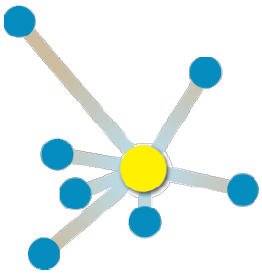
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.



European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.21
Task 3

ELRC Workshop Report for the Czech Republic

Author(s): Jan Hajič, Pavel Pecina, Stanislava Gráf

Dissemination Level: Public

Version No.: V1 - Final

Date: 2022-07-07

Contents

1 Executive Summary

The 3rd ELRC workshop in the Czech Republic took place on 3rd May 2022, in a form of a half-day online event from 9:30 to 12:30, which was organized by the team from the Institute of Formal and Applied Linguistics by the Faculty of Mathematics and Physics of Charles University in Prague.

The event was topically focused on the current state of language technologies and Artificial Intelligence (AI) – from the perspective of its state of the art as well as in the context of the Czech region and language. The keynote presentations were delivered by experts from public and private sector. The target audience were current and prospective language technology users, developers and researchers, and policymakers and representatives from the public sector, research and academia.

The main message of the workshop was to present and discuss ways for transforming digital interaction in our multilingual Europe with the use of language technology and how the advancement in AI can boost this transformation. Participants could learn about the experience from research, development and implementation of solutions involving machine translation, natural language processing and speech transcription, related major obstacles preventing proliferation of such technologies in our daily lives, but also trends and open issues in related research, its history and current expertise in the Czech context, as well as success stories of Czech research, from both public research organizations and private companies operating in the LT/AI field. First-hand experience with using the CEF Language Tools (https://language-tools.ec.europa.eu/) at the Czech Supreme Audit Office were presented and demonstrated current strengths and weaknesses of AI-powered language technologies in real-life scenaria.

The workshop language was Czech for most of the sessions. The ELITR (European LIve Translator)[1] machine translation tool was used to provide the transcript of speech to English and Czech.

Even though research and development of LTs and AI have come a long journey, there is still a long way to go. Their importance increases every day – in order to keep up with the pace and address future needs, continued support in this area is needed. Such support is provided by the new funding programme Digital Europe for the period of 2021-2027 that was also presented at the workshop. The programme will provide funding for deployment of new infrastructure for sharing data across European states, domains, and sectors. This will be done not only by creating and developing new capacities but also by using and connecting existing ones. Investing into these key areas is an important next step towards the diverse, multilingual and truly unified Europe.

The presentations are available at the official event website: https://lr-coordination.eu/czech3rd.

[1] Development of ELITR was supported by H2020 funds (https://elitr.eu/).

2 Workshop Agenda

09:30 – 09:40	**Welcome and introduction** [Czech] *Prof. Jan Hajič, UFAL, MFF UK (https://ufal.mff.cuni.cz/)*
09:40 – 10:00	**The potential of Language Technology and AI – where we are, where we should be heading** [Czech] *Ing. Jan Kleindienst, Ph.D., The MAMA AI (https://themama.ai/)*
10:00 – 10:30	**Language Technologies in Czech Republic / for Czech language** [Czech] *Prof. Jan Černocký, Speech@FIT, VÚT Brno (https://speech.fit.vutbr.cz/)*
10:30 – 10:45	***Coffee Break***
10:45 – 11:15	**The CEF AT Platform** [English] *Francois Thunus, DGT*
11:15 – 11:35	**The new Digital Europe Programme and the Language Data Space** [English] *Philippe Gelin, DG CONNECT*
11:35 – 12:05	**The use of the language technologies at the Czech Republic Supreme Audit Office** [Czech] *Jaroslav Rucký, NKÚ (https://www.nku.cz/)*
12:05 – 12:30	**Final discussion and workshop conclusion** [Czech] *Prof. Jan Hajič, UFAL, MFF UK (moderator)*

3 Summary of Content of Sessions

3.1 Welcome and introduction

The workshop began with the presentation given by Professor Jan Hajič, the ELRC Technology NAP for the Czech Republic. Prof. Hajič welcomed the workshop attendees and introduced the Institute of Formal and Applied Linguistics (UFAL) by the Faculty of Mathematics and Physics the Charles University that organized the 3rd Czech ELRC Workshop. The Institute's research topics with the relevance to the workshop were presented: machine translation (projects funded by European Union: EuroMatrix, EuroMatrixPlus, T4ME, QTLeap, QT21, Faust, HimL, ELG), language technology research and development (CUBBITT translation system featured in Nature Communications), language resources (Prague Dependency Treebanks, Universal Dependencies, 300+ other datasets), and data collection and distribution via the Research Infrastructure LINDAT/CLARIAH-CZ hosted at the Institute.

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

The next presentation was given by Dr. Jan Kleindienst, co-founder of two European technology start-ups that are focused on building personalized, reusable and sustainable AI for businesses and individuals (MAMA AI and TELMA AI). Former head of IBM Watson AI R&D Lab with 25 years history of inventing advanced speech and conversational solutions for many global customers, Dr. Kleindienst is an enthusiast and seasoned professional in managing international research and engineering around AI, natural language processing, speech recognition & synthesis, and conversational technologies. With a Ph.D. in Mathematics and Computer Science from Charles University in Prague, he has co-authored 40+ international patents and many scientific publications around AI. He is a co-founder of the Platform for Artificial Intelligence at the Czech Confederation of Industry and the vice-chairman of the Research Board of the Technology Agency of the Czech Republic.

The presentation focused on the state of the art of language technologies and AI and their practical applications. Dr. Kleindienst started his talk by providing some statistics on enterprises using artificial intelligence per European country – counting enterprises with at least 10 people employed in AI, excluding the financial sector. The highest percentage (23%) of such enterprises are in Ireland, whereas the Czech Republic is placed 14th with only 6% of enterprises using AI. As from the perspective of the areas where the AI is used – analysing big data internally using natural language processing (NLP), generation or speech recognition, or using service robots, chatbots and automated chat services – the rate is even smaller, that is only 1-2% of enterprises use AI this way. The data is from 2020.

Despite significant progress in the NLP field over the past several decades, grasping the human language by artificial intelligence remains a hard problem (AI-complete actually). During this talk Dr. Kleindienst discussed practical achievements and challenges of applying AI to NLP in real-world settings, including Conversational AI, AI for IT Operations, neural text-to-speech, bioinformatics – from data collection to model building. Also selected emerging trends that may lead to further advances in understanding our language by AI were mentioned.

Challenges and trends in applied NLP & AI

The key parameters of success on the market of applied AI to NLP are: time to market, cost of deployment, cost of maintenance.

The current key challenges in the area of NLP and AI discussed during the presentation are:

- **Wider automation** of the entire process of "producing" the AI (the ML data lifecycle) – from the data collection to training the models, its testing and deployment. A special importance has the area of data governance that overlaps the entire process of AI application.
- **Better integration** to existing customer infrastructure and processes – integrating with the customer data, processes, and information system is where the application of AI adds value, while the most challenging part is the time to implement such solutions for the customer. Since this time is not negligible, it usually represents the biggest barrier for companies to implement their own AI solution.
- **Zero/low-code tooling** to manage the implemented solution – an important aspect related to the previous point. Company's end-users prefer simple-use tooling to manage their AI solution that can be easily understood and does not behave like a black box to them.
- **Efficient testing**, monitoring, alerting – it is important to be able to efficiently test the implemented solution. Perceived inconsistent behaviour impacts the users' trust in the solution. To gain and maintain trust, it is recommended that testing is meticulously performed.

Current trends in applied NLP & AI discussed during the presentation are:

- **Cross-domain transfer learning** – one of the broader challenges is the cross-domain transfer learning is to apply the ML/DL techniques and approaches from one domain to another (e.g. inspiration in NLP from bioinformatics and vice versa).
- **Cross-model learning** – ability to learn from several modalities at the same time (text, video, audio).
- **Applied ethics** – privacy, explainability, accessibility, de-biasing. How to collect data for learning, storing and generating new data, and how to efficiently protect it (sensitive information, personal information).
- **Circular AI, Sustainable AI** – how to address the ever-increasing size of NLP models (resource demand and operations-wise), centralization of large NLP models (requirements on data, equipment, and energy limit their accessibility), circular AI economy (operational models allowing sharing, leasing, reusing AI – to create building blocks to be reused for further applications)

3.3 Language Technologies in the Czech Republic for the Czech language

The presentation on language technologies in the context of the Czech Republic and the Czech language was delivered by Professor Jan Černocký. Prof. Černocký is the Head of the Department of Computer Graphics and Multimedia at the Faculty of Information Technology, Brno University of Technology (FIT BUT). He founded the BUT Speech@FIT research group in 1997 and serves as its executive director. He graduated from BUT (Ing.) and from Université Paris Sud, France (Ph.D.) and was with ESIEE Paris, France and OGI Portland, Oregon, USA. His research interests include artificial intelligence, signal processing and speech data mining (speech, speaker and language recognition). He was Primary Investigator of several national, European and US (DARPA and IARPA) funded projects. He served as co-chair of major speech and signal processing conferences: IEEE ICASSP 2011 in Prague, IEEE ASRU 2013 in Olomouc, and Interspeech 2021 in Brno. At FIT BUT, he is responsible for signal and speech processing courses. Prof. Černocký is a Senior Member of IEEE and member of International Speech Communication Association (ISCA). In 2006, he co-founded Phonexia which is currently one of world's most important players in speech technologies. He is also active in popularization of speech data mining and science in general in the media.

In his presentation, Prof. Černocký focused on text and speech technology capacities in the Czech region. The Czech Republic has probably the world's highest concentration of quality academic labs and innovative companies per capita. This is true not only for the academic sector (important academic labs are at six Czech universities: Charles University, Czech Technical University, Technical University of Liberec,

University of West Bohemia, Masaryk University, Brno University of Technology) but also for the industry sector (both international and locally founded companies). For example, Charles University hosts an ERC-grantee dr. Ondřej Dušek who was just recently awarded with the grant for a project on Next Generation Natural Language Processing, and the Brno Technical University has been recognized among the five most influential organizations in the area of the speech recognition (alongside Google, Facebook, Carnegie Mellon University, and IBM). During his talk prof. Černocký debates the reasons for these "phenomena" from a historical and educational perspective.

One of the reasons mentioned might be the complexity of the Czech language itself and that analysing the language and its structure is taught from the early grades at schools. Linguistics as a scientific discipline has a quite long tradition in our region that spreads almost over a century. To support that, a few notable personalities that defined the speech and text processing R&D in the last decades were mentioned:

- **Vilém Mathesius**, **Roman Jakobson**, **Petr Sgall** and their students who started to form the "structural linguistics" school in 1930s before Noam Chomsky, and which was renewed again after 1990 at Charles University
- **Bedřich Jelínek**, the founder of statistical approach to speech and language processing
- **Hynek Heřmanský**, the author of the anthropomorphic speech processing, proponent of neural networks, author of RASTA-PLP feature extraction
- In recent years **Tomáš Mikolov**, the author of recurrent neural networks for language modelling and embeddings

Some recent success stories of Czech research and industry were presented as well:

- **Seznam.cz,** the Czech internet portal that was created in the 1990s and even back then was known and popular for its full-text search capabilities, or its more recent map-search tool called Mapy.cz that allows for multilingual search by names of locations around the world.
- **Phonexia,** a company that focuses on speech mining and has introduced the first neural network-based models for speaker recognition which are used by companies across the world in more than 60 countries.
- **MAMA AI**, the former IBM Watson R&D group, currently a Czech-based start-up specializing on AI-powered dialogue systems.
- **Parrot,** a start-up company founded by the BUT alumni that specializes on court report transcription and is very successful on the US market.

Despite the success of personal voice assistants, automatic translators and "AI" applications in general, speech and NLP research is far from complete – the current open challenges are worth mentioning: multilingualism, robustness, far-field microphones, as well as linking speech, NLP, machine translation and other downstream tasks in end-to-end systems.

Some of the possible near-future applications include contact centres with multilingual voicebots and chatbots, agenda of offices for foreign nationals (reception, management and integration of foreign nationals), or national security, intelligence and defence. However, to truly advance in the local context, it is necessary to cooperate with other big players in the field and share experience and resources.

More support to advance in the language technology and AI is still needed. It starts with acknowledgement that these are long-haul research topics that should be considered as key areas for EU economic development. As well as ongoing support for related projects and necessary infrastructure, and legislation that allows for using data for training AI models are also crucial for further progress.

3.4 The CEF AT platform

The eTranslation platform was introduced by Mr. Francois Thunus, from the Commission's Directorate-General for Translation (DGT). Mr. Thunus started as a linguist and interpreter at the Court of Justice, then moved on as informatician in machine translation at Systran. Later on, he switched to the Publication Office as head of IT team, and joined DGT two years ago for the speech project.

The presentation was focused on the European Commission's Automated Translation platform (eTranslation) and NLP tools (https://language-tools.ec.europa.eu/) which are offered to European public administrations, local and regional authorities, small and medium-sized enterprises (since March 2020), EU Freelance Translators, universities, non-governmental organisations and Digital Europe Programme projects.

The history of EC's automated translation dates to the 20th century and started with a rule-based system. In circa 2000, the EC switched to an open-source system called Moses (a statistical system, data mostly from legal texts). In 2013 the Commission decided to switch to a system based on a different approach – that is machine translation using neural networks – called eTranslation. Recently, with the help of CET.AT (Connecting Europe Facility programme), more tools were added (not only translation tools), whereas supporting actions for data collection (ELRC) and other funded projects were possible. The initiatives and supporting actions within the CEF programme are to be continued within a new programme called Digital Europe.

The CEF eTranslation services can be accessed via web upon registration (EU login needed) or via API.

Services currently available through the language tools website: eTranslation, Multilingual Tweet, Speech-To-Text (currently English, French, German, and Spanish), NLP Tools, Interactive Terminology for Europe, European Language Resource Coordination (ELRC), and services to integrate the eTranslation with an organization's online services.

Translation is available for all 24 European languages and Russian, Mandarin, Japanese, Turkish and Ukrainian. The goal is to add more non-EU languages in the next phases.

There are two important features that differentiate EC's eTranslation services from other, commercial services:

1. **Confidentiality** – whatever goes through the system is safe, meaning the EC is not using the input provided for translation in any other way. In addition to that, the majority of the services are running from the Commission's data centers (which is also one of the reasons the services are not offered to the wide general public).
2. **Domain coverage** – in addition to general text, following domains are included: EU formal language, Court of Justice Case Law, Cultural, Deutsche Bundesbank, IP Case Law, Ministère des Finances (France), Public Health, Technical Regulation Information Systems, Valtioneuvoston Kanslia

The best eTranslation results can be expected for texts related to EU policies. Translation of non-standard, new or creative text, single words or expressions, and basically anything that highly depends on the context has usually lower quality.

In the next phases of developing and broadening the eTranslation services the plan is to extend domain coverage (e.g. adding scientific text, social media), language coverage, and add more language technologies (add more languages for Speech-to-Text, anonymization, and a basic CAT tool).

3.5 The new Digital Europe Programme and the Language Data Space

The new Digital Europe Programme and the latest language technologies deployment efforts of the European Commission were introduced by Mr. Philippe Gelin from DG CONNECT. Within the European Commission, Mr. Philippe Gelin is responsible for the Multilingualism sector within the Directorate General – Communications Networks, Content and Technology (DG CONNECT), where he develops European wide policies including the research and deployment funding programme related to language technologies.

Language technologies progressed extensively during the past few years – mainly due to steady progress of AI, CPU, and data collection. The tools and their deployment are becoming cheaper and faster, and more available to general public (e.g. via mobile devices). Language technologies can unify people across different languages. In the online world, many European languages are under-presented. In order to support more diversity and unification, European Commission's tools include: supporting legislation, coordination of actions, and funding.

The new **DIGITAL programme** started in 2021 – it is composed of **five strategic objectives (SOs**):

- SO1 high-performance computing
- SO2 cloud, data and AI
- SO3 cybersecurity
- SO4 advanced digital skills
- SO5 best use of digital technologies

The DIGITAL programme is not a research programme but rather a deployment programme to support mainly the industry. Its focus is on building strategic digital capacities of the European Union and facilitating the wide deployment of digital technologies. The overall programme goal is to bridge the gap between digital technology research and market deployment, while supporting the EU twin objectives of a green transition and digital transformation.

Common European Data Spaces – the data spaces are the centrepiece of the European strategy for data. The vision is to securely collect the data that is being used by organizations across the EU and make it back available for exchange. The data spaces are meant to harness the value of the data, to overcome existing legal and technical barriers, and allow for data-driven innovation.

The idea is to create the data spaces to be sector specific. A specific data space for languages is planned as well. Since text, audio and video files represent circa 50% of data that is produced daily, in order to use them (process and extract relevant information from them), complex AI-based language services are needed. In order to develop such services, a big amount of data that is aggregated, organised in a comprehensive way is needed again to train the services. Thus, the data space will focus on collection, creation, sharing, and re-use of language data and models. The sectoral data spaces will be built on the European Data Space Technical Framework (consisting of Cloud-to-Edge Infrastructures and Services, including smart middleware solutions, marketplace) and will connect to the AI on-demand platform (that will connect providers and consumers of the AI resources) and TEFs (= Testing and Experimentation Facilities) (for experiments with AI-based solutions in real-world environments).

The language data space will be built based on several design principles:

- The owners of the data need to keep control over it (including licensing and price)
- Governance framework that is defined as fair, transparent, non-discriminatory
- Respect of EU rules and values – personal data protection, consumer protection legislation, competition law
- Technical data infrastructure – usage of defined building blocks, coordinated construction of the data spaces

- Interconnection and interoperability – the goal is to progressively interlink the data spaces with the aim to create a fully integrated ecosystem to allow for the exploitation of data across domains
- Openness – for all actors.

A Centre of Excellence for Language Technologies (CELT) will be established in order to facilitate the deployment of the LDS. On a strategic level, a CELT to coordinate the efforts of collection and creation of multimodal language data and models across the EU member states will be established, making use of existing EU initiatives and language data collections such as ELRC, EURAMIS, SCIC repositories, IATE, CLARIN, META-SHARE, ELG, and CEF Automated Translation. The primary function of such CELT will be networking and governance of the LDS. On the operational level, the CELT+ will be established to include also private actors from industry to plan in more details the LDS deployment.

The envisioned architecture of the LDS is based on connecting distributed datasets of participating stakeholders who provide access to their data in exchange for the access to other datasets in the LDS. In order to make the data accessible to users, the LDS will be connected to the AI-on-Demand platform bringing to consumers (users) a variety of additional services of language technologies.

3.6 The use of language technologies at the Czech Republic Supreme Audit Office

The last presentation was delivered by Mr. Jaroslav Rucký from NKÚ (Nejvyšší kontrolní úřad ČR) / SAO (the Czech Republic Supreme Audit Office). Mr. Rucký is the Head of Team for EUROSAI Presidency at the Department of International Relations at the Supreme Audit Office.

The presentation introduced the use of machine translation tools in the normal operation of the Supreme Audit Office (SAO), the activities of the SAO and its international cooperation within EUROSAI (The European Organisation of Supreme Audit Institutions), and some of the auditors' needs from the perspective of translation services and examples of practical issues they face with machine translation.

The main activity of the Supreme Audit Office of the Czech Republic is to review the state's management of public revenue and expenditure. It is a member of INTOSAI (International Organisation of Supreme Audit Institutions) and EUROSAI (one of the Regional Organisations of INTOSAI). Since 2021, the Czech Republic Supreme Audit Office has been presiding the EUROSAI. There are five official languages within EUROSAI (English, French, German, Russian, and Spanish), however, the Organisation has 51 members as of 2022. The permanent EUROSAI secretariat resides in Spain.

The main activities of the Department of International Relation of the SAO are communication with the foreign institutions and translations of documents to Czech – internally used documents are translated by the Department, external professional translation services are used for official reports.

Mainly due to its participation in EUROSAI activities, the SAO work with many documents (official audit reports by member organisations) in various languages on a daily basis. The only effective mechanism to conveniently, quickly, and efficiently translate documents from and to languages of EUROSAI members is to use machine translation, in particular eTranslation. Currently, EUROSAI operates four databases with the official documents and is planning to consolidate all of them into one. At the same time, EUROSAI is preparing a new website with the access to its database together with a service of automated translation for its users.

The SAO is also participating in other translation related projects like: BIEP – Benchmark Information Exchange Project for auditors (https://biep.nku.cz/), or ELITR – European Live Translator, a H2020 project coordinated by the Institute of Formal and Applied Linguistics at Charles University (https://elitr.eu/) that was also used at the 3rd ELRC workshop for the Czech Republic.

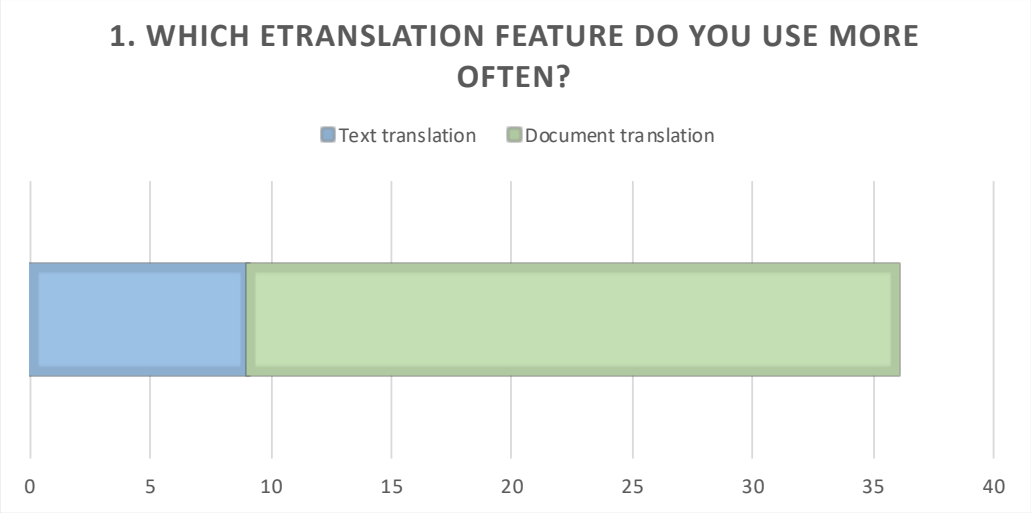
The idea to use eTranslation at SAO emerged from a Czech ELRC workshop that was held in Prague a few years ago, which the SAO representatives attended. The eTranslation tools are used for quick and

comprehensible unofficial translation of official documents, in order to get a general understanding of such document (for official documents, professional translation services are still used). The SAO organized also internal training on use of eTranslation in 2020, over 70 employees attended. The eTranslation is also propagated on the website of the BIEP project to make the information exchange among auditors from different countries more accessible.

An internal survey on eTranslation for the purpose of this presentation was conducted at the SAO, asking the following questions:

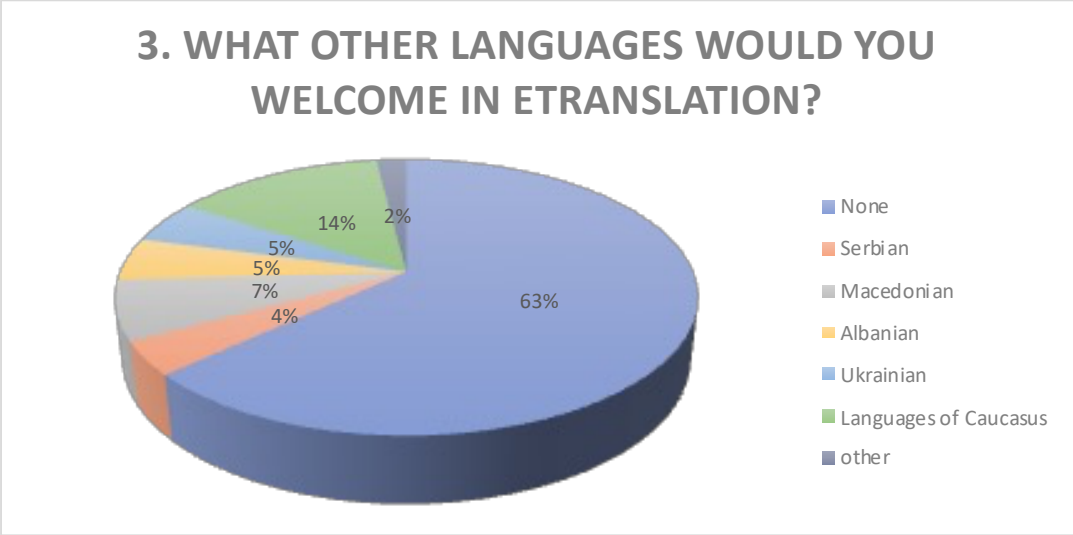
1. Which eTranslation feature do you use more often?
2. From which languages do you most often translate into Czech?
3. What other languages would you welcome in eTranslation?
4. What prevents you (if anything) from using eTranslation more often?
5. What is your overall satisfaction with eTranslation?
6. Do you use other translators to translate shorter texts?
7. What do you like or dislike about eTranslation? (strengths and weaknesses perceived)

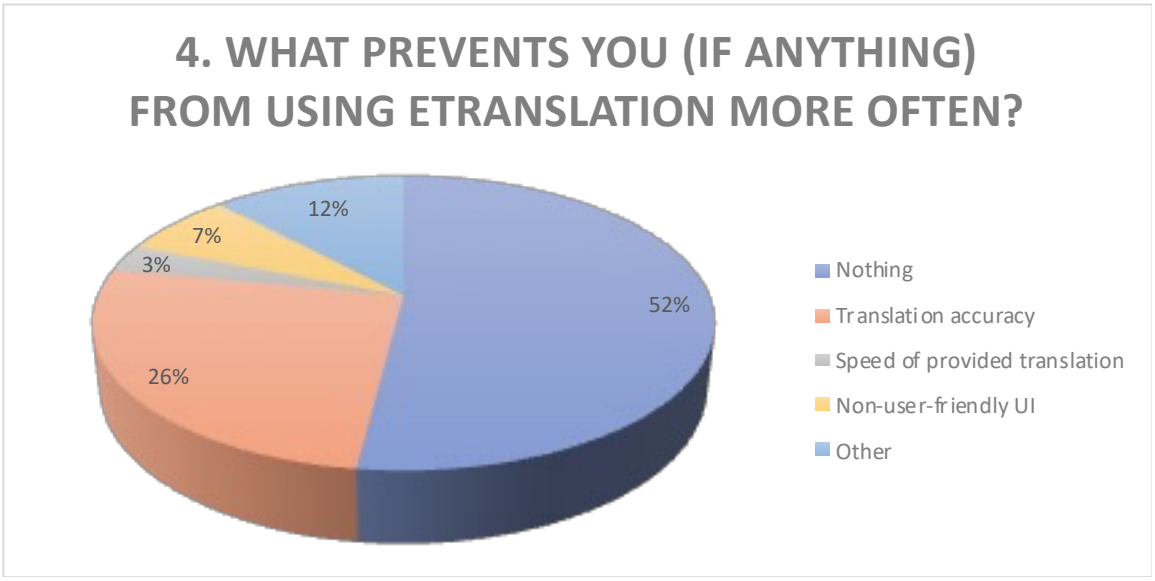
From the 37 participating respondents, the following feedback was collected:

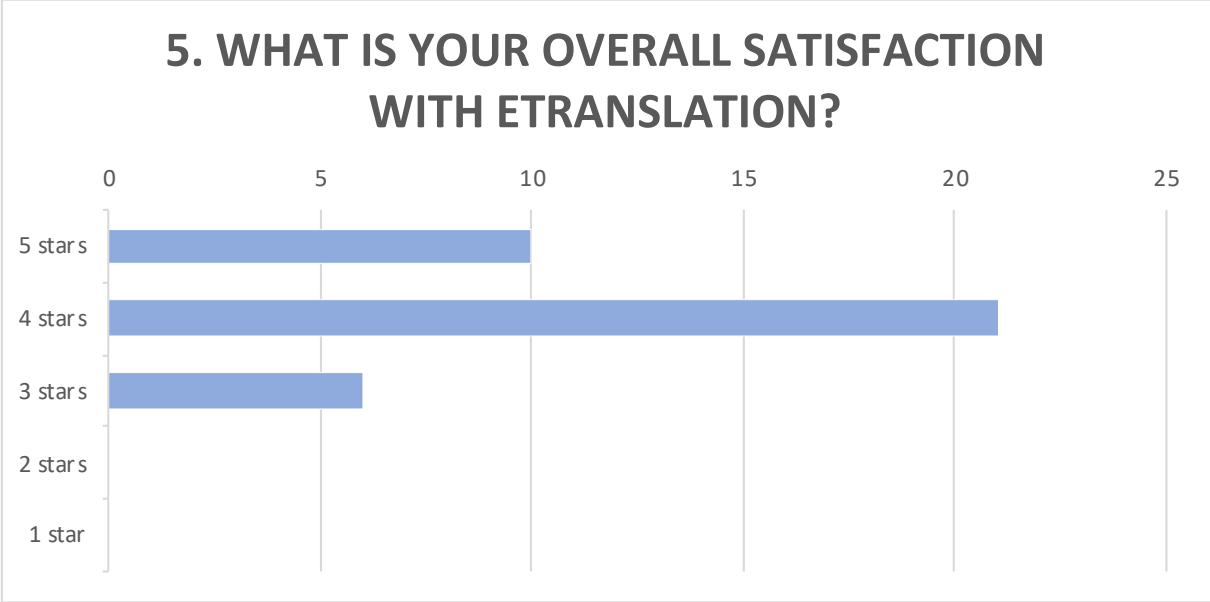


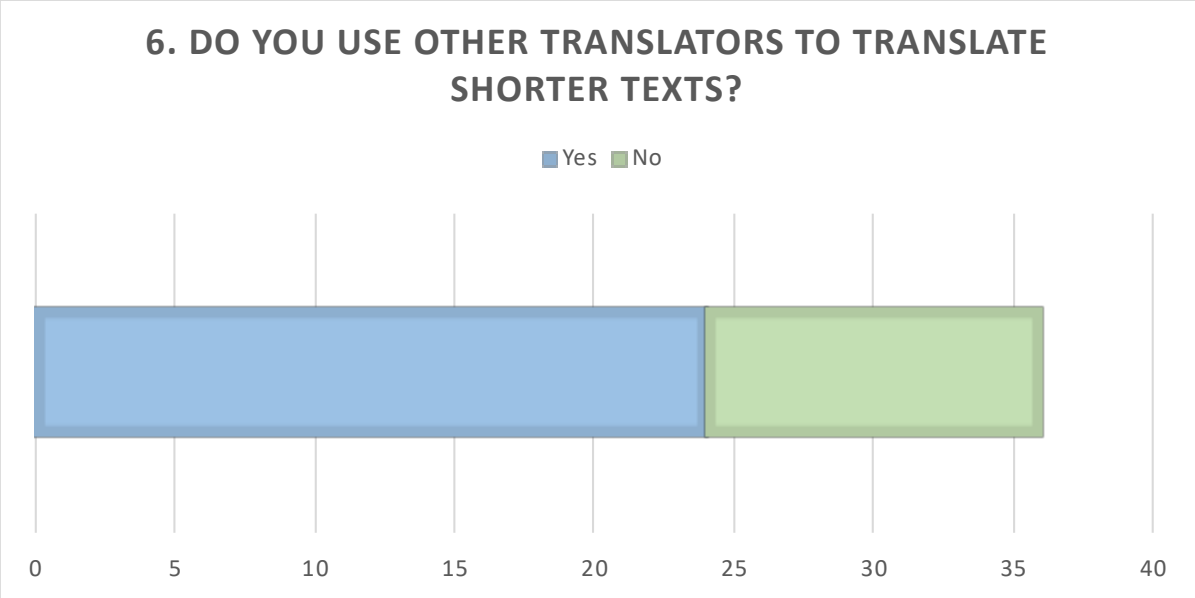
Notes to results: translation is very easy and convenient

Notes to results: missing mainly languages of Caucasus or west Balkan – the vision is to work on inclusion of all countries within EUROSAI, but the language barrier is usually the biggest problem




4. WHAT PREVENTS YOU (IF ANYTHING) FROM USING ETRANSLATION MORE OFTEN?
12%
7%
3%
52%
26%
Nothing
Translation accuracy
Speed of provided translation
Non-user-friendly UI
Other


5. WHAT IS YOUR OVERALL SATISFACTION WITH ETRANSLATION?
0
5
10
15
20
25
5 stars
4 stars
3 stars
2 stars
1 star



Note to results: Yes because of the requirement to log in into the eTranslation (it is more convenient to use an alternative translator for shorter texts that works without login)

Feedback to the last question was summarized in the following list of perceived strengths and weaknesses of the eTranslation:

Strengths of eTranslation:

- Translates entire documents and preserves the text layout
- Translates images and graphs
- Legal terminology (law translation) and legal style
- EU formal language feature
- Translation within a few minutes

Weaknesses of eTranslation:

- Frequent use of synonyms for the same word – it is confusing and undesirable for a terminus technicus (e.g. for auditor, synonyms like checker, controller, inspector, are used)
- Occasional incorrect translation of homonyms – e.g. 've středu' translated as 'Wednesday', sometimes as 'in the centre'
- Nominalization of sentences
- Correct sentence structure (incorrectly identified subject, object, and the actor – especially problematic with person names, or due to declension)
- Non-verbal text elements (footnotes, Roman numerals)
- Idioms, figurative language
- Hyperlinks from the original text are not preserved in the translated text
- Differences in the output translation when using Text translation and Document translation
- Translator is 'confused' when a new line is inserted in the middle of a sentence

3.7 Take-home message and conclusions

The main issues that were mentioned revolved around the quality and accuracy of the translation, ease of use or features of services that are currently available:

- inaccurate translations – typically related to the names, inflexion, figurative language, context, domain
- options to 'teach' the model – how to influence the models for better consistency (example provided by the eTranslation users who would welcome, if the tool could learn from manual corrections of translation by users, e.g. for undesirable use of synonyms)
- ease of use – users of eTranslation prefer alternative tools for short translations that are more convenient/ more easily accessible

The typical technical issues related to the language technology mentioned:

- data – insufficient data to train the models in general, domain-specific – and the problems related to its collection: access to data, sharing, and protection of sensitive data (ethical issues, anonymization)
- better automation and integration with existing systems of organisations
- sustainability – increasing complexity with the requirements on wider and faster accessibility

4 Synthesis of Workshop Discussions

The final discussion and was moderated by prof. Hajič and it summarized a few key topics from throughout the day. The take-home message could be summarized as follows:

- A significant progress has been made in the area of language technologies in recent years, mainly thanks to the progress in technology (AI, computational capacities, data resources);
- The Czech Republic has a long history of linguistics research, and over the years it has developed great expertise in the area of language technology: individual contributors, research groups and labs – both in academia and industry, LT services, resources and infrastructure;
- Despite the use of AI-based tools for machine translation, natural language processing, or speech recognition is on the rise, it has still a long way to go;
- The need for language technology in multilingual Europe is undeniable – tools enabling translation between different languages are necessary to progress in international cooperation, unification and standardization across the countries, and their further development should be recognized as the key areas for EU economic development;
- One of the most valuable assets for LT development is the data;
- To further support research and development for language technology, the ongoing funding for related projects and infrastructure is very much needed, as well as legislation that enables using data (creating, sharing, protecting) for training the AI models.

5 Country Profile: Language data creation, management and sharing

The situation related to LT development, digitalisation and data collection has changed over the last three years. The single biggest difference is the use of large language models, and the transfer of practically all MT development to Deep Learning. In addition, speech technology (ASR/TTS) is used more and more in research – not only at the Institute of Formal and Applied Linguistics but also country wide – as the ASR and TTS quality improves. More data is available in general; for example, in the Universal Dependencies collection the number of languages grew from 70 to 130+ and the number of treebanks to more than 200; for Czech three more treebanks have been added in the past three years. With respect to language resources and tools, the major improvement is a much larger morphological dictionary of Czech (Morfflex CZ), now fully compatible with the PDT-C 1.0 treebank, fully manually morphologically annotated with almost 4M tokens. Some new services have also been implemented and made freely available in the LINDAT/CLARIAH-CZ research infrastructure services list. In the same infrastructure, many Digital Humanities and Arts (DHA) datasets have been made available (such as digitized news clips and speeches released by the National Film Archive). A nationwide catalogue of language and DHA metadata and data is being created and will be available soon. It includes catalogued data from all national libraries in the Czech Republic. However, only some of them are available for LT development because of copyright issues.

The following action items would be most relevant to facilitate data sharing:

- To tackle legal problems – current legislation mentions language data marginally without any detailed plan (finance-wise including)
- Raising awareness of language data as open data and valuable asset
- Increasing interest in MT/ LT in public services and SMEs as part of the national digital policy
- Establish good data management practices in public services and SMEs
- Identify and gain access to outsourced translations

Availability of data in the Czech Republic is still the biggest concern, also in connection with the fact that the 2019 CD has not been transformed into the national legal system yet. Also, the size of Czech resources, even if covered quite well overall, is still well below the major languages. However, the quality of MT (EN-CS specifically) has improved dramatically in the past three years.

The overall objective is to have methods, algorithms and ready-made system(s) for full Natural Language Understanding. Whether it is done by Deep Learning alone or in combination with symbolic methods and/or databases is not that important, but data is certainly important. Identifying gaps in technology and data is the next important goal. It is still not clear which applications are possible now and in the next decade with current technology, or which improvements are possible with incremental development, and which will need breakthroughs. Availability of high quality, clean data is the next big thing. Czech needs to have larger Language Models than those currently available (BERT, GPT-like, TMs).

In order to secure sustainable development of language technologies on the national level, funding should be available long-term with longer perspective, which is currently not the case in the Czech Republic, since all research, including infrastructural support, is project-based only. Nationwide public funding in the form of a language programmeme, like the one in Spain, would be a relevant approach.