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

European Language Resource Coordination

Connecting Europe Facility

Deliverable D3.2.9
Task 8

ELRC Workshop Report for Portugal

Author(s):	António Branco, Tânia Sofia Oliveira
Dissemination Level:	Public
Version No.:	V1.0
Date:	2018-08-30

Contents

1 Executive Summary

This document reports on the second ELRC Workshop in Portugal, which took place in Lisbon, on June 27, 2018, at the premises of the Representation of the European Commission in Portugal (Largo Jean Monnet 1-10º, 1269-068 Lisboa). It includes the agenda of the event (section 2) and provides details on the content of each individual, interactive and panel workshop session (sections 3 & 4). The event received over 50 registration requests and had over 35 attendees spanning a wide range of government departments and public organizations.

The dedicated event webpage can be found at http://lr-coordination.eu/pt/portugal .

2 Workshop Agenda

08:30 – 09:30	**Registration**
09:30 – 09:45	**Welcome and introduction** *Paulo Mauriiti, AMA, Member of the Board of Directors*

Session 1. Connecting a multilingual Europe: European context and local needs

09:45 – 10:05	**Connecting public services across Europe: ambition and results so far** *Aleksandra Wesolowska, DG CONNECT, EC*
10:05 – 10:20	**National initiatives for digital public services and (open) data** *André Lapa, AMA, Dados.gov*
10:20 – 10:50	**CEF in Portugal: an outlook into current and future challenges – Panel session** Moderator: *Paulo Vale, AMA, ELRC Public Services NAP* Panelists: • *André Lapa, AMA* • *Danilo Furtado, Directorate-General for the Territorial Development*
10:50 – 11:30	*Coffee-break*
11:30 – 12:30	**The CEF eTranslation platform @ work** *Michael Jellinghaus, DGTranslation, EC* *Paulo Batista, DGT Field Officer PT, EC*
12:30 – 13:30	*Lunch*

Session 2. Engage: hands-on data

13:30 – 13:50	**The European Language Resource Coordination (ELRC) action** *Khalid Choukri, ELRC consortium*
13:50 – 14:05	**ELRC in Portugal** *António Branco, FCUL, ELRC Technology NAP*
14:05 – 14:30	**Preparing and sharing data with the ELRC repository** *Khalid Choukri, ELRC consortium*
14:30 – 14:50	**Can language data be shared and how?** *Andre Lapa, AMA, PSI Group* *Khalid Choukri, ELRC consortium*
14:50 – 16:30	**Identifying and managing your data: Questions & Answers** Moderator*:* *Paulo Batista, DGT Field Officer PT, EC* *All available for answers (NAPs, EC representation, Local FO, ELRC consortium representative)*
16:30 – 16:40	**Discussion and Conclusions** Moderators*:* *António Branco, FCUL, ELRC NAP* *Paulo Vale, AMA, ELRC NAP*
16:40 – 17:00	*Coffee Break and networking*

3 Summary of Content of Sessions

3.1 Welcome and introduction by the EC and AMA

Paulo Batista, the DGT Field Officer for Portugal, opened the second ELRC workshop by thanking all the participants for their attendance. He also informed the audience that the folders that were handed out contained both information on the workshop and feedback forms that participants should fill out and return at the end of the day.

This second workshop was presented as a follow-up of the first one held in 2016 in the same room, and Paulo Batista reasserted the workshop's specific objectives namely the collection of Language Resources for Machine Translation within the framework of CEF and ELRC and the involvement of Digital Services Infrastructures at a national level.

Paulo Batista then gave the floor to Paulo Mauritti, member of the Board of the Administrative Modernization Agency (AMA).

On behalf of AMA, **Paulo Mauritti** began by expressing his gratitude to the Representation of the European Commission in Portugal for hosting the workshop, and then welcomed all participants, including the representatives from AMA, thanking them for their presence. He stated that AMA is supporting the ELRC initiative, in the person of its representative Paulo Vale, the ELRC Public NAP for Portugal, and that Portuguese, one of the most widely spoken languages in the world, is an area of concern for AMA. He acknowledged that technology can support languages in the context of the digital world, eTranslation being one of these tools, stressing the need to encourage the engagement and to raise awareness for linguistic data sharing as the more these are shared, the better the language technology tools will be. The objective is twofold: to support citizens by facilitating translations, not replacing the translators, and to set up a national repository of digital services available for all the services of the public administration. To conclude, Paulo Mauritti stated that everyone's commitment and motivation is needed and will be much appreciated

Paulo Batista took the floor again to thank Paulo Mauritti for his introduction, for AMA's endorsement of the workshop and launched the first session.

3.2 Connecting public services across Europe: ambition and results so far

A video presentation from **Aleksandra Wesolowska** (DG CONNECT) was played for the audience with live interpretation into Portuguese.

3.3 National initiatives for digital public services and (open) data

André Lapa, Manager of the Open Data portal project (dados.gov) from AMA, started his presentation by saying he was going to give a national perspective on Open Data and open databases. He explained that one of AMA´s aim was to simplify administrative management quoting the *Portal do Cidadão*, the SIMPLEX program or the e-government, as examples. AMA´s main objective is to support an open government, achieved through participation, collaboration and transparency. The open data definition means that anyone can use these data and do whatever they want with them, which leads to the development of open software, the availability of open sources and the sharing of knowledge. The State shares reusable data under the 2013/37/U Public Sector Information (PSI) directive which has been transposed into Portuguese legislation in 2016 by LADA, *Lei de Acesso aos Documentos Administrativos* (Lei n.º 26/2016 de 22 de Agosto). This law is even more relevant for environmental data. However, the State has not opened all data, privacy can remain, even if data can be open.

According to the G8 Open Data Charter which recognises areas of high value, both for improving democracies and encouraging innovative re-use of data, there is a fair amount of information that should be accessible without the need for permission. Already a good number of applications use open data, which not only creates value, but also brings transparency and innovation. One can refer to initiatives such as the Pordata portal, where detailed statistics help draw a portrait of Portugal through a range of statistics in various areas, even if not using open data.

André Lapa showed several examples of existing local and sectorial initiatives, such as the SNS, the justiça.gov portals or the Lisbon Municipality website, whose independent existence is important. Such initiatives need to be harmonized and this is where the dados.gov page plays a role. It works as a hub, making it easier for the user to locate the relevant catalogue(s) accessible from this central portal. The dados.gov page was re-launched in May 2018 as dados.gov+, the improved version of the National Open Data portal. At the time of the workshop, the available resources are: 1690 datasets, 4297 resources, 4 reuses, 91 users, 60 organizations and 1 discussion. The Portuguese portal was developed based on the structure of the French administration portal (ETALAB) built in open-source, developed with the use of collaborative tools and open upload of public interest data. André Lapa finished his presentation stating that one of the portal´s objectives is to harmonize data at the European level.

3.4 CEF in Portugal: an outlook into current and future challenges – Panel session

Paulo Vale started by introducing the two panellists: **Danilo Furtado,** Head of the Geographic Information Division, Directorate-General for the Territorial Development and **Isabel Baía,** Head of ICT Projects and Policies Unit at AMA. Then, highlighting the importance of the CEF financing mechanism available to interconnect cross-borders digital services in Europe, he gave a brief overview of the five CEF building blocks, and provided statistics on the CEF Telecom grants financing 18 projects in Portugal which cover 3 building blocks (eID & eSignature, eInvoicing and eTranslation) and 8 specific sector DSIs between 2014 and 2017. With respect to Open Data, Portugal is participating in two projects: the Open Waste Compliance, with the Portuguese Environmental Agency and the Cross-Nature project, with the Directorate-General for the Territorial Development (DGT). As for the eTranslation building block, Portugal is also actively involved in the ELRI project with a joint participation from AMA, the Faculty of Sciences of the University of Lisbon and Linkare TI.

Danilo Furtado started by thanking AMA for their invitation for the workshop and then introduced his institution, the Directorate-General for the Territorial Development, the national agency responsible for public policies on spatial and urban planning, production of cartography, creation and management of the National Spatial Data Infrastructure. As part of its research activities and within the INSPIRE[1] framework, DGT has taken part in a FP7 SmartOpenData project whose aim was to link environmental and geospatial data to improve biodiversity protection while making resources available. Danilo Furtado concluded his presentation by noting that the INSPIRE Directive was making multilingualism possible, even if not yet implemented in Portugal.

Isabel Baía mentioned that Portugal counts several digital services, like the electronic notification and the single digital address, but doesn't have any multilingual digital service yet. She noted that services like the Entrepreneur Desk should be made multilingual as it doesn´t make sense to propose such a

[1] The Inspire Directive aims to create a European spatial data infrastructure, which will enable the sharing of environmental spatial information among public sector organizations and better facilitate public access to spatial information across Europe.

service only in Portuguese in a country with so many foreigners (and entrepreneurs are not necessarily locals).

Paulo Vale stated that the contributions are related to the work areas of the public administration and then asked Danilo Furtado if he sees any advantages in the project. Danilo Furtado answered that there are some advantages for the services for which only the main context of the translated documents is necessary, but in other cases, more technical summaries and descriptions can be complicated to translate. However, he reasserted that the challenge of the Inspire Directive was to lead to multilingual data.

Paulo Vale then turned to André Lapa regarding the dados.gov platform and asked him how willing the public administrations were to supply data. AMA has been talking with the public administrations and **André Lapa** said that despite some cautious reactions from some services, overall the initiative had been perceived as positive. The biggest barrier to sharing data seems to be related to the availability of skilled staff (computer scientist) who can actually handle the data upload on the CEF platform. The Directorate-General for the Territorial Development is used to sharing data and there are calls that can be used to answer internal needs and open data, including the connection to the European portal.

Paulo Vale highlighted the main financial issues in relation to preparing the data for eTranslation to implementing process.

Isabel Baía gave an example: AMA has received a request for an academic paper in relation with the datasets from the first SIMPLEX programme, the Portuguese administration simplification programme. Accessing the datasets was not an issue, however, the datasets were in Portuguese and therefore needed to be translated, especially now that the country wants to expand the SIMPLEX programme to other countries (Morocco, Egypt).

Ana Sommer, from the National Company Registry (RNPC), informed the participants that an English version of the commercial registry exists since 2008. She then detailed the process and insisted that the translated output needs to be gone through with a fine-tooth comb. For what concerns their data, the generic information is automatically translated using Systran, the translation engine they chose, but the specific information must always be reviewed, in particular when it comes to what should or should not be translated (for instance, person and street names, among others, should remain in the source language). For accuracy purposes, this can only be performed by human translators.

Paulo Vale reiterates the message that the idea is to improve and support the translators' work, not to replace them or jeopardize their job.

Renata Margarido, from the Ministry of Justice, enquired about the next steps. According to her, a compelling message on eTranslation needs to be conveyed to convince public services to share data and use the eTranslation platform. Currently, a lot of money is spent on translations by courts and potential mutualisation is hindered by the lack of information flow among the entities that have competences. There is a lot of willingness to participate, but impediments remain. There have been some success with European regulations, like the General Data Protection Regulation (GDPR), but just good will is not enough.

Paulo Batista thanked all the participants for their input and called for the coffee-break.

3.5 The CEF eTranslation platform @ work

Michael Jellinghaus, Machine Translation expert at the DGT and eTranslation Project manager gave some background on automated translation at the EC. MT@EC, the EC statistical machine translation system, was launched in 2013 and mostly based on the EU legislation texts (Euramis database). The system then evolved to what is now eTranslation, the neural MT system. With the CEF.AT, the purpose is to set up a secure service on the European cloud infrastructure. But how can we use it? The platform is available for public administrations, system suppliers, Digital Service Infrastructures and EU institutions.

He continued his presentation with a quick demo on how the platform operates. He showed how to upload the source material and then receive the translated output by e-mail, with the same formatting as in the input document. He added that one source text or document can be translated into all 24 EU languages and that several documents can be sent simultaneously. The user can request the translation of a full document without affecting its formatting. Also, the translated information remains confidential, and all document processed are deleted from the server, meaning that the intellectual property rights are not transferred to a third party over the translation. He went on with the e-Translation on the N-Lex portal showing it was possible to look for words in the original language rather than look for translated words in a translated document. The same is also possible on the European Data Portal, making cross-language information research easier. He stressed that eTranslation was open to DSIs and public administrations in all the Member states, plus Iceland and Norway. All interested institutions can contact the DGT and ask for the access credentials.

Next, Michael Jellinghaus emphasized the importance of the language resources in the performance of a Machine Translation system. eTranslation, unlike MT@EC, is a Neural Machine Translation (NMT) system. Since the machine learns differently, it can be used for translating texts from a wider domain with better results than statistical systems. At the moment, not all languages are covered by eTranslation and Portuguese for instance will be available at the end of the summer 2018. He then showed some examples of news translated using SMT and NMT and highlighted the differences. Even with out-of-domain data, the NMT system gave better results. When using in-domain data such as legislation, the translation document can be used almost with no further corrections.

Michael Jellinghaus added that key to success is to gather more data. By data he means any electronic text, both monolingual (Portuguese) and parallel (Portuguese and English, for instance); more training data and more domain-specific training data. He encouraged all those present to share any data that they might have. He then presented some future improvements that result from users' requests, namely more languages and ended with the idea that the CEF eTranslation platform goal was to solve language problems.

When asked by the representative of the Foreign Affairs ministry about the volume of data needed, Michael Jellinghaus answered “the more the better”, specifying however that the bare minimum would be 100K sentence pairs. He also made clear that eTranslation is available for all 24 languages, except Gaelic (by derogation for Ireland) and Croatian. Another participant asked whether the quality of the output could instantly be improved by adding more resources for certain languages. Michael Jellinghaus answered that by default the engines are renewed twice a year (which means including new data sets), but for specific demands, it can be done within an average 2-month time.

Paulo Batista ended the session with the reminder that sharing creates value.

3.6 The European Language Resource Coordination (ELRC) action

Khalid Choukri introduced the ELRC initiative, starting with the four consortium organizing partners: DFKI, ELDA, ILSP and TILDE, supported by National Anchor Points (NAPs): the Technical NAPs – António Branco in Portugal – and the Public Services NAPs – Paulo Vale in Portugal – and the DGT Field Officers. In total, there are about 60-65 people that represent the EU Member states, plus Norway and Iceland.

“What does ELRC do?” To this question, Khalid Choukri answered that the action is organized around 5 main actions: **Collect** (Language Resources), **Identify** (Needs of public services), **Engage** (With the public sector in the identification of Language Resources), **Help** (With any technical or legal issues) and **Act** (Observatory for Language Resources across Europe). The central action is to identify the multilingual needs of the public services with the aim of setting up a continuous collaboration between the European Commission services across EU Member states. He also explained that, to support the Help action, a Helpdesk has been set up and deals with all technical and legal issues.

The answer to the next question “why ELRC?” was: to facilitate multilingual cross-border communication and exchange of information in key public service scenarios. For instance, the Online Dispute Resolution has to be multilingual. To make MT work there is a need for in-domain text, properly translated, in order to improve the quality of the eTranslation outputs. As far as Portuguese data already collected are concerned, there are not many: 12 corpora, 1 lexicon and 1 tool, so additional efforts should be done to increase this number. Other languages such as Dutch and French have low scores in terms of collected Language Resources.

Khalid Choukri showed how to access the LR-SHARE repository and ended his presentation by reminding that there is the ELRC Technical and Legal Helpdesk (http://www.lr-coordination.eu/helpdesk) available for anyone who has data and needs help.

3.7 ELRC in Portugal

António Branco, from the Lisbon University and ELRC Technology NAP in Portugal, started by presenting the other colleagues relating to the Portuguese Language: Paulo Vale and Paulo Batista. In 2014 he supported the winning submission (ELRC) to the call for proposals and in 2016 the first ELRC workshop was held exactly in the same place of this second workshop, which now has a few of the same faces as in the first workshop and many new ones, which was considered a good sign. The first workshop had a different agenda: the objective was to do some networking and set the challenge on data sharing, to be embraced by AMA.

From the currently publicly available 266 resources on the ELRC repository, 166 are in English and 8 in Portuguese, most of them in the field of law. If we consider all the contributions so far (455 in total), there are 377 in English and 12 in Portuguese. António Branco took the opportunity of the low numbers for Portuguese to remind the audience that all contributions are acceptable: monolingual data, texts, and lexicons, among others.

Then, António Branco presented the ELRC twin-project ELRI (European Language Resource Infrastructure), with 7 participants and a budget of 2,4 million Euros. ELRI’s objective is to validate and prepare data to train Machine Translation systems. The delivered data is used in translation memories that can be used locally, so it´s a win/-win situation that can connect with the ELRC project. The ELRI consortium, that includes 3 Portuguese partners (AMA, Linkare and the Faculty of Science of the University of Lisbon) and for which there are plans to expand to other countries, was set up on the basis of the feedback from the first ELRC workshop participants revolving around the key questions: “How to obtain the authorization to share data?” and “To whom can the data be confidently delivered”. The ELRI project aims at responding to those questions by creating share points – the national relay stations - that serve as interface with ELRC/eTranslation. Two years have passed since the first round of workshops. What was achieved during that period of time? António Branco answered by showing 2 pictures: one of a sports convertible car with a “for sale” sign and one of the convertible car with someone behind the wheel. Then he explained the meaning of the images: 2 years ago this was an attractive business for sale and now, 2 years passed, AMA is behind the wheel strongly committed to the initiative and everything is moving forward. Automatic translation on the public administration is a challenge that is being faced by AMA and there is a great need for this service.

3.8 Preparing and sharing data with the ELRC repository

Khalid Choukri made a presentation illustrating how to share language data. He showed an example of a shared resource, so that everybody can have some notion of what kind of data he was talking about. He emphasized the need for language data from the EU Member states, since there is already a connection with the EU itself and all its institutions and bodies. He gave examples of types of data that are useful for eTranslation, also pointing the accepted and preferred data formats. He also gave some instructions on the data preparations, listing the Do’s and the Don’ts. For instance, source and translated texts should be submitted in separate files, with similar file names so that they can be easily identified. Criteria for grouping the data and the preferred domains were also presented and then a step-by-step guide was showed to demonstrate how to upload data to CEF eTranslation. As far as the contribution mode, Portugal is a key player in eDelivery, but it is not expected that this is the chosen way to deliver data. Khalid Choukri then took some time to describe the On-site Assistance services which provide support to data cleaning, anonymization, format conversion, among other processes, all of them offered on-site and free of charge. It is important to remind that the processed data are always returned to the data contributor.

He then showed how to request assistance on the website and ended his presentation with some key messages: data is important and we need data, ELRC can support you in the data sharing process, there is a budget to be used for this situation: we have on-site services for support, we can clear data for you to re-use after and we can open data for you.

Finally, Khalid Choukri considerately answered some questions from the floor regarding the resource articulation between the EC and the repository and the interest over NATO/military data.

3.9 Can language data be shared and how?

André Lapa, from AMA, although not a law expert, gave an overview of the European legislation on the reuse of Public Sector Information (PSI) and explained that the Directive was reviewed in 2013 (2013/37/UE) and transposed in 2016 to the Portuguese Law of Access to Administrative Documents (LADA). He stated that PSI focusses on data reuse but does not specify what types of data should be open aside from the environmental ones. He also talked about the cases where the LADA does not apply and illustrated it with a simple example. How can we identify data than can be open or not when it comes to people identification? Sharing data at the parish level does not seem problematic, but in a parish with only 30 or 40 people, data sharing can lead to an easy identification.

Regarding data sharing in practice, André Lapa pointed that it all comes down to open licenses. The Creative Commons Attribution 4.0 (CC BY 4.0) is the license used by default and the only obligation attached to it is to give appropriate credit when the data is reused. On the other hand, with the Creative Commons Zero (CC 0), the "no copyright reserved", the data can be used without asking for any permission, which is equivalent to putting data in public domain. This license type is the one used by Lisbon municipality and it should be the one used more often. However, there are always safeguard limits, so in case of doubt, the participants were advised to contact AMA for help. He ended his part of the presentation by saying that a new path is being treaded and there are countries that have created some new and more specific licenses. He then gave the floor to Khalid Choukri to talk about the Legal Framework for Contributing Data.

Khalid Choukri gave a brief overview of the structure of rights, starting with PSI rules, then focusing on the key messages: there are experts in the ELRC consortium who can help; in PSI everybody has to be involved; the shared datasets are used to train the neural networks, but selected parts of the data are used, not long sections of text. He then described the 5 steps necessary for releasing data and ended with another reminder to use helpdesk whenever needed.

3.10 Identifying and managing your data: Questions & Answers

Paulo Batista thanked again everybody for their presence and started the Q&A part of the session. **Paulo Marrecas Ferreira** raised a question about the CC 0 license. He mentioned that quoting is mandatory on academic papers and it is not the first time university teachers have problems with missing quotes on dissertations with the added point that the digital format can make quotations even more difficult. André Lapa said that the CC 0 license is still recent and is associated with data from the Lisbon Municipal Council so he didn´t had an answer to give. He then asked if assignment is a sine qua non condition to re-use the information. He reminded that assignment can be more reassuring to the user, but there is an open data portal and whoever uploads the data must be able to choose the license type to be used. However, this is a discussion that is still open. Paulo Vale added that when it comes to language resources, the purpose is not to share the text, the texts help in the translation process. In practical terms, how do you make a citation operational in these situations? He stated that this is a very interesting discussion, one to be maintained at all times, but the main point for the workshop was that the institutions are already used to share, but what is wanted is an increase of this sharing.

Conceição Henriques, from the Foreigners and Border Service (SEF), stated that this is an important issue for SEF, that this institution really feels the need for sharing. She added that SEF owns an enlarged and specific linguistic corpus and that it can contribute and further asked if AMA is open to a SEF contribution. She then mentioned the existing funds for the creation of multilingual platforms and asked if this was only for European languages. She stated that there are a lot of relevant languages regarding immigration in Portugal; large language groups like Chinese, Russian, Ukrainian and African languages, but those are very fragmented. Paulo Vale answered that AMA is present in a lot of events and is committed in and actively searching for relevant data. Although all data is important, those are really important and AMA is available to collaborate with SEF and is going to actively look for contacts in the scope of AMA´s projects. He reinforced that without everybody's contribution, the work is meaningless. The technology is already available, what it is necessary is to mobilize the public administration. ELRC is more focused on European languages, but the digital services provide a whole wide world opening. The idea is to start with the languages already available and leave the door open for other languages. Khalid Choukri complemented this by stating that CEF is the acronym for Connecting Europe Facility, so the scope is more on Europe, but the issue of expanding to new languages is already present and being taken to the EU. The hope is that there is an opening for non-European languages in this program or another program, but for the moment, officially, the focus is only on EU languages. Michael Jellinghaus stated that it is important to define the existing needs. There is already data for Chinese, Russian and Arabic and translations can be established for these languages, but first the language needs must be identified. Khalid Choukri took the opportunity to remind people that everybody has the opportunity to complete a survey regarding languages and technologies people are actually using. The forms are distributed inside the participants' folders. Paulo Batista wrapped up this discussion saying that migration leads to new language requirements and the need to adaptation from the part of the Foreigners and Border Service (SEF), Social Security (SS) and the Ministry of Justice (MJ), so the available resources can be identified and the helpdesk contacted so that a contribution to ELRC can be made.

Renata Margarido from the Ministry of Justice said that it is difficult for the entities to acknowledge what they have. Internal initiatives tend to be more difficult, it is easier to reply to an external invitation to participate. Besides, technology is expensive. Paulo Batista stated that a technological approach should be made before reaching the resources, but Renata Margarido opposed that resources are more restrictive. A transversal approach initiated by the government can lead to better results than just wait for singular entities initiatives. Reports can be made to decision-making positions, there can be interest on the initiative, but other internal priorities remain. Paulo Vale then asked: "What is the best strategy to approach the public administration? Is it best to enter via international affairs or is it best to enter via translators and then go up?" He said that the technological door has not been yet considered, that it is a new situation. He stated that the idea is to first "spread the word" to the different services using all the participants of the workshop. After that, AMA is going to actively get in contact with the different institutions, trying to get in by the top and then reaching the resources. He added that what was important was to use the eTranslation tool and assess whether it is useful. He asked the participants if they knew that this tool existed and that it was free for everybody inside the public administration. He gave the answer himself stating that for many of the presents this was the first contact with the platform, so he encouraged everybody to test it with official documents and see if it meets their expectations. He added that, in the future, the quality was going to improve with the increase of data contributions to the platform, so he encouraged the potential users not to abandon the platform if first the results were not satisfactory. We said that the platform could always be used as a complement and that it was going to evolve. In the end of summer there is going to be an evolution on the Portuguese language, so he made another appeal for the platform use, because the EC wants and needs potential users' feedback. He ended by saying that there is a lot more to be done in this area, Portuguese is the 4th most spoken language in the world, but according to António Branco´s presentation, there are only 12 Portuguese contributions on the platform.

Graça Tomé from the Portuguese National Laboratory of Civil Engineering (LNEC) said that she had already tested the platform, but that for her specific area of work, it was very poor. She added that there are many civil engineering institutions interested in the tool, but that there are many copyright limitations for the documents that can be shared. The institution (LNEC) has, however, terminology data, descriptors and repository that can be shared. Khalid Choukri stated that if there is interest, this information should be shared.

Paulo Batista ended the session by summarizing that there are available resources to be shared and Michael Jellinghaus reminded that without training data, good results cannot be accomplished, so supplying terminology to the ELRC was a good starting point.

3.11 Discussion and Conclusions

Khalid Choukri thanked all the other speakers for their presence in the workshop and also all the participants for their attendance. He also acknowledged the work of the interpreters throughout the day and asked for a round of applause.

António Branco started by saying that he could guess what was on everybody's mind: "If I am from the Lisbon University, what am I doing here?" He explained that the focus of his research was natural language, and his mission and passion are to develop technology using Portuguese language. At the moment, the ELRC-Share platform only counts 8 Portuguese resources, as opposed to hundreds available for other languages. He reminded that there is a need to be competitive and that his job as a scientist/researcher is to contribute to the development of Portuguese language in technologies, even if the sovereignty of the national language relies solely on the public administration. He ended with an appeal on the importance of helping to bring Portuguese on to the global digital society.

Paulo Vale called the meeting to an end thanking everybody for their presence and reminding that he is available for any contact, networking or idea sharing. The audience was reminded to fill out the feedback and engagement forms and that certificates of attendance were available.

4 Synthesis of Workshop Discussions

In general, all the participants were well impressed by the project, the eTranslation platform and all its potentialities. The presentation of the eTranslation platform was well received by the participants and triggered a lot of interest. The need for data was seen as the fundamental point, but the participants expressed their concerns on the actual data sharing through many questions: "Is a relevant dataset shareable and how?", "Who is entitled to authorize the sharing?". Some of the other concerns pertain to how the information is passed on and how it reaches the higher authorities in a way that this issue is seen as important and can become part of the priorities. There was also some interest regarding the helpdesk and the possibility of financial help from the EU in the process of data sharing. The workshop raised the awareness at the level of the persons attending, but there is a need to reach upper levels inside the institutions to unlock the data sharing. In this regard, a lot of support can be expected from AMA.

4.1 ELRC and Open language Data in Portugal

- The PSI Directive (2003/98/EC) has been transposed in Portugal, since October 1st of 2016, by the LADA Law number 26/2016 (published on August 22nd). This law allows the new access regime to administrative and environmental information and reuse of administrative documents.
- https://dados.gov.pt/pt/ is the Portuguese Open Data portal, the "open data central catalogue in Portugal"
- The Portuguese Open Data portal has a significate number of available data, but the number of reuses is still small and the number of organizations represented is still very limited.
- The reuse license is one of the essential conditions to share open data and the one that is used by default and the one of recommended use at the dados.gov portal, is the Creative Commons Attribution 4.0 – CC BY 4.0

4.2 Success stories and lessons learnt

- The participants got to know the existence of the ELRC portal and helpdesk and of the eTranslation, that was considered as important
- The word should be spread and reach the relevant people, since the flow of information on the Portuguese public administration may not always be easy
- The video presentation from Aleksandra Wesolowska was considered too technical and therefore not very easy to follow.
- Suggestions were made to increase Portuguese examples in the presentations and to open the initiative to public universities, where the translation platform could be very useful for research

4.3 Sessions questions

4.3.1 Questions from the session *National initiatives for digital public services and (open) data*

Paulo Vale from AMA asked, in a provocative manner, how many linguistic datasets are available among all the other data on the Open Data portal. André Lapa answered that he thinks that none is available. Paulo Vale said that it is important to connect subjects and encourage the sharing through all initiatives. Sharing creates value and if data is available, it can be shared and there are benefits for all. The ELRC gathers language data and provides services built using the data to citizens, companies and public administration. AMA undertakes the initiative and is creating a project in this area, but needs support for the Portuguese language.

Albertina Duarte from the Portuguese Translators Association asked who decides on the interest, usefulness or adequacy of the uploaded data on the Open Data Portal. André Lapa said that those who upload data are liable and that there are 2 agents in control: the community and AMA itself. There is no monitoring and no licencing policy is enforced which means that anybody can upload data. This is similar to the process implemented in France (ETALAB) which of course cannot prevent spam. AMA checks that no copyrighted data is uploaded but the responsibility remains with those who upload the data onto the portal.

4.3.2 Questions from the session *The CEF eTranslation platform @ work*

Paulo Vale asked Michael Jellinghaus to give the participants an idea of the processing time needed for the platform to translate one document. Michael answered that it depends on the amount of work on the platform, but generally speaking, translated outputs are sent back by e-mail within minutes and even when multiple languages are requested, there is not competition between them.

Paulo Marrecas Ferreira from the Documentation and Comparative Law Office of Portugal's Attorney General's Office asked if the process was automated. Michael answered that like with every machine translation, the output needs to be checked. Also, there is document confidentiality, which means that the uploaded documents and the translations are not seen by third-parties. Paulo Batista added that there is always the need for the human review after the result is received by e-mail. João Coelho from DGT stated that the automatic translation is based on memories made by human translators and that human translator will always be needed. Every tool gets “approved”, but the person who receives the translation is responsible for it. The MT complements the translation process; it is not a substitute to the translation process. Michael Jellinghaus gave an example: if there is a long document, the MT can give a first draft translation that shows the text parts that are relevant and give an indication of the parts the human translator needs to focus on.

When asked by Isabel Okatha, from the Ministry of Foreign Affairs, about the volume of data needed to train the machine, Michael Jellinghaus answered “the more the better”, specifying however that the bare minimum would be 100K sentence pairs, but the idea was to collect tens of millions for each of the 24 EU languages, which can easily be reached, except for Gaelic (by derogation for Ireland) and Croatian.

Paulo Batista said that the translation results are compatible with Computer-Assisted Translation tools and that protection takes time, which is why Google takes less time. He reminded that all it takes is to send an e-mail to get access to the eTranslation platform and that the idea is to shift the translators work from just translating to reviewing the translations and that the more data, the better the translations adaptation and Michael Jellinghaus complemented that by stating that all data is useful, even in small amounts.

Paulo Vale asked how long it takes for the uploaded data to have an impact on the translation quality. Michael Jellinghaus answered that by default the engines are renewed and re-trained twice a year (which means including new datasets), but for specific demands, it can be done within an average 2-month time.

4.3.3 Questions from the session *The European Language Resource Coordination (ELRC) action*

Paulo Marrecas Ferreira made preliminary comments on intellectual property and the evolution of the European Commission position in this matter. He asked whether ELRC initiative would really open the way to sharing (he referred to what India is doing on the domain of vaccines) or if it is only a way to facilitate the work of administrations. Khalid Choukri replied that we cannot do without the diversity of the European copyright legislations which the European Commission has tried to mitigate with the PSI directive, adding that Text and data mining copyright exception in the UK and Fair Use Act for Research in the US already permits the unlicensed use of some copyright-protected research works. Similar legal mechanisms are also implemented in Israel, Korea or Japan.

4.3.4 Questions from the session Preparing and sharing data with the ELRC repository

Paulo Marrecas Ferreira asked about the articulation of resources between the EU institutions and the repository. Khalid Choukri went back to the presentation and showed all the EU institutions and bodies that are connected to the DGT.

Fátima Cruz from the Ministry of National Defense/General Staff of the Armed Forces asked if there was a portal to help with translations. She stated that they work only internally but the institution is related to NATO. She informed the audience of a standardization conference on language learning which could be of interest (all presentations along with relevant content and links can be accessed from the Bureau for International Language Coordination (BILC) website). She ended stating that there is a lot of NATO/military documentation and asked whether it could be of interest for the ELRC initiative. Khalid Choukri answered positively, reasserting that all data from human translators from any EU country, Iceland and Norway can be of interest.

4.4 Workshop Presentations

The presentations are published on the Portuguese workshop agenda webpage (http://lr-coordination.eu/l2portugal_agenda)