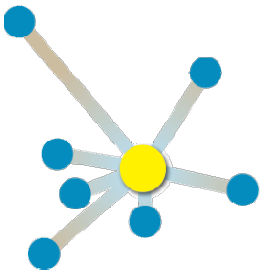
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1091 programme.



European Language Resource Coordination

Connecting Europe Facility

Deliverable D3.2.20
Task 8

ELRC Workshop Report for the Czech Republic

Author(s): Jan Hajič, Pavel Pecina

Dissemination Level: Public

Version No.: <V1.0>

Date: 2019-01-10

Contents

1 Executive Summary

This document reports on the 2nd ELRC Workshop in Czech Republic, which took place in Prague on Nov. 28, 2018 at the Faculty of Mathematics and Physics, Computer Science School in the Lesser Town in Prague 1. It includes the agenda of the event and briefly informs about the content of each individual talk (including those presented remotely from the EC and panel workshop session. The event was attended by 34 participants spanning a wide range of ministries and public organisations, as well as some freelance translators. The dedicated event webpage can be found at http://www.lr-coordination.eu/l2czech.

The workshop has been organized by the Institute of Formal and Applied Linguistics, Computer Science School, Faculty of Mathematics and Physics, Charles University in Prague. There was a great help by the Czech DGT representative, who helped secure a place for the workshop in the Europe House in the center of Prague. The organizers invited several new speakers (compared to the first workshop in 2015), as well as some speakers from then to the panel to get a time-development perspective. The highlight of the workshop was the panel, which offered perspective from both the Czech government (Mr. Hrabě, Ministry of Interior, responsible for the Czech government Digital agenda) as well as from the Czech parliament (house) (Mr. Michálek, Member of Parliament, proponent of open access to public data).

2 Workshop Agenda

2nd ELRC Workshop in the Czech Republic

08:00 – 09:00 Registration
09:00 – 09:10 Welcome and introduction
Jan Hajič, Institute of Formal and Applied Linguistics, Charles University, ELRC Technology Anchor Point
09:10 – 09:15 Welcome by the EC
Jan Faber, DG Translation, Prague

Session 1. Connecting a multilingual Europe: European context and local needs

09:15 – 09:35 Connecting public services across Europe: ambition and results so far
Aleksandra Wesolowska, Directorate-General for Communications Networks, Content and Technology, European Commission
09:35 – 09:55 National initiatives for digital public services and (open) data
Pavel Hrabě, Ministry of the Interior of the Czech Republic
09:55 – 10:40 CEF in the Czech Republic: an outlook into current and future challenges – Panel session, moderator: ELRC Public Services NAP (Jan Hajič)
Panellists:
Jiří Kotouček, National contact point (NCP) for legal issues of the Framework Programme Horizon 2020, Technology Centre ASCR
Jakub Michálek, MP, House of Parliament of the Czech Republic
Jan Faber, DG Translation, Prague
Stelios Piperidis, ILSP / Athena R.C, CEF/ELRC
10:40 – 11:00 The CEF eTranslation platform @ work
Szymon Klocek, Directorate-General for Translation, European Commission
11.00 – 11:30 Coffee Break

Session 2. Engage: hands-on data

11:30 – 11:55 The European Language Resource Coordination (ELRC) action
Stelios Piperidis, ILSP / Athena R.C, ELRC
11:55 – 12:15 ELRC in the Czech Republic
Jan Hajič, ELRC Technology Anchor Point
12:15 – 12:40 Can language data be shared and how? National and European legal framework
Jiří Kotouček, National contact point (NCP) for legal issues of the Framework Programme Horizon 2020, Technology Centre ASCR
12:40 – 13:40 Lunch Break
13:40 – 14:15 Preparing and sharing data with the ELRC repository – and what happens next
Maria Giagkou, ILSP / Athena R.C, ELRC
14:15 – 14:45 Identifying and managing your data: Questions & Answers
Moderator: Jan Hajič
14:45 – 15:00 Conclusions (Jan Hajič)
15:00 – 15:30 Coffee Break and networking

3 Summary of Content of Sessions

3.1 Welcome and introduction

Mr. Jan Hajič, ELRC Technical NAP in the Czech Republic, of Charles University, and Stelios Piperidis, the ELRC representative for the region, opened the event by welcoming the audience and introducing the key persons in organizing the event, namely the ELRC consortium and the local Czech EC/DGT representative, and also introduced Pavel Pecina, who took responsibility for organizing the event this year.

3.2 Welcome by the EC

Mr. Jan Faber, the local representative of the DGT in the Czech Republic and the local host of the event (in the Europe House, Jungmannova 24, Prague 1) welcomed the participants. In his talk, he stressed the usefulness of eTranslation (formerly MT@EC) and stressed the importance of quality translation in the public sector and in the communication between the EU, specifically the European Commission and the European Parliament and all DGs. He also cited certain challenges of the use of automatic translation and stressed the education aspect for new generation of translators, not only at the university level.

3.3 Connecting public services across Europe: ambition and results so far

Ms. Aleksandra Wesolowska, DG Connect, EC, presented her message regarding eTranslation in the context of CEF by videoconference. She has stressed that the data availability is a key point on making the service of high quality to gain widespread use. She has described the process of creating the service and running it for the EC and EP as well as for all public services at all levels of administration in the Member States.

3.4 National initiatives for digital public services and (open) data

Mr. Pavel Hrabě, Ministry of the Interior of the Czech Republic, presented the “Digital Czech” agenda of the Czech government for 2018+. He pointed out that the government provides 700+ services to the public (both to organizations as well as directly to citizens). While some of them are already provided in digital form to a certain extent, they have to be upgraded to cover more services, to allow for third party applications, to integrate data across all branches of the government, and to give access to open data whenever possible. He has also shown the interconnection of “Digital Czech” programme, as one of the four Pillars to the government digital agenda, to the other three - namely the “Information strategy of the Czech Republic”, the “Digital Economy and Society” (a.k.a. “Society 4.0”) agenda, and the international cooperation related “Czech Republic in the Digital Europe” agenda. The main focused goal of the initiative is to make online services efficient and user-friendly for citizens to use directly. The broad impact and long/term sustainability is part of the “Digital Economy and Society” pillar, which covers all aspects of government involvement in the digitization process, from legal aspects to direct support to research, development and innovation in the economy.

3.5 CEF in the Czech Republic: an outlook into current and future challenges – Panel session

Moderator: ELRC Technical NAP (Jan Hajič)
Panellists:

- Jiří Kotouček, Technology Center of the Czech Academy of Sciences, National contact point (NCP) for legal issues of the Framework Programme Horizon 2020
- Jakub Michálek, deputy (member of parliament), house of the parliament of the Czech Republic
- Jan Faber, DG Translation, Prague
- Stelios Piperidis, ILSP / Athena R.C, CEF/ELRC

Mr. Jan Hajič introduced the panel and raised four questions the panellists were requested to address:

a) what is the most important topic (from the individual panellist's and his institutional environment point of view) in the area of public services digitization effort,
b) which technologies can help to achieve this goal and what are the biggest obstacles (legal, technical, human-resources-related),
c) please explain CEF and its task(s) in the context of public services
d) have you used some of the CEF DSIs, specifically, eTranslation (and what is your experience).

The panellists briefly stated their points (obviously, not all questions have been addressed in full by all panellists – e.g. the CEF questions has been most thoroughly addressed by Stelios Piperidis as the ELRC representative); quite open and lively discussion followed, with a very active audience. As a reaction to Mr. Hrabě's morning presentation, Mr. Michálek (MP) asked which of the many services are a good example (or use case); Mr. Hrabě reacted that in his view, none of them is really a modern, fully functioning service with all aspects that such a service should have. However, it has been mentioned that there are at least several services in the area of taxation, for both individual taxpayers as well as companies that do exist and are in widespread use. Additionally, the citizen ID service (called "mojedatvaschranka" -"mydatamailbox") administered by the government-run Czech Post Office, despite of all its shortcomings, does a good job in communication between government offices and the citizens, even though its ID capability is severely underused. Some people from the audience mentioned that they do have data (translations) to transfer to ELRC, but that legal aspects will have to be checked due to various legacy issues.

3.6 The CEF eTranslation platform @ work

Szymon Klocek, representing the Directorate-General for Translation, European Commission, presented the eTranslation platform via a live video link. Starting with a short retrospective he explained that eTranslation is the successor to MT@EC, and it is a cloud-based neural machine translation system. The CEF Automated Translation building block is a platform which includes eTranslation but is also intended to integrate other NLP tools (e.g. named entity recognition, transliteration, etc. based on user needs). Related projects and grants (e.g. ELRC, Translator for the Council presidency, generic services projects) are also under the umbrella of the CEF Automated Translation building block.
The eTranslation platform is available for individuals and machine-to-machine use. Users are Digital Service Infrastructures (EESSI, ODR, Open Data Portal, Europeana, etc.), System suppliers (EURLex, N-Lex, Internal Market Information system) and individuals in public administrations. Benefits of using

the platform are: increase speed and productivity, reduce costs and facilitate information exchange. The eTranslation user interface is available since November 2017. Some of its features include:

- User interface – All EU Languages,
- Supports common document formats (Office formats, ODT, PDF, html, xml),
- Quality output

Mr. Klocek went on to discuss the recent shift to neural machine translation (NMT). Large Artificial Neural Networks are trained on existing translations. NMT training induces hidden representations for textual data and exploits hidden generalisations. It constitutes a radical departure from the phrase-based SMT approaches. NMT can generate more fluent and grammatical translations and it improves its performance on previously unseen data. CEF.AT brings reliable and trustworthy translations for EU and National Public Administrations, support for lesser resourced languages with fewer speakers, opportunities for the private sector through grants, higher quality language technologies, thus fostering demand and public availability of data collected by ELRC.

3.7 The European Language Resource Coordination (ELRC) action

Mr Piperidis presented the ELRC, an action funded by CEF. CEF is a key EU funding instrument that supports the development of high performing, sustainable and efficiently interconnected trans-European networks in the fields of transport, energy and telecommunications. The Telecom strand, among other things, funds the deployment of digital service infrastructures (DSIs) – this part of CEF Telecom is called CEF Digital. One of these Digital Service infrastructures is eTranslation. ELRC supports the eTranslation DSI, by coordinating the collection of language resources that are necessary to enhance the system.

The ELRC Consortium is made up of:

- DFKI – the German Research Centre for Artificial Intelligence;
- ELDA – the Evaluations and Language Resources Distribution Agency;
- TILDE – a Latvian Language Technology and Services Provider
- ILSP – the Institute for Language and Speech Processing.

Additionally there are 30 ELRC Technological NAPs (one per CEF affiliated country), 30 ELRC Public Services NAPs (one per CEF affiliated country) and relevant legal experts.

The aims and objectives of the ELRC action include:

- collecting language resources;
- identifying public services in need of multilingual functionalities and their corresponding multilingual needs;
- engaging the public sector in the identification and continuous sharing of such language resources;
- helping with legal and technical issues associated with the collection and/or provision of language resources;

- acting as observatory for language resources across all EU Member States and CEF-affiliated countries.

Mr. Piperidis went on to explain that the way to enhance the performance of eTranslation is with the corresponding "training data" that is "fed" into the system (i.e. language resources, such as bilingual corpora, multi-lingual corpora, terminologies, etc.). In-domain training data (i.e. translations from the target domain) are essential for achieving high-quality translation.

3.8 ELRC in the Czech Republic

Mr. Jan Hajič, ELRC Technology NAP for the Czech Republic, explained the task of the NAPs in the ELRC project, and the task of identifying and collecting data from public sources for the CEF eTranslation development (and beyond). He presented the Institute of Formal and Applied Linguistics and specifically, the Research Infrastructure it hosts (LINDAT/CLARIN), as the institutional anchor point for such a collection, which then can pass the data on to ELRC. LINDAT/CLARIN hosts many resources itself and provides also many software language technology tools and remote services, accessible both programmatically as well as by human users for direct use, for example for machine translation. He emphasized cooperation with other European and overseas partners on the development of language technology (not only automatic translation) and the need for data for such development.

3.9 Can language data be shared and how?

Presented by Mr. Jiří Kotouček, Technology Center of the Czech Academy of Sciences, National contact point (NCP) for legal issues of the Framework Programme Horizon 2020.

In his talk, Mr. Kotouček summarized all aspects of copyright as applied to textual (and other) data of interest to both developers and users of language technology, including the specific database rights. In the context of ELRC, it was important that he specifically explained the PSI directive, which is already part of the Czech legal system even before the CR has joined the EU (since 1999); he has also explained the new proposal under preparation, which should extend the directive to public transport and research data. The Open Research Data Pilot from H2020 and its successive regulation in Horizon Europe was also explained. Data sharing under various circumstances was been explained next, as well as the relevant issues regarding GDPR, another important aspect of many LT datasets.

3.10 Preparing and sharing data with the ELRC repository – and what happens next

Ms Giagkou started her presentation with a discussion of the value of data. She emphasized that, like numerical data, language data are a valuable asset as well, especially for language technologies, such as machine translation. She went on to explain the types of data that are useful for training eTranslation, i.e. any piece of text in a natural language and its equivalent in another language. Data residing in local public organisations, produced in-house or outsourced, e.g. Reports, Communication, News, Web Content that is managed for several languages, Policies, Terminologies, Archives, Forms, and FAQs are useful for improving the platform. She explained ways to provide language resources and appropriate formats about this. She also presented the basic steps through which language resource

can be shared through the ELRC-SHARE repository. Dr. Giagkou provided examples of processing services applied to the contributed data in order to convert them to MT-ready training datasets and invited the participants to use the freely available ELRC of technical and legal helpdesk and on-site assistance.

3.11 Identifying and managing your data: Questions & Answers

Mr. Jan Hajič engaged the audience in a discussion about possible data contributions by the institutions whose representatives were present at the event. It became clear that there are various collections of texts available, but that the legal issues involved might not be easy to solve. The representative of the Czech Social Security Office, who has already provided some data in the past, as well as representatives of translation agencies were actively engaged in the discussion, raising mainly the issue of the conditions of use of the eTranslation service, especially in cases where freelance translators are working for the public sector. Some part of the discussion also touched upon the possibility of providing automated translation services directly to the LSPs by research departments at the universities (on commercial basis). Part of the discussion also involved the required data processing to convert raw language data into MT-ready datasets, that could not only feed the eTranslation system, but that could also reach research communities for developing high quality automatic translation.

3.12 Conclusions

The previous session flowed naturally into conclusions; Mr. Jan Hajič thanked the ELRC representatives for their help with the organization of the workshop, as well as to Mr. Faber, as the local DGT representative, and Mr. Pavel Pecina for actually organizing the event. He also thanked all the speakers and panellists for their time and engagement, and he closed the workshop.

4 Synthesis of Workshop Discussions

In the Czech Republic, the situation has not changed much since the previous workshop in December 2015 in terms how translation services are organized in central public government offices. There is no central repository or translation office that would be shared by the ministries; nor have the ministries themselves specifically committed to managing and performing translations, except for two. The rest simply contract LSPs, whenever a need arises, or have them contracted under a framework agreement (since public money have to be tendered).

However, as opposed to the situation two years ago, the current government is interested in the Digital Agenda, and several initiatives exist, such as the "Digital Czech Republic" (presented by Mr. Hrabě of the Ministry of Interior). Whereas the public services are not controlled centrally yet, which is reflected in the fact that the few fully functioning digital services that are in place do not meet high standards when compared to more advanced Member States in terms of digitization of the government and public sector agendas, as has the panel discussion revealed.

In addition, an important point from the ELRC and CEF/eTranslation point of view is that commercial LSPs cannot use the eTranslation service even if they translate for the public sector. This fact might prevent the public sector bodies themselves to use eTranslation, in part because inconsistencies could arise if only part of the translations are performed through it, while other parts are done solely through the LSPs. This is an important issue raised in the discussions by the LSPs as well as by the representatives of the Ministries and other public service institutions, such as the Czech Social Security Office.

The existence of the Open Data Portal, introduced recently, also raises hope that more (textual) data will be available soon, both for ELRC and in general for the development of language technology applications suitable for the public sector in the Czech Republic.

4.1 ELRC and Open language Data in the Czech Republic

The Open Data Portal of the Czech Republic is available at https://data.gov.cz/. However, there are almost no textual data (except for textual fields in various otherwise mainly numerical datasets, such as various code tables, election results, census data etc.). However, the portal openly solicits people and organizations to share data under open licenses; thus there is hope that truly textual data will appear in the future and that they will be available for both research and innovation in language technology services and software tools.

The country is fully compliant with the EU copyright-related issues. The PSI directive has been adopted in 1999 (Law 106/1999 Sb.), and the Copyright Law (Law 121/2000 Sb., as amended) contains standard definitions and exceptions, including a broad research exception (included later in one of its amendments), which in fact might be jeopardized if the current Art. 3 and Art. 3a of the new EU Copyright Directive is adopted as proposed. The Copyright Law also defines what is not subject to copyright law, which is important for the distribution of various language-technology-related models for various language analysis and synthesis tasks, which are typically not subject to copyright law. It also fully respects the PSI Directive (embodied in another Law, as referenced above).

4.2 Success stories and lessons learnt

Highlights of the workshop include:

- situation in public sector data (translated texts) availability has not improved and there is very little data available and clear of copyright issues, exemptions aside (Czech Social Security Office);
- the use of eTranslation by commercial entities when serving the public sector should be considered and conditions sketched to be possibly adopted for such use; the practical workflow of the government offices and the central ministries in the Czech Republic would be better supported by such arrangement;
- centralization of the Digital Agenda of the government at the Ministry of Interior seems to be a step ahead in managing and advancing the country's digitization of public sector services;
- as has been mentioned by the MP in the panel discussion, creating a best practice example (i.e., an example application with all the attributes of open access, technical parameters, and efficient use) would be an important achievement to promote such level of digital service.

5 Workshop Presentations

The workshop presentations are available at http://www.lr-coordination.eu/l2czech_agenda.