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 Croatia

Author(s):	Marko Tadić, Croatian Language Technologies Society
Dissemination Level:	Public
Version No.:	V1.1
Date:	2016-6-20

Contents

1 Executive Summary

The Croatian ELRC Workshop took place in the premises of the Representation of the European Commission in Croatia (Augusta Cesarca 4, 10000 Zagreb) on the 21st of April 2016. It was organized jointly by the Croatian Language Technologies Society and the DGT Local Office in Croatia. The list of invited persons encompassed over 400 addresses from different bodies of public administration and public institutions. The workshop was attended by 56 participants from 35 different organizations, while the 13 sessions organized within the predefined format, were presented by 14 speakers. The whole event was video-recorded and the recordings with presentation slides are available online at the workshop web page http://lr-coordination.eu/hr/croatia.

The workshop was welcomed by the head of the Representative of the European Commission in Croatia, Mr. Branko Baričević. The most interesting parts of the workshop were two panels,that initiated lively discussions, and the presentation by the information commissioner of the Republic of Croatia Ms Anamarija Musa on the EU legal framework regulating the public data sharing.

2 Workshop Agenda

08:30 – 09:00	**Registration**
09:00 – 09:20	**Opening and Welcome** **Branko Baričević** (Head of the EC Representation in Croatia)
09:20 – 09:30	**ELRC Workshop Aims and Objectives** **Stelios Piperidis** (ILSP/ELRC)
09:30 – 10:00	**The EU and Multilingualism** **Marina Petrić** (DGT-Zagreb Office)
10:00 – 10:30	**Language and Language Technology in Croatia** **Marko Tadić** (University of Zagreb, Faculty of Humanities and Social Sciences)
10:30 – 11:00	**Panel: Multilingual Public Services in Croatia** **Moderator:** **Marina Petrić** (DGT-Zagreb Office) **Panelists:** **Gordana Štampar** (Croatian Bureau of Statistics) **Miljenka Prohaska-Kragović** (Ministry of Foreign Affairs & European Integration) **Maja Bratanić** (Institute of Croatian Language and Linguistics) **Romana Sinković** (Croatian National Bank)
11:00 – 11:30	Coffee Break
11:30 – 12:00	**How Automated Translation works?** **Marko Tadić** (University of Zagreb, Faculty of Humanities and Social Sciences)
12:00 – 12:30	**How can Public Institutions benefit from the CEF.AT Platform?** **Spyridon Pilos** (European Commission - video conference)
12:30 – 13:30	Lunch Break
13:30 – 14:00	**What Data is needed?** **Sanja Seljan** (University of Zagreb, Faculty of Humanities and Social Sciences)
14:00 – 14:30	**Legal Framework for contributing data** **Anamarija Musa** (Information Commissioner of the Republic of Croatia)
14:30 - 15:00	**Panel: Data and Language Resources in Croatia: Where to Look for What** **Moderator:** **Marko Tadić** (University of Zagreb, Faculty of Humanities and Social Sciences) **Panelists:** **Marija Brčić** (Digital information and communication office of the Government) **Damir Pavuna** (Community of Translators, Croatian Chamber of Economy) **Marina Orešković** (Ciklopea d.o.o.)
15:00 - 15:30	Coffee Break
15:30 - 16:00	**Data and Language Resources: Technical and Practical Aspects** **Stelios Piperidis** (ILSP/ELRC)
16:00 - 16:30	**Interactive session: How can we engage?** **Marko Tadić** (University of Zagreb, Faculty of Humanities and Social Sciences) **Marina Petrić** (DGT-Zagreb Office) **Stelios Piperidis** (ILSP/ELRC)
16:30 - 16:45	**Concluding remarks and discussion** **Stelios Piperidis** (ILSP/ELRC) **Marko Tadić** (University of Zagreb, Faculty of Humanities and Social Sciences) **Marina Petrić** (DGT-Zagreb Office)

3 Summary of Content of Sessions

3.1 Session S1: "Opening and Welcome"

The Croatian ELRC Workshop was opened by the organizers: Marko Tadić, president of the Croatian Language Technologies Society and Marina Petrić, DGT representative in Croatia. The workshop participants were warmly welcomed by Branko Baričević, the head of the Representative of the European Commission in Croatia with several remarks on the importance of the 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 to appear in such a number and wished the successful workshop.

3.2 Session 2: "ELRC Workshop Aims and Objectives"

In this session Stelios Piperidis (ILSP, Athens) presented the broad workshop context: the fundamental multilinguality of EU and the proposed solution how to assure the equal opportunities to all official languages (+2 in CEF) and how languages should be regarded as assets and not problems. The EU economy should use them as valuable source instead of discarding their role. The translation has been regarded as the main operation that is used to overcome the existing language barriers within EU and EU is taking all the necessary steps to ensure that from the translation of all official documents to all official languages of the EU, until the offering of the freely available automated translation systems to different bodies of public administration and institutions of member states. In this respect the CEF.AT is presented as an online CEF Digital Service Infrastructure that focuses on citizens, public services, administrations, ministries, public institutions etc. In order to make CEF.AT more reliable and useful for needs of different clients, more textual data should be collected. In this respect the ELRC has been established as the permanent mechanism that will feed CEF.AT with relevant language resources. The NAP for Croatia were introduced and their contact points communicated.

3.3 Session 3: "EU and Multilingualism"

In this session Marina Petrić presented the very multilingual nature of the EU starting from the fundamental treaties until the newest directives. The translation services that assure the availability of all official EU documents in all official languages today don't translate manually any more, but are strongly supported by different language technologies. In different EU bodies more than 4300 translators and more than 1000 interpreters work on the daily basis, while DGT alone encompasses more than 2500 employees situated in Brussels, Luxembourg and representatives in member states. In 2014 more than 2.3 million pages (690 million words) were translated predominantly from English (81%) and other languages (12.5%), French (3.7%) and German (2.8%). The participants were informed on the existing CAT tools and language resources (like EURAMIS, EUR-Lex, IATE, DGT TM…) used by EU translation services as well as MT@EC system for automated translation. Also, the role of translation in the Digital Single Market has been pointed out where the Digital Public Sector has been mentioned as the part of the DSM strategy for overall use of pan-European public services, such as: e-justice, e-procurement, e-health, Europeana, open data portal etc. For these services there is not a single operational language (no lingua franca), but all EU official languages should be represented. In this services the CEF.AT plays unavoidable role of multilingual mediator between different bodies of public administration and between these bodies and EU citizens.

3.4 Session 4: "Language and Language Technology in Croatia"

In this session Marko Tadić presented the situation with different languages and language technologies in the Republic of Croatia. The article 12, alinea 1 of the Constitution of the Republic of Croatia the "the official language in the Republic of Croatia is the Croatian language and the official script is the Latin script." However, in practice there are fields where other languages are used such as tourism, trade, cross-border cooperation, national minorities. The article 12, alinea 2 declares that "in individual local communities beside the Croatian language and Latin script, some other language and Cyrillic or some other script can be introduced in the official use according to the law." By the census of the Republic of Croatia from 2011, mother tongues for 4.285 million of inhabitants were: Croatian (95.6%), Serbian (1.23%), Italian (0.43%), Albanian (0.40%), Bosnian (0.39%), Roma (0.34%), other (<0.3%). Most of inhabitants of Croatia speak at least one foreign language since the first foreign language is obligatory from the primary school and often a second is added in the high school. Very important, but often neglected cause of good knowledge of foreign languages in Croatia is that there are no foreign movies dubbing, but subtitles are used (apart in the movies for children under 6). So all inhabitants of Croatia are constantly exposed to multimodal parallel corpora processing while following the broadcasts originally recorded in a foreign languages. In Croatia there are several LT centres: University of Zagreb (Faculty of Humanities and Social Sciences, Faculty of Electrical Engineering and Computing), Institute of Croatian Language and Linguistics, University of Rijeka (Department of Information Sciences). The Croatian Language Technologies Society has been established in 2004 and functions as the coordinating association for the activities in the field of Croatian LT. By the end of this presentation the most important language resources and tools for processing Croatian were presented.

3.5 Session 6: "How Automated Translation works?"

Marko Tadić presented in this session the basics of the Machine Translation in a schematic way in order to make the workshop participants aware of the importance of large quantity of multilingual data needed by the MT systems. The template predefined by ELRC coordinator was translated and modified to accommodate the Croatian particularities and examples that made the notion of MT more present and closer to the audience. The difference between human translation and MT was stressed and explained and the difficulties of translating process were illustrated by description of steps that are needed to be undertaken for the preparation and usage of the Statistical Machine Translation systems. The principle of SMT systems were explained in a clear way first on a single-word and later on the phrase-based translation approaches. The role of SMT approach in CEF.AT was presented at the end of this presentation.

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

Spyridon Pilos held a video-conference presentation of CEF.AT platform and its usage by public institutions. He situated the CEF.AT within the framework of pan-European public services that provide information in any member state in any place in a requested mother tongue. In this framework MT is the only viable solution for accessing information, for receiving information, as well as for search analytics. This role makes MT critically important for multilingual EU. MT@EC services were opened on 2013-06-26 and today offer translation in all directions between all 24 official EU languages. MT@EC operates as web user interface, but also as web services with lot of additional features such as preservation of the source document formatting and indication of quality for individual language pairs (BLEU scores). MT@EC is today used by EU institutions and bodies, but also by online services supported by EU and public administrations of all EU/EEA countries. The primary source of data for this SMT is EURAMIS that by the end of 2015 measured more than 940 million aligned sentences. This MT@EC service represents a technological fundament for CEF.AT platform that not only provides MT service, but also enables a full multilingualism in EU. The main reason for this is more and better data that will be collected through ELRC initiative.

3.7 Session 8: “What Data is needed?”

In this session Sanja Seljan presented what kind of data is needed for enhancing the existing MT systems and why. She listed the possible sources of multilingual data that could be used for training SMT systems and explained the differences between parallel corpora and monolingual corpora. She also presented how aligned text sequences represent the most valuable source of training data. The second part of her presentation for focused on the types and formats of texts are needed (including their metadata descriptions). Also, she pointed out the difference between the “Public Web” and the “Deep Web” where a lot of valuable multilingual data is hidden and needs to be taken out to the surface and used for massive enhancement of SMT systems. She particularly illustrated how the quality of translation by SMT systems increases with the grow of available training data and this made a clear argument for supporting the ELRC initiative.

3.8 Session 9: “Legal Framework for contributing data”

The Information Commissioner of the Republic of Croatia, Anamarija Musa, presented the European legal framework for public information and data sharing. The reusage of publicly available information and data is vital for the development of single european market for innovative applications based on data generated/available in the public sector. The main legal instrument that regulates this is the Public Sector Information Directive (PSI) from 2013 that sorted out many weaknesses from the previous version of this Directive. The Information Commissioner explained to the audience that PSI takes into account four main sets of rules, namely, 1) rules of PSI directive itself; 2) rules of IPR; 3) rules of data protection; 4) rules of data exclusion (for security or privacy protection reasons). One of the main advantages of reusing of the public data is that they are already in the digital format and this should boost the digital market since particularly open data are freely available under different licensing mechanisms. These licensing schemas and the mechanism of quality/accessibility scoring were also explained by Anamarija Musa during her presentation. The presentation ended with a series of examples of sites where different kinds of publicly produced and accessible data were available.

3.9 Session 11: “Sharing Data and Language Resources: Technical and Practical Aspects”

In this session Stelios Piperidis described the workflow from data to language resources with identification of all steps in this process. These steps included identification and selection of data including the basic documentation about it. Legal status and licensing mechanisms are where ELRC can help the data providers with. Also, ELRC can help with the metadata description of data sources. The next step is cleaning data from the overhead that introduces noise in the process of training the SMT systems: boilerplate removal, data anonymization, etc. Validation of data quality is the next step and uploading the data in the LR repository the following step needed for the processing. What is needed from the data providers are identification of data sources, of raw data sets, processing that can be carried out in collaboration with the ELRC and data provider. ELRC portal already provides many supporting services in this process: technical and legal helpdesk, forum and repository. Stelios Piperidis then illustrated the procedure of uploading the data to the ELRC repository.

3.10 Session 12: “Interactive session: How can we engage?” and “Concluding remarks”

In this final session Stelios Piperidis, Maja Petrić and Marko Tadić again explained the purpose of the ELRC initiative and invited the audience for their final remarks. The discussion followed an optimistic mood since a lot of potential data sources have been detected and many useful applications of the CEF.AT platform were announced and expected by the translators working in different bodies of public administration, institutions, but also private translation companies. The workshop was recognized as very successful at the end.

4 Synthesis of Workshop Discussions

4.1 Panel 1: “Multilingual Public Services in Croatia”

In the panel “Multilingual public services in Croatia” the participants were Gordana Štampar (Croatian Bureau of Statistics), Miljenka Prohaska Kragović (Ministry of Foreign and European Affairs), Maja Bratanić (Institute of Croatian Language and Linguistics) and Romana Sinković (Croatian National Bank) while the moderator was Maja Petrić (EC, DGT). The panelists (Štampar and Sinković) first presented the existing multilingual public services and how they practically use them in their everyday work, while Kragović presented the ten-year process of translation Acquis Communautaire into Croatian, and where and how these translations can be used today. Bratanić presented the valuable Croatian terminological online resource, namely, the Croatian terminological infrastructure, nationally funded through the project STRUNA. It is realized as a system of terminological databases for 18 professional fields (currently completed) that cover more than 31,000 prescribed Croatian terms with additional 15,000 synonyms and more than 100,000 English translational equivalents (obligatory). This terminological resource is included in the metasearch engine Quest used by EC translation services. The discussion that followed concentrated on different aspects of multilingual public services in Croatia: their very existence, their availability and conditions of their accessibility (how, when, to whom etc.).

4.2 Panel 2: “Data and Language Resources in Croatia: Where to Look for What”

In the panel “Data and language resources in Croatia” the participants were Marija Brčić (Digital Information-documentation Office of the Republic of Croatia), Damir Pavuna (Integra d.o.o. & Translators Union at Croatian Chamber of Commerce), Marina Orešković (Ciklopea d.o.o.) and the moderator was Marko Tadić. Marija Brčić presented the role and the repository of the Digital Information-documentation Office of the Republic of Croatia. This Office collects all the official documents of any body of public administration in Croatia and enables permanent access to these documents online. Several document collections are composed of: 1) Collection of legal regulations (over 30,000 valid and non-valid Croatian legal regulations, 2990 EU regulations, 1780 unofficial translations from Croatian into English of different legal regulations; 2) Collection of 4736 international treaties that Croatia signed (valid and not-valid); 3) Collection of the official journals of bodies of local and regional administration: 29,000 volumes of journals (all in digital form); 4) Collection of other documents and publications: 35,000 monographs, almost 2000 periodicals and 880 publications translated into English. This presentation undoubtedly positioned the Digital Office as one of the most valuable sources of data for CEF.AT and ELRC initiative when the Croatian language is concerned. Damir Pavuna presented the problems that the individual translators and translation offices meet when trying to work in translation and localization business in an organized way. He criticized the fragmentation of the production and usage of different language resources and tools, since they are produced either by LT research institutions (and predominantly for research purposes), or by private companies/freelance translators (and predominantly kept in house), or by public institutions and bodies of public administration (and also kept in house). Marija Orešković presented the usage of language resources in a commercial environment and all details connected to these (IPR, issues of reusabiliy, availability of public resources etc.). The discussion in this panel was seen as one of the most fruitful and vivid in many panels so far.

5 Workshop Presentation Materials

All presentation slides and videos are available online from the Croatian ELRC Workshop web site: http://lr-coordination.eu/hr/croatia.

The Croatian ELRC workshop has also been widely covered in the e-media:

1) http://eu.hina.hr/content/9196129
2) http://eu.hina.hr/content/9195046
3) http://direktno.hr/en/2014/domovina/45627/Prikupljanje-jezičnih-resursa-za-kvalitetnije-strojno-prevođenje.htm
4) http://ec.europa.eu/croatia/news/2016/20160420a_hr.htm
5) http://www.hgk.hr/zajednice/zajednica-za-prevoditeljstvo/europska-koordinacija-jezicnih-resursa-european-language-resources-coordination-elrc?category=569
6) http://www.akademija-art.hr/globus/eu-komisija/36257-europska-komisija-razvija-djelotvorniji-sustav-racunalnog-prjevodenja
7) http://www.pristupinfo.hr/radionica-elrc-a-pravni-okvir-za-ponovnu-uporabu-informacija-javnog-sektora-zagreb-21-travnja-2016/
8) http://www.prs.hr/index.php/suradnja/druge-suradnje/1942-radionica-europske-koordinacije-jezicnih-resursa
9) https://www.facebook.com/integra.doo.7
10) https://www.facebook.com/poslovnadogadanja/posts/472732259592956