



## **Deliverable D3.2.14**

### **Task 3**

# **ELRC Workshop Report for Romania**



<b>Author(s):</b>	Laura Mihăilescu, Dan Tufiş
<b>Dissemination Level:</b>	Public
<b>Version No.:</b>	V1.0
<b>Date:</b>	2021-11-10



## Contents

<b>1</b>	<b><a href="#">Executive Summary</a></b>	<b>3</b>
<b>2</b>	<b><a href="#">Workshop Agenda</a></b>	<b>4</b>
<b>3</b>	<b><a href="#">Summary of Content of Sessions</a></b>	<b>5</b>
<b>3.1</b>	<b>Welcome and introduction</b>	<b>5</b>
<b>3.2</b>	<b>The potential of Language Technology and AI – where we are, where we should be heading</b>	<b>5</b>
<b>3.3</b>	<b>Language Technologies for Romanian (Panel session)</b>	<b>5</b>
<b>3.4</b>	<b>The CEF AT platform</b>	<b>6</b>
<b>3.5</b>	<b>Language Technologies for the public sector; requirements and offerings (Panel session)</b>	<b>6</b>
<b>3.6</b>	<b>Language data creation, management and sharing; the value of data for the development of top-quality LT: existing practices and challenges (Panel session)</b>	<b>8</b>
<b>3.7</b>	<b>Demo session</b>	<b>8</b>
<b>3.8</b>	<b>Take-home message and conclusions</b>	<b>9</b>
<b>4</b>	<b><a href="#">Synthesis of Workshop Discussions</a></b>	<b>10</b>
<b>5</b>	<b><a href="#">Country Profile: Language data creation, management and sharing</a></b>	<b>12</b>

## 1 Executive Summary

The third ELRC workshop in Romania took place as an online event via Zoom on 29th September 2021 at 9.30-13.00 local time. The COVID-19 pandemic situation, which was pretty bad in Romania, did not allow a face-to-face workshop as the previous two events.

The event was organized by the Research Institute for Artificial Intelligence “Mihai Drăgănescu” (ICIA) of the Romanian Academy and European Institute of Romania (IER). ICIA is a member of the ELRC consortium, and its representative (Dan Tufiş) is the Technical National Anchor Point (TNAP). IER is a member of the ELRC consortium, and its representative (Laura Mihăilescu) is the Public National Anchor Point (P-NAP). The local organization committee consisted of Dan Tufiş, Elena Irimia, Marius Şucu, Maria Carp, and Verginica Mititelu. Support and advice with the practical arrangements were additionally provided by the ELRC office (Eileen Schnur and Stefania Racioppa) and ELRC coordinators (Khalid Choukri, Andrea Lösch and Hélène Mazo) and gratefully acknowledged.

All invited speakers spoke in Romanian, but there were interpretation services offered, so that the participants could follow the event either in English, or French. We particularly thank IER for contracting the interpretation services.

96 participants registered on the ELRC website and 72 attended the online workshop. The participants were mostly from public institutions, public administrations, and ministries of Romania, as well as universities, research institutes, industry partners/SMEs, national radio, and freelance translators.

The following section includes the agenda of the event (Section 2). The section 3 briefly informs about the content of each presentation of the workshop (Subsections 3.1-3.7). In Section 4, a summary of the discussion raised during the workshop is provided. In the last section (Section 5), we focus particularly on Romania and the current situation on language data creation, management, and sharing.

All presentations by the speakers are available at <https://lr-coordination.eu/romania3rd>.

## 2 Workshop Agenda

---

<b>09:30 – 09:35</b>	<b>Welcome address</b> Laura Mihăilescu (P-NAP, IER)
<b>09:35 – 09:50</b>	<b>Introduction</b> Dan Tufiş (T-NAP, ICIA)
<b>09:50– 10:15</b>	<b>The potential of Language Technology and AI – where we are, where we should be heading</b> Stefan Trăuşan-Matu, (UPB)
<b>10:15 – 11:00</b>	<b>Language Technology in Romania – Panel session</b> Moderator: Dan Tufiş (ICIA) Panelists: <ul style="list-style-type: none"> <li>• George Bara (RWS/SDL)</li> <li>• George Suci (Beia)</li> <li>• Monica Florea (Siveco)</li> <li>• Radu Ion (ICIA)</li> </ul>
<b>11:00 – 11:30</b>	<b>The CEF AT Platform Presentation/Demo</b> Cristian Braşoveanu (DGT)
<i>11:30– 11:45</i>	<i>Coffee Break</i>
<b>11:45 – 12:30</b>	<b>Language Technologies for the public sector; requirements and offerings - Panel and discussion</b> Moderator Laura Mihăilescu (IER) Panelists: <ul style="list-style-type: none"> <li>• Costel Ciuchi (Secretariat-General of the Government)</li> <li>• Eliza Vaş (IER)</li> <li>• Corina Bîca (UEFISCDI)</li> <li>• Liviu Badea (ICI)</li> </ul>
<b>12:30 – 13:15</b>	<b>Creating, managing, and sharing language data; the value of data for developing leading language technologies: Existing practices and challenges - Panel and discussion</b> Moderator Corneliu Burileanu (UPB) Panelists: <ul style="list-style-type: none"> <li>• Liviu Dinu (Univ. Bucureşti)</li> <li>• Dan Cristea (ITI-Iasi)</li> <li>• Mircea Giurgiu (UTCN)</li> <li>• Verginica Mititelu (ICIA)</li> </ul>
<b>13:15 – 13:20</b>	Conclusions
<b>13:20 – 14:00</b>	<b>Demo Session: Public Romanian LT services</b> Vasile Păiş (ICIA)

---

## 3 Summary of Content of Sessions

### 3.1 Welcome and introduction

In her welcome address, **Mrs. Laura Mihăilescu** (P-NAP in Romania), recollected the previous two ELRC workshops and invited the participants to take an active part in the workshop. She introduced the housekeeping rules, e.g., to switch the camera and microphone off when the speakers are talking, and to ask questions either directly after each presentation or in the chat. After the welcome, she explained how participants can select their preferred language (English, or French) to listen to the talks. There were no technical difficulties, and participants got settled very quickly.

**Mr. Dan Tufiş** (T-NAP in Romania) was the main moderator for the whole event. He made a brief introduction to the topics to be discussed during the workshop, invoking the process of reshaping multilingual Europe with its focus on Artificial Intelligence centered on language. He then described the role of CEF and its program ELRC, mentioned the new program DIGITAL EUROPE and its budget for the period 2021-2027. Finally, Dan Tufiş introduced the agenda and reminded the participants that their opinions matter and were expected. The participants were encouraged to ask for a personalized participation certificate. Dan Tufiş concluded the introduction by suggesting all the participants to have a look at the Country profile and make whatever suggestions they might have and gave the virtual floor to Mr. Ştefan Trăuşan-Matu, the speaker of the following session.

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

The session was held by **Mr. Ştefan Trăuşan-Matu**, Professor at University Politehnica of Bucharest. The presentation offered an overview of the potential of LT and AI in both international and national levels. Both important examples of the use of AI in everyday life and the impact that these systems already have, have been presented. Another key point of the talk was the presentation of the degree of digitization both at national and European levels. The importance of language data for developing LTs and AI systems was emphasized by the presentation of the resources created for this purpose for the Romanian language.

### 3.3. Language Technologies for Romanian (Panel session)

Participants in this panel session were **Mr. George Bara**, VP, Strategic Partnerships at RWS, **Mr. George Suci**, Director CDI at BEIA Consult International, **Ms. Monica Florea**, Head of Unit European Projects at Software Imagination & Vision (SIMAVI) and **Mr. Radu Ion**, Senior researcher II at Research Institute for AI "Mihai Drăgănescu", Romanian Academy. **Mr. Dan Tufiş**, Director of Research Institute for AI "Mihai Drăgănescu", Romanian Academy, moderated the session.

The panel discussion considered the state of the art in Romanian Language Technologies (Romanian LTs) in both the academia and the industry. It was concluded that, in academia, Romanian LT are very well represented, both in terms of publications and computational resources (e.g., corpora, lexicons, treebanks, wordnet, etc.) and algorithms available online. These resources and algorithms are usually free for research purposes but their ownership and usage in industry and public administration is unclear. Furthermore, there is a shortage of people with the required skills who could bring these

## ELRC Workshop Report for Romania

algorithms to the public, in a form that is useful for the general public and public administration (it was estimated that the transformation of a research LT model into a production model requires 5,000 hours of work). While there are master's degrees and PhD programs in Natural Language Processing in all main universities in Romania, more machine learning and AI courses are needed at the undergraduate level.

It was observed that Romanian legislation is prohibitive in making AI reach public administration. That is, for various Romanian LT research to make it into the public administration, specific laws are required which do not exist yet. The ongoing Romanian national project aiming at a „National Strategy in AI” (<https://strategie-ia.utcluj.ro/>) tries to answer the present questions and provide solutions for the adoption of AI (Romanian LTs included) in Romania.

### 3.4. The CEF AT platform

This session was presented by **Mr. Cristian Braşoveanu**, IT Product Officer – eTranslation at the European Commission (DGT). The presentation started with an historical perspective on the MT-focused initiatives and projects at the European Commission, MT@EC, eTranslation and, finally, CEF.AT. The distinction between the two target audiences of the CEF system was stated: CEF eTranslation is developed to provide Pan-European digital public services to public administrations in Member States, Iceland, and Norway, while DGT eTranslation is tuned for translators and staff of the EU Institutions and for digital services of the EU Institutions. It was underlined that eTranslation has been opened to European SMEs since March 2020, with specific conditions and limitations. It was explained that eTranslation is both a Web interface that provides a draft of automatically translated text which can be reviewed by human translators and a Machine-to-Machine API to be integrated in workflows, websites, digital services, etc. The domains' coverage and the translation quality of the system was discussed and Mr. Braşoveanu explained that texts related to EU policies are very well translated because the system was trained on translated EU documents. However, a lower level of quality is to be expected for non-standard, new, or creative text. He also exemplified how words or expressions can be wrongly translated when their proper translation depends on a missing context. Mr. Braşoveanu went on to discuss future aims for the project such as extending domain and language coverage and incorporating additional language technologies. The major facilities of the translation platform were demonstrated, and as a conclusion, he provided practical information about registering and using the system, underlining who the eligible individual users are.

### 3.5. Language Technologies for the public sector; requirements and offerings (Panel session)

The participants in this section were **Mr. Costel Ciuchi**, Counsellor Expert at the Secretariat-General of the Government, **Ms. Eliza Vaş**, Expert within the European Studies Unit of the European Institute of Romania, **Ms. Corina Bîca**, Expert of the Executive Agency for Higher Education, Research, Development, and Innovation Funding (UEFISCDI) and **Mr. Liviu Badea**, Senior Researcher I Head of AI Laboratory at the National Institute for Research and Development in Informatics (ICI).

The session was moderated by **Ms. Laura Mihăilescu**, Head of Translation Coordination Unit of the European Institute of Romania.

## ELRC Workshop Report for Romania

**Ms Vaş** presented a policy brief and a study available only in Romanian, on the site of the European Institute of Romania: [From digitalisation to digital transformation in Romania](#) (policy brief, July 2021) and [Romania and the Digital Single Market of the European Union. Opportunities and Challenges](#) (a study published in 2017).

In the study on the Digital Single Market, the research team argued that Romania should pay particular attention to expanding creative industries and improving access to digital technologies for small and medium-sized enterprises so as to become an exporter of digital goods and services to the EU single market. It concluded that the necessary conditions are met in Romania to create and export several types of services that are based on language technology, e.g., voice recognition, automatic speech transcription, automatic speech translation, virtual assistants based on artificial intelligence, etc.

The team of authors of the policy brief provided an overview of the digital transformation in Romania and came up with several conclusions for policy makers:

- The COVID-19 pandemic is expected to become the catalyst for digital transformation, especially in the business environment.
- There is a territorial fragmentation of digital public services in Romania, which separates the performing authorities and administrative units from those that do not enjoy resources and have no access to technology.
- There is a need to understand the importance of data to invest in data infrastructure and increase the capacity to analyse and report data.
- We need to standardise digital skills development programmes — based on the European Framework of Digital Competences for Citizens

**Ms Bica** presented the four areas of action in the National Strategy on the 2020 Digital Agenda:

- a) e-Government, Interoperability, Cyber Security, Cloud Computing, and social media,
- b) ICT in education, health, culture and eInclusion,
- c) eCommerce, R & D, and innovation in ICT,
- d) Broadband and digital infrastructure services.

Unfortunately, none of these priorities includes or explicitly mentions AI and NLP.

She further presented Romania's digital status in the EU and gave some examples of LT integrated in several applications used by UEFISCDI (the major Romanian funding agency for research, development, and innovation projects). She emphasized that the missing infrastructure, the lack of interoperability of databases at national level for public services, the resistance to change and the unclear legislation are among the main obstacles in using LT in public services.

**Mr Ciuchi** talked about the standardisation of the sites of the Romanian public institutions in terms of domain (gov.ro), way of presenting information and even common design elements. He also mentioned the efforts made to further digitalise the public sector. The possible roles of the government in the development of AI and LT, several LT used in e-government and a few examples of MT and LT-based services were quoted: Romania2019.eu, e-justice, eTranslation, etc.

### 3.6. Language data creation, management and sharing; the value of data for the development of top-quality LT: existing practices and challenges (Panel session)

The panelists of this session were **Mr. Liviu Dinu**, Professor at University of Bucharest, **Mr. Dan Cristea** Principal Researcher at ITI-Iasi, Romanian Academy, **Mr. Mircea Giurgiu**, Professor at Technical University of Cluj-Napoca (UTCN) and **Ms. Verginica Mititelu**, Senior researcher II at Research Institute for AI "Mihai Drăgănescu". **Mr. Corneliu Burileanu**, Professor and Vice-rector at University POLITEHNICA of Bucharest moderated the session.

The discussions in this panel tackled both language data from a general point of view and the specific issue of data for Romanian. The importance of language data for creating services and products was acknowledged by all participants. Despite their ubiquity, access to them is not easy. Beyond data access, collection, storage, processing, cleaning, publishing, sharing, (re)usage are also needed. Given the deep learning techniques used nowadays for various NLP tasks, access to large amounts of data is of utmost importance; general language data as well as domain specific data are needed for the development of applications and for further improving their performance.

Romania lacks the legal frame for language resources creation and sharing. Researchers, with no background in law, cope with the challenges of creating such resources mainly on their own, but never certain that they are within the bounds of legality. Adding to this that development of language resources is not considered an achievement for a researcher by the national body for researchers' evaluation (except for the Commission on informatics), it could be seen as a wonder that more and more researchers have been investing their efforts in such endeavors. However, anchored in today's international research methods, Romanian researchers are one big step ahead the authorities and legislative body and are contributing to the effort of ensuring their language lives in the digital era.

All resources that are accessible are by no means the result of a national initiative in this direction, but the result of individual institutional initiatives or of European infrastructures: thus, ICIA has provided or has helped other creators to provide resources and information about them within META-SHARE, European Language Grid, ELRC-SHARE, LLOD Cloud. Stakeholders who have understood that the value of a resource is increased by its reuse by others have also made resources available on their institutional websites (see <https://www.racai.ro/tools/>, [nlptools.info.uaic.ro](http://nlptools.info.uaic.ro)).

A national infrastructure containing all Romanian language resources, accessible to both private and public bodies would definitely boost the LT market, with benefits for all Romanian consumers of products and services.

### 3.7. Demo session

The demo session "Public Romanian LT services" was presented by **Mr. Vasile Păiș**, Senior Researcher III at the Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy (RACAI). The presentation focused on the RELATE platform (<https://relate.racai.ro>). This platform hosts several LT services and LRs specific to Romanian language. The LT services were developed by a team of researchers at RACAI and at partner institutions throughout different research projects. During the demo, a brief overview of the platform architecture and available services was



## ELRC Workshop Report for Romania

given. Following this, the different technologies were showcased with live demonstrations. This included:

- basic language toolkits: TEPROLIN and UDPipe, providing lemmatization, part-of-speech tagging, dependency parsing
- named entity recognition: general entities, biomedical entities, legal domain entities
- text anonymization
- Romanian WordNet querying directly and aligned with the English WordNet
- EuroVoc classification
- textual machine translation
- speech processing: automatic speech recognition (ASR), text-to-speech (TTS), speech-to-speech machine translation
- RELATE platform functionality for large corpora management and manual annotations mechanisms useful for creating gold corpora with both textual and bimodal (text+audio) components

In addition to the presentation of services, real use cases involving the RELATE platform, from real research projects were discussed. Furthermore, LRs produced using the platform were presented, such as the corpus LegalNERo providing named entity recognition in the legal domain and the RTASC bimodal (text+audio) corpus providing aligned text and speech sentences.

### 3.8. Take-home message and conclusions

The workshop included a short session on conclusions and various administrative announcements. It was presented by Dan Tufiş who summarized the discussion as follows:

- 1) Data is essential to the success of AI/LT and thus they are extremely valuable.
- 2) The LT have a huge potential to change the governance, administration, commerce, and society.
- 3) The technological level of Romanian language in the AI era is improving, but efforts are still needed to improve it. The Romanian language can be hardly called now an under-resourced language.
- 4) The position of our authorities with respect to the development and adoption of the language technologies is uncertain and hesitant.
- 5) There were reviewed costs/dangers of inaction.
- 6) There were invoked several actions at the level of the EU for a larger, faster development and adoption of the linguistic technologies and AI-language centric.

In his concluding remarks, Dan Tufiş reminded the participants some useful links for contacting ELRC helpdesk service: <http://www.lr.coordination.eu/helpdesk>, [help@lr-coordination.eu](mailto:help@lr-coordination.eu), skype ELRC Helpdesk, and telephone number +33 970 440 522

At the end of this session, Khalid Choukri, member of the ELRC board responsible for Romania, asserted the quality of the workshop and congratulated the local organisers for the successful scientific event.

## 4 Synthesis of Workshop Discussions

The panels offered quite consistent views on the topics raised by the moderators.

The panel discussion on the **Romanian Language Technologies** (Romanian LTs) in both the academia and the industry concluded that, in academia, Romanian LT are very well represented, both in terms of publications, computational resources (e.g., corpora, lexicons, treebanks, wordnet, etc.) and algorithms available online. These resources and algorithms are usually free for research purposes but their ownership and usage in industry and public administration is unclear. Furthermore, there is a shortage of people with the required skills who could bring these algorithms to the public, in a form that is useful for the general public and the public administrations (it was estimated that the transformation of a research LT model into a production model requires 5000 hours of work). While there are master's degree and PhD programmes in Natural Language Processing in all important universities in Romania, more machine learning and AI courses are needed at the undergraduate level. Although industry is present in LT, its commercial offers on the Romanian market for public institutions and public in general is limited, and more often than not prohibitive due to lack of funding.

It was observed that Romanian legislation is prohibitive in making AI reach public administration. That is, for various Romanian LT research to make it into the public administration, specific laws are required which do not exist yet. The ongoing Romanian national project aiming at a „National Strategy in AI” (<https://strategie-ia.utcluj.ro/>) tries to answer the present questions and provide solutions for the adoption of AI (Romanian LTs included) in Romania.

The participants of the panel on **Language Technologies for the public sector; requirements and offerings** spoke about the possible roles of the government in the development of AI and LT, several LT used in e-government and a few examples of MT and LT-based services were mentioned: Romania2019.eu, e-justice, eTranslation, etc. The expert from the General Secretariat of the Government raised the issue of the necessary standardisation of the sites of the Romanian public institutions in terms of domain (gov.ro), ways of presenting information and even common design elements and about efforts made to further digitalise the public sector.

The findings of two reports published by IER ([From digitalisation to digital transformation in Romania](#) and [Romania and the Digital Single Market of the European Union. Opportunities and Challenges](#)) were briefly presented and also the conclusions of the reports' authors for policy makers were reminded:

- The COVID-19 pandemic is expected to become the catalyst for digital transformation, especially in the business environment.
- There is a territorial fragmentation of digital public services in Romania, which separates the performing authorities and administrative units from those that do not enjoy resources and have no access to technology.
- There is a need to understand the importance of data, to invest in data infrastructure and increase the capacity to analyse and report data.
- We need to standardise digital skills development programmes — based on the European Framework of Digital Competences for Citizens.

## ELRC Workshop Report for Romania

- Romania should pay particular attention to expanding creative industries and improving access to digital technologies for small and medium-sized enterprises so as to become an exporter of digital goods and services to the EU single market. It concluded that Romania has the necessary conditions to create and export several types of services that are based on language technology: voice recognition, automatic speech transcription, automatic speech translation, virtual assistants based on artificial intelligence etc.

Unfortunately, the four areas of action in the National Strategy on the 2020 Digital Agenda which was presented (a) e-Government, Interoperability, Cyber Security, Cloud Computing and social media, b) ICT in education, health, culture and eInclusion, eCommerce, c) R & D and innovation in ICT and d) Broadband and digital infrastructure services) does not include explicitly AI and NLP. However, there were some examples of LT integrated in several applications used by UEFISCDI (the principal funding agency for research, development, and innovation). Among the main obstacles in using LT in public services are the missing infrastructure, the lack of interoperability of databases at national level for public services, the resistance to change and the unclear legislation.

The discussions in the panel on **Language data creation, management and sharing; the value of data for the development of top-quality LT: existing practices and challenges** was acknowledged by all participants, and they outlined the importance of language data for creating services and products. Despite their ubiquity, access to them is not easy. After accessing, data needs collection, storage, processing, cleaning, publishing, sharing, (re)usage. Again, it was emphasized that Romania lacks the legal frame for language resources creation and sharing. Researchers, with no background in law, cope with the challenges of creating such resources mainly on their own, but never certain that they are within the limits of legality. Adding to this the fact that development of language resources is not considered an achievement for a researcher by the national bodies for researchers' evaluation, it could be seen as a wonder that more and more researchers have been investing their efforts in such endeavors. However, anchored in the nowadays international research methods, Romanian researchers are one big step ahead the authorities and legislative body and are contributing to the effort of ensuring their language lives in the digital era. All resources that are accessible are by no means the result of a national initiative in this direction, but the result of individual institutional initiatives or of European infrastructures: thus, ICIA has provided or has helped other creators to provide resources and information about them within META-SHARE, European Language Grid, ELRC-SHARE, LLOD Cloud. Actors who have understood that the value of a resource is increased by its reuse by others have also made resources available on their institutional websites (see <https://www.racai.ro/tools/>, [nlptools.info.uaic.ro](http://nlptools.info.uaic.ro)). A national infrastructure containing all Romanian language resources, accessible to both private and public bodies would definitely boost the LT market, with benefits for all Romanian consumers of products and services.

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

Laura Mihăilescu, Dan Tufiş

### State of Play:

#### Translation practices in ministries and public administrations in Romania:

In Romania, there are different scenarios how translation needs are met in the public administration. In general, translations are regarded as a secondary activity and most translations are outsourced when a need arises. The Ministry of Culture, the Superior Council of Magistracy and the Ministry of Justice for example outsource most of the needed translation to language service providers.

When other public administrations need translations, they are either done in-house by whoever knows a foreign language and without CAT tools, or they are outsourced. Only very few institutions in Romania have dedicated translation departments or employees whose main task is to translate, these institutions are the European Institute of Romania – EIR – (4 translators, 3 legal revisers, 1 terminologist), the National Bank of Romania (7 translators), the Constitutional Court and the Romanian Standards Association (ASRO). The most common language combination for in-house translations is Romanian <> English and sometimes Romanian <> French (French is mainly used for the legal field). Although the EIR operates under the coordination of the Ministry of Foreign Affairs, translations in the ministry itself are mainly outsourced because of the volume and the variety of languages needed. At the ministerial level, there is no administration with their own in-house translation service.

Most translations are done from English and French into Romanian, and from Romanian into English and Hungarian.

Some professional translators are still worried about their job and are skeptical about the quality of machine translations. Their opinions are propagated upstream to the decision makers and thus low interest for this approach.

#### Public procurement and the use of CAT tools

Public institutions that outsource translations were obliged by law to search the electronic public procurement system first (SICAP: <http://sicap-prod.e-licitatie.ro/pub>). If public institutions could prove that they could not find a suitable offer or if the offers were more expensive on SICAP than on the open market, they were free to choose any LSP on the market. However, this is now no longer an obligation but a recommendation.

A regulation that is still in effect is a fixed price for the translation of one page in the legal field. According to Order no 772/2009, the Ministry of Justice and other institutions operating in the legal field pay a certified translator with 33,56 lei per page (that is approx. 6,71 euro) – at this price, the translation is usually accepted by inexperienced language service providers and the result usually lacks in quality. Even when the translator can provide a certificate for translating legal texts, this is not necessarily a seal for quality because obtaining certificates for translating legal texts is a formality and does not require specific legal training. The certificates can be obtained upon request to the Ministry of Justice by any person who graduated from any language university regardless of whether the

### **ELRC Workshop Report for Romania**

respective student has followed any course for translation of legal texts. In addition, by taking an exam organised by the Ministry of Culture, a person without a university diploma in foreign languages may obtain such a certificate. This causes a serious issue when it comes to the translation quality of legal texts.

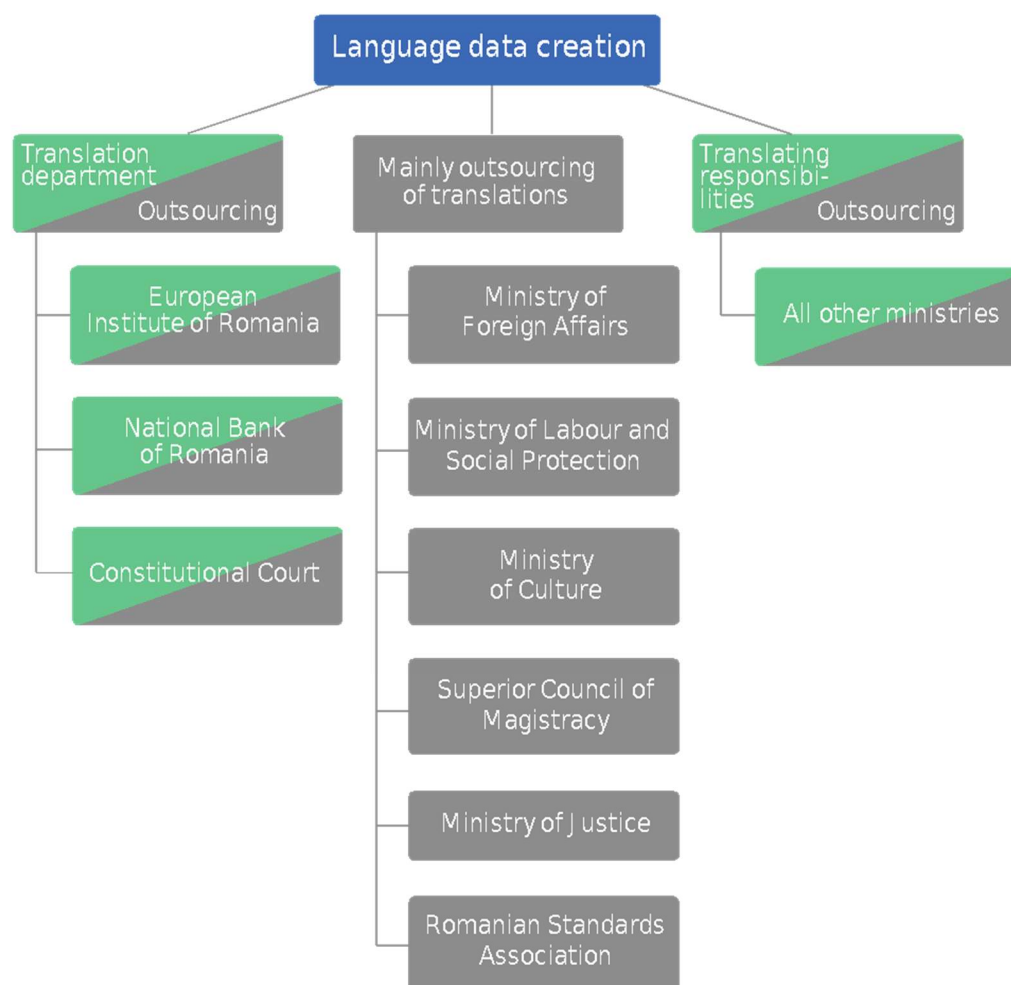
Interesting fact:

It is not yet standard practice for Language Service Providers to use CAT tools. Usually, only LSPs working for foreign clients invest in CAT tools to meet clients' requests.

Another challenge arises from the fact that the majority of those who request translations (public institutions or not) usually do not request additional services like revision/review, terminology lists, glossaries, or translation memories. The awarding of contracts is based on the number of pages, pair of languages, sometimes the domain, the deadline and certificates of the translators. CAT tools are considered to be very expensive and sometimes too complex. Therefore, only very few LSPs can afford them. LSPs that work with European institutions or foreign clients, however, usually do use CAT tools. CAT tools are also used by the translators in the in-house translation services mentioned above, whereas only the European Institute of Romania uses a server-based translation memory system.

**ELRC Workshop Report for Romania**

**The current language data sharing infrastructure in Romania public bodies looks as follows:**



In Romania, there is no proactive exchange of translation memories, terminology, or expertise on the national or inter-ministerial level. During the translation of the acquis, a terminology network was set up and although it was considered very useful, it proved to be difficult to manage for a number of reasons. The terminology experts in the various ministries were not paid for this activity and therefore could only allocate limited time to this task. In addition, a high fluctuation of human resources in the ministries and their constant reorganisation led to information loss. Another initiative that was started by the Romanian Language Department in the DGT was the RO+Network, a Linguistic Network of Excellence for Institutional Romanian that allowed for information exchange between the DGT and experts of the Romanian language regarding linguistic or terminological questions. The experts provided advice pro bono, but the network is no longer active.

Currently, there are no concrete plans for the organisation of a terminology network on the national level.

## ELRC Workshop Report for Romania

### Open Data in Romania

Most people in public institutions do not consider or do not know that language resources can also be Open Data. Hence there are no bi- or multilingual data sets in TMX format provided by public institutions on the national Open Data Portal (<http://data.gov.ro/>). Language data is also not sought after as only numerical data is considered useful, especially for decision makers.

### Language policy in Romania

Romania's official language is Romanian, which belongs to the group of Romance languages. It is spoken by about 25 million people in Romania and abroad. The most spoken minority languages are Hungarian, German, and Romany although 20 languages have minority status in Romania. Education is also provided in the languages of the minorities and learning foreign languages is included in the compulsory school curriculum. According to Law no 500/2004, the Romanian language is to be used in all official documents. The law does not address the use or planned use of language technologies to protect and support the Romanian language in the digital age. Some other provisions state that any technical manual or instructions regarding the use of a foreign product must also be translated into Romanian and that all TV productions in a foreign language must be subtitled into Romanian.

### Stakeholders

Although the use of machine translation and computer aided translation tools are not a common practice in Romanian public administration and in the public sector yet, an increased interest in past ELRC events was shown. The second ELRC workshop was attended by almost 100 participants representing various ministries and language service providers. Among the data donors of multilingual language data to ELRC are the Romanian Parliament, the European Institute of Romania, and RACAI. The latter two institutions are represented by the ELRC National Anchor Points and are critical institutions for language data collection in Romania. With the big push from media and international organizations on AI and LT there appeared some signs of interest from the Romanian authorities and private companies (see the ongoing Romanian national project aiming at a „National Strategy in AI” (<https://strategie-ia.utcluj.ro/>)).

### Main challenges for sustainable data sharing

- Legislation regarding the use of the Romanian language is not observed by all public institutions (according to Law no 500/2004, public institutions have to use the Romanian language in all official documents, that is with special characters (ă, â, î, ș, ț) and observing the directives of the Romanian Academy (“sunt/sînt”, writing with “â/î”).
- There is a general tendency to disregard the quality of the Romanian language for a number of reasons (e.g., the message is considered more important, lack of time, low speed when typing with Romanian special characters), which also affects the quality of the Romanian language on the internet. This low quality also affects the collection of language resources in various ways. Ignoring the official orthography, i.e., not using diacritics, or using non-standard ones (or combining these practices) can result in a different meaning which turns textual documents into low-quality data that is not accurately representing the Romanian language and is thus avoided by data collectors.
- **Educational issue:** Poor use of CAT tools (they are considered to be very expensive and more of a luxury) and of computers (still not a norm to use Spell check, Track Changes, advanced text

## ELRC Workshop Report for Romania

formatting); Proper education on data management is a major challenge.

- **Quality issue:** It is still not a standard for public institutions to ask LSPs to provide revision/review for their translations and to return the translation memories (the decision criterion for contracting LSPs is usually the lowest price not the quality).
- **Financial issue:** CAT tools and the respective training are very expensive for the public sector.
- **Interoperability issues:** For example, Romanian characters were not initially supported by CAT tools.
- **Fundamental issue:** Language data is not considered a valuable asset and is not managed adequately.
- **Continuity issue:** Decision makers change frequently, however, proposed changes must be top-down, creating an even bigger challenge.
- **Political leadership:** The Secretariat General of the Government could/should lead the reform consequently, as the institutional level is not very relevant.

### Action plan:

To support the Romanian language, to improve the translation workflow and to make data sharing in the future easier, the following actions are recommended:

- **Provide comprehensive access to CAT tools:**

This objective addresses the need to raise awareness on the productivity gain through CAT tools and MT but also the procurement of necessary funding to make them available to staff translators. Special attention needs to be paid to facilitating training for efficient and purposeful use of CAT tools, including managing TMs in a way that allows for a smooth sharing of future language resource.
- **Raising awareness of language data as Open Data and a valuable asset:**

To achieve this objective, EU legislation would be most effective and would help increase interest in language technology related issues.
- **Establishing good data management practices in public services:**

This goal includes actions such as creating databases with translated documents and their metadata (e.g., date of translation, information on whether the document was translated in-house or outsourced, IPR holder, etc.).
- **Identifying and gaining access to outsourced translations:**

This objective could be addressed through a high-level decision with a clear mandate for public institutions to collect language data and make it available respectively.