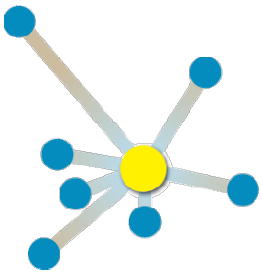
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1091 programme.



European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.6
Task 8

ELRC Workshop Report for Slovenia

Author(s): Simon Krek and Špela Sitar

Dissemination Level: Public

Version No.: <V1.0>

Date: 2018-06-19

Contents

1 Executive Summary

The 2nd ELRC workshop in Slovenia took place on 24 April 2018 at the EU House in Ljubljana. It was organised by the European Language Resource Coordination (ELRC) consortium and Jozef Stefan Institute. The event was supported by the representation of European Commission in Slovenia.

The programme included an opening speech by Boris Koprivnikar, Minister of Public Administration and Vice Prime Minister of the Government of the Republic of Slovenia, and a welcome message by Zoran Stančič, head of the EC Representation in Slovenia and former Deputy Director General at DG CONNECT.

The programme included representatives from the Ministry of Public Administration, Ministry of Culture and the Ministry of Economic Development and Technology, mainly in connection with the ongoing Connecting Europe Facility activities in Slovenia, i.e. the national programmes on Europeana, e-Justice, e-ID (SI-PASS), e-Delivery and Online Dispute Resolution. The programme additionally included other CEF-funded projects at the Jozef Stefan Institute, and applications relevant to the ELRC action objectives, i.e. the local implementation of the Nematus NMT system, which includes the English-Slovenian pair, and the Tacita anonymisation software. The rest of the programme included standardised presentations of CEF in general, the eTranslation platform, the ELRC action and the ELRC-SHARE repository.

2 Workshop Agenda

When:	**24 April 2018**
Where:	**EU House, Dunajska 20, 1000 Ljubljana, Slovenia**
Organisation:	• **The European Language Resource Coordination (ELRC) consortium** • **Jozef Stefan Institute**
09:10 – 09:30	**Registration**
09.30-09.40	**Welcome and introduction** *Minister of Public Administration Boris Koprivnikar*
09.40-09.45	**Welcome by the EC** *Head of EC Representation in Slovenia: Zoran Stančič*
Session 1: Connecting a multilingual Europe: European context and local needs	
09.45-10.00	**Connecting public services across Europe: ambition and results so far** *Aleksandra Wesolowska, Project Officer, Directorate-General Communications Networks, Content and Technology, European Commission (* video link, interpretation in Slovenian*)*
10.00-10.20	**National initiatives for digital public services and (open) data** *Ministry of Culture: Simona Bergoč*
10.20-11.00	**CEF in Slovenia: an outlook into current and future challenges – Panel session** Moderator: ELRC Technology NAP Simon Krek Panelists: • *Ministry of Public Administration:* Alenka Žužek Nemec • *Ministry of Economic Development and Technology:* Barbara Miklavčič • *Ministry of Culture:* Skender Adem
11.00-11.20	**The CEF eTranslation platform @ work**

	Markus Foti, MT@EC/eTranslation Project Manager, Directorate-General for Translation, European Commission (live video link, interpretation in Slovene)
11.20-11.45	***Coffee Break***
Session 2: Engage: hands-on data	
11.45-12.10	**The European Language Resource Coordination (ELRC) action** Stelios Piperidis, *Institute for Language and Speech Processing/ R.C. Athena, ELRC*
12.10-12.30	**ELRC in Slovenia** *ELRC Technology Anchor Point* Simon Krek *eTranslation TermBank project Andraž Repar, Matjaž Rihtar, both Jozef Stefan Institute* Tacita anonymizer Matej Kovačič, *Jozef Stefan Institute*
12.30-13.00	**Can language data be shared and how? National and European legal framework** Slovenian open data portal OPSI Mateja Prešern, Ministry of Public Administration
13.00-14.00	***Lunch break***
14.00-14.30	**Preparing and sharing data with the ELRC repository – and what happens next** *Maria Giagkou, Institute for Language and Speech Processing/ R.C. Athena, ELRC*
14.30-15.00	**Identifying and managing your data: Questions & Answers** *Simon Krek, Stelios Piperidis*
15.00-15.35	**Conclusions and coffee break with networking** *Simon Krek*

3 Summary of Content of Sessions

3.1 Welcome and introduction

Boris Koprivnikar – Minister of Public Administration, Vice Prime Minister of the Government of Republic of Slovenia.

Mr. Koprivnikar emphasised the importance of reusing and integrating CEF building blocks in public digital services. He especially highlighted the function of the CEF Automated Translation building block as an enabler of cross-border communication in the digital era. If the state or the nation does not enable inclusion of Slovene in emerging technologies, the Slovene language will face digital extinction. Slovenia has a lot of expertise in the area, what is lacking is more efficient collaboration between stakeholders.

3.2 Welcome by the EC

Zoran Stančič – Head of EC Representation in Slovenia.

Mr. Stančič described the need for MT at the European Commission. There are two requirements for MT to work: technology and data. It is important not to become entirely dependent on technological giants outside the EU. Mr. Stančič also emphasised the importance of the of Slovene government's support to the workshop and to ELRC in general.

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

Aleksandra Wesolowska

Aleksandra Wesolowska, Project Officer DG CONNECT, presented the Connecting Europe Facility, with special emphasis on the Digital Service Infrastructures (DSIs) and particularly the CEF Automated Translation building block. She concluded with the need for the involvement and connection of the national public administrations with eTranslation and the current funding opportunities.

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

Simona Bergoč – Head of Slovenian Language Service, Ministry of Culture

Mrs. Bergoč presented the Resolution on the National Programme for Language Policy 2014–18, and in particular, the work of the Council for Continuous Monitoring of the Development of Language Resources and Technologies for Slovene. The council comprises representatives of the Ministry of Culture, Ministry of Education and Sport, Ministry of Public Administration, as well as experts from various Slovene research institutions. The Council defined key areas that should be financed in the coming period: speech technologies, development and upgrade of language corpora, semantic technologies, machine translation, terminological portal and grammar/spelling checking applications. At the end, some examples of opening data for machine translation were mentioned, in particular the Evroterm terminological database produced by the Department of Translation at the Secretariat-General of the Slovene government.

3.5 CEF in Slovenia: an outlook into current and future challenges – Panel session

The panel session hosted representatives of the Ministry of Culture, the Ministry of Economic Development and Technology, and the Ministry of Public Administration.

Skender Adem – Ministry of Culture (Europeana - DSI)

Mr. Adem presented the scope and goals of the Europeana DSI, the plans for its development and the importance for multilingual or cross-lingual search. He discussed the financing of Europeana through CEF, and the foci of relevant funding opportunities.

Barbara Miklavčič – Ministry of Economic Development and Technology (Online Dispute Resolution)
Mrs. Miklavčič presented the Online Dispute Resolution platform, with special emphasis on the automated translation functionalities available through the platform. She pointed out that up to date the performance of machine translation systems for less spoken languages, such as Slovene, is not as high as for widely spoken languages. This fact alone advocates the necessity for intensified efforts in collecting more Slovene language data. However, she expressed a concern regarding personal data protection, which can hinder data sharing for machine translation. Mrs. Miklavčič promised that this issue will be investigated by the Ministry in the near future.

Alenka Žužek Nemec – Ministry of Public Administration
Mrs. Žužek Nemec is the person responsible for the Digital Single Market and CEF at the Ministry of Public Administration. She presented the relevant national activities, and more specifically the development/integration of e-Justice, SI-PASS (Web application and signature service) e-ID, and e-Delivery in the Slovene public administration. She explained the project NOBLE ("NO Barriers in eDeLivEry"). The objective of the NOBLE project is to provide cross-border delivery of e-mail delivery, e-documents) in which the ministry was involved.

3.6 The CEF eTranslation platform @ work

Markus Foti
Mr. Foti, MT@EC/eTranslation Project Manager, participating via teleconference (with live video link), presented the eTranslation platform, its principles and mode of operation, its intended and current users (in terms of DSIs), the steps required to connect with the platform and the benefits of using MT@EC/eTranslation.

3.7 The European Language Resource Coordination (ELRC) action

Stelios Piperidis

Mr. Piperidis, representative of the ELRC consortium, presented the consortium and its goals, its activities and the current situation as regards data collection at the European level, the repository developed and the services offered by the helpdesks to data contributors and users.

3.8 ELRC in Slovenia

Simon Krek – Jožef Stefan Institute

Mr. Krek presented the ELRC achievements at the national level since 2015, in particular the Slovenian data collected in the framework of ELRC, and the problems and issues faced by ELRC representatives during the collection process. Among them a general reluctance to sharing data by public sector representatives was mentioned, which was attributed to the lack of incentives and to the shortage in human resources and time. Additionally, as one of the main concerns of the Slovene data holders is the protection of personal data, Mr. Krek specifically focused on efficient anonymisation software for

Slovene (introducing one of the subsequent presentations, that of Mr. Rihtar). The presentation concluded with some examples of translations produced by the Nematus NMT system at Jozef Stefan Institute, and compared the translations of the same sentences presented at the first ELRC workshop in 2015, translated by the PMBT system used at the time. The difference between the results was explained.

Andraž Repar – Jožef Stefan Institute

Mr. Repar presented another CEF-funded project eTranslation TermBank which exclusively collects terminological resources for eTranslation. Slovenian partner (Jožef Stefan Institute) covers three languages: Slovenian, Croatian, Bulgarian, and three domains: eHealth, eJustice, Online Dispute Resolution.

Matjaž Rihtar – Jožef Stefan Institute

Mr. Rihtar presented the technical side of the installation of the Nematus NMT system at Jožef Stefan Institute. He explained which data was used for the training (OPUS corpus - http://opus.nlpl.eu/), and how translation models were built.

Matej Kovačič – Jožef Stefan Institute

Mr. Kovačič presented the Tacita anonymiser developed at Jozef Stefan Institute for the Ministry of Justice. The features used and other technical details of the software were presented. The Tacita anonymizer has been trained on 2840 court decisions of the Supreme Court and High Court in Ljubljana: non-anonymised documents in OpenOffice (.odt) format; anonymised documents available through HTTP access in JSON format (http://sodnapraksa.si). The precision of the system is 72,8 % and recall is 91,8 %. In the future, the system can be used also on parallel data contributed by public bodies to ELRC, which are considered sensitive by the data holders and cannot be shared in their current form.

3.9 Can language data be shared and how?

Mateja Prešern – Ministry of Public Administration

Mrs. Prešern presented activities of the ministry in making Slovenian public administration data publically and openly available. The concept of open data produced in public administration was presented, together with the legislation dedicated to copyright, privacy protection etc. She pointed out that also language data are defined as public data in Slovene legislation (as part of the PSI directive transposition). Existing applications based on open access data were presented, as well as the national open data portal: https://podatki.gov.si/.

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

Maria Giagkou

Mrs. Giagkou focused on the technical aspects of preparing and sharing data through the ELRC repository: types of data needed by eTranslation, appropriate data formats and domains were explained and exemplified, tips for good and bad practices were given as regards data preparation and management. The presentation also included a detailed, step-by-step guided tour of the procedures of registration to the repository and contribution of data. Finally, the presentation described the services offered by the ELRC consortium, aiming to facilitate public sector data contributors with all stages of data extraction, conversion to appropriate formats, data clean-up, alignment, anonymization, and

metadata curation and validation, and explained the services of on-site assistance and the help desks' function.

3.11 Questions & Answers - Conclusions

Simon Krek, Stelios Piperidis

See below Section 4 (Synthesis of Workshop Discussions)

4 Synthesis of Workshop Discussions

In the questions and answers session it was emphasised that the Ministry of Public Administration is very important for ELRC activities and should be involved as much as possible in the future. A representative of the Ministry of Foreign Affairs explained that only their web pages can be used for ERLC purposes since the Ministry mostly deals with sensitive data. The next Slovenian presidency in 2021 was mentioned as an important event, which will sensitize government officials with respect to the need for machine translation in cross-border communication, having in turn a possible positive effect on the data collection activities of ELRC. With regard to concerns expressed about the right to share copyrighted material or sensitive data, it was also explained that what is interesting to a machine translation system is not the actual content of possibly sensitive or copyrighted (e.g. ISO standards) data. It is rather the linguistic realization of this content, e.g. the terminology used. Furthermore, anonymisation was discussed as one of the most important parts of the pipeline. The Slovene participants considered this as essential in order to be able to contribute more language data. The representative of the Translation Service at the Secretariat-General explained that they had complaints about making available their parallel corpus online, due to disclosing private data, and now they only have selected data publicly available in the Evrokorpus corpus. Therefore, if anonymisation is solved more data could be provided.

4.1 ELRC and Open Language Data in Slovenia

Mateja Prešern, representative from the Ministry of Public Administration, presented activities of the ministry in making Slovenian public administration data openly available. The general concept of open data produced in public administration was presented, together with the legislation dedicated to copyright, privacy protection etc. Existing applications based on open access data were presented, as well as the repository of open access data produced by the ministry: https://podatki.gov.si/.

On this portal, called 'OPSI - Odprti podatki Slovenije', Open Data of the whole Slovenian public sector is brought together and made available to the public. The portal, built on Open Source software, is an indirect successor of the National Interoperability Framework Portal. While since 2013 the NIO portal presented an extensive number of Open Datasets, providing Open Data was not its core activity. OPSI portal has a dual function. The first one is providing a central catalogue of all the records and databases of Slovenian public bodies. In this catalogue the metadata about all the Open Data from state authorities, municipalities and other public sector bodies is made available. The second function of the portal is to be the single access point for data in a machine-readable format and with an open data licence. This includes open data collections which had already been published on different websites.

The portal does not contain multilingual data.

4.2 Success stories and lessons learnt

- The workshop provided an opportunity for representatives from various ministries to learn about CEF and eTranslation. During the workshop informal discussions it was apparent that even people working at the same ministry are not aware of the full spectrum of CEF DSIs used in their organization, e.g. translators knew MT@EC, but they were not aware that their ministry is also reusing a number of other CEF DSIs. In this respect the event was considered a valuable one, as it was focused on a particular horizontal issue (machine translation) that brings together otherwise non-connected people.
- The fact that full Slovene legislation with more than 100 million tokens was made available in the Slovene Open Data portal as an open access database in JSON format is considered as an important achievement of the ELRC data collection task in Slovenia.

- Solving anonymisation issues was considered as crucial in order to overcome the concerns for data protection.
- The upcoming Slovenian EU presidency is considered as a great opportunity to further raise awareness of eTranslation and the value of public language data for MT systems.