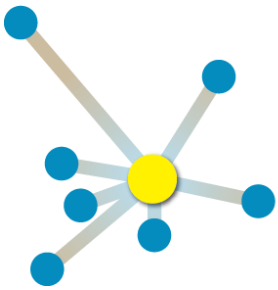
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.



European Language Resource Coordination

Connecting Europe Facility

ELRC Workshop Report for Czech Republic

Author(s): Jan Hajič (Charles University in Prague, MFF)
Kateřina Bryanová (Charles Univeristy in Prague, MFF)

Dissemination Level: Public

Version No.: <V1.1>

Date: 2016-03-02

Contents

1 Executive Summary

This document reports on the ELRC Workshop in Czech Republic, which took place in Prague, on the 15th of December 2015 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, interactive and panel workshop session. The event was attended by 35 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://lr-coordination.eu/czech.

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, and by the Government Office. Help was also provided by the Institute of Translatology, at the Faculty of Arts at Charles University in Prague.

2 Workshop Agenda

ELRC Training Workshop - programme, 15.12.2015

Prague, Czech Republic

08:00 – 09:00 Registration

09:00 – 09:05 Opening and Introduction (Jan Hajič, ÚFAL MFF UK, Stelios Piperidis, ELRC)

09:05 – 09:15 Welcome by local representatives (Vítězslav Zemánek, DGT in CR, Marie Černíková, MEYS CR)

09:15 – 09:25 Project goals (Stelios Piperidis, ELRC)

09:25 – 09:40 Europe and multilinguality (Vítězslav Zemánek, DGT in ČR)

09:40 – 10:10 Language and language technologies in the ČR (Tomáš Svoboda, ÚTRL FF UK and Jan Hajič, ÚFAL MFF UK)

10:10 – 10:50 Panel 1: Translation in Public sector in the Czech Republic (moderator: Tomáš Svoboda,ÚTRL FF UK), other panel members:
Marie Černíková, MEYS CR
Vítězslav Zemánek, DGT in CR

10:50 – 11:20 Coffee Break, Networking

11:20 – 12:00 Automated Translation – How does it work? (Ondřej Bojar, ÚFAL MFF UK)

12:00 – 12:30 How can public institutions benefit from CEF.AT? (Daniel Kluvanec, EC DGT)

12:30 – 13:30 Lunch Break, Networking

13.30 – 14.00 What data are needed in Machine Translation? (Ondřej Bojar, ÚFAL MFF UK)

14:00 – 14:30 Legal framework (Jakub Michálek, Deputy, City of Prague)

14:30 – 15:00 Panel 2: Data and Language Resources in the Czech Republic and in Slovakia (moderator: Jan Hajič), other panel members:
Tomáš Svoboda (ÚTRL FF UK)
Daniel Kluvanec (DGT EK)

15:00 – 15:30 Coffee Break, Networking

15:30 – 16:00 Data and LRs: practical aspects and “best practice” (Stelios Piperidis,ELRC)

16:00 – 16:30 Interactive: how can we engage? (Jan Hajič, ÚFAL MFF UK)

16:30 – 16:45 Wrap-up and next steps (Stelios Piperidis, ELRC, Jan Hajič, ÚFAL MFF UK)

3 Summary of Content of Sessions

3.1 Session 1: "Opening and Introduction"

Stelios Piperidis, the ELRC representative for the region, and Jan Hajic of Charles University in Prague, the local ELRC Technology National Anchor Point representative, opened the event by welcoming the audience and introducing the key persons in conceiving and organizing the event, namely the ELRC consortium and the EC/DGT representatives.

3.2 Session 2: "Local representatives' welcome"

The local representative of the DGT in the Czech Republic, Mr. Vitezslav Zemanek and the Public Sector National Anchor Point of ELRC in the Czech Republic, Ms. Marie Cernikova, welcomed the participants and stressed the importance of translation and quality translation in the public sector and in the communication between the EU, specifically the European Commission and the Parliament and all DGs, and the state governments and relevant offices and the public.

3.3 Session 3: "The Goals of the Project"

Stelios Piperidis, the ELRC representative, stressed first the multilingual aspect of Europe and reiterated that there is one important barrier, namely the language barrier, which prevents the EU's market to become a really single market. He also introduced the CEF scene from the perspective of such a multilingual Digital Single Market strategy. He identified current multilingual challenges in the European public services and in the business sector in general and stressed the support of the EC to digital multilingualism. He reported on the nature and the objectives of the CEF Digital and explained the rationale behind the CEF.AT platform and the expected benefits for public services in the individual countries. He also went through the workshop objectives and its logistics. He introduced ELRC and explained its relation to CEF and CEF.AT, while he briefly presented the main stakeholders, principles and goals of this endeavor. He stressed the main points of potential collaboration between ELRC and the public sector in view of the multilingualism support within the EU.

3.4 Session 4: "Europe and multilinguality"

Vitezslav Zemanek, the DGT representative in the Czech republic, showed in numbers and examples the nature of European languages and the challenges of translation, both within the DGT and in general. He also reported on the Czech translation services (translators and interpreters in the Commission, the amount of translation, etc.) at the Commission, and workflows they use and face.

3.5 Session 5: "Languages and Language technologies in the Czech Republic"

Tomas Svoboda, professor at the Translatology Institute of the Charles University presented the situation with regard to languages and language use in the Czech Republic. He stressed that while the situation seems to be straightforward from the outside (only one official language), it is more complicated in reality (minorities, esp. the Gypsy and Roma, also Ukrainian, Russian, and other workforce, high volume of tourists, etc.). Jan Hajic, professor at the Institute of Formal and Applied Linguistics at the Charles University, then continued with an overview of software tools available for Czech in the industry as well as in research institutes and organizations, which he listed and described. Specifically, he pointed to the

Language White Paper series of analyses and evaluations for all EU languages in terms of technology coverage and quality, prepared by the META-NET project and published by Springer.

3.6 Session 6: "Automated Translation – How does it work?"

Prof. Ondrej Bojar, from the Institute of Formal and Applied Linguistics at the Charles University in Prague, explained the mechanics of Machine Translation. He described the basic algorithms (in a simplified way), and how a state-of-the-art machine translation system is being created, by machine learning methods using parallel and monolingual corpora. He also showed examples of good practice in translation, and presented current results for Czech. He stressed that translation is not and will not be perfect, but that it is getting better every year. For Czech, research needs to continue on additional techniques for handling the rich inflection which with low data availability (compared to English) results in so far lower quality of machine translation output than for English and other languages which have substantially more textual resources at their disposal.

3.7 Session 7: "How can public institutions benefit from CEF.AT"

This session's topic was presented by Daniel Kluvanec, the director of Business Development at the DGT in Brussels. He described the CEF.AT platform, and elaborated on the difference of the MT@EC platform which is already available and the CEF.AT being built, and the differences between them.

3.8 Session 8: "What data are used in machine translation?"

Prof. Bojar of UFAL at Charles University presented the types of data needed for building a state-of-the-art machine translation system, using current popular and good quality toolkits, such as the Moses toolkit (developed with the help and partial support from the EC in the past 10 years; prof. Bojar is one the early and major contributors to the Moses toolkit). He explained the use of parallel data (i.e., previously manually and high-quality translated texts) as well as the use of very large corpora of monolingual data, which are need for reliable modeling of the correct sequences of words on the target side of the translation. He also stressed the importance of collecting and using large domain-oriented data, which are always scarce.

3.9 Session 9: "The Legal framework"

Jakub Michálek, a holder of both a law degree as well as a college level education in physics, and a supporter of open access to texts and other material subject to copyright and IPRs in general, gave an overview of possibilities of sharing textual and other similar data, especially in the context of the public sector, which he knows well also due to his current position as a deputy in the Prague City Council.

3.10 Session 10: "Data and LRs: practical aspects and "best practice""

Stelios Piperidis explained the typical workflows and sharing possibilities, identification of resources and aspects of storing, licensing and distributing public language resources. He focused on issues such as identification of the data sources and datasets, the basic metadata documentation, data cleaning and privacy and ethics management as tasks in which the public sector providers will collaborate with ELRC. The presenter encouraged the audience to participate in these activities and work together with ELRC, and he showcased the mechanisms with which ELRC will fully support the providers throughout the whole process, i.e. the helpdesk and user forum mechanism, the ELRC repository and the ELRC website.

3.11 Session 11: "Interactive: how can we engage?"

Jan Hajic engaged the audience in discussion about possible engagement of the institutions whose representatives have been present. It became clear that there are various collections of texts available, but that the legal issues involved might not be easy to solve, e.g. at the Czech Patent Office. Discussion focused on ways of legal release of data, and in part also on technical processes needed to get those data to the research community for developing higher quality machine translation.

3.12 Session 12: "Wrap-up and next steps"

The last session was jointly led by Stelios Piperidis and Jan Hajic. They summarized the workshop, the discussions and the possibilities and advantages of sharing texts in the public sector. Help was offered to all participants with any aspects of the process(es) should they decide to help and engage in CEF.AT and in sharing the data for at least research purposes in general.

4 Synthesis of Workshop Discussions

4.1 Panel 1: Translation in Public sector in the Czech Republic

The first panel concentrated on the issues of public sector translation in the Czech Republic. It was moderated by Prof. Tomas Svoboda, of the Translatology Institute at Charles University in Prague. He is also engaged in DGT activities in the Czech Republic, and is an experienced court expert on translation. Members of the panel were the public sector ELRC representative, Ms. Marie Cernikova, of the Ministry of Education, Youth and Sports of the Czech Republic, and Vitezslav Zemanek, the DGT representative in the Czech Republic.

The panelists had a short time for their introductory presentations, in which they presented the problems (and questions) form their everyday perspective.

The panel discussed several issues concerning translation in the public sector. It became apparent that there is no common ground, network nor common mechanism, neither technically nor organizationally in the Czech Republic to translate in the public sector, not even at the Ministry (government) level. Only six central government organizations, among them only two ministries out of 20, have a person overseeing their document translation efforts. In all other cases, the translations are done on as-needed basis, either by internal people (usually though not hired as translators) or by external contracted Language Service Providers, or Translation Agencies. Quality control differs institution by institution, and ranges from strict to virtually non-existent.

4.2 Panel 2: Data and Language Resources in the Czech Republic and in Slovakia

The second panel was moderated by Jan Hajic, professor at the Institute of Formal and Applied Linguistics at the Charles University in Prague. Panelists were again Tomas Svoboda, of the Translatology Institute at the Charles University in Prague, and Daniel Kluvanec, of the DGT. Slovak language has been added to the discussion due to the fact the two languages are very close (even though there will be a separate ELRC event in Slovakia), and also because Mr. Kluvanec knows a lot about the Slovak language resources, being native Slovak.

The panel again gave the opportunity to the panelists to have a short introduction, with the discussion following the introduction. Due to the uneven state of affairs in the public sector in the country, as already discussed in the first panel, the panel concluded that the current availability of language resources for Czech is low and that it is not going to change soon, unless a substantial pressure is generated from the outside.

Several possible paths were identified by some members of the audience who posed questions. As an example, the Institute for standards and measurement (UNM) seem to be willing to at least check their resources and their availability to at least researchers.

5 Workshop Presentation Materials

The workshop presentations and videos can be accessed at the event webpage, at http://lr-coordination.eu/czech_agenda.