European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.

Deliverable D3.2.3
Task 3

ELRC Workshop Report for Greece

Author(s):	Maria Gavriilidou, Maria Giagkou, Stelios Piperidis
Dissemination Level:	Public
Version No.:	<V1>
Date:	2021-01-21

Contents

1 Executive Summary

The 3rd ELRC workshop in Greece took place on December 15, 2020 as a virtual event. It was organised by the Institute for Language and Speech Processing of the “Athena” Research Centre, member of the ELRC consortium.

The workshop sought to engage participants in a fruitful discussion on the digital readiness of the Greek language in the framework of multilingual Europe and of the digital society in general. Developers, integrators and users of Language Technology, both from the private and public sector shared experiences, requirements and ways for transforming digital interaction in multilingual Europe with Language Technologies. Finally, the value of language data was extensively discussed.

The workshop agenda was structured on three main topics: a) the state of the art of language-centric AI, with a special focus on the availability and maturity of systems for Greek, b) the demands and needs of the public sector with regard to language technologies and, c) the availability and management of public language data.

The main finding of the workshop regarding digital readiness of Greek is that, due to the widely recognised lack of Greek language data, research and industry providers may have to rely on adapting language-independent systems. The public sector in Greece is in need of various language technologies, the most emergent being text processing and analytics tools for legislation codification, chatbots and machine translation. Finally, regarding public language data, it was realised that existing public open data infrastructures that implement open government policies are not designed for -and thus do not prioritise- publishing language resources. As a result, the language data they host are scarce, they are not described with appropriate metadata that would make them discoverable and they are not readily available for the purposes of language development research and applications. Different approaches, workflows and infrastructures are required, in order to unleash the potential of the vast size of language data produced by the public sector, the most critical being the formal inclusion of language data sharing and publication in the defined procedures, workflows and organisational charts of public bodies.

The workshop was attended by 166 participants, mainly from the public sector and the research community.

2 Workshop Agenda

9:30 – 9:50	***Welcome - Frame and objectives of this webinar*** **Stelios Piperidis**, Athena R.C./ILSP, ELRC
9:50 – 11:00	***The potential of Language Technology and AI – Language Technologies in Greece*** Moderator: **Stelios Piperidis**, Athena R.C./ILSP, ELRC Panelists: **Ion Androutsopoulos**, Athens University of Economics and Business **George Giannakopoulos,** ''SCIENCE FOR YOU'' N.G.O. - SciFY **Vassilis Katsouros,** Institute for Language and Speech Processing / “Athena” Research Centre (Athena R.C./ILSP) **Giorgos Petasis**, National Centre for Scientific Research “Demokritos” **Spyros Raptis**, Innoetics/Samsung S.A. **Themos Stafylakis**, Omilia Natural Language Solutions Ltd
11:00 – 11:15	***Short break***
11:15 – 12:00	***Language Technologies by/for the public sector*** Moderator: **Nancy Routzouni**, Min. of Digital Governance, ELRC Public Services National Anchor Point Panelists: **Iraklis Varlamis,** Harokopio University **Thodoris Papadopoulos,** Min. of Digital Governance **Giannis Charalambidis**, Research Centre for eGovernment, University of the Aegean
12:00 – 12:30	***The CEF Automated Translation Platform*** **Szymon Klocek**, Directorate General for Translation, European Commission
12:30 – 12:45	***Short break***
12:45– 13:30	***Language data creation, management and sharing: existing practices and challenges*** Moderator: **Maria Gavriilidou**, Athena R.C./ILSP, ELRC Technology National Anchor Point Panelists: **Dimitris Gioutikas,** National School for Public Administration and Local Government **Dimitris Kapopoulos,** Min. of Digital Governance **Alexandros Nousias**, National Centre for Scientific Research “Demokritos” **Athanasios Sklapanis,** Min. of Digital Governance **Vassilis Papavassiliou**, Athena R.C./ILSP
13:30 - 13:45	***Conclusions***

3 Summary of Sessions Contents

3.1 Welcome and introduction

After a short introduction to the workshop practicalities and its agenda by Maria Giagkou, the workshop started with Stelios Piperidis outlining the subject and frame of the event. The Greek workshop was entitled “Artificial Intelligence for multilingual services to citizens and businesses”. To exemplify what such services may entail, an everyday use case of a mobile phone assistant was presented, as it utilizes and integrates a number of speech and text technologies, from speech recognition and synthesis to natural language understanding and generation. This introductory presentation additionally provided a short overview of the ELRC objectives and framework, i.e. CEF.

3.2 The potential of Language Technology and AI - Language Technologies in Greece (Panel session)

The first panel discussion addressed the AI/LT supply side in Greece and for the Greek language. It hosted a number of panellists from the research community and the industry. The moderator, Stelios Piperidis, Researcher at ILSP/Athena R.C. and ELRC Representative, initiated the session with some examples of AI applications that have already affected our everyday lives in various domains, such as marketing, health services, finance, smart cars etc. He went on to hint why language-centric AI is challenging: language is ambiguous, meaning depends on context, a meaning can have many signifiers and a signifier can instantiate various different meanings. To trigger discussion, he continued with a comment on the state of digital readiness of European languages. According to the 2012 META-NET study, many European languages, Greek among them, are in the Language Technology danger zone, i.e. they don’t meet adequate technology support, thus facing the risk of digital extinction. At this point the main discussion points and the panellists were introduced:

- **Ion Androutsopoulos**, Professor of Artificial Intelligence, Department of Informatics, Athens University of Economics and Business
- **George Giannakopoulos**, Researcher and co-founder of SciFy NGO
- **Vassilis Katsouros,** Researcher, Director of the Institute for Language and Speech Processing (ILSP/”Athena” R.C.)
- **Giorgos Petasis**, Researcher, Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos" (N.C.S.R. Demokritos)
- **Spyros Raptis**, Head of Text to Speech Synthesis R&D, Innoetics/Samsung S.A.
- **Themos Stafylakis**, Head of Machine Learning and Voice Biometrics, Omilia Natural Language Solutions Ltd

Main discussion points:

What is AI and how is it related to language? How important is language understanding towards the goal of “Artificial Intelligence”?

AI is a field of computer science which is also inspired by other scientific fields, such as economics and biology. It tries to develop systems that solve difficult problems, by employing some type of intelligence. Such systems are considered to think like humans, but the truth is that we don’t know how humans think. Other definitions focus on behaviour: intelligent systems behave like humans when solving difficult problems, but humans make mistakes too, e.g. they have car accidents. Intelligent systems finally can be described as those that behave rationally. But this approach has theoretical problems on the one hand and on the other not even humans always behave rationally, e.g. riding a bicycle is not achieved through rational thinking. Nowadays the term AI is often used interchangeably to the term machine learning, i.e. systems that are not explicitly coded, but learn from training examples.

AI is nowadays more and more expanding to other scientific fields, it is interdisciplinary. Law, ethics, sociology and anthropology need to investigate how it affects our lives.

AI is tightly coupled with language understanding**.** One of best known tests, the Turing test, (i.e. a test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human) is based on language understanding. A digital agent needs knowledge and knowledge is mainly transferred through natural language. For instance a biomedicine Question Answering system, should be able to discover and understand texts in the respective field. It goes without saying that we need language understanding and generation in order to have AI systems that are at least useful.   Still, it is important to note that humans communicate not only with text and speech. They use other modalities as well, feelings and sentiment. So, when talking about language, we should also refer to all its modalities and a truly intelligent system needs to be able to handle them.

What is the state of the art? What are the machine's language abilities and which languages do machines "know" better?

To illustrate the progress made in language understanding, let's take digital agents, as they integrate a number of LTs. Speech recognition and synthesis are impressive. Where we need to work on is the background language understanding abilities of such systems, i.e. on the machine's ability to not only understand meanings, but to also have more meaningful dialogues, to interact deeper, maintain a conversation, to be more "cognitive". While synthetic speech is now almost indistinguishable from human speech, we still need machines to understand spontaneous speech, to react in silence, to be more expressive and emotional. So expressivity and subtle interaction are the tasks that research on speech will focus on in the years to come. Another promising task is multilingual speech synthesis, i.e. speech systems that switch between languages. Last, with regard to devices, except for the mobile phone, soon we will also be interacting with our home devices, smart watches etc. This poses a challenge with respect to not only the quality and effectiveness of such systems, but also with respect to their computing requirements, i.e. they should be able to run on smaller and smaller devices. As these technologies become more and more ambient, what we need to additionally work on is regulation. We are often astonished with what has been achieved, to such a degree that we now have started considering ways to mark boundaries to how much AI intrudes our lives.

Methodologically, language understanding tasks have experienced a revolution in recent years. Deep learning approaches have boosted AI and provide great potential. In the last two years language understanding has been boosted due to the self-supervised approach, i.e. training models without having a specific task in mind. Such systems are trained with a small dataset and then they are adjusted to different tasks.

While most of the developments start with, or focus on, English, or some big language (e.g. Spanish), recent approaches that transfer knowledge from English to other languages for which not enough training data are available are very promising.

This makes things easier for smaller languages. We should underline, however, that the prioritization of languages does not have to do exclusively with their technological readiness and the availability of data, but mainly with their market size.

What are the mature systems that are available for Greek? Is Greek digitally ready? Are we dependent on the big international players?

While Greek is already supported in a number of commercially available platforms, the level of quality that has been achieved is not comparable to that of other languages such as English or Spanish. From a research and commercial point of view, some tools are available for content analytics, sentiment and polarity analysis mainly for social media and news texts, entity extraction and linking, information extraction, e.g. arguments extraction, QA, language generation and corporate fame. In terms of infrastructures, META-SHARE and CLARIN:EL host a considerable number of resources for Greek. In terms of readiness and maturity, some tools, especially for Java and Python are readily available, Spacy and Apache Tika also support Greek to an extent, but still this is fragmented. As a result, effort and expertise are required to use these tools for Greek. To build end-to-end systems for Greek still requires development of specific components and this in turn requires language resources. The use of embeddings makes our work easier, since knowledge can be transferred from training on large datasets, however the balance between using language-specific resources and embeddings is a question worth investigating. To achieve better accuracy, you need resources, not just pre-trained embeddings.

While algorithmically we are there, especially due to deep learning methods, which are language independent, sizeable data, annotated or raw, as well as computational power are required.

At the level of government policies for AI and LT, what is happening in Greece and what should we do as a country?

Greece, as part of the EU ecosystem, follows the EU policies for AI, high-performance computing and cyber security. The Digital Transformation Book and the national AI Strategy are now being prepared. The public sector is one of biggest “customers”. Many projects are being implemented for the public sector, e.g. codification of legislation and of decisions of the State Legal Council, digitisation of court decisions etc. Other projects focus on data, specifically on data interoperability, to tackle data siloing per ministry. Another ongoing project is Syzeuxis II which has to do with the telecommunications infrastructure of the public administrations, for instance 5G which will enable better cloud based services and IoT in different domains, e.g. in agriculture, fishery, meteorology etc.

Developing and maintaining AI requires involving all stakeholders, at the national, European and international levels. AI should be human-centric, fair and accessible to all.

The panel session concluded with a reference to the CEF AT Catalogue of Tools and Services (https://cef-at-service-catalogue.eu/), where mature LT tools and services developed in Europe are listed and, by appropriately applying the faceted search functionalities, anyone can investigate tools and services that support Greek and/or have been developed by providers based in Greece.

During the panel discussion several comments and questions were posted through the chat. We list here some of the most relevant contributions:

- With regard to the Turing test: see also the Chinese Room, a relevant cognitive test by John Searle: https://en.wikipedia.org/wiki/Chinese_room.
- With regard to pre-trained language models: this is the publication on the Greek BERT: http://nlp.cs.aueb.gr/pubs/greek_bert_setn_2020.pdf
- See also the work on the “Pepper” robotic platform by George Petassis at NCSR Demokritos
- An example of multimodal applications is the generation of medical diagnoses texts from medical images: http://nlp.cs.aueb.gr/pubs/sivl2019_survey_biomedical_image_captioning.pdf

- Q: Is there any application for special education?
  A: For special education and accessibility in general, see scify.org, http://langaware.com/ and the European project SHAPES.
- Q: What is available for processing Ancient Greek (Hellenistic Koine)?
  A: See http://cltk.org/. See also these publications on processing of Historical Greek Polytonic Scripts https://users.iit.demokritos.gr/~bgat/OldDocPro/05_paper_305.pdf; https://users.iit.demokritos.gr/~bgat/DAS2016_katsouros.pdf; https://dl.acm.org/doi/10.1145/3322905.3322926

At the end of the panel discussion, we asked the participants to answer 3 poll questions to investigate their experiences as users of language technologies. 73 participants answered these questions and the answers are reported in the following tables:

Q1: Which of the following language technologies do you use more? (you can select up to 3)

	#answers
Automated translation	59
Information search and retrieval (e.g. web search)	60
Digital assistants (e.g. on mobile phones and call centers)	20
Speech recognition (e.g. dictation of messages on your mobile)	12
Text to Speech synthesis (e.g. your mobile phone assistant reads out your messages)	1
None of the above	1

Q2: If you use some of these technologies, you usually use them in (single choice):

	#answers	%
English	47	64,4%
Greek	22	30,1%
Other language	4	5,5%

Q3. As a simple user, how satisfied are you with the quality and reliability of these technologies for Greek? (single choice)

	#answers	%
Excellent	2	2,7%
Good	43	58,9%
Fair	23	31,5%
Poor	3	4,1%
Very Poor	1	1,4%
I don't know / No answer	1	1,4%

Two main points should be highlighted, based on the participants' answers above. First, the workshop attracted a significant number of participants that are at least interested in machine translation and use the respective systems. Second, the fact that more than half of the respondents use language technologies in languages other than Greek is a strong indication of the level of predominance of English as lingua franca in digital interactions. Still, however, the quality of these systems in the Greek language is considered to be good or fair enough.

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

The second panel session addressed the LT demand side, specifically the demands and needs of the Greek public sector for LT-enhanced digital services. The panel was moderated by Nancy Routzouni, advisor to the Min. of Digital Governance and ELRC Public Sector National Anchor Point.

Nancy Routzouni started the session by announcing two important developments on the policy level: First, the publication of the "Digital Transformation Bible 2020-2025" (https://digitalstrategy.gov.gr/), which was under public consultation at the time of the workshop. The Digital Transformation Bible is a record of the necessary interventions in the technological infrastructure of the state, in the education and training of the population for the acquisition of digital skills as well as in the way Greece utilizes digital technology in all sectors of the economy and public administration. Its main role is to describe the vision, philosophy and goals of the national strategy for the digital transformation of the country. It describes the guiding principles, the model of governance and implementation, but also the strategic axes of digital transformation. Furthermore, it describes more than 400 specific projects, classified into short-term and medium-term, horizontal and sectoral, which implement the strategy for Digital Greece.

The second development announced is the upcoming National Artificial Intelligence Strategy. The text is currently being drafted by the Ministry of Digital Governance and it will be submitted to public consultation in January 2021.

As triggers for discussion, an overview of the roles of governments with regard to AI and LT development were presented. Governments can act as:

- Financiers and investors
- Regulators
- Conveners and standards-setters
- Data stewards
- Smart buyers and co-developers
- Users and service providers

The panel discussion aimed to touch upon the last two points, i.e. to investigate the role of the Greek public sector as buyer/co-developer of LT services and as user and provider. In this direction, some indicative business cases of LT in eGovernment are the following:

- Processing of the huge amount of data dealt with daily by public administrations in view of developing applications for Automatic Message Answering, Case Routing, Phone Call Summation etc.
- Policy making through analysis of public consultations or public opinion, utilising technologies such as Competitive intelligence analysis and Sentiment analysis

Relevant language technologies are already mature and used by public services in Europe, some examples being the chatbot service by the Finnish portal for foreign entrepreneurs, a public consultation and analytics platform used by Belgian local administrations, the Spanish machine translation system for public administrations (PLATA), the Latvian free LT services (Hugo.lv).

Nancy Routzouni then introduced the panellists:

- **Iraklis Varlamis,** Associate Professor, Department of Informatics and Telematics, Harokopio University of Athens
- **Thodoris Papadopoulos,** Coordination Unit of the Public Administration Common Digital Gateway, Min. of Digital Governance
- **Giannis Charalambidis**, Professor at the Department of Information and Communication Systems Engineering, University of Aegean, and Director of the Greek Research Centre for eGovernment

Main discussion points:

What are the most emerging needs and requirements of the Greek public administration for language-centric AI technologies?

The central governmental portal for digital services, gov.gr, should investigate the use of chatbots, for instance, an intelligent agent that can decide, given the profile of a citizen, which financial aid he/she is eligible for. A second important application is text processing for the benefit of society. For instance, currently a CEF project, ManyLaws (www.manylaws.eu), is implemented, which processes legal texts in many languages in view of automatic codification. Another important application of LTs is the analysis of public opinion, including social media analytics for stance and sentiment analysis.

A very important aspect in this discussion is the wealth of unexploited data that reside in the public administration front- and back-offices, and they comprise anything that is written by the administration, which is by default open. These data can fuel interesting applications based on short-term decision algorithms that automate simple decision making.

Which public services do you think can help and fuel the development of LT?

Public administration is one of the biggest creators and providers of language data. Since now everything is digitised, it is easier for the public sector to act as provider of language data. The same is true for the European administration: with 27 member states and 24 official languages, the EU produces a wealth of language data. In Greece, the Transparency portal (diavgeia.gov.gr) is a noteworthy case. It currently hosts 42 million documents, i.e. approximately 100 billion pages of text. These are available under Creative Commons licences and can be used by the research community.

Another source of public data is the Government Gazette. A few years ago, NLP techniques were piloted to automatically codify the Government Gazette, so that citizens can easily retrieve the most updated version of a law. This work is planned to be extended under the auspices of the Ministry of Digital Governance and to be exploited for the development of the National Portal for Law Codification (which is actually anticipated for in the Digital Transformation Bible). A source for public open language data are the public procurement texts at www.promitheus.gov.gr and the Central Registry of electronic public procurements.

Finally, a noteworthy source of public data are the court decisions and minutes, given, of course, that they are (pseudo)anonymised, as well as the incoming citizens' applications to public administration. Such language data can train systems, such as chatbots and automated case routing.

An additional interesting source could be the new National Registry of Procedures (Diavlos) which is being developed with the aim to record and present the physical and digital administrative procedures of Public bodies and services, in order to have a unique reference point, from which reliable and valid information is extracted. In this way the administrative procedures become transparent, documented, correctly and uniformly structured, allowing their easy search and identification. Note that all the information will be provided in both Greek and English. The registry integrates several different existing registries and resources, such as catalogues and lexica, all available in Greek and English. These can be data sources for training very interesting applications, including machine translation and intelligent dialogue systems that can guide citizens through the administrative procedures.

Which problems of the administration can LT address? Which LTs could enhance the Common Digital Gateway? Are there any services hosted in the Gateway with multilingual needs?

Gov.gr, the Common Digital Gateway, hosts and presents all the public digital services, without however implementing them in a common platform. In essence, it integrates metadata for the existing services and redirects to the hosting public organisation. At the moment we are investigating the integration of eTranslation to the Greek Common Digital Gateway. The use of eTranslation, in combination with specialised thesauri to handle domain-specific vocabularies will be a very interesting approach to making it usable by EU citizens who don't speak Greek or English.

A second plan for the gov.gr platform is the design of a Helpdesk common to all services hosted. This is a great opportunity to try a chatbot service that answers citizens questions, both Greek nationals but also foreigners. The system could additionally classify the citizens' requests and route each case appropriately. For instance, an issue about the social security number, should be automatically classified and forwarded to the respective organisation's helpdesk. Several other ideas are investigated, such as speech recognition for serving refugees and foreigners.

What should a national strategy for AI aim at and what guidelines should it include with regard to Language Technologies?

In our strategic design we should aim not only at better services, but we should also aim at openness, participation and collaboration throughout the process.

A strategy should define both the frame and the actions. With respect to the frame, our existing model for services offering is not appropriate anymore, because it does not anticipate the use of AI systems, for instance AI-enabled decision making. We need a new model for service offering. An AI strategy should not be just a strategy for AI, but instead a comprehensive redesign of our operational models. We additionally need to address the availability of technical infrastructures.

With regard to actions, we need sectoral plans, i.e. to investigate which AI systems are needed in different sectors, e.g. health, education etc. Experimentation should be anticipated. We need to be willing and ready to experiment with solutions and abandon what is not fit for the purpose. Forward looking is also necessary; the public sector should participate in emerging innovative actions.

If the AI strategy hopefully includes a section on language technologies, two main points should be emphasised: Open data and sharing economy. Our language is a small one. It is our responsibility to share what we build, e.g. datasets, lexica, algorithms etc., and the state should help towards this sharing culture. Otherwise we will not keep pace with the advancement achieved for other languages.

The panel session was concluded at this point. During the discussion comments and questions were posted through the chat. We list here some of the most relevant contributions:

- Given the opportunity of the reference to the Finnish chatbots, please check out the COVID chatbot developed by ILSP/Athena R.C. and send us your feedback: https://apps.ilsp.gr/covid-va/chatbot/
- Q: A number of laws are available as scanned images which raises the need for OCR technologies. How is this handled?
  A: Indeed, this is true especially for older laws. In our use case we used Tesseract, a Google library for text extraction from scanned documents (https://opensource.google/projects/tesseract). Since that was a pilot problem, such issues were not important. But they definitely need to be resolved when deploying a production-ready system for law codification.
  A: The problem starts at the Parliament and is rooted in the way the laws are written. The European Commission, having realised this, made available the LEOS application which assists at the legislation editing phase.
- Are the principles of good legislation in line with the requirements for processing legislative texts? See https://www.hellenicparliament.gr/Nomothetiko-Ergo/Anazitisi-Nomothetikou-Ergou?law_id=b5ffa7ff-4d09-4251-a536-4e2b85e17998; https://www.hellenicparliament.gr/Nomothetiko-Ergo/Anazitisi-Nomothetikou-

Ergou?law_id=b5ffa7ff-4d09-4251-a536-4e2b85e17998 and https://www.kefim.org/deiktis-poiotitas-nomothetisis-2019/

- For "virtual judges" have a look here https://www.aclweb.org/anthology/P19-1424/
- For an overview of systems for legislation processing have a look at the SETN 2020 workshop: https://altws.mashup.gr/about/
- These documents [referring to the Transparency and Open Data portals] are not corpora. They need heavy processing in order to be used in LT.
- For procurement of AI systems see https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/890697/Guidelines_for_AI_procurement__Mobile_version_.pdf
- For challenges and obstacles in the adoption of AI by public services see: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

At the end of the panel discussion, we asked the participants to answer 2 poll questions to investigate the degree of trust to language technologies when interacting with public administrations. The results are presented in the following tables:

Q1: As a citizen, would you trust a digital assistant to communicate with a public administration and request information?

	#answers	%
Yes	50	68,5%
No	17	23,3%
I don't know / No answer	6	8,2%

Q2: Would you trust a machine translation system to communicate with a public administration in a language you don't speak?

	#answers	%
Yes	42	57,5%
No	24	32,9%
I don't know / No answer	7	9,6%

3.4 The CEF AT platform

The CEF AT platform was presented by Szymon Klocek, DGT, EC. He presented the evolution of the EC's machine translation system from the statistical to the neural paradigm and its development to cover more language technologies through the CEF AT platform. The target users of eTanslation are:

- Translators and staff of the EU Institutions
- Digital services of the EU Institutions
- CEF Digital Service Infrastructures
- Pan-European digital public services
- Public administrations in Member States, Iceland and Norway
- European SMEs, as of March 2020

eTranslation can be accessed through either

- a web user interface to automatically translate documents and text snippets or
- an API to integrate machine translation in workflows, websites, digital services, etc.

eTranslation supports all official EU languages, Norwegian, Icelandic, Russian, Chinese (Mandarin) and Turkish and provides not only a general language engine, but also domain-adapted engines, such as the EU formal language engine, health, culture etc. Szymon Klocek subsequently commented on the

translation output quality, underlining that, because eTranslation has been trained on a huge database of translated official EU texts, it is very good in translating formal EU language and may not be as good when it comes to non-standard or creative texts. However, the availability of the general language engine which is trained on respective non-official texts, delivers high-quality output. The need to select the appropriate domain-adapted engine according to the text type to be translated was highlighted.
Regarding future plans for development of the CEF AT platform, Szymon Klocek noted that the EC is working on extending the domain coverage (e.g. scientific texts); on supporting additional non-EU languages of social & economic importance, and regional languages; on developing more language technologies, such as speech recognition, anonymization, named-entity recognition and a basic Computer-Aided Translation tool. Some of these tools have already been made publically available at https://language-tools.ec.europa.eu/.
Finally, an overview of the eligible users was provided (Public Administrations, Universities, CEF-funded projects, SMEs) and the steps and links to self-register and use eTranslation were presented.

- Self-registration via https://webgate.ec.europa.eu/etranslation/public/welcome.html
- Web service (API) Technical documentation: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/How+to+submit+a+translation+request+via+the+CEF+eTranslation+webservice
- eTranslation Service Desk: help@cefat-tools-services.eu
- Access to eTranslation web user interface: https://webgate.ec.europa.eu/ETRANSLATION

A number of questions were addressed to Mr. Klocek through the chat. Some of them concerned the eligibility of user types (universities and freelancers, other H2020 funded projects) and the availability of CEF tools like the Twitter translator and were positively answered. Other questions are listed here:

- Q: What is the advantage of using eTranslate as opposed to Google Translate, Microsoft Bing Translate etc?
  A: Using eTranslation is free and safe, and you keep ownership of your data at all times.
- Q: Are there plans to make eTranslation a paid service? It is free for now...
  A: According to current policies eTranslation is and will continue to be free.
- Q: Small-scale studies that I have carried out indicate that eTranslate does not perform better than Google Translate or PureNMT in the EN-EL language pair even for texts in the EU domain. What do your studies indicate?
- A: Larger-scale studies indicate that eTranslation is performing very well.
- Q: But isn't it true that the more we use it we are contributing to the improvement of a European Platform rather than to a privately owned platform, right?
  A: To improve the system when using it, wouldn't you need to collect the translation queries of the users and any corrections they make? Wouldn't this be a confidentiality risk? User feedback is also a way to poison the system.
  A: eTranslation does not collect user feedback to use it for retraining the system.
- Q: is there a way to provide EC Translation tool with Greek documents?
  A: Yes, through https://elrc-share.eu
- Q: Do you see eTranslation to be used as a tool to help translate websites of EU Agencies that have yet to be translated in other EU languages?
  A: Yes, the API is used for this. The output should be edited by humans.
- Q: Why is the Commission developing a CAT system of its own?
  A: Initially to address its own internal needs.

Right after the presentation and QA, the participants were presented with three poll questions regarding their experiences with machine translation systems. 68 participants answered the questions as follows:

Q1. Have you ever used a machine translation system?

	#answers	**%**
Yes	65	95,6%
No	3	4,4%

Q2. If yes, which one?

	#answers	**%**
eTranslation	4	5,9%
Other commercial or freely available system (e.g. Google translate)	41	60,3%
All of the above	19	27,9%
I don't know/No answer	4	5,9%

Q3. Based on your experience with such machine translation systems, how satisfied are you with the translation quality to / from Greek?

	#answers	**%**
Excellent	0	0%
Good	14	20,6%
Fair	40	58,8%
Poor	9	13,2%
Very Poor	2	2,9%
I don't know / No answer	3	4,4%

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

The final panel session of the Greek workshop focused on language data and sought to investigate the policies, legal framework and infrastructures for sharing language data in Greece. The panel was moderated by Maria Gavriilidou, Researcher at the Institute for Language and Speech Processing/ Athena RC and ELRC Technological National Anchor Point. She started the session by underlining the importance of data for the development of AI and in particular of language data for language-centric AI. Language data are a special type of data; they are produced constantly as humans communicate, but they are not readily available for processing and use. LT systems are trained on processable language data in digital form, in vast quantities and, ideally, covering a wide range of types, domains and languages. To reach the readiness level required in order to be used in language technology development, language data need to go through the following steps: Creation or discovery and collection, storing, processing and curation, publication, sharing, and finally use and reuse.

Maria Gavriilidou continued by presenting the main findings of the ELRC White Paper and specifically the findings for the language data creation and management practices in Greece (https://www.lr-coordination.eu/sites/default/files/Documents/ELRCWhitePaper.pdf). According to this study, the practices for the creation of multilingual data (translations) in the Greek public sector are fragmented. There is no formal translation procedure, common to all public administrations and digital translation culture is rather marginal.

Regarding the available infrastructures, Greece participates in well-established European initiatives and maintains infrastructures for language data addressed mainly to the research community: CLARIN.EL and the European Language Grid platform. ELRC-SHARE, ELRC's data repository, addressed both to the research community and to public bodies is also available and it hosts language data pertinent to machine translation tasks.

The data produced by the public sector are hosted in two main public infrastructures:

- Transparency portal (https://diavgeia.gov.gr/en), where all government institutions are obliged to upload their acts and decisions (in pdf format), and
- Open Government Data Portal (http://data.gov.gr), which hosts the central catalogue of Open Government Data and offers open access to digital resources of the Greek government institutions to citizens, services and information systems for reuse for any purpose. The latter is being harvested by the European Data Portal.

Concluding her introduction, Maria Gavriilidou noted that digitisation culture is not adequately endorsed by the Greek public sector. Language Technology is not part of the country's language policy and the importance of AI has only very recently been recognized. She additionally commented that, in order to manage our language data the following aspects should be considered and catered for:

- The institutional and legal framework for managing and distributing language data
- The technical infrastructures for hosting and sharing them
- Awareness-raising and training of data owners/providers.

In the light of these issues, she introduced the panellists:

- **Dimitris Gioutikas,** Deputy Director of the National School for Public Administration and Local Government
- **Dimitris Kapopoulos,** Head of the Department for Information Systems for Open Government, Min. of Digital Governance
- **Alexandros Nousias**, Legal expert, Ethics officer, N.C.S.R. Demokritos
- **Athanasios Sklapanis,** Head of the Open Government and Transparency Department, Min. of Digital Governance
- **Vassilis Papavassiliou**, Research associate, Athena RC/ILSP

Main discussion points:

What is the framework for sharing language data in Greece?

Legislation for open data is in place, the main one dating back in 2014, while a circular specifically encouraging public administration to share language data with ELRC had also been published in 2016 by the Ministry of the Interior.

Have these regulatory acts been adequately implemented? Are there any newer official acts regarding sharing of language data?

Law 4305/2014 on the re-use of public information and its implementing circular as of 2015, in general, have been fairly implemented in Greece. It tried to communicate to public administration some very difficult notions of openness and to introduce the procedures that implement this policy.

The implementation of the 2016 circular for ELRC has not been as effective as we had hoped for. The reason for this is the lack of an openness and sharing culture in public administration. The second reason is that the value of this endeavour has not been recognized. Another reason why it is difficult to collect multilingual texts to feed ELRC and eTranslation is the workload and the low prioritization of this task; leaving the task to public servants' personal interest and awareness of the value of the task. Such tasks and responsibilities should be included in each public body's organizational chart. An important step in this direction is the Presidential Decree 40/2020, which assigns responsibilities for supporting ELRC to the Department of Information Systems for Open Government of the Min. of Digital Governance. This is a best practice example and should be included in the organizational charts and responsibilities of other ministries as well.

Are there any updates to the legal framework? Are language data explicitly mentioned?

The discussion on open and restricted data is not new. Law 4305 was an ambitious and pertinent, at the time, legal tool. Due to the lack of sharing culture, it did not convey results to its full potential. We are now more mature. This is why Law 4305 has been updated in the light of Directive 1024/2019 on open data, which is integrated in the Greek legislation with the law on digital governance. The Digital Transformation Bible also discusses open data. So, the framework is there. The question is its implementation. The Data Governance Act, a legislative proposal that paves the way in practical terms for sharing open data, will greatly alleviate currently attested difficulties.

Are the concerns for privacy and personal data valid with respect to sharing language data?

The basic principle is: open as far as possible, restricted as far as necessary. Such principles and the legal frame in general do not prohibit development of AI. On the contrary, they explicitly encourage sharing of information. The argument that data cannot be shared because of personal data and IPR issues is not valid. To protect IPR and personal data, one needs to design the appropriate formulas and operational procedures for sharing, instead of not sharing at all.

If the appropriate licence is applied, data can be shared as open, even if some restrictions apply, e.g. attribution, share-alike or non-commercial.

It is true that EU and Greek legislation explicitly state various types of data but not language data. This may be due to the fact that language data are difficult to define. Law 4304 refers to "documents, information or data" and the recent law refers to "documents and data". If we consider language data as "documents", it can be claimed that they are covered by the existing legislation. The reason why language data are not included in data.gov.gr is that the national open data portal, which is the operational means that implements the policy for public information reuse, is designed and intended to present tabular data.

This is very important, as it proves that the value of language data is not adequately recognised. As a result, a wealth of public textual data is not published and thus not usable by the LT community. A second issue is the metadata of the datasets on the open data portal. According to the European Data Portal, only a small percentage of the Greek open data are sufficiently described with metadata. The issue of quality of data and metadata is critical.

This demands a change in culture through awareness raising activities. It also needs political patronage, i.e. support and endorsement of the task by high-level officials. Quality of data is very important for machine translation tasks. Press releases, terminologies and budget and public debt implementation bulletins are fit candidates for the task, because they are thoroughly edited. A dedicated infrastructure for public language data might be needed, because the existing ones, Diavgeia and data.gov.gr, have not been designed for this purpose.

The importance of training should also be underlined. We need public servants that are aware of these topics and appropriately trained. The curriculum of the National School for Public Administration and Local government includes courses on personal data and public open data, digital skills and more specialized courses in eGovernment, such as Data science, data management and statistical processing etc.

What difficulties did you have to address in your attempt to collect language data for ELRC in your organization? Based on your experience, what are the main challenges?

Engagement with ELRC was made feasible because of the aforementioned circular of the Ministry of the Interior, which encouraged all directorates to assign one person per directorate as operationally responsible for collecting datasets for ELRC. Of course, not all directorates reacted, depending on whether their activities included the creation and management of language data, especially multilingual data. The fact that this type of involvement is not mandatory, nor monitored with deadlines and that no responsibilities are defined within the organisations' structures, in addition to the fact that the value of the endeavour is not widely recognized, hinder further involvement of the public administrations. It also undermines its sustainability.

A repository of language data produced by the public sector, whether as a subdomain of data.gov.gr or a separate domain, would boost LT development, which in turn would be for the benefit of the public, through relevant services.

The quick development of LT services for the public sector and for citizens in the framework of crisis management could prove the value of language data. Can we report one such example from the current COVID-19 pandemic?

Imagine a dialogue system that informs citizens about the Covid testing and vaccination centres or about the relevant financial aids available. Or an information platform enhanced with machine translation which can thus provide information to residents who are not speakers of the national language. Given that the algorithms and the computational infrastructure are available, what is missing is the data, in this case language data pertinent to the COVID pandemic. The LT community has reacted to address such needs. An initiative in this direction is the COVID-19 Multilingual Information Access (MLIA) initiative (http://eval.covid19-mlia.eu/) which seeks to collect in a short time language data for the pandemic, which can then be used to train and adapt relevant language technologies, in particular machine translation and information extraction. Currently, 5 million documents have been collected for information extraction tasks in seven languages, including Greek, and 5 million parallel sentence pairs for the machine translation tasks. The first results are expected in January.

This contribution concluded the panel session. Some interesting reactions have been recorded in the chat, especially with regard to the comment made during the panel discussion that the Greek language policy does not include LT. For instance "is there actually a language policy in Greece?", "The newly established Centre of Excellence for Multilinguality and Language Policy of the University of Athens will organise a workshop on this issue soon."

Finally, the audience was requested to answer three additional poll questions, to investigate data sharing practices and challenges. 41 participants answered the following questions:

Q1. Do you / your organization have language resources / collections of translated texts in digital form?

	#answers	%
Yes	24	58,5%
No	7	17,1%
I don't know/No answer	10	24,4%

Q2. Does your organisation employ a data management plan, i.e. guidelines and/or standards for making the data created by the organisation findable, accessible, interoperable and reusable?

	#answers	%
Yes	11	26,8%
No	12	29,3%
I don't know/No answer	18	43,9%

Q3. What, based on your experience, are the main difficulties that may prevent the sharing of language data? (up to 3 choices)

	#answers
Legal issues	23
I/my organisation do not/does not see any value in sharing language data	3
It is not my responsibility / I am not authorized to do so	12
Lack of time	4
Other	5
I don't know/No answer	5

3.6 Conclusions

The workshop was concluded with a short wrap-up session by Maria Giagkou. The main topics and findings discussed were summarised. As take-home messages, Maria Giagkou underlined once again that lesser spoken languages, like Greek, are threatened with digital extinction. The development of Language-centric AI for the Greek language is the way through. And since AI needs data, making our language data available for reuse is our small contribution to the digital preservation of our language. Finally, participants were encouraged to contribute their language data through the ELRC-SHARE repository and to contact the ELRC Helpdesk for support.

4 Synthesis of Workshop Discussions

The workshop agenda was structured along three main topics: a) the state of the art of language-centric AI, with a special focus on the availability and maturity of systems for Greek, b) the demands and needs of the public sector with regard to language technologies and, c) the availability and management of public language data.

With regard to the first topic, LT in Greece and for the Greek language, it became apparent from the various contributions that the picture for Greek is fragmented. Some tools and services do exist, but the research and industry providers mainly rely on adapting language-independent systems, because of the widely recognised lack of Greek language data. Three main factors were identified as prerequisites for developing language-centric AI: data, trained human experts and access to powerful computing infrastructure.

The Greek public administration is currently working towards expanding digitisation of all public services and centralising their availability through the Common Digital Gateway. This portal will be available in a number of languages and the integration of eTranslation is currently being investigated. In addition to machine translation, the need for chatbots arose from the contributions of the public sector representatives. Chatbots are envisaged for providing information to citizens on administrative procedures and case routing, as part of the Gateway's central Helpdesk. Finally, language technologies for text classification and information extraction are considered to be valuable additions for a constant demand of the Greek public sector and all of stakeholders involved, i.e. the automatic codification of legislation. Such technologies are already additionally used for building some of the public administration core registries, such as the central common registry of administrative procedures and the registry of public organisations.

With regard to the availability of public language data, a number of sources have been indicated, such as the Government Open Data portal (data.gov.gr), the Transparency portal (https://diavgeia.gov.gr/) and the Public eProcurement portal (http://www.promitheus.gov.gr/). However, it was widely argued that these public infrastructures are designed as instruments that implement open government policies and not as language resources repositories. Thus, the language data they host are scarce, they are not described with appropriate metadata that would make them discoverable and they are not readily available for the purposes of language development research and applications. For the same reason, it should not be expected that the public sector will soon be updating its policies and administrative procedures in order to explicitly include the publication of language data produced by the administration through the established workflows for publishing open data on the Government Open Data portal. The current infrastructures are not fit for the purpose and it has been extensively argued that, in order to effectively address the need for publishing public language data, a new infrastructure (e.g. a sub-domain of data.gov.gr) would be required, along with new administrative procedures that would explicitly mandate all public organisations to publish their language data, such as bilingual press releases and bulletins, as they do for their tabular datasets, decisions and acts. Both from the panellists' contributions and from the participants' responses to the respective poll questions, it became apparent, that, although the public sector is one of the major language data creators and holders, and that these data are by default open, they remain unexploited. Data management plans are not often employed. Collecting and sharing such public data is hindered by the complex administrative procedures for explicit assignment of responsibilities, the lack of legal support which could ease the legal concerns and help public servants decide on the sharability of a dataset, in addition to the fact that the value of the endeavour is not widely recognized.

An important step towards tackling the administrative obstacles and facilitating the procedures for collecting language data within a public organization was the Presidential Decree 40/2020, which assigned specific responsibilities to the Department of Information Systems for Open Government of the Min. of Digital Governance for supporting ELRC. This support is explicitly detailed as collecting language data from the Ministries of Finance and Digital Governance. This is a best practice example and should be included in the organizational charts and responsibilities of other ministries, as well.

5 Country Profile: Language data creation, management and sharing

The situation in Greece with regard to language data creation, management and sharing practices has not changed during the last year, i.e. since the publication of the Country profile as part of the ELRC White Paper in 2019. The practices for the creation of multilingual data (translations) in the Greek public sector are fragmented. There's no formal translation procedure, common to all public administrations. Some authorities maintain in-house translation departments, others outsource translations. In all cases, the outputs are not managed according to a data management plan, translation memories are not created or not requested when translation is outsourced, while in some cases the translation output is not even stored in digital form.

Two important developments on the policy level, however, should be reported: The first development is the establishment of the Ministry of Digital Governance, which brings together all the critical IT and telecommunications structures related to the provision of electronic services to citizens and the wider digital transformation of the country, previously scattered in different public organisations. As a result most of the critical ELRC stakeholders, including the units involved with the management and publication of public open data, are now part of the Ministry of Digital Governance.

Second, on the policy level, the "Digital Transformation Bible 2020-2025" (https://digitalstrategy.gov.gr/) has been published (under public consultation at the time of the workshop). The Digital Transformation Bible is a record of the necessary interventions in the technological infrastructure of the state, in the education and training of the population for the acquisition of digital skills as well as in the way our country utilizes digital technology in all sectors of the economy and public administration. Its main role is to describe the vision, philosophy and goals of the national strategy for the digital transformation of the country. It describes the guiding principles, the model of governance and implementation, but also the strategic axes of digital transformation. Furthermore, it describes more than 400 specific projects, classified into short-term and medium-term, horizontal and sectoral, which implement the strategy for Digital Greece. The Bible includes special provisions for the release and exploitation of public data. Among the anticipated provisions and actions is the establishment of Thematic Data Repositories in selected vertical sectors. A number of data that are considered of high-value are mentioned, e.g. geodata, meteorological, environmental and cultural. Unfortunately, language data are not explicitly mentioned nor there seems to be any special provision for their inclusion in a specialized Thematic Data Repository.

Staying at the policy level, the, much anticipated for, Hellenic National Strategy for Artificial Intelligence was announced at the ELRC workshop. The text is currently being drafted by the Ministry of Digital Governance and it will be submitted to public consultation in January 2021.

Regarding the challenges identified when it comes to sustainable language data management and sharing by and in the public organisations, the discussions at the 3rd ELRC workshop in Greece have confirmed previous findings. The main obstacles that prevent public administrations from effectively adopting sharing practices and integrating them in their workflows are: the lack of openness and sharing culture; the lack of appreciation of the value of this endeavour; the lack of explicit endorsement of the task by the political or high-level managerial personnel and of a subsequent inclusion in the public bodies' organizational charts and structures. Legal concerns constantly appear to be present in the list of conceived challenges, although the national legal and institutional frame is considered to be in place and it provides the theoretical framework and the guidelines for making public data as open as possible.

An important step towards tackling the administrative obstacles and easing the procedures for collecting language data within a public organization was the Presidential Decree 40/2020, which assigned specific responsibilities for supporting ELRC to the Department of Information Systems for Open Government of the Min. of Digital Governance. This support is explicitly detailed as support in collecting language data from the Ministries of Finance and Digital Governance. This is a best practice example and should be included in the organizational charts and responsibilities of other ministries as well.