ELRC Workshop Report for Slovenia

Author(s):	Simon Krek (Jozef Stefan Institute)
Dissemination Level:	Public
Version No.:	<V1.1>
Date:	2016-02-29

European Language Resource Coordination

ELRC Workshop Report for Slovenia

Contents

1 Executive Summary

This document reports on the ELRC Workshop in Slovenia, which took place in Ljubljana on 8 December 2015 in the Great Hall at the **Jožef Stefan Institute**, Jamova 39, 1000 Ljubljana, Slovenia. It includes the agenda of the event (section 2) and briefly informs about the content of each individual, interactive and panel workshop session (sections 3 & 4). The event was attended by 38 participants from the Slovene public sector, mainly from ministries and government offices, as well as independent data experts and freelance translators, and 10 speakers. The list of participants is included in the document. The dedicated event webpage can be found at http://lr-coordination.eu/slovenia.

2 Workshop Agenda

08:00 – 09:00   Registration

09:00 – 09:10   **Opening and welcome**

(Simon Krek – "Jožef Stefan" Institute (JSI) and Stelios Piperidis - ILSP/ELRC)

09:10 – 09:25   **Local welcome and welcome by the European Commission**

(Bronka Štraus – Ministry of Education, Science and Sport, Maja Pavlović - EC representation, Slovenia)

09:25 – 09:35   **Aims and Objectives** (Stelios Piperidis – ILSP/ELRC)

09:35 – 09:50   **Europe and Multilingualism** (Peter Jakša - EC representation, Slovenia)

09:50 – 10:20   **Languages and Language Technologies in Slovenia** (Simon Krek - JSI)

10:20 – 11:00   **Panel 1: Multilingual Public Services in Slovenia** (Evroterm/Evrokorpus)

Marčela Novljan Lovrinčić, Miran Željko, Adriana Sedej, Marja Adamič (Secretariat-General of the Government of the Republic of Slovenia, Translation and Interpretation Division)

11:00 – 11:30   Coffee Break and Networking

11:30 – 12:00   **Automated Translation: How does it work?**

(Stelios Piperidis – ILSP/ELRC and Simon Krek – JSI)

12:00 – 12:30   **How can Public Institutions benefit from the CEF.AT Platform?** (Spyridon Pilos - EC DGT)

12:30 – 13:30   Lunch Break

13.30 – 14.00   **What Data is needed?**

(Špela Vintar - Department of Translation, Faculty of Arts, University of Ljubljana)

14:00 – 14:30   **Legal framework for Contributing Data** (Maja Bogataj Jančič - Intellectual Property Institute)

14:30 – 15:00   **Data and Language Resources in Slovenia**

Miran Željko (Secretariat-General of the Government of the Republic of Slovenia, Translation and Interpretation Division)

15:00 – 15:30   Coffee Break and Networking

15:30 – 16:00   **Data and Language Resources: Technical Aspects and Best practices to Capitalize on your Valuable Data** (Stelios Piperidis – ILSP/ELRC)

16:00 – 16:30   **How can we engage?** (Simon Krek - JSI)

16:30 – 16:45   **Wrap-up, On site Conclusions and Commitments** (Simon Krek – JSI and Stelios Piperidis – ILSP/ELRC)

3   Summary of Content of Sessions

3.1   Session 1: "Opening and welcome",

Simon Krek ("Jožef Stefan" Institute), the ELRC National Anchor Point in Slovenia, 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. Welcome on behalf of the ELRC consortium was expressed by **Stelios Piperidis** (ILSP/ELRC) who briefly described the purpose of this event and previous events from ELRC perspective.

3.2   Session 2: "Local welcome and welcome by the European Commission"

Bronka Straus (Ministry of Education, Science and Sports) welcomed the participants of the workshop and emphasised the integration of machine translation into Slovene national language policy documents – the Resolution on the National Programme for Language Policy 2014-2018, as well as the Action Plan for Language Education and the Action Plan for Language Infrastructure, adopted as government plans in November 2015. **Maja Pavlović** (EC representation Slovenia), welcomed the participants and elaborated on the importance of multi-linguality in Europe. She briefly described the aims of MT@EC service in the context of Digital Agenda for Europe and Connecting Europe Facility.

3.3   Session 3: "Aims and Objectives"

Stelios Piperidis analysed the multilingual character of the EU and the role of translation in achieving communication within Europe. He then focused on the CEF and CEF.AT, and briefly presented the main stakeholders, principles and goals of this endeavour, as well as its value towards serving the needs of European citizens. He stressed the fact that machine translation depends on language data which need to be provided by the users in order for the benefits to return to them. He then went through the workshop objectives and the agenda and concluded by stressing the importance of the project to the creation of a multilingual Digital Single Market.

3.4   Session 4: "Europe and Multilingualism"

Peter Jakša (EC representation Slovenia) briefly presented the history of official languages in the European Union. He continued with the presentation of the strategy of the Digital Single Market and the new instrument to connect the EU – the Connecting Europe Facility (CEF). He presented the EU multi-lingual challenge from the perspective of public services and the EC platform for machine translation. The role of member states in this endeavour was emphasised, with the stress on the advantages that the CEF.AT platform brings to particular states, such as the possibility for EU citizens to access information that is available only in languages they do not understand, or the possibility of companies to participate in calls in different countries.

3.5 Session 5: "Languages and Language Technologies in Slovenia"

Simon Krek briefly presented the history and the current developments in the field of language technologies for Slovene. He summarised the results of the Slovene META-NET White Paper, particularly in relation to machine translation and the resources needed to make advancements in that field. He presented the national language policy documents, in particular the Action Plan for Language Infrastructure, comparing these documents with strategic documents in other countries, particularly for Estonian, Lithuanian and Portuguese. Based on the analysis, he made predictions about the development of language technologies for Slovene by 2020, with some suggestions on which steps are needed to ensure better prosperity of Slovene in the digital age.

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

Stelios Piperidis gave a presentation on how statistical machine translation works. Examples were given of how a machine translation system can learn from a suitable corpus of data and explained that how the translation component (translation model) and the target language component (language model) of a machine translation system can fruitfully collaborate to produce a successful translation output. He further talked about how it is possible to make a computer "learn" natural language and vocabulary. **Simon Krek** explained how it is possible to teach a computer to align segments of different languages on the basis of data which are chosen carefully and in a targeted way. **Stelios Piperidis** concluded the presentation by stressing that volume and suitability of data will exponentially reflect on the quality of the translation output by giving further examples of what machine translation systems can afford today.

3.7 Session 7: "How can Public Institutions benefit from the CEF.AT Platform?"

Spyros Pilos (EC DGT) described in detail the MT@EC system, i.e. the languages it supports, the technologies upon which it is based, its user interface and other technical features, including input data format, delivery of results and security in document transfer. He also presented statistics about its usage and invited the public sector representatives to use it. He stressed that the CEF.AT platform will build upon MT@EC and aims to address the CEF DSIs. Also, he presented the benefits that CEF.AT will bring to its users in terms of translation of better quality and security, adapted to the user needs and specific domains.

3.8 Session 8: "What Data is needed?"

Špela Vintar (Department of Translation, Faculty of Arts, University of Ljubljana) briefly described how translation industry evolved and the current state of professional translation in Slovenia. She showed two use cases of machine translation – what happens if a translation agency buys a machine translation service and if students try to work with machine translated text. She explained the difference between Computer-Aided Translation and Machine Translation and mentioned some services providing online

machine translation tools. Then she focused on the data that need to be collected, the required format of the data so that they can be fed into MT systems, and the importance of the data that the public sector holds and with which it can contribute to the CEF.AT platform.

3.9   Session 9: "Legal framework for Contributing Data"

Maja Bogataj Jančič (Intellectual Property Institute) elaborated on the directives 2003/98/EC 2013/37/EU, particularly in the context of Slovene legislation. She described the Decree on communication and re-use of information of public character, and then focused on the reasons which can lead to denial of re-use of information of public character. The role of Slovene Information Officer was depicted in the context of IPR vs. open access legal issues. At the end of the presentation she elaborated on the changes of the Public Information Access Act which can be expected, as the Act is currently in the legislation process. The changes mainly encourage open access and re-use of information of public character following the 2013/37/EU directive.

3.10   Session 10: "Data and Language Resources: Technical Aspects and Best practices to Capitalize on your Valuable Data"

In this session **Stelios Piperidis** made a summary of afternoon sessions starting with the technical aspects, followed by the types of data (i.e. technical reports, webpages, speeches, glossaries, terminological databases or lists of words in one or more languages etc.) that is needed for enhancing the performance of machine translations systems, and legal issues. He stressed the importance of the data that the public sector holds and with which it can contribute to the CEF.AT platform. He encouraged the participants from public administration to contribute to the project by providing or identifying the required textual content.

3.11   Session 11: "How can we engage?"

See Session 12 – the two sessions were joined into a general wrap-up and conclusions session.

3.12   Session 12: "Wrap-up, On site Conclusions and Commitments"

Simon Krek started the session by summarising workshop messages. He began with the basic aim of the workshop which was to inform local community in Slovenia that the EC is working on machine translation technology and services, emphasising that to be successful in this endeavour all partners in this project must work together, including public administration in Slovenia, and also in other member states. The analysis of the state of MT for the Slovene-English pair showed that the results are currently not optimal, but this is exactly the reason why Slovene public administration, and even broader – all institutions in Slovenia, should help bring the service to performance level close to best performing language pairs.

In the discussion, one of the participants suggested that ELRC should be in contact with the Public Administration Ministry since the ministry is currently working on IPR questions related to the work of public administration in Slovenia. Another participant emphasised that it cannot be expected that translators working in public administration in Slovenia would be able to work on collecting and preparing data specifically for MT, as they are overwhelmed by their daily assignments.

4   Synthesis of Workshop Discussions

4.1   Panel 1: Multilingual Public Services in Slovenia

The first panel was moderated by **Peter Jakša**, the panellists were four representatives from the Secretariat-General of the Government of the Republic of Slovenia, Translation and Interpretation Division. They presented the work of the Secretariat-General, especially in relation to their role as the coordinator of translation services for all the ministries, not only the government in narrow sense. Two important collections of data were presented: *Evroterm* and *Evrokorpus*, the online concordancer for the corpus of translated legislation and a corresponding online terminological database of the Translation and Interpretation Division. In the debate, the panellists described their experience with machine translated texts and also stressed that MT is not really used in their daily work.

4.2   Panel 2: "Data and Language Resources in Slovenia"

The second panel slot started with a presentation of the current performance of the MT@EC service for the English-Slovene language pair prepared by **Miran Željko** who is in charge of the *Evroterm* and *Evrokorpus* services at Secretariat-General of the Government, Translation and Interpretation Division. The speaker discussed a set of indicative results elaborating that at the moment the system is not really ready for production level performance which could be integrated in daily work at the Secretariat-General. Following this presentation the discussion with the audience was opened reaching the conclusion that for Slovene a particular effort will be needed in terms of data collection and sharing so that the language pairs involving Slovene reach the necessary performance level.

5   Workshop Presentation Materials

The workshop presentations can be accessed at the event web page, at

http://lr-coordination.eu/slovenia_agenda.

The video recordings of the workshop can be accessed at Videolectures.not:
http://videolectures.net/elrc2015_ljubljana/?q=elrc.