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 D3.2.25
# Task 3

# ELRC Workshop Report for Bulgaria

**Author(s):** Hristina Dobreva, Kalina Ivanova, Ministry of Transport and Communications of the Republic of Bulgaria

**Dissemination Level:** Public

**Version No.:** V3.2 - Final

**Date:** 2023-01-10

# Contents

# 1 Executive Summary

This document reports on the third European Language Resources Coordination (ELRC) workshop in Bulgaria. The event was held online via Webex on November 24, 2022.

More than 80 participants registered for the workshop from 13 ministries, 8 commissions and agencies, the administration of the Council of Ministers, the Bulgarian Academy of Sciences, the Supreme Judicial Council, 6 universities, 7 companies working in the field of professional translation, information and language technologies, as well as from the European Commission and foreign participants interested in the Bulgarian experience. According to the participants' list from Webex, 74 participants took part in the meeting, which does not include the 2 interpreters.

The Secretary General of the Ministry of Transport and Communications Mr. Ivan Markov officially opened the event. The workshop moderator was Mrs. Hristina Dobreva, National Anchor Point from the Public Administration, and a member of the ELRC Language Resource Board.

The main presentations at the forum were made by her and the Technology National Anchor Point - Prof. Dr. Svetla Koeva from the Institute of Bulgarian Language at the Bulgarian Academy of Sciences. The main topics of the workshop were the importance of data, the possibilities of automatic translation and the eTranslation platform, machine learning, the potential of artificial intelligence, technological components and language technologies in Bulgarian, details on the creation of a European Language Data Space, etc.

The participants had the opportunity to get more information from the representatives of the European Commission about the new initiative for the European Language Data Space (Mr. Dhafer Lahbib, DG CNECT) and about the development and new applications of the automatic translation platform eTranslation (Mr. Bogomil Kovachev, DG Translation).

In the panel session "Language technologies by/for the public sector" Mrs. Annie Rusinova from the National Revenue Agency presented her personal experience of using eTranslation and sharing data with the ELRC-SHARE repository. Mr. Tsvetan Deyanov from the Ministry of e-Governance demonstrated the implementation of eTranslation functionalities in the Single Portal for Access to Electronic Administrative Services.

In the panel session "Language data creation, management and sharing: existing practices and challenges" Mr. Stanimir Minkov from the Administration of the Council of Ministers presented the Portal for public consultations - strategy.bg and the Platform for access to public information - pitay.government.bg, and Mrs. Reny Borisova from the Ministry of e-Governance presented the Portal for open data - data.egov.bg. The fruitful discussion was complemented by the participation of Dr. Orlin Kouzov, the head of the Education Laboratory at the Big Data for a Smart Society Institute (GATE), who shared his views on the application of various innovative technological solutions, and specifically language technologies in the public sector, the role of artificial intelligence and the use of big data in detecting misinformation and disinformation. With a video presentation, Dr. Temenuzhka Spasova from the Ministry of e-Governance presented the National Spatial Data Portal - inspire.egov.bg.

During the workshop, there was also a demonstration session of language and artificial intelligence technologies in Bulgaria by Prof. Dr. Svetla Koeva, who presented the CURLICAT project - Curated multilingual resources for CEF.AT, and Dr. Yordan Kralev from the Technical University - Sofia, who presented applications and services using the Multilingual Image Corpus.

## 2 Workshop Agenda

| | |
|---|---|
| 09:30 - 09:40 | Welcome and introduction<br>Mr. Ivan Markov, Secretary General of the Ministry of Transport and Communications<br>Mrs. Hristina Dobreva, State expert in the Ministry of Transport and Communications and National Anchor Point |
| 09:40 - 10:10 | The Language Data Space<br>Mr. Dhafer Lahbib, European Commission, DG CNNECT |
| 10:10– 10:40 | Language Technologies for Bulgarian language<br>Prof. Dr. Svetla Koeva, IBL, BAS and National Anchor Point |
| 10:40 – 11:00 | The potential of Language Technology and AI – where we are, where we should be heading<br>Mrs. Hristina Dobreva, State expert in the Ministry of Transport and Communications and National Anchor Point |
| 11:00– 11:15 | Break |
| 11:15– 11:45 | The CEF AT Platform<br>Mr. Bogomil Kovachev, European Commission, DG Translation |
| 11:45– 12:15 | Language technologies by/for the public sector<br>Moderator:<br>Mrs. Hristina Dobreva, State expert in the Ministry of Transport and Communications and National Anchor Point<br>Participants in the panel sessions:<br>Mrs. Annie Rusinova, National Revenue Agency<br>Presenting personal experience with sharing language resources and using the eTranslation<br>Mr. Tsvetan Deyanov, Ministry of e-Governance<br>Presenting the machine translation of text in the egov.bg portal |
| 12:15– 12:45 | Language data creation, management and sharing: existing practices and challenges<br>Moderator:<br>Mrs. Hristina Dobreva, State expert in the Ministry of Transport and Communications and National Anchor Point<br>Participants in the panel sessions:<br>Mr. Stanimir Minkov, Council of Ministers Administration<br>Presenting the Portal for public consultations - strategy.bg and the Platform for publicly available information - pitay.government.bg<br>Mrs. Reny Borisova, Ministry of e-Governance<br>Presenting the Open data portal - data.egov.bg<br>Dr. Orlin Kouzov, GATE<br>Presenting the application of technologies in administration<br>Dr. Temenuzhka Spasova, Ministry of e-Governance<br>Presenting the National Spatial Data portal - inspireportal.egov.bg (video)<br>Mrs. Hristina Dobreva, State expert in the Ministry of Transport and Communications and National Anchor Point<br>Presenting the Single information portal - COVID-19 - coronavirus.bg |
| 12:45– 13:30 | Demo session of LT and AI technologies available for Bulgaria |

| | |
|---|---|
| | Prof. Dr. Svetla Koeva, IBL, BAS and NAP: Collection of multilingual resources for CEF.AT<br>Dr. Yordan Kralev, Technical university of Sofia: Multilingual image corpus for multimodal content processing |
| 13:30 – 13:45 | Conclusions<br>Mrs. Hristina Dobreva, State expert in the Ministry of Transport and Communications and National Anchor Point |

# 3 Summary of Content of Sessions

## 3.1 Welcome and introduction

The event started at 9:30 (EET) and was opened by Mrs. Hristina Dobreva, Public Services National Anchor Point, and a member of the Language Resource Board. She welcomed the participants in the third ELRC Workshop in Bulgaria, organized by the Ministry of Transport and Communications of the Republic of Bulgaria and the Institute for the Bulgarian Language at the Bulgarian Academy of Sciences. More than 80 participants registered for the seminar from 13 ministries, 8 commissions and agencies, the administration of the Council of Ministers, the Bulgarian Academy of Sciences, the Supreme Judicial Council, 6 universities, 7 companies working in the field of professional translation, information and language technologies, from the from the European Commission, as well as foreign participants interested in the Bulgarian experience.

The Secretary General of the Ministry of Transport and Communications Mr. Ivan Markov officially opened the event. He briefed the participants on the aim of the ELRC project, namely, to bridge the gap between the real needs and requirements of EU public administrations and the capabilities of machine translation systems, and to support all national languages in their quest for affirmation and full rights. Languages are an important part of our European cultures and identities. Linguistic diversity is a great wealth, but also a serious challenge. In Europe, national borders are open, but linguistic ones still exist. The need for administrations to develop digital administrative services in more than one language can make this process more expensive. In this respect, modern language technologies such as machine translation can help to overcome language barriers while at the same time contributing to preserving linguistic diversity. As administration representatives, it is important to know that modern language technologies are based on machine learning. Machine learning is a process in which machines improve by learning from enough high-quality training data. Therefore, all documents, materials, and data that are created and stored by administrations are invaluable for training such systems. Public administrations and services in Europe have a great need for translation and produce a lot of linguistic data in different languages. This data is inherently public and subject to reuse and sharing. Since 2015, The Ministry of Transport and Communications, through its National Anchor Point, is part of the of ELRC project. The ELRC successfully collects language data resources in all official European languages, with a special focus on bilingual and multilingual language data from different domains. As a good practice, Mr. Markov noted that the Ministry has a common registration of its domain in the automatic translation platform of the European Commission eTranslation, and employees can use it without the need for personal registration.

Mr. Markov wished a fruitful workshop and discussion to all participants and expressed hope that the forum will further raise awareness about the importance of language data collected and stored by public administrations and the usefulness of their sharing.

The workshop continued with an introduction by the National Anchor Point - Mrs. Hristina Dobreva about language technologies, the ELRC project, and funding under the Digital Europe programme in different areas, including artificial intelligence.

Mrs. Dobreva said that this year the scope of the workshop would be broadened beyond machine translation and language data. The true digital transformation of Europe and the country needs more technological components, more language technologies that, combined with machine translation, can effectively contribute to the multilingual Digital Single Market.

Mrs. Dobreva began her presentation with an example of everyday language technology interaction with mobile phones, in cases when we ask them simple and more complex questions. She explained that to receive an answer from the mobile device, it automatically recognises speech, i.e. it understands our natural language. In order to provide an answer, the device analyses our text,

searches, and selects relevant information, then it generates natural language and provides us with the appropriate answer suited to our query.

Mrs. Dobreva provided general information about the European Language Resource Coordination (ELRC) project, which has been running since April 2015, collecting language resources. Most of the resources are actually existing translations or translated texts, translation memories (i.e. a database where translated segments are stored), multilingual corpora (i.e. large collections of text in several languages), but also terminologies (i.e. a collection of terms and their usage). She informed that the project ends in early 2023.

Mrs. Dobreva drew attention to the Commission's AI strategy and said that maximizing resources and coordinating investments is vital and a key element for the success of the set objectives. The National Anchor Point explained that through the Digital Europe programme, which is the EU's first financial instrument targeting digital technologies, and through the Horizon Europe programme, the Commission plans to invest €1 billion per year in AI. According to Mrs. Dobreva, the Digital Europe programme aims to bridge the gap between digital technology research and market deployment. It will benefit European citizens and businesses, especially small and medium-sized ones. Digital Europe investments support the European Union's twin goals of a green transition and digital transformation and enhance the Union's resilience and technological sovereignty. The two programmes aim to mobilize additional investment from the private sector and Member States and achieve an annual investment volume of €20 billion over the next ten years. Mrs. Dobreva also said that the Recovery and Resilience Facility is the largest stimulus package ever funded by the EU budget, allocating €134 billion to digital technologies. It will fundamentally change the landscape, enabling Europe to step up its ambitions and become a world leader in developing cutting-edge reliable AI.

The participants in the workshop were asked to complete the ELRC country survey for Bulgaria as well as their evaluation form (workshop evaluation form) for the event. They were informed that if anyone needed a certificate of attendance, they should contact the organizers and they would send them one. After the workshop 5 certificates of attendance were issued upon request.

## 3.2 The European Language Data Space

In the agenda of the Third National ELRC Workshop the participation of Mr. Philippe Gelin was foreseen to give a detailed presentation of The European Language Data Space, but as he was not able to take part in the event the presentation was given by Mr. Dhafer Lahbib, European Commission, DG CNNECT.

Mr. Lahbib introduced the Digital Europe programme, which provides funding in five key areas: supercomputing; artificial intelligence; cybersecurity; deep digital skills, and best use of digital technologies. Funding for data is also provided through this programme. Participants were briefed on the Commission's 2030 goals and the Digital Decade objectives. The main focus of the programme is to build the EU's strategic digital capacity and facilitate the widespread deployment and use of digital technologies. The main objective of the programme is to bridge the gap between digital technologies, research, and market deployment while supporting the EU's green transition and digital transformation objectives.

Mr. Lahbib drew attention to the origins of the common European data spaces that are the center of the European Data Strategy. Their main objective is to harness the value of data for the benefit of the European economy and society. The creation of common, interoperable data spaces across the EU in strategic sectors should overcome existing legal and technical barriers to data sharing and thus release the potential of data-driven innovation. The Commission's aim is to create a true single market for data.

The main principles for creating data spaces are data control, mainstreaming into government policies, respect for European rules and values, provision of technical infrastructure for data, interconnection, interoperability, and openness.

The purpose of creating data spaces is to increase the volume of data exchange. AI-based complex language services need large datasets to be trained. Text, audio, and video files are a large part of the data we create daily. To process and extract relevant information from this data, we need to have AI-based complex language services.

Mr. Lahbib went into detail on the structure of the European data space ecosystem. As a guiding part of this ecosystem, a European Data Innovation Board (EDIB) will be set up to assist the EC in issuing guidelines to facilitate the development of data spaces as well as to define relevant standards and interoperability requirements for cross-sectoral data sharing.

He drew attention to the working directions of language data spaces. Mr. Lahbib said that as a first step, an Institutional Center of Excellence for Language Technologies (CELT) should be established. The goal is to create a framework for collecting, creating, sharing, and reusing language data and models, including legal aspects.

The next step is to create a language data space. This requires creating the appropriate architecture, promoting data sharing, and supporting the deployment of large multimodal language models and a wide range of AI-based language technologies on the AI-on-demand platform.

At a strategic level, the CELT will coordinate the creation and collection of multimodal language data and models across Member States. In terms of governance, a multi-stakeholder data and services system and data governance scheme will be developed.

At the operational level, CELT+ will be established with participation by mainly industry. CELT+ will oversee the development of a language data plan, i.e. blueprint, based on existing EU/national legislation, including a detailed roadmap for the implementation of the language data space. The role of the European Commission is to support all these processes.

Mr. Lahbib showed a detailed diagram of the distribution architecture of the language data space and the possible stakeholders that can be involved. Stakeholders could be from the public and private sectors, industry, media, research institutes, cultural associations, and public administration.

According to Mr. Lahbib, the language data space is a platform and a marketplace where demand for and supply of language resources and services are matced. The benefits of the language data space could be easy access to data, data flows/streams, more data available, more data trading, data effort monetisation, and data revaluation.

New data-driven products and turnkey solutions will be available throughout the data space, e.g. multilingual search engines, Q&A, etc. Customised language technologies and services will be possible, e.g. anonymization, business intelligence through data, data analytics, sentiment analysis, etc., and access to international markets through multilingual solutions.

The space will facilitate finding partners and collaboration opportunities, sharing best practices and insights, receiving and providing feedback from/to other stakeholders, and introducing targeted improvements to reduce time to market.

The data space will contribute to sharing data against remuneration e.g. multilingual and/or multimodal data, to the development, sale, and purchase of language-based applications, tools, and models e.g., automatic speech transcription, automatic text summarisation.

Mr. Dhafer Lahbib presented details of the European Digital Infrastructure Consortium (EDIC) which is in the process of being legally established. This is a new legal framework to support multi-country projects. The concept is inspired by the success of the European Research Infrastructure Consortium

(ERIC). The advantages of establishing an EDIC could: be easier coordination of funding between Member States, rapid establishment, and flexible implementation, ensuring common standards and interoperability, and sustainability.

Significant milestones and next steps for the European Digital Infrastructure Consortium are/have been: a political agreement in July 2022; informal call for expressions of interest: November/December 2022; expected entry into force in January 2023 (to be confirmed); EC guidance on the most appropriate mechanism; initiation of an EDIC early 2023, but in line with the language data space.

## 3.3 Language Technologies for Bulgarian language

Prof. Dr. Svetla Koeva made a presentation on language technologies for the Bulgarian language based on the readiness for computer processing of the Bulgarian language in the era of artificial intelligence, the availability of language technologies for the Bulgarian language, and the language resources and technologies for Bulgarian that can improve digital services.

Language technology includes a broad scientific field that deals with the development of systems capable of processing, analysing, reproducing, and 'understanding' human languages, whether in written or spoken form. Within the framework of technological development, language technologies develop more slowly. Language and human thinking are inextricably linked. Language technologies seek to analyze both text and speech. An important area of language technology is the transformation of text into speech and the transformation into another kind of text, for example, translation into another language. More sophisticated applications of language technologies are focused on answering questions or executing voice commands, detecting misinformation and fake news, authorship detection, media monitoring (i.e., news analysis and summary), trend analysis and prediction, market research, and attitude analysis of goods and services. Of great importance is the application of language technologies in computer-assisted learning, which can help not only pupils and students but also people with visual or hearing impairments and those who are lagging in their studies. Language technologies offer great opportunities for cross-lingual communication and collaboration, providing equal access to information and knowledge for speakers of different languages (especially in a Digital Single Market).

Prof. Dr. Koeva presented several studies based on different methodologies and focused on the official languages of the EU - a study in 2012 by the Multilingual Europe Technology Alliance Network, which resulted in the creation of the so-called "META-NET White Papers" of the languages in Europe, in which the technological level of different languages is assessed in several categories (machine translation, the ability of language technologies to convert speech into text and vice versa, the availability of language resources, etc.). As of 2012, the Bulgarian language has been defined as fragmented. In 2022, a new study was carried out within the European Language Equality project, analyszing the technological support for European languages and assessing it on the basis of different language data repositories, including that of the ELRC and European Language Grid. As a recommendation of the study to achieve digital language equality for all EU languages by 2030, there should be a targeted effort towards the collection, provision, and processing of large and diverse data from the public sector, broadcasters, etc., flexible access to high-performance computing based on GPUs and more qualified experts in academic research centers and research units of companies.

The linguistic technologies we currently have at our disposal for Bulgarian include various tools for text pre-processing: tokenization, sentence splitting, spell checking, semantic analysis, etc.; text analysis; language models; machine translation systems, etc.

Also Prof. Dr. Koeva presented the new ELRC White Paper on Artificial Intelligence for a Multilingual Europe which focuses on the need to understand the importance of language data.

## 3.4 The potential of Language Technology and AI – where we are, where we should be heading

Mrs. Hristina Dobreva gave the next presentation under the title "The potential of artificial intelligence and language technologies - where are we and where should we go". She focused on the opportunities of deploying intelligent systems to achieve sustainable economic growth and prosperity in Europe.

Mrs. Dobreva stressed that AI represents one of the greatest opportunities for global societal and economic progress and drew attention to the economic benefits of data processing. Mrs. Dobreva explained that AI systems use symbolic rules or learn numerical patterns and can adapt their behavior after analyzing how their previous actions have affected the environment in which they operate.

According to Mrs. Dobreva, at the heart of the European approach to the development and use of AI is the notion that technological advances must be accompanied by a legal and ethical framework to ensure the security and rights of citizens, as well as measures to collect accessible, high-quality data, disseminate information widely and ensure equal access to the benefits of AI technologies.

Mrs. Dobreva expressed the opinion that data is an important raw material for AI and an essential precondition (along with computing infrastructure) for the development of new algorithms and applications.

Mrs. Dobreva presented the proposals for a new regulatory framework for AI from 2021, aimed at ensuring the protection of fundamental rights and safety of consumers and confidence in the development and deployment of AI.

Attention was drawn to the Coordinated Plan with an update from 2021, aimed at creating a stronger link with the European Green Deal, emerging markets, and in response to the coronavirus pandemic. Environmental and health actions are strengthened in the updated plan. It proposes a concrete set of joint actions by the European Commission and Member States to achieve EU global leadership on trusted AI. The Commission also calls for and proposes concrete actions supported by funding instruments around coordination and pooling of resources in five other areas: the public sector, robotics, mobility, home affairs, and agriculture. Participants were also briefed on the New Regulation for safer machinery, which ensures that the new generation of machinery ensures the safety of users and consumers and promotes innovation. There was also a focus on the creation of centers of excellence for artificial intelligence in Europe - for collaboration and networking.

Mrs. Dobreva then gave the main points of the Concept for the Development of Artificial Intelligence in Bulgaria until 2030, adopted by Protocol No. 72.4 of the Council of Ministers on December 16, 2020. Access to open public data and unrestricted cross-sectoral traffic of non-personal data will allow creation of high added value products and services for the benefit of citizens, businesses, the public sector, and academia.

Mrs. Dobreva gave concrete examples of AI in everyday life, such as virtual personal assistants, programs for automatic spelling correction and grammatical accuracy of texts, programs for automatic translation, automatic subtitling of movies; smart cars, smart farming, getting medical information or advice, e-security, Internet of Things, etc.

Mrs. Dobreva illustrated in a timeline the development of AI, machine learning, and deep learning and gave details on each of these units.

She also noted the existence of some problems. For example natural languages are rich, diverse, flexible, and complex. A word can have many meanings, there are different ways of saying the same thing, there are different linguistic structures, the dependence of meaning on context, and the existence of literal and figurative language (metaphor, irony).

Mrs. Dobreva gave the example of computational linguistics, which supports the performance of judgment processes and semantic analysis of speech and audio, and video data, resembling human thought processes. She noted the ability of AI to capture cause-and-effect relationships, understand patterns between concepts and phenomena, and extract knowledge about the world through language.

The presentation showed examples of deep learning in different language technologies: acoustic speech recognition; text-to-speech systems, dialogue systems/chatbots, question-answering systems, named entity recognition, relation extraction, text summarisation, and sentiment analysis.

Mrs. Dobreva provided participants with information on the Gartner Hype Cycle™ for Artificial Intelligence (AI) for 2022. According to the analytics company, many innovations are expected to enter mainstream use in two to five years, including composable AI, decision AI, and edge AI, while digital immune systems and trusted AI are among the strategic technologies for 2023.

At the end of her presentation, Mrs. Dobreva focused on the Concept for the Development of Artificial Intelligence in Bulgaria untill 2030, noting the main objective of the document, the priority sectors, the main areas of impact, the specific measures, etc.

## 3.5 The CEF AT platform

Mr. Bogomil Kovachev from the Directorate-General for Translation of the European Commission presented the EC's automatic translation platform - eTranslation. The system is based on neural networks and is adapted to different subject areas. Initially, the MT@EC platform was limited in scope to legal texts and was mostly used for internal EC needs, for translators, etc. Since 2017, the automatic translation platform is called eTranslation, it has been upgraded and is based on neural networks, adapted to different subject areas and the possibility is provided to integrate the system with other digital services. The platform supports several translation-enabled file formats (word, pdf, pptx, etc.) and translation into several languages simultaneously. The system is addressed to the following groups of users: staff of the European institutions, including translators, MS public administrations and, following the inclusion in the Connecting Europe Facility, the system is open to CEF-funded projects, NGOs, academia, SMEs, etc.

The system supports translation to and from any of the 24 official EU languages, Arabic, Icelandic, Chinese, Norwegian, Russian, Turkish, Ukrainian, and Japanese. Initially, the system worked as a website for text and/or document translation, but now it is also provided as an API for integration into websites and digital services, etc. Mr. Kovachev gave concrete examples of eTranslation integration to different systems - e.g. RE-OPEN EU, and for the quality of translation.

The eTranslation system is quite actively used in Bulgaria compared to the European average, with 7 institutional registrations for access to the web service. As of 2018 the Ministry of Transport and Communications (then the Ministry of Transport, Information Technology and Communications) is the first institution in the country to make a general registration of its domain in the eTranslation platform and all employees can use it without the need for personal registration. There are 826 active users, over 18 thousand requests for eTranslation, and over 141.000 translated pages. The use of the Bulgarian language system through the web client has increased, with nearly 1,5 million pages translated from English into Bulgarian and over 250.000pages translated from Bulgarian into English by the end of the third quarter of 2022.

As regards the quality of the translations, the system performs better on texts related to European policies or texts from the Official Journal. Difficulties for machine translation, not only for eTranslation, come when there is non-standard text related to emerging topics (e.g., Coronavirus), or literature, when the texts contain isolated words or fragments whose translation depends on the context. Mr. Kovachev also presented other language technologies offered by the EC, such as Multilingual Tweet, anonymisation and speech recognition.

Regarding eTranslation, Mrs. Dobreva clarified that the Ministry of Transport and Communications and BAS promote the platform by holding national workshops and publishing news on their websites, including whenever the system is opened to new users, which has contributed to the widespread use of the platform in Bulgaria.

## 3.6 Language technologies by/for the public sector (Panel session)

The workshop continued with a panel discussion on "Language Technologies by/for the Public Sector" moderated by Mrs. Hristina Dobreva. Up to this point the agenda of the event had focused on the opportunities offered by language technologies and machine translation, while this panel discussion focused on the demand side, specifically who uses language technologies and the eTranslation tool. Administrations are in the AI era and have different roles with respect to language technologies, including regulators, data and standards creators, users, and service providers, etc. Examples of language technologies in e-government are the processing of large amounts of data from daily work such as responses to messages, phone calls, etc., and policymaking through analysis of public consultations.
Mrs. Dobreva presented some examples from Europe of the use of language technologies in e-government, as well as examples of services using language technologies other than machine translation.

All administrations are the creators and owners of large amounts of data from different domains, which is often simply stored without realising the benefit of sharing and reusing it. The panel invited experts from two institutions that have specific experience with the eTranslation platform and the opportunities it provides through its deployment in different administrative services and portals, as well as their experience with sharing language resources in the ELRC repository.

The first panelist, Mrs. Annie Rusinova from the National Revenue Agency (NRA), shared her personal experience over the years with the use of the eTranslation and the sharing of NRA language resources in the ELRC-SHARE repository. Since 2011 she has been part of the Bulgarian Language Network team, which was established by the Bulgarian Language Department at DG Translation of the European Commission as an informal structure for dialogue and cooperation between experts from the Bulgarian government and research institutions and translators from the European Union institutions. The NRA is one of the few public institutions in the country with its own translation unit, which in turn leads to the accumulation of a huge database of language resources. Mrs. Rusinova said that the submission of language resources to the ELRC-SHARE repository was very easy, but the main difficulties were the quality of the materials that could be shared, the quality of the translation, and the confidentiality of the information. Mrs. Rusinova said that she uses the eTranslation platform daily in her work, highlighting that it is timesaving as the main benefit is that it provides the most appropriate terms when translating administrative texts. The main recommendation Mrs. Rusinova made is that after receiving a translation through the platform, it is imperative to review it for inaccuracies. Mrs. Dobreva raised a question that was also raised by participants at the Second National ELRC Workshop in 2018 as a major issue, namely, was the approval of the NRA management taken for the submission of their language resources. Mrs. Rusinova clarified that for the NRA specifically, as an institution that works mainly with personal data, it is mandatory that the decision to share data is made by the senior management of the agency, but also for generally available data such as strategies, regulations, etc.

The second panelist, Mr. Tsvetan Deyanov from the Ministry of e-Governance, shared his practical experience from the integration of eTranslation in the Single Portal for Access to Electronic Administrative Services - https://egov.bg , which is a great facilitation for citizens, users of electronic administrative services. He mentioned that they have selected eTranslation not only because it has been developed by the EC and it supports multiple formats, but also because the tool provides a standardized translation that is in line with EC rules. In the Single Portal for Access to Electronic Administrative Services, a "Text translation" option has been developed in the "My Space" section, which allows users to translate text (files) from and into the official languages of all EU Member States through the integration with eTranslation.
Mrs. Dobreva asked an interesting question for the audience, whether there is information on the number of requests for translation of text from the "My Space" section. This would help to understand to what extent the translator has helped to increase the interest for the service. Mr. Deyanov clarified that there has been an increase in traffic to this section of the portal since the integration with eTranslation.

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

The workshop agenda continued with a panel session “Language data creation, management and sharing existing practices and challenges” moderated by Mrs. Hristina Dobreva. She gave a brief introduction to the topic, specifying that language technologies (and AI in general) need data that implicitly encode what we want to do. Linguistic data is a special kind of data: language is constantly generated, but to train AI systems with linguistic data, we need digital and processable data in large quantities and for different languages, domains, and thematic areas, depending on our tasks and goals. However data collection is a complex task. The main obstacles for collecting and sharing language translation data identified at the Second National ELRC Workshop and published in the 2019 ELRC White Paper for Bulgaria, are: difficulties in identifying and convincing high-level officials to authorize data sharing;the lack of established procedures for document translation at the administrative level, leading to legal problems, e.g. on data ownership, privacy and copyright issues; technical difficulties related to data processing, etc.

The first participant in the panel was Mrs. Reny Borisova from the Ministry of e-Governance, who presented the Open Data Portal - https://data.egov.bg. The portal is a centralized national web portal that provides mechanisms for automatic and manual publishing, management, and searching of datasets provided in an open, machine-readable format, along with relevant metadata, making the data easily discoverable. The themes of the portal are aligned with those of the official European data portal (https://data.europa.eu), and the requirements of the Open Data Directive are implemented in our national legislation. Each organization should have its own registration in order to publish its open data, and since 2019 this data is subject to verification and the information is available to users without registration. Directive (EU) 2019/1024 stimulates the reuse of information, promotes access to dynamic data, increases high-value data, engages the provision of a limited number of high-value datasets in all countries, extends the scope of open data access to public enterprises, etc., and any information that is not specifically protected should be publicly available. The Ministry of e-Governance has the obligation to validate the sites for compliance with the eGovernment Act, i.e. the law that transposed Directive (EU) 2016/2102 on the accessibility of websites and mobile applications of public sector bodies. In this regard, the Ministry’s main recommendation to public sector bodies is to have at least bilingual websites. Currently, the Ministry of e-Governance is an active user of the eTransation platform, and it is integrated into the Single Portal for Access to Electronic Administrative Services.

The panel discussion continued with Mr. Stanimir Minkov from the Administration of the Council of Ministers, who presented the Portal for Public Consultations - https://strategy.bg and the Platform for Access to Public Information - https://pitay.government.bg. On the Public Consultation website, all draft legal acts are published, together with the accompanying documents such as motives, impact assessment, etc., and they are made available to the public for a certain period to share their opinion. Mr. Minkov noted that the documents published in the Publications and Strategy Documents sections of the Public Consultation Portal would be of interest to experts working in the field of machine translation and they could be fed into the ELRC-SHARE repository. The presentation of the Platform for Access to Public Information provided an example of the reuse of information. Mr. Minkov had also participated to the Second ELRC Workshop in Bulgaria in 2018, and he had presented the topic of open data. Thanks to his strong and long experience in administering the Portal for Public Consultation and now the Platform for Access to Public Information, he presented the information in an accessible and interesting way to the participants, and also gave ideas and guidance on language data sharing. Both portals also contain information in English.

The moderator asked if there are any barriers (e.g. regulatory) to sharing the documents from the portals as language resources given that they are public and even though they are monolingual, as monolingual corpora are also valuable for training language technology systems. Mr. Minkov clarified that there are no obstacles in sharing them, but not through automated exchange. He also pointed to the short lifetime of the documents as a possible problem, as draft acts are published on the portal for public consultation, not as final versions. This does not apply to strategic documents.

The discussion continued with the participation of Dr. Orlin Kouzov, Head of the Learning Lab at the Big Data for Smart Society (GATE) Institute, who shared his views on the application of various innovative technological solutions and specifically language technologies in the public sector, the role of artificial intelligence and the use of big data in detecting misinformation and disinformation. Dr. Kouzov said that after the invitation to participate in the workshop, he tested the capabilities of eTranslation platform by translating into Bulgarian and Italian one presentation, and the translation was on a very good level, even in the pictures. However, he pointed out that it was not a good practice to send the translations by email instead of sending a link to download the finished translation. Regarding the application of AI in the field of language technology in administration, technology is developing dynamically. He gave as an example that in MS Word it is now possible to translate documents, which most people do not know about, or by dictation, MS Word types the text itself. Dr. Kouzov stressed the need to take concrete measures to improve the level of proficiency of the "technological" generation. The administration uses language technology such as translation, note-taking, emailing, writing reports, presentation synthesis, etc. Of interest to the administration is the detection of fake news and fighting misinformation.

The programme continued with a specially recorded video by Dr. Temenuzhka Spasova from the Ministry of e-Governance presenting the National Spatial Data Portal - https://inspire.egov.bg and the opportunities that spatial data provide for multilingualism. The portal supports a bilingual version, but not all resources are uploaded in both languages because the initial input of information is in English. The portal does not use AI because each of the resources is pre-generated as text. The translations of the resources are done manually, there is no automated translation and continuous improvement of each of the resources is needed.

Mrs. Hristina Dobreva had prepared a presentation of the Single Information Portal - COVID-19 - https://coronavirus.bg, but due to the time limits, she briefly introduced the participants with the main highlights - the portal has a built-in chatbot and during the pandemic, the Bulgarian government has given the opportunity to send updates on the morbidity via Viber channel.

## 3.8 Demo session of LT and AI technologies available for Bulgaria

Prof. Dr. Svetla Koeva from BAS presented the project Curated multilingual resources for CEF.AT - CURLICAT (https://ibl.bas.bg/en/kolektsiya-ot-mnogoezikovi-resursi-za-cef-at-curlicat/), funded by the Connecting Europe Facility. The project's main objective was to collect selected monolingual data in seven languages (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian) in predefined thematic areas, which are relevant for the European Digital Service Infrastructures (DSIs), also called building blocks, e.g. eDelivery, eID, eSignature, and eTranslation. The initial source of data is the national corpora for Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, and Slovenian, which are freely accessible so that they can be distributed via the ELRC-SHARE repository. The provision of seven monolingual, large-size data collections aim to support the improvement of the eTranslation. This will allow users from different European countries to access information in a language they understand well. The Bulgarian National Corpus was created at the Bulgarian Language Institute and consists of Bulgarian texts and 47 parallel corpora with different sizes and contains 1.2 billion words. The material in the corpus reflects the state of the Bulgarian language (mostly in its written form) from the mid-twentieth century (1945) to the present day and provides the possibility to retrieve specialized or general sub-corpora according to certain criteria (subject, author, year/period of publication, source, etc.). One of the activities involves corpus expansion and copyright clarification, and another is data anonymization through the development of anonymization solutions. The MAPA (Multilingual Annotation for Public Administration) programme for anonymization of names and other sensitive data (https://mapa-project.eu) and supplementing the MAPA results by identifying additional names in Bulgarian and replacing them with random names and by identifying dates, phone numbers, and other identifiers with specific spellings in Bulgarian. The results of the project are: seven monolingual corpora (for Bulgarian, Polish, Romanian, Slovak, Slovenian, Hungarian, and Croatian) of more than 2 million sentences and more than 20 million words, containing texts with free distribution and modification rights (total more than 14 million sentences and more than 140 million words) from selected subject areas; the texts are provided with rich and harmonized metadata between the languages; the texts are enriched with detailed linguistic information: morphological (part of speech and grammatical features), syntactic (syntactic dependencies, name groups), terminological (terms from Interactive Terminology for Europe and from selected subject areas), lexical (names for people, organizations and locations).

Dr. Yordan Kralev from the Technical University of Sofia presented applications and services using the Multilingual Image Corpus (MIC 21) for multimodal content processing. The Multilingual Image Corpus consists of an ontology of visual objects (based on WordNet) and a collection of thematically related images whose objects are annotated with segmentation masks and labels describing the ontology classes. The data collection has applications for image classification and objects detection in images as well as for tasks involving semantic segmentation. The main contributions of the project team are the provision of a large collection of high-quality copyright free images; the formulation of a Ontology of visual objects based on WordNet noun hierarchies; the precise manual correction of automatic objects segmentation within the images and the manual annotation of object classes; and the association of objects and images with extended multilingual descriptions based on WordNet inner- and interlingual relations. The goal of the project is to create models that perform semantic processing on images, i.e., extract information from an image with respect to what objects are contained in them and what their contours are, and at the same time translate these annotated objects into multiple languages. To create such models with machine learning methods, it is necessary to have a database, the so-called multilingual corpus, which must be developed so that the model can be tuned or trained by an algorithm. The good quality of model training depends on the good quality of the database (the multilingual corpus should be more developed). The project has developed an ontology with more than 700 classes, translated into 25 languages; the images are grouped into 130 thematic areas with a high level of similarity between the areas, models have been developed for automatic processing to each of the thematic areas and software tools for visualisation and development. The developed

models allow further expansion of the multilingual corpus of images. Dr. Kralev gave a demonstration of how the Multilingual Image Corpus system works for multimodal content processing, which is publicly accessible at https://mic21.dcl.bas.bg. The models are publicly accessible for download at: https://dcl.bas.bg/MIC-21/models. Dr. Kralev also gave a demonstration of another tool used for the manual annotation of images and/or for forming the initial annotation sample from images - COCO Annotator.

## 3.9 Conclusions

At the concluding sessions, Mrs. Dobreva gave an overview of the topics covered during the event in terms of the potential of language technologies to transform governance, administration, and society, the technological readiness of our language in the age of artificial intelligence, actions and developments at EU level for wider development and adoption of language technologies at the national level.

Mrs. Dobreva argued that lesser spoken languages are facing a digital time bomb. She repeated once again that research and industry in language technology can facilitate digital interaction in a multilingual Europe. One of the main messages of the workshop was that language technology and artificial intelligence need more data.

Even though this is the end of the ELRC project, Mrs. Dobreva urged the workshop participants to share their data. She explained that the project continues offering help on technical or legal aspects related to using, creating, collecting, processing, and sharing language resources.

Mrs. Dobreva thanked everyone for their participation and noted that the steady number of participants until the end of the event shows their interest in the topics presented. Participants were encouraged to fill in the event feedback forms.

# 4 Synthesis of Workshop Discussions

Even though the workshop was held online, the number of registered participants and attendees indicate significant interest in the workshop topics. Based on the presentations given and the discussions held at the workshop, the main message is that data is very important, and it is a driver of the economy. As a result of the ELRC workshops, there are seven institutions in the country that have followed the lead of the Ministry of Transport and Communications and have made a common registration of their domains, allowingtheir staff to use the eTranslation platform without having to make a separate registration. Thanks also to the dissemination of information and promotion through the MTC website about the opening of the platform to different users, and the introduction of new functionalities and applications, the country ranks among the top in terms of the number of its users.

During the event, the potential of AI for global social and economic progress and its key role in the development of language technologies was highlighted. Interesting examples of their use in the Bulgarian public administration were presented. During the forum, participants were introduced to many government portals that have a huge number of resources, and the use of different communication channels and technologies such as chatbots for official communication with the public was demonstrated.

Workshop participants stressed the need for authorisation at the leadership or policy level for sharing resources owned by the respective organisation. Mrs. Dobreva informed once again that she and the Technical National Anchor Point, Prof. Dr. Svetla Koeva, had a high-level meeting in the office of one of the Deputy Prime Ministers in a previous political cabinet to raise the importance of data sharing by administrations and the need for a decision at the political level in this regard. Unfortunately, no concrete outcome has come out of the meeting. As an effort of the two NAPs, several series of official letters have been sent throughout the lifetime of the project, by the respective minister to all ministries and their subordinate agencies with information about the project and the possibilities of sharing language resources by the different administrations.

As a result of the Second ELRC Workshop in Bulgaria held in 2018, the functionalities of the eTranslation platform have been implemented in the country's Single Portal for Access to Electronic Administrative Services. This was demonstrated as a functionality to the participants in this workshop. The "Text Translation" section of the portal allows users to translate text (file) from and into the official languages of all EU Member States, with representatives of the Ministry of e-Governance reporting increased interest in the service.

The participation in the panel discussions of an expert from the administration of the National Revenue Agency with a presentation of her personal experience in the provision of resources in the ELRC-SHARE repository was welcomed with interest by the participants. The expert also shared her experience with the eTranslation platform and the significant improvement in translation quality it has demonstrated.

The Participants also had the opportunity to receive information on the development of the Open Data Portal, data policy, and national developments.

There were representatives from other countries who sent e-mail messages to Mrs. Dobreva to express that the event was very interesting and useful for them. In this regard, it is considered good practice to live stream the workshops. It is also helpful to have a simultaneous interpretation from/into the local language and English.

# 5 Country Profile: Language data creation, management and sharing

An updated profile for Bulgaria 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)