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 D3.2.23
Task 8

ELRC Workshop Report for Croatia

Author(s): Marko Tadić

Dissemination Level: Public

Version No.: V1.0

Date: 2019-04-03

Contents

1 Executive Summary

The 2nd Croatian ELRC Workshop took place on the premises of the European Commission Representation in Croatia (Augusta Cesarca 4, 10000 Zagreb) on the 12th of February 2019. It was organized jointly by the Croatian Language Technologies Society and the DGT local Field Office in Croatia. Participants from 49 different organizations, mainly public administrations and public institutions, attended the workshop. In the 12 sessions organized within the predefined format, 10 presenters appeared. The video recordings and presentations are available online at the workshop web page http://www.lr-coordination.eu/l2croatia.

The participants were welcomed by the head of the Representative of the European Commission in Croatia, Mr. Branko Baričević. The most interesting parts of the workshop were the panel session on CEF in Croatia that initiated lively discussion, and the presentation of the CEF eTranslation platform by the EC DGT representative Michael Jellinghaus, that achieved the highest evaluation score from the participants.

The presentations raised a number of questions by the participants who were especially interested in three issues related to Automated Translation: 1) the control of the quality of the language data used in training machine translation systems (in terms of both the relevance of the documents and the quality of translations), 2) the impact of the type of the text on the quality of machine translation, 3) the accessibility of the eTranslation system by a wider audience (universities, translation agencies).

The importance of open data was emphasized in several presentations, in order to raise awareness of the value of data and of the benefits of sharing.

2 Workshop Agenda

09:00-09:30	***Registration***
09:30-09:40	**Welcome and introduction** *Presenter:* ***Marko Tadić*** *(ELRC T NAP, Faculty of Humanities and Social Sciences, Zagreb)*
09:40-09:45	***Welcome by the EC*** *Presenters:* ***Branko Baričević*** *(EC Representative in Croatia)*
09:45-10:05	***Connecting public services across Europe: ambition and results so far*** *Presenter:* ***Aleksandra Wesolowska*** *(Directorate-General for Communications Networks, Content and Technology, EC) (video message)*
10:05-10:25	***National initiatives for digital public services and (open) data*** *Presenter:* ***Božo Zeba*** *(Central State Office for the Development of the Digital Society)*
10:25-11:10	*Panel session* ***CEF in Croatia: an outlook into current and future challenges*** Moderator: ***Marko Tadić*** *(ELRC T NAP, Faculty of Humanities and Social Sciences, Zagreb)* Panellists: ***Maja Radišič-Žuvanić*** *(Ministry of Economy, Entrepreneurship and Craft, CEF Telecom coordinator for Croatia)* ***Božo Zeba*** *(Central State Office for the Development of the Digital Society)* ***Mladen Stojak*** *(Ciklopea)*
11:10-11:30	***The CEF eTranslation platform @ work*** *Presenter:* ***Michael Jellinghaus*** *(European Commission, DGT)*
11:30-12:00	***Coffee break***
12:00-12:25	***The European Language Resource Coordination (ELRC) action*** *Presenter:* ***Stelios Piperidis*** *(Institute for Language and Speech Processing / Athena R.C., ELRC)*
12:25-12:45	***ELRC in Croatia so far and projects MARCELL and NEC MT Data*** *Presenters:* ***Marko Tadić*** *(ELRC T NAP, Faculty of Humanities and Social Sciences, Zagreb)* ***Mladen Stojak*** *(Ciklopea)*
12:45-13:10	***Re-usable information: Connection to open data. National and European legal framework*** *Presenter:* **Dubravka Bevandić** *(*Service for the Protection of the Right of Access to Information, Information Commissioner's Office*)*
13:10-14:10	***Lunch break***

14:10-14:45	***Preparing and sharing data with the ELRC repository – and what happens next*** *Presenter:* ***Maria Giagkou*** *(ILSP / Athena R.C, ELRC)*
14:45-15:15	***Identifying and managing your data: Questions & Answers*** *Moderator:* ***Marko Tadić*** *(ELRC T NAP, Faculty of Humanities and Social Sciences, Zagreb)*
15:15-15:30	***Conclusions*** *Moderator:* ***Marko Tadić*** *(ELRC T NAP, Faculty of Humanities and Social Sciences, Zagreb)* ***Stelios Piperidis*** *(Institute for Language and Speech Processing / Athena R.C., ELRC)*
15:30-16:00	***Coffee break and networking***

3 Summary of Content of Sessions

3.1 Welcome and introduction

The Croatian ELRC Workshop was opened by the local organizer Marko Tadić, president of the Croatian Language Technologies Society and ELRC Technology National Anchor Point in Croatia. Professor Tadić presented the overall aims and objectives of the workshop and briefly introduced the participants to the context of the workshop and the agenda of the second ELRC workshop in Croatia.

3.2 Welcome by the EC

The workshop participants were warmly welcomed by Branko Baričević, the head of the European Commission Representation in Croatia with several remarks on the importance of European multilinguality and the sustained ability to communicate between the citizens of the EU using their own languages and yet to be able to fully understand each other. He also stressed the importance of support for all official EU languages in the Single Digital Market since this will allow it to grow steadily. Mr. Baričević thanked the participants for their attendance and concluded with wishes for a successful workshop.

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

The session *Connecting public services across Europe: ambition and results so far* was presented as a video message from Aleksandra Wesolowska (Directorate-General for Communications Networks, Content and Technology, EC). In her talk, Ms. Wesolowska presented the CEF platform eTranslation. Apart from stressing the importance of data for the eTranslation improvement, she gave an overview of the Digital Services Infrastructures (DSIs) and Digital Single Market sectorial platforms/websites and their characteristics. Finally, she described the reuse of building blocks within CEF. In the context of the ELRC Workshop in Croatia, what was of particular interest was the fact that eTranslation is reused by the cross-border eProcurement platform between Slovenia, Slovakia and Croatia.

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

Mr. Božo Zeba, from the Central State Office for the Development of the Digital Society, presented the National initiatives for digital public services and open data. His presentation consisted of three parts. In the first part of his talk, Mr. Zeba presented the efforts made by the Croatian Government for the development of digital society. He presented the currently available public services on the eCitizen (eGrađanin) platform, which offers 685 services and has 626.667 unique users. Second, he elaborated on open data in Croatia, emphasizing the importance and benefits of open data. In this part of his talk, he presented the Open Data Portal, administered and governed by the Ministry of Government and the Central State Office for the Development of the Digital Society. In the third part of the talk he presented the initiatives regarding the establishment of the Single Digital Gateway. Finally, he concluded with several remarks on planned projects regarding digital public services and open data in the Republic of Croatia.

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

Moderator:

Marko Tadić, ELRC Technical NAP

Panellists:

Maja Radišič Žuvanić, Ministry of Economy, Entrepreneurship and Craft, CEF Telecom Coordinator

Božo Zeba, Central State Office for the Development of the Digital Society

Mladen Stojak, Ciklopea, CEF project NEC TM Data

The introductory part was presented by Maja Radišič Žuvanić, CEF Telecom Coordinator representative. She briefly presented the Connecting Europe Facility, with special emphasis on CEF Telecom. Ms. Radišič Žuvanić presented CEF Telecom projects, the consortia of which Croatian organisations are represented (18 projects so far in three DSIs: eIdentification & eSignature, eInvoicing, eTranslation). She also presented the CEF eTranslation call for proposals 2019, which focuses on projects dealing with 1) shared LRs for CEF AT via the ELRC-SHARE repository, 2) collaborative language tools, 3) integration of CEF eTranslation to national or European digital services.

After the introductory presentation, Prof. Tadić started the discussion by asking Mr. Stojak, as a representative of a translation agency, about the impact of CEF initiatives in supporting and enhancing digital public services and the role of eTranslation as a multilingualism enabler of digital public services. He elaborated on the market demands for fast, as cheap as possible, but effective and high-quality translations, and how the related CEF projects can help addressing these demands. Ms. Radišič Žuvanić stressed that the public sector adapts very slowly to these needs when compared to the private sector. However, Mr. Zeba commented that, although he is the public sector representative, he participated in the first CEF projects in Croatia, stressing that some of the public bodies are more eager to implement digital innovations than others. He continued by elaborating on the importance of open data, and Mr. Stojak emphasized the benefits of open data. Prof. Tadić added that we have to build the culture of open data – we are not used to share our data with others (although we share our data on social networks on a daily basis).

The panellists discussed about the users of public services – public administration, citizens, businesses and research community, and concluded that the best services are those which reach out to the widest possible user groups.

In the Q&A part of the panel, one of the participants emphasized the importance of open data for scientific communities and raised the question of 1) public television data, which should be open to the citizens, and 2) digitization of cultural heritage. Mr. Zeba answered that the Ministry of Culture invests in cultural heritage digitization and there are several projects in that direction. Another participant asked whether there is some kind of research on the suitability of types of texts to be machine translated (e.g. weather report vs. scientific article). Prof. Tadić answered that there are two endpoints when it comes to MT: literature on the one hand, which is impossible to automatically translate, and informative texts on the other, e.g. declarations of nutritional ingredients in food products, which are appropriate for machine translation.

3.6 The CEF eTranslation platform @ work

Mr. Michael Jellinghaus, eTranslation Engines Project Manager, DGT, elaborated on the use of the eTranslation platform, its availability, users and benefits and gave practical examples of its actual use. Furthermore, Mr. Jellinghaus underlined the security and privacy features of the eTranslation service, as documents are deleted after 24 hours or upon demand immediately after delivery. He also explained the use of the eTranslation API in machine to machine scenarios, providing the examples of

the N-Lex and the European Data portals. Mr. Jellinghaus argued that machine translation, especially neural machine translation, performs best when trained on large volumes of text pairs (source-translated). Finally, Mr. Jellinghaus talked about Neural Machine Translation and explained how artificial neural networks are trained. He also said that the key to success for eTranslation is more language resources for all languages with better lexical coverage. Mr. Jellinghaus underlined the need for future improvements such as: speed, transliteration, more formats (PDF output, JSON), named-entity recognition and new languages (AR, RU, TR, ZH).

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.

The Governance Body of ELRC is the Language Resources Board which consists of the ELRC Technological National Anchor Points (NAPs, one per CEF affiliated country), and the ELRC Public Services NAPs (one per CEF affiliated country).

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 Croatia: projects MARCELL and NEC MT Data

The overview of the ELRC in Croatia was presented in two parts. First, Prof. Tadić presented the situation with Croatian language resources collected in the ELRC-SHARE repository so far. Currently, there are 21 LRs, including 18 bilingual or multilingual corpora, 1 monolingual corpus and two terminological databases, which is not enough. Professor Tadić also emphasized the problem of Public Services NAP in Croatia, which has not been appointed yet. Prof. Tadić identified the lack of awareness on the part of higher lever officials as a main obstacle for data sharing. Lower lever officials that manage language data and would be willing to contribute them to ELRC require official authorization by their superiors. Among the main achievements of the ELRC action in Croatia, the growing number of contributors was identified, as well as the implementation of two CEF projects with Croatian partners in the consortia. In the second part of the presentation, these two projects were presented.

Marko Tadić presented MARCELL – Multilingual Resources for CEF.AT in the legal domain. The overall objective of this action is to provide the texts that form the body of national legislations (laws, decrees, regulations) in seven countries: Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia. At present, national legislation texts are not automatically available to CEF.AT and present Machine Translation (MT) systems could be improved if they had access to national legislative texts. The main aim of MARCELL is to provide CEF.AT with the existing texts of national legislations, but also to provide a pipeline that would feed CEF.AT with newly emerging national legislation texts in the respective seven languages.

Mladen Stojak from Ciklopea presented the NEC TM Data project, which aims to investigate the amount and value of parallel data created through procurement of language services at the national level; collect parallel data from contracting translation companies; increase the volumes of parallel data available to the CEF eTranslation platform; facilitate the flow of Translation Memories from translation companies to public bodies; organize bilingual Big Data at a national level and among Member States; enable PAs to leverage TMs and save on translation contracts by applying industry practices (fuzzy matching) and support translator´s work by enabling online translation (collaborative TM).

3.9 Re-usable information: Connection to open data. National and European legal framework

The talk delivered by Dubravka Bevandić, head of the Service for the protection of the Right to Access to Information from the Information Commissioner's Office, explained the importance of the Public Sector Information Directive (PSI) and re-usable data in the context of European and Croatian legal framework. Ms. Bevandić stressed the importance of the re-usage of information, which leads to innovation, the creation of new jobs, transparency and cooperation. She also introduced a four-step procedure for releasing data under the PSI Directive, which pertain to data protection, data exclusion (confidential information, personal data) and compliance with the PSI and the national PSI transposition rules.

In the second part of her talk, Ms. Bevandić emphasized the importance of the Open Data Portal as a means of re-using the data. She presented the Croatian and the European Open Data Portal, as well as the EC's priorities when it comes to open data (e.g. geolocation data, Earth and environment observation and monitoring, transportation data, statistics).

She introduced the Open Licence of the Republic of Croatia and the practice of the Information Comissioner's Office related to the right to access to Information. Ms. Bevandić finished her talk with

a few examples of re-used data in different applications (e.g. search engine for kindergartens, HAK's digital interactive map of Croatia, etc.).

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. The presentation focused on the appropriate file formats and domains of language resources that could be considered useful for training the eTranslation system. She also presented the basic steps for sharing language resource 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, e.g. tag removal, data extraction, alignment etc. Such data processing services are applied to the raw/unprocessed texts contributed to the ELRC-SHARE repository. The output of processing can be returned to the contributor, if so wished. Dr. Giagkou closed her presentation by inviting the participants to use the freely available ELRC services of technical and legal helpdesk and on-site assistance.

3.11 Identifying and managing your data: Questions & Answers

Prof. Tadić started this session with a brief overview of the Data Management Plan and related concerns. Then he gave the floor to participants. They raised the issue of the current relevance of the potential data (e.g. laws which are currently not in force), as well as the issue of data quality. Prof. Tadić stressed that the quantity of texts is of the outmost importance. Moreover, the participants raised the question of the intellectual property of translations, in particular if the translators are authorized to share their translations if they are not the authors of the source text. It was clarified that both the author of the source text and the translator are the intellectual property rights (IPR) holders of the parallel documents. However, the intellectual property right should not be confused with the right to share and distribute the documents. Granting the right to share the data, does not affect the IPR. In any case of doubts or concerns, ELRC can provide legal advice to potential data contributors. Participants also asked if the documents sent to the Central State Office for the Development of the Digital Society is shared with the ELRC, or they have to send them twice. The main issue recognized by the participants is lack of interest and awareness of the importance of sharing data among the higher level officials, who are actually the ones authorized to decide to contribute data to the ELRC-SHARE repository. Raising awareness among the decision makers is emphasized as the most important future step of the ELRC action in Croatia.

3.12 Conclusions

In the concluding session of the ELRC workshop Marko Tadić summarized the topics covered, emphasizing the importance of engagement with the ELRC action. In his concluding remarks Stelios Piperidis emphasized two issues:

1) the issue of coordination between different bodies and initiatives in Europe dealing with the collection and sharing of language resources: Mr. Piperidis, assuring the participants that all the collected data are eventually stored in the ELRC-SHARE repository through which they are forwarded to the EC and, depending on their conditions of use, to the European Open Data Portal. Thus, it is of minimal importance if a data holder decides to share his/her data through one project or channel or through another. The bottom line is that the shared language resources eventually reach their target.

2) the issue of data quality: he underlined that, although neural machine translation algorithms are robust enough not to be affected by a few poor-quality translations, quality does matter, as much as quantity does. To avoid reusing low quality translations, he asked potential data contributors to share translations even if they have concerns for their quality, but to explicitly mark them as such. In parallel, he clarified that there is no such thing as “gold” translation. There are, however, equally acceptable alternative translations. He thanked the participants, Marko Tadić, Matea Filko and Marina Petrić for organizing the workshop and Branko Baričević, the head of the European Commission Representation in Croatia, for his warm hospitality.

4 Synthesis of Workshop Discussions

In the Republic of Croatia, the situation regarding the culture of sharing the data has not changed much since the previous ELRC workshop. Namely, the main problem is the lack of interest of the decision makers, which can be illustrated by the fact that Croatian ELRC Public Services NAP has not been appointed yet. The participants of the workshop are mainly lower level officials that cannot make an official decision on sharing their language resources, and they pointed out that this type of workshop should be organized for higher level officials in order to raise awareness about ELRC efforts. The other issue highlighted at the workshop was the connection between different actions and bodies related to the collection of (language) data and language resources. Participants were interested in the relation between ELRC and other infrastructures and platforms (e.g. CLARIN and META-SHARE) on the one hand, and the Central State Office for the Development of the Digital Society on the other hand. These connections should be made clearer at the events organized in the future, or even at the ELRC website.

Another issue pertains to the publicly financed institutions which are not always willing to publicly share their data (namely, Croatian national television, which has large collections of translational data, e.g. in the form of subtitles). However, this problem is not easy to solve without strong political will.

Finally, the participants were concerned with three things related to the AT systems: 1) the control of the quality of the language data used in training the systems for machine translation (in terms of both the relevance of the documents and the quality of translations), 2) the impact of the type of the text to the quality of machine translation, 3) the accessibility of the eTranslation system to wider audience (universities, translation agencies).

4.1 ELRC and Open language Data in Croatia

The country is fully compliant with the EU copyright-related issues. The PSI directive is transposed in Croatia by the following legislative instruments: Act Nr. 403/13 of 8 March 2013 on the right of access to information (Zakon o pravu na pristup informacijama) and extracts from the General Administrative Procedure Act (Zakon o općem upravnom postupku - Act Nr. 1065/2009 of 1 April 2009), but there are several other regulations, directives and criteria related to the re-use of information as well. In order to facilitate and encourage the re-use of information, all public bodies are obliged to make the information publicly available and accessible in a digital and open format, equipped with metadata and in compliance with open standards. They are also obliged to appoint the information officers, who are in turn obliged to resolve request for information within 15 days. The central body responsible for the implementation of the PSI Directive is the Information Commissioner's Office.

The central place for the re-use of information in the Republic of Croatia is the Open Data Portal of the Republic of Croatia, available at https://data.gov.hr/. It offers the list of sets of data (geolocation data, transportation data, meteorological data, environmental data…) equipped with metadata. Users can thus find the particular data set, or search across different data distributors, topics, data formats (XLS, CSV, HTML, aspx, PDF, XML are the most common formats) and the openness degree (5 stars of open data).

4.2 Success stories and lessons learnt

Highlights of the workshop include:

- the 2nd workshop attracted interest from a much wider range of institutions in comparison to the 1st ELRC workshop in Croatia
- there is still lack of interest at the level of decision makers, obvious from the fact that ELRC Public Services NAP in Croatia has not been appointed yet
- consortiums with Croatian partners are very successful in applying for CEF funds, and this is one of the ways to encourage efforts in digitization, LRs collection and improvement of MT systems related to the Croatian language
- data holders/contributors are sometimes confused with the relation between ELRC and other data initiatives or projects (META-SHARE, CLARIN, CEF-funded projects), and don't fully understand the differences between them
- data contributors are sometimes reluctant to give their LRs due to concerns about their relevance and/or quality – this is the point that has to be included in future ELRC workshops.