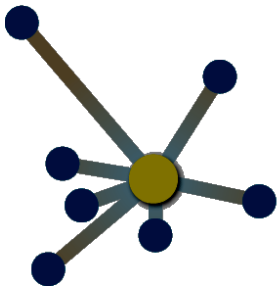
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.



European Language Resource Coordination

Connecting Europe Facility

ELRC Workshop Report for Slovakia

Author(s): Miroslav Zumrík (Ľ. Štúr Institute of Linguistics)
Jana Levická (Ľ. Štúr Institute of Linguistics)

Dissemination Level: Public

Version No.: <V1.1>

Date: 28.04.2016

Contents

1 Executive Summary

This document reports on the ELRC Workshop in Slovakia, which took place in Bratislava at the European Information Centre, the seat of EC Representation in Slovakia. 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 67 participants spanning a wide range of ministries and public organizations, as well as some scientists and freelance translators. The dedicated event webpage can be found at: http://lr-coordination.eu/sk/slovakia.

The workshop was organized by the Slovak National Corpus, the department of Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, together with the Ministry of Culture of the Slovak Republic. The event was supported by the Slovak DGT representation, which provided the premises, technical equipment as well as contacts of potential participants and speakers.

2 Workshop Agenda

8.30 – 9.00	**Registration**
9.00 – 9.15	***Opening session*** ***Dušan Chrenek*** (*Head of EC Representation*) ***Zuzana Komárová*** (*Ministry of Culture of the Slovak Republic*) ***Mária Šimková*** (*Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences*) ***Stelios Piperidis*** (*ELRC*) ***Miroslav Zumrík*** (*Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences*))
9.15 – 9.30	***Aims and Objectives*** ***Stelios Piperidis*** (*ELRC*)
9.30 – 9.45	***Europe and Multilinguality*** ***Nataša Procházková*** (*Local DGT Officer*)
9.45 – 10.15	*Panel Discussion 1* ***Language Services in Slovak Public Sector*** Moderator: ***Nataša Procházková*** (*Local DGT Officer*) ***Panel members: Michal Kmeť (****Association of Slovak Translation Companies****), Zuzana Mrvová*** (*National Bank of Slovakia*)***, Barbora Maliarová*** *(Ministry of Justice of the Slovak Republic*)
10.15 – 10.45	***Coffee break***
10.45 – 11.15	***Machine Translation – How does it work?*** ***Ondřej Bojar (****Institute of Formal and Applied Linguistics, MFF UK, Czech Republic****)***
11.15 – 11.45	***What data is needed?*** ***Ondřej Bojar*** (*nstitute of Formal and Applied Linguistics, MFF UK, Czech Republic*)
11.45 – 12.15	***Languages and Language Technologies in Slovakia*** ***Jozef Juhár*** (*Institute of Electronics and Multimedia Communication, TUKE*)
12.15 – 13.15	***Lunch break***
13.15 – 13.45	***CEF.AT – How can public institutions benefit from this platform?***

	Daniel Kluvanec *(Business Manager Adviser for MT, EC)*
13.45 – 14.15	*Panel Discussion 2* ***Language Resources in Slovakia*** Moderator: ***Mária Šimková*** (*Ľ. Štúr Institute of Linguistics*) *Panel members:* ***Ladislav Hluchý*** *(Institute of Informatics, Slovak Academy of Sciences),* ***Jozef Juhár*** *(Institute of Electronics and Multimedia Communication, TUKE),* ***Daniel Kluvanec*** *(Business Manager Adviser for MT, EC)*
14.15 – 14.45	***Legal framework for contributing data*** ***Róbert Dobrovodský*** *(Ministry of Justice of the Slovak Republic and University of Trnava),* ***Magdaléna Miklošová*** *(Ministry of Culture of the Slovak Republic)*
14.45 – 15.15	***Coffee break***
15.15 – 15.45	***Data and language resources: practical aspects and „best practice“*** ***Stelios Piperidis*** (*ELRC*)
15.45 – 16.15	***How can we engage? (discussion and conclusions)*** ***Jana Levická*** (*Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences*), ***Stelios Piperidis*** (*ELRC*)

3 Summary of Content of Sessions

3.1 Session 1: "Opening"

In the opening session, Miroslav Zumrík, the local Technology NAP, welcomed the audience and introduced representatives from key institutions, organising or co-organising the workshop: Stelios Piperidis from ELRC, Dušan Chrenek, Head of the EC Representation in Slovakia, Zuzana Komárová, general director of the National Language and Arts Section from the Ministry of Culture of the Slovak Republic, and Mária Šimková, Head of the Slovak National Corpus, a department of Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences. Each of them shortly presented their respective institutions and stressed the importance of Slovak language, quality translation and employing up-to-date translation solutions in the Slovak public sector.

3.2 Session 2: "Aims and Objectives"

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 emphasized 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.

3.3 Session 3: "Europe and Multilinguality"

Nataša Procházková, the local DGT officer, showed in numbers and examples the nature of European languages and the challenges of translation, both within the DGT and in general. She reported on the aims of Single Digital Market strategy, CEF programme, CEF.AT platform, as well as on the platforms' benefits for both Slovakia and other member states.

3.4 Session 4: "Machine Translation – how does it work"

Ondřej Bojar, from the Institute of Formal and Applied Linguistics at Charles University in Prague, explained the mechanics of Machine Translation. He described the basic algorithms (in a simplified way), and development of a state-of-the-art machine translation system, by employing 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 automatic translation is not and will not be perfect, but that keeps improving every year. Moreover, he stated that Czech and Slovak as morphologically rich languages require the continuation of research in terms of additional techniques for handling the rich inflection which together 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.

3.5 Session 5: "What Data are used in Machine Translation"

Ondřej Bojar then 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 of 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 needed 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.

3.6 Session 6: "Languages and Language Technologies in Slovakia"

This topic was presented by Jozef Juhár from the Department of Electronics and Multimedia Communication at the Technical University in Košice. Juhár outlined basic principles of language technologies in Slovaki are and then presented some language tools, corpuses, projects and software solutions within, e.g., automatic speech recognition, transcription and indexing of audiovisual records at the Technical University in Košice.

3.7 Session 7: "CEF.AT – how can Public Institutions benefit from this platform"

The session's topic was presented by Daniel Kluvanec, the director of Business Development at the DGT in Brussels. Daniel Kluvanec talked about interactions between actors in the Member States and the EU, the role of Machine Translation and outlined various groups of MT users. He then elaborated on the MT@EC platform which is already available and the CEF.AT that is being built.

3.8 Session 8: "Legal framework for contributing Data"

This session included two legal experts, Róbert Dobrovodský from the Ministry of Justice of the Slovak Republic and research affiliate at the Faculty of Law, the University of Trnava, as well as by Magdaléna Miklošová from the Ministry of Culture of the Slovak Republic, who concentrated especially on the amended Copyright Act. Both experts focused on complex legal issues of open data, open access and public sector. At the same time, Róbert Dobrovodský, the PSI Directive expert, stressed the acknowledgments, which the Slovak open database of acts and other legal documents within public sector has earned amongst foreign experts.

3.9 Session 9: "Data and Language Resources: 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.

3.10 Session 10: "How can we engage (discussion and conclusion)"

In this concluding session, Stelios Piperidis and Jana Levická, scientific affiliate at the Slovak National Corpus, Ľudovít Štúr Institute of Linguistics, sought to wrap-up the workshops' aims and conclusions and, most importantly, they tried to engage the audience in a broader discussion, feedback and consideration with respect to creating a network of future data contributors from the public sector.

4 Synthesis of Workshop Discussions

4.1 Panel 1: Language Services in the Slovak Public Sector

The first panel was moderated by the local DGT officer, Nataša Procházková, and attended by Michal Kmeť from the Association of Translation Companies of Slovakia, Zuzana Mrvová, the Head translator at the Department of International Relations, National Bank of Slovakia, and Barbora Maliarová from the Department of Expertise, Interpretation and Translation Services at the Ministry of Justice of the Slovak Republic. The panelists commented on the peculiarities and challenges of language services in the Slovak public sector, e. g. the lack of specialized language and translation departments in majority of public institutions, which makes the use of outsourcing quite inevitable and which consequently leads to lower quality and consistency of translated texts.

4.2 Panel 2: Language Resources in Slovakia

The second panel was moderated by the Head of the Slovak National Corpus, Mária Šimková, and attended by Ladislav Hluchý from the Institute of Informatics, Slovak Academy of Sciences, Jozef Juhár from the Technical University in Košice and Daniel Kluvanec from DGT EC. The discussion focused on public textual and language resources in the National Corpus (primary, speech, dialect corpus, parallel corpora and specialized databases), as well as resources in other, mainly scientific institutions. The participants pointed out both achievements and challenges within the field (resulting from the IPR situation, weaker financial and technical support, complicated cooperation with commercial sphere), sketched the plan for building a corpus of public administration, and stressed the need for more resources.

5 Workshop Presentation Materials

The workshop presentations are available at the dedicated event webpage: http://lr-coordination.eu/sk/slovakia_agenda.