reached human parity?" Samuel Läubli, Universität Zürich, Institut für Computerlinguistik
11:00	"How the neural machine translation exactly works?" Mark Fišel, University of Tartu, Institute of Computer Science, Professor in natural language processing
11:45	"eTranslation – latest developments" Hugo-Tanel Kaasik, DGT ET translator
12:15	Panel discussion on pros and cons of the machine translation from the users perspective Moderator - PAEMURD Elis
13:00	Lunch Break
14:00	Estonian Translators' Survey 2021 Kerli Ilves, University of Tartu Skytte Institute of Political Studies
14:30	State of play of the Estonian Language Technology Development program Kadri Vare, Project Manager (Estonian Language Technology) at The Ministry of Education and Research / Susanna Oja, Head of Competence Centre for NLP at Institute of the Estonian Language
15:00	The central translation environment – results of the analysis and prototype procurement

	Mari Peetris, Translation Adviser at the State Gazette Division of the Ministry of Justice / Maarja Ottis, Business Analyst at the Department of Supporting Information Systems of the Centre of Registers and Information Systems
15:30	MTee MT engines project Susanna Oja, Head of Competence Centre for NLP at Institute of the Estonian Language
16:00	National Language Technology Platform (NLTP) project and its impact for Estonia Martin Luts, Language Engineering Specialist for Tilde Eesti

3 Summary of Content of Sessions

3.1 Welcome and introduction

The workshop was opened with introductory words by Elis Paemurd, the language advisor of Estonian Language Department, DGT, Leila Anupõld, the Acting Head of Estonian Language Department, DGT and Indrek Hallik, Sales manager of Tilde Eesti.

Ms. Paemurd and Ms. Anupõld opened the day with remarks about International Translators Day, DGT in Estonia and the main theme of the conference, while Indrek Hallik introduced the topics of the second half of the day, ELRC's role in language technology in general – development and popularizing of machine translation and data gathering – and its role in Estonia. The cooperation between ELRC and Tilde and this being the 3rd seminar in Estonia were highlighted.

It was also noted that participants can ask questions and respond to polls in Sli.do and that the conference will be covered by English interpretation.

3.2 Has machine translation really reached parity with professional human translation? The impact of document-level context on quality evaluation and translator performance (Samuel Läubli)

The presenter gave an overview of MT history, outlining technological paradigms and shifts in developers' hopes on reaching "ideal" MT systems.

The importance of document-level context, both for evaluation of MT systems and for human translators while translating was discussed. The main conclusion is that the current approach (what is considered "best practice"), using only sentence pairs for evaluating MT quality is not sufficient. How context could be useful (in terms of speed, accuracy) and how to present context for professional translators via next generation CAT tools was introduced.

Other aspects of MT evaluation with an impact on the results and conclusions on parity:

- Evaluating fluency, accuracy, adequacy
- Raters. Evaluation by professional translators or "crowd" (students, researchers)
- Source text original of (machine) translations, "translationese".

Why is the context important? Some MT errors are not visible in sentence-level evaluation. The proposal is that MT quality should not be evaluated with isolated sentences.

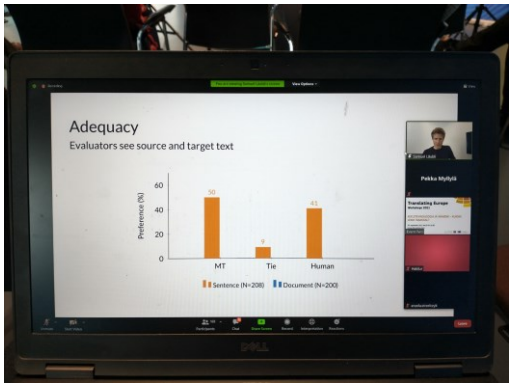
More precisely, MT makes more mistakes in:

1. Wrong words: semantics, grammaticality
2. Omissions
3. Word order errors

4. Mistranslated named entities

In summary, the conclusions are: claims about human–machine parity were exaggerated, but MT is improving fast. Let's focus on finding out and focusing on what humans and machines are best at.

The research the author introduced implies that CAT software with sentence segmentation should implement top–bottom orientation for source and target text boxes (rather than left–right) and if assumed that main activity in human translation will shift (even more) from text production to revision (reading), CAT software should not use sentence segmentation.



3.3 How does neural MT work? (Mark Fishel)

Mark Fishel focused on 3 "claims" related to (machine) translation:

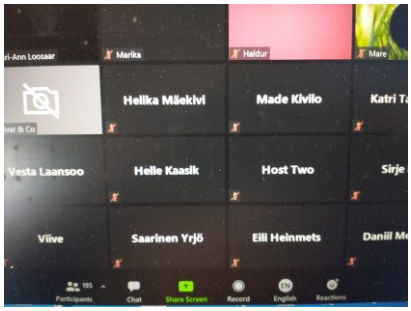
1. Translation is easy
2. (Creation of) Machine translation systems is a task already completed.
3. Machine translation/Machine learning is just imitating training material.

Some simple examples in Estonian and English were discussed, illustrating the need to understand the context in order to translate ambiguous words and phrases.

Next, the presenter explained how current state-of-the-art neural networks can learn from examples and how systems with self-attention work.

He also briefly touched the topic of under-resourced languages e.g. Võro[1] and approaches to tackle the training data scarcity problem: using languages close to the one, synthetical data (that is, creation of parallel data from monolingual using MT and PE).

Lastly, Mark Fishel discussed some ideas on how to apply MT in the task of gramcheckers.



[1] https://en.wikipedia.org/wiki/V%C3%B5ro_language

3.4 MTee – Estonian MT project (Susanna Oja)

Susanna Oja introduced an Estonian public sector project MTee, which aims to increase the speed and availability of information by training Estonian <> English, German and Russian MT systems for general, legal, public health/crisis and military domains.

The project has been initiated and supported by the Estonian Ministry of Education and Research and is led by the Institute of the Estonian Language. The project will end in December 2021 and is implemented by the University of Tartu and Tilde.

The system features automatic detection of a domain, translating webpages and documents while preserving the layout. In addition to that, novel approaches for translating speech and applying gramcheckers are part of the project. Some examples of translating different domain texts into various languages were given.

Lastly, Susanna unveiled the next steps beyond the current project: adding new language pairs and domains and creating an Estonian central translation platform.

Particular call to actions were to follow the project newsfeeds and if relevant, donate data sources for MT trainings.

3.5 eTranslation – latest developments (Hugo-Tanel Kaasik)

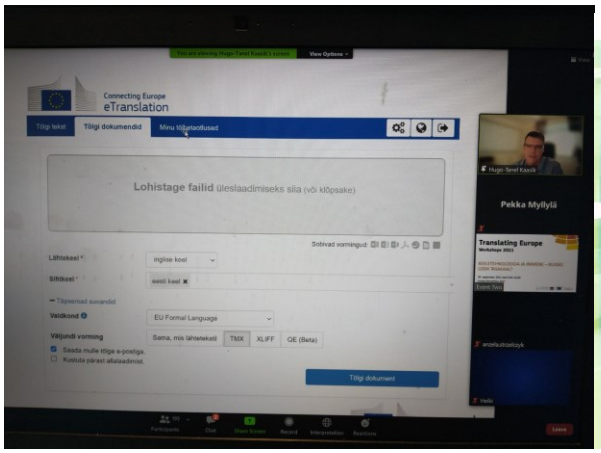
Hugo-Tanel Kaasik shared the latest news on eTranslation, its availability for a broader audience: SMEs, public sector, universities. He gave a walk-through how to start using the eTranslation system, including registering, using the system for text and document translation, selecting domains and their relative quality of translation.

The strengths and weaknesses of the systems were discussed, related to security (privacy, confidentiality), language and file type coverage, pricing, translation quality for different language pairs and domains/ engines.

Hugo described the ISO18587 post-editing standard, the differences between full and light-PE and which to apply for different use cases like legal acts, emails, etc.

He introduced the state of MT usage among DGT's Estonian translators and how it has changed over the last couple of years.

Hugo also presented some advancements in MT in general, including document-level context-sensitivity, machine-translating from and into the Estonian language using DeepL, and others.



3.6 Discussion panel (moderated by Elis Paemurd)

The following topics were discussed:

1. How language technologies influence human language. It was pointed out that LT already is heavily used in our daily communications: automated text and voice translation, automated corrections of input texts, and more. This trend is supported by the spread of smart devices like earplugs, glasses, virtual and augmented reality. In this respect, an important question came up, i.e. “Do algorithms based on statistics have an impact on our languages towards a more “standardized”, non-personalized direction”? Avoiding this is the role of a post editor.
2. Technology will not only increasingly serve as facilitator of communications with other humans, but also become a communication partner, as it is the case with e.g. chatbots.
3. The motivation to learn foreign languages in the era of automated translation. The technology was seen as a near-future language teacher. Computer translation and learning foreign languages complement each other.
4. One of the machine translation sources of quality loss is a poorly authored original text.
5. Practices to use MT in daily translator tasks: Professional translators may use many different MT engines depending on their clients’ rules. Most translation tasks nowadays include MT output. Translators’ attitude has changed over the last 5 years – at that time, it was not useful for a translator. Nowadays, the use of MT is mandatory, MT is “passable” with some minor modifications and MT helps a translator to perform better and to earn more per timeslot.
6. Usage of context in MT versus segmentation of the source text. The translator may be “trapped” by the MT suggestions.
7. Domain-specific MT engines have shown better results for professional translators.
8. The changing nature of human translators was discussed. Nowadays, translators are considered more like editors who are post editing MT output. Consequently, know-how about common mistakes in MT and about errors that are hard to detect, is required.
9. How would the work of a translator change over the upcoming years? Depending on the translation tasks and goal, it may be important to build technical skills (even programming) but also more soft skills e.g. having creative mindset.

3.7 State of play of the Estonian Language Technology Development program (Kadri Vare, Susanna Oja)

Kadri and Susanna gave an overview of the Estonian Language Technology Programme https://www.keeletehnoloogia.ee/en?set_language=en.

The position of Estonian language technology in the languages with the same number of speakers is stable. As a result of consistent work, important basic technologies have been developed, and applications that are actually used by the end-user for speech recognition, speech synthesis, and machine translation have been created. Applications are based on extensive language resources and text analysis tools.

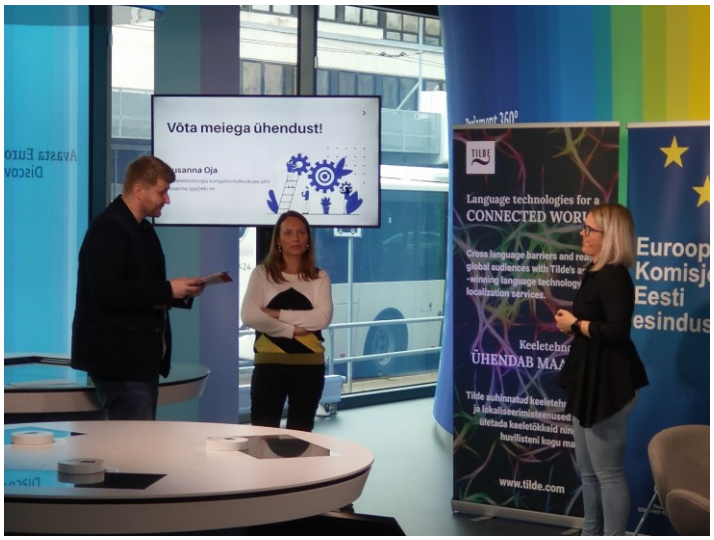
Through the programme, the state supports a field where it is not always profitable for the private sector to take on the risks associated with the development of technology for a language with a small number of speakers - as a small number of speakers also means a small market.

The activities of the research and development program "Estonian Language Technology 2018-2027" supporting the development of language technology will implement the objectives of two sectoral strategies: the research and development and innovation strategy "Knowledge-based Estonia 2014-2020" and "Estonian Language Strategy 2018-2027".

The programme focuses on LT-based technologies: speech, machine translation, text analytics tools, and corpuses. The outcomes of the programme are freely available.

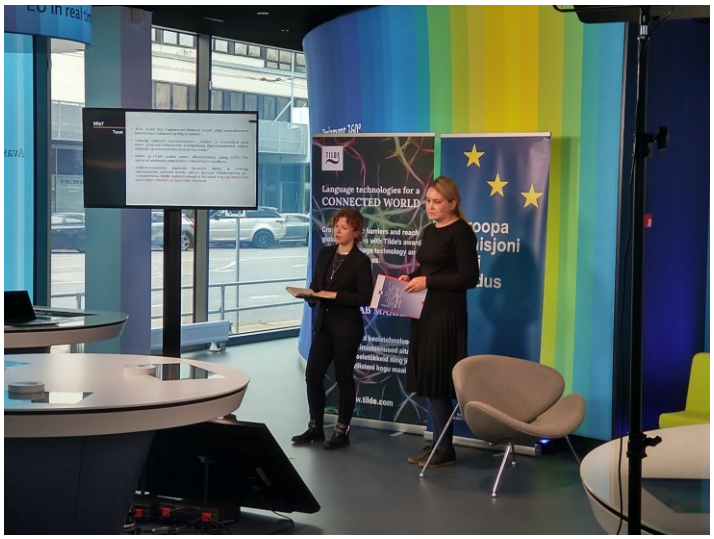
The current main projects were presented: subtitling broadcasts, the machine translation MTee project (see below section 3.9), gramcheckers and the public sector virtual assistant #bürokratt.

An LT competence center has been established to speed up the work in the area.



3.8 The central translation environment – results of the analysis and prototype procurement (Mari Peetris, Maarja Ottis)

Mari and Maarja presented the results of an analysis and prototyping of the central translation environment. The main factors driving the initiative were given: translation costs, translation quality and speed, terminological consistency, data security, among others and the studies which covered the issues.



The goals of the project are to ease the before mentioned problems. Special attention goes to management of translation memories.

The next steps of the project will be to carry out a detailed analysis and to develop the system in 3 phases, first to deliver an MVP in a year.

The analysis report is available at https://wiki.rik.ee/display/KT

3.9 MTee MT engines project (Susanna Oja, Kadri Vare)

Susanna presented the project MTee – a machine translation project for Estonian, English, Russian and German, translating in general, legal, crisis (public health) and military domains.

The goal is to speed up the spreading of time-critical information among non-Estonian speakers, and to make such information available. Also, increasing terminological consistency is a goal of the project.

The project has been initiated by the Estonian Ministry of Education and Research and is managed by the Institute of the Estonian Language and implemented by the University of Tartu and Tilde. The project duration will be from May to December 2021.

The system features automatic domain detection, translation of text, documents, and webpages, translating speech (by means of Estonian ASR), and an integrated gramchecker.

Susanna also introduced the main sources of training data, including ELRC.

The presentation was concluded by giving first insights into the UI and providing some translation examples, explaining how domain-specific engines can help in this context to achieve better translation quality.

In the coming years, more language directions and domains will be added and the system will be integrated into the central translation platform.

3.10 National Language Technology Platform (NLTP) project and its impact for Estonia (Martin Luts)

Martin Luts introduced the NLTP – National Language Technology Platform project and its impact on Estonia.

The project commenced in April 2021 and the first results of the platform will be available in September 2022.

It is an effort of many participating countries and organizations including University of Tartu and Tilde.

The platform includes neural machine translation engines, eTranslation, translation memory management facilities, data repository, speech tools and a CAT tool. It allows to translate text, documents and webpages (a website translation widget) as well as light translation task management. The platform is targeted at the general public and eGovernment and egov services.

Potential use cases for Estonia were introduced: public broadcasts, tourism information and legal acts.

4 Country Profile: Language data creation, management and sharing

Based on the input from the workshop session, and the participants' answers and feedback, the following highlights could be given

- Language resources for training domain-specific engines are scarce and require much effort to collect or create manually or synthetically
- There is a legal framework soon to be established to collect public sector translations and TMs which is expected to help in MT training tasks.

5 Workshop Participants

The Estonian ELRC workshop received more than 300 registrations spanning a wide range of ministries and public organizations and covering LSPs and academia. The workshop was well attended – there were more than 20 onsite participants and more than 200 participants via the internet. Before the workshop, we organized a brainstorming session with potential participants. As a result, we were able to extend our contact base and to collect suggestions for suitable candidates considering the theme of the workshop. Furthermore, in this way, we were able to make sure that important public administration institutions were represented.