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

Deliverable D3.2.24
Task 3

ELRC Workshop Report for Croatia

Author(s):	Marko Tadić, Daša Farkaš, Matea Filko
Dissemination Level:	Public
Version No.:	V2.4 - Final
Date:	2022-12-08

Contents

1 Executive Summary

The 3rd Croatian ELRC Workshop took place in the premises of the European Commission Representation in Croatia (Augusta Cesarca 4, 10000 Zagreb) on the 15th of November 2022. It was organized locally jointly by the Croatian Language Technologies Society and the DGT Field Office in Croatia. The list of invited persons encompassed over 400 addressees from different bodies of public administration and public institutions. The workshop was attended by 56 participants on-site and 45 participants online from 74 different organizations, while in the 12 sessions organized within the predefined format, 10 different presenters appeared. The whole event was video-recorded. The recordings and presentations are available online at the workshop web page https://lr-coordination.eu/croatia3rd.

2 Workshop Agenda

08:30-09:00	***Registration***
9:00-9:05	**Welcome and introduction** *Presenter:* ***Marko Tadić*** *(ELRC TNAP, University of Zagreb, Faculty of Humanities and Social Sciences)*
9:05-9:10	***Welcome by the EC*** *Presenter:* ***Andrea Čović Vidović*** *(Head of Media at the European Commission Representative in Croatia)*
9:10-9:15	***Welcome by the Central State Office for the Development of the Digital Society*** *Presenter:* ***Lidija Suman*** *(Head of the Unit)*
9:15-9:45	***The potential of Language Technology and AI – where we are, where we should be heading*** *Presenter:* ***Ana Meštrović*** *(University of Rijeka, Faculty of Informatics and Digital Technologies)*
9:45-10:15	***Common European Language Data Space*** *Presenter:* ***Philippe Gelin*** *(Head of Unit G3: Accessibility, Multilingualism and Safer Internet, DG CONNECT, EC)*
10:15-11:00	***Language Technologies by/for the public sector*** *Moderator:* ***Željka Motika*** *(ELRC PNAP, Central State Office for the Development of the Digital Society)* Panelists: ***Petra Bago*** *(University of Zagreb, Faculty of Humanities and Social Sciences)* ***Tamara Horvat Klemen*** *(Central State Office for the Development of the Digital Society)* ***Matea Filko*** *(University of Zagreb, Faculty of Humanities and Social Sciences)*
11:00-11:30	***Coffee break***
11:30-12:10	***The CEF eTranslation platform/CEF LT services*** *Presenter:* ***Szymon Klocek*** *(DG Translation, European Commission)*
12:10-12:40	***Language Technologies for Croatian: Ten years after*** *Moderator:* ***Matea Filko*** *(University of Zagreb, Faculty of Humanities and Social Sciences)* Panelists: ***Jan Šnajder*** *(University of Zagreb, Faculty Electrical Engineering and Computing)* ***Marko Tadić*** *(University of Zagreb, Faculty of Humanities and Social Sciences)* ***Ana Meštrović*** *(University of Rijeka, Faculty of Informatics and Digital Technologies)*
12:40-13:30	***Lunch break***

13:30-14:10	***Language data creation, management and sharing: existing practices and challenges in HR*** *Moderator:* ***Marko Tadić*** *(ELRC TNAP)* Panelists: ***Božo Zeba*** *(Central State Office for the Development of the Digital Society)* ***Marko Kovačić*** *(Information Commissioner's Office)* ***Mladen Stojak*** *(Ciklopea)*
14:10-14:25	***Best Practices Examples in LT: hugo.lv*** *Presenter:* ***Jānis Ziediņš*** *(Culture Information Systems Centre, Riga, Latvia)*
14:25-14:40	***Best Practices Examples in LT: EU Council Presidency Translator*** *Presenter:* ***Marko Tadić*** *(ELRC TNAP, University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb)*
14:40-15:10	***National Language Technology Platform (NLTP)*** *Presenter:* ***Artūrs Vasiļevskis*** *(Tilde, Riga, Latvia)*
15:10-15:30	***Conclusions – LT Requirements and Offerings: Do They Converge? (Q/A)*** *Moderator:* ***Marko Tadić*** *(ELRC TNAP, University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb)*
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 for Croatia. He 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 third ELRC workshop in Croatia.

The workshop participants were warmly welcomed by **Andrea Čović Vidović**, Head of Media at the Representation of the European Commission in Croatia, and **Lidija Suman**, head of the Unit at the Central State Office for the Development of the Digital Society.

3.2 The potential of LT and AI – where we are, where we should be heading?

The session *The potential of Language Technology and AI – where we are, where we should be heading?* was presented by Professor **Ana Meštrović** (Faculty of Informatics and Digital Technologies, University of Rijeka). In her talk she first presented the overall context of AI in Europe, followed by several examples of AI usage that already exist (personal assistants, chatbots, financial advice in investments, intelligent cars, diagnostics in medicine, etc.). According to the Digital Economy and Society Index (DESI) for 2022, Croatia is at the 21st position in EU and somewhat below the EU average score. She described the position of machine learning and deep learning as the methods in AI that are commonly large contributors to the development of LTs. She also explained the role of large language models that have replaced in many cases the traditional rule-based or statistical methods of natural language processing. At the end of her talk she pointed out the need for policies for data management, that would entice the institutions and companies publish and share their data, create datasets of public interest and support the legal grounds and methods to share data.

3.3 Common European Language Data Space

The session *Common European Language Data Space* was presented by **Philippe Gelin** (Head of Unit G3: Accessibility, Multilingualism and Safer Internet, DG CONNECT, EC). In his online talk he presented the latest developments at the end of the Connecting Europe Facility Programme and at the beginning of the Digital Europe Programme. The concept of data spaces, stemming from the realisation of the value of data and integrated to initiatives for a data-based EU economy, was introduced and particular emphasis was put on the Language Data Space and its future role in European LT. Also, the Centre of Excellence in LT (CELT) and Centre of Excellence in LT Plus (CELT+) were presented, where the CELT will be formed by Member States ministries representatives and CELT+ will comprise research and industry representatives. The evaluation of the call for tenders for a Common European Language Data Space had been complete at the time of the workshop, however no particular details were revealed until the official announcement. Calls for grants are envisaged for 2023 and we all hope they will provide further funds to develop LTs for under-resourced languages as Croatian certainly still falls into that category.

3.4 Language Technologies by/for the public sector (Panel session)

Moderator: **Željka Motika**, ELRC P-NAP, Central State Office for the Development of the Digital Society

Panelists:

Petra Bago, Faculty of Humanities and Social Sciences, University of Zagreb

Tamara Horvat Klemen, Central State Office for the Development of the Digital Society

Matea Filko, Faculty of Humanities and Social Sciences, University of Zagreb

The introductory part was presented by Željka Motika, ELRC P-NAP, from the Central State Office for the Development of the Digital Society. She emphasized the role of the Central State Office for the Development of the Digital Society in the data collection for the LT development and named several EU funded projects. Such projects illustrate the successful collaboration between her office and academic institutions that were responsible for the development of the LT for Croatian.

The panelists then presented some of these projects in more detail. Petra Bago presented the CEF project **PRINCIPLE** – *Providing Resources in Irish, Norwegian, Croatian and Icelandic for the Purposes of Language Engineering*. The project aimed at producing high-quality LRs to improve translation quality of eTranslation, with specific focus on three DSIs: eJustice, eProcurement and eHealth. She elaborated on the challenges in data collection from the public administration documents and emphasized that they had a really good cooperation with five data contributors from the public sector, namely 1) the Central State Office for the Development of the Digital Society, 2) the Central State Office for Central Public Procurement, 3) the Ministry of Foreign and European Affairs, 4) the State Commission for Supervision of Public Procurement Procedures, and 5) Faculty of Humanities and Social Sciences, University of Zagreb. Ciklopea d.o.o. acted in the project as the data contributor from the private sector. She offered several solutions to the problems they encountered: to incorporate data donation process into workflows of data creators; to identify language data collection as a priority in Croatia, and to build infrastructure and (financial) support at the national level to serve as a hub for collection and processing of language resources and tools as well as a centre for training stakeholders interested in contributing, developing and/or using Croatian language technologies.

Matea Filko presented two CEF projects: **MARCELL** – Multilingual Resources for CEF.AT in the Legal Domain and **CURLICAT** – Curated Multilingual Language Resources for CEF.AT. The main aim of the MARCELL project was to enhance the eTranslation system developed by the EC in the legal domain by collecting large corpora of national legislation (laws, degrees, regulations) in seven countries and in seven EU-official languages from Central and Eastern Europe. These corpora were additionally annotated with EUROVOC descriptors and IATE terms. As a separate task, a Croatian-English parallel corpus was built. It consists of 1.800 legislative documents (where Croatian is a source language) aligned at a sentence level. Corpora built through the CURLICAT project cover domains relevant for some of the CEF DSIs, such as eHealth, Europeana and eGovernment in general. The Action delivered more than 20 million sentences (containing more than 470 million words) from several domains including culture, education, health and science. Moreover, the action addressed the gap in machine translation technology, which crucially depends on the provision of domain-specific quality language resources for the under-resourced languages. She emphasized the great role of the Central State Office for the Development of the Digital Society in data collection for these projects. Finally, she discussed the importance of ELRC workshops in Croatia in raising awareness on the importance of data sharing for building LTs for Croatian.

Tamara Horvat Klemen discussed the importance of LTs in the public administration. She also emphasized the benefits of their usage, especially when it comes to the increase in efficiency. She briefly presented the National Language Technology Platform (**NLTP**) - another EU funded project in which the Central State Office for the Development of the Digital Society and the Faculty of Humanities and Social Sciences participate as partners. She presented the results of a survey addressed to public administration employees. The results have shown that there is a rising interest in LTs and the employees are aware of some of the benefits of LTs, but they are still not familiar with most of the LTs and they lack knowledge about data management and the importance of language resources they create. However, she emphasized that this is not a problem of lacking the individual training, but a problem of the community that expressed the need for LTs in total. This certainly opens a possibility for additional, more focused workshops where particular topics in LTs would be covered.

Finally, the panellists discussed the importance of explicitly recognising that LT is an important area of the AI strategy of the Republic of Croatia.

3.5 The CEF eTranslation platform/CEF LT services

The session *The CEF eTranslation platform/CEF LT services* was presented by **Szymon Klocek** (DG Translation, European Commission). In his talk he presented the current state of eTranslation system and CEF LT services. The original eTranslation system, that was developed for 24 official languages of EU plus Icelandic and Norwegian, was expanded to include other important languages for Europe such as Japanese, Arabic, Russian and Ukrainian. Also, he explained how the target user base for these services has been expanded from public administrations to European SMEs, universities and NGOs, in order to assist in the development of a data-based economy by easing the online access with instant machine translation of contents that could never be translated by humans.

3.6 Language Technologies for Croatian: Ten years after (Panel session)

Moderator: **Matea Filko**, Faculty of Humanities and Social Sciences, University of Zagreb

Panelists:

Jan Šnajder, Faculty of Electrical Engineering and Computing, University of Zagreb

Ana Meštrović, Faculty of Informatics and Digital Technologies, University of Rijeka

Marko Tadić, Faculty of Humanities and Social Sciences, University of Zagreb

The introductory part was presented by Matea Filko from the Faculty of Humanities and Social Sciences, who emphasized the problem of developing LTs for smaller/under-resourced languages such as Croatian in the AI era as the main topic of the panel. She presented the panelists who represented Croatian LT providers.

Each panelist briefly introduced the area of the LTs and the AI they work on and the moderator started the discussion by asking prof. Tadić to give a brief overview of the development of LTs for Croatian from 2012, when the *Croatian Language in the Digital Age* META-NET white paper was published, to 2022, when he authored a *Report on LTs for Croatian* as a part of the European Language Equality (**ELE**) project. He stated that a lot has been achieved in the past 10 years, but there is still much more work to do, especially with the emerging new technologies. However, we managed to build some state-of-the-art LTs, among which he singled out the EU Council Presidency Translator, which showed better results than Google Translate at the time of its launch in early 2020. However, one of the main problems is visibility of our resources, and we should make try to present them to the wider audience.

When it comes to the problem of the availability of experts, Prof. Šnajder argued that sometimes it is better to keep things small, i.e. to work in small groups and that he is overall satisfied with the development of LTs for Croatian. As Croatian is a language with less data available, he thinks that traditional models can be more accurate than the models relying on large quantities of language data. However, he elaborated that the bigger problem is money – public research institutions such as faculties, which are mainly responsible for LT development in Croatia, cannot compete with the salaries in the private sector and it is hard to keep the experts in the field.

Prof. Meštrović emphasized the importance of the good collaboration between experts and institutions. As she started to work in the field of AI and LT, the support and help of other researchers from Croatia was invaluable.

In the Q&A part of the panel, one of the participants asked what panellists think about crowdsourcing and sharing data with big companies, such as Google or Facebook, that in turn build language models for Croatian as well. The panellists shared some concerns on data privacy, and when it comes to crowdsourcing, they think that it is harder to collect large quantities of data when the language doesn't have many speakers that are digitally literate, although they completely support crowdsourcing as a concept.

3.7 Language data creation, management and sharing: existing practices and challenges in HR (Panel session)

Moderator: **Marko Tadić** (ELRC T-NAP)

Panelists:

Božo Zeba (Central State Office for the Development of the Digital Society)

Marko Kovačić (Information Commissioner's Office)

Mladen Stojak (Ciklopea)

The introductory part was presented by Marko Tadić from the Faculty of Humanities and Social Sciences, who pointed out that current practices for language data collection in Croatia and the legal framework for such activities are challenging. Also, the question about the positioning of the new Digital Europe Programme, a successor to CEF Programme, in certain Ministries (Economy or Science) was raised. The panelists shed a light to these problems from three different perspectives.

Božo Zeba from the Central State Office for the Development of the Digital Society explained what the role of this institution in overall open data policies and practices in Croatia is. He pointed out that this Central State Office maintains the portal for open data in Croatia (https://data.gov.hr) and that a number of datasets are freely accessible there, including language datasets. Also, the Central State Office collects and publishes all official documentation of the Republic of Croatia and provides a portal for free access to this documentation with language technologies support (i.e. lemmatized search) to document retrieval. The Central State Office served as a source of language data in a number of CEF projects (MARCELL, NEC-TM, PRINCIPLE, EU Presidency Translator, CURLICAT) and it is a partner in the NLTP project.

Marko Kovačić, from the Information Commissioner's Office, briefly described the current legal framework that allows sharing of language data. He mentioned the existing differences between the previous EU PSI directive and new Open Data Directive and what will be its role in new Digital Europe Programme.

Mladen Stojak, from Ciklopea, a private translation and localization company, shared with the audience insights of the problems that a SME, a language data provider and language data user, is confronted with. He pointed out the necessity of introducing a data sharing culture, such as delivering the generated TMs with the translations that were outsourced to them. Ciklopea participated as a partner in project NEC-TM, so he also shared an experience of SME in a CEF project.

3.8 Best Practices Examples in LT: hugo.lv

The session *Best Practices Examples in LT: hugo.lv* was opened by **Jānis Ziediņš** (Culture Information Systems Centre, Riga, Latvia). His talk opened the block of presentations with examples of best practices in LT for the public sector. First, the hugo.lv platform, that encompasses several LT services, was described in detail including its development process. The experience collected with launching and deploying the platform was also presented. This platform served as one of two predecessors to a current National Language Technology Platform (NLTP) project.

3.9 Best Practices Examples in LT: EU Council Presidency Translator

The session *Best Practices Examples in LT: EU Council Presidency Translator* was presented by **Marko Tadić** (University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb). In his talk he presented the other predecessor to the NLTP project, the EU Council Presidency Translator that was developed from 2019 to 2020. This machine translation system for English<->Croatian was deployed in January 2020 and at the time it generated higher quality translations for these two language pairs than Google Translate. Since then, it became very popular in many public sector institutions. From January 2020 until October 2022 it translated 43.6 Mw in Hr->En and 178.3 Mw in En->Hr directions.

3.10 National Language Technology Platform (NLTP)

The session *National Language Technology Platform (NLTP)* was presented by **Artūrs Vasiļevskis** (Tilde, Riga, Latvia). In his talk he described the NLTP project, the collection of mono- and multilingual data from five partnering countries (Latvia, Estonia, Croatia, Iceland and Malta), the development of the platform that will encompass a number of LT services, predominantly MT and CAT, but also terminology management and speech processing modules (AST, TTS) for selected languages. He also invited the audience to a launching event of NLTP in Croatia that will take place in late February or early March 2023.

3.11 Conclusions – LT Requirements and Offerings: Do They Converge? (Q/A)

In the concluding session of the ELRC workshop **Marko Tadić** summarized the topics covered at the workshop, emphasizing the importance of engagement with the ELRC action that provided a role model how to start the language data collecting campaign in Croatia using the experience and best practices from other member states.

4 Synthesis of Workshop Discussions

After two similar ELRC workshops in Croatia, we can claim that we have managed to establish a community of not only local LT experts, who are gathered around the Croatian Language Technologies Society anyway, but also of public sector representatives and private companies. This is clearly illustrated in the numbers of participants rising steadily form the 1st ELRC Workshop in Croatia until the 3rd. We believe that, even if not fully established, a language data sharing culture in Croatia has been proposed as an expected *modus operandi* with no legal consequences. This is already an important step forward.

The points raised in panels and discussions can be sorted in two main groups: 1) Achievements of Croatian LT in the last ten years, including the previous ELRC workshops in Croatia as well as best practice examples; 2) Plans for future development of Croatian LT and outlook to the European LT landscape. In this respect, the technical and legal issues involved in in collecting and processing language data were also discussed, and possible solutions were proposed.

Additionally we would like to highlight the following topics, which we consider important:

- Beside the existing research institutions in LT in Croatia, the role of the Central State Office for the Development of the Digital Society has become of crucial importance, not just for the access to the governmental or state open data, but for language data from these sources as well. In fact, this Central State Office could even serve as a role model to countries with several millions of inhabitants for a concentrated method of dealing with eGov services and its digital/digitalisation processes. These services can then support advancement in and deployment of LT in public administration and economy.
- Thanks to this role of the Central State Office for the Development of the Digital Society we can say that we have established a community of Language Data users and providers, who are also perspective LT users. The address book of this community in Croatia has more than 400 items already.
- The role of ELRC campaign for collecting Language Data has been proven to be important as well. Namely, concentrated and coordinated European initiatives for collecting and sharing language data have a positive effect in convincing and mobilising the local public sector, they make the LT field more visible, and they help the existing European LT community grow.

5 Country Profile: Language data creation, management and sharing

An updated profile for Croatia with respect to language data creation and management has been recently published in the 2022 ELRC White Paper (see: https://lr-coordination.eu/sites/default/files/LRB/LRB-12/ELRC-White-Paper.pdf).