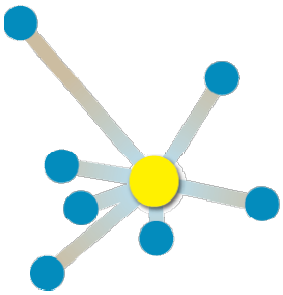
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.



European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.14
Task 8

ELRC Workshop Report for France

Author(s): Thibault Grouas, François Yvon, Hélène Mazo

Dissemination Level: Confidential

Version No.: V1

Date: 1.08.2019

Contents

1. Executive Summary

This document reports on the ELRC+2 Workshop in France, which took place in Paris, on June 26th, 2019 at the Ministry of Culture. It includes the agenda of the event (section 2) and briefly sums up the content of each presentation and of the panel workshop session (sections 3 & 4).

The ELRC+2 Workshop in France was attended by43 people, among which: 16 people came from Public Administration, 9 from academia and research institutions, 5 from DSIs and Open Data Portal, 2 from EC, and 11 from other organizations.

The dedicated event page can be found at http://lr-coordination.eu/l2france_agenda.

2. Workshop Agenda

09:30 – 10:00 Registration

10:00 – 10:20 **Welcome Addresses by the Ministry of Culture representatives**
Paul de Sinety, General Delegate for the French language and the languages of France
Alban de Nervaux, Head of the Legal and International Affairs Department

10:20 – 10:30 **Welcome Address by the European Commission**
Philippe Gelin, Head of multilingualism Sector, DGCONNECT, European Commission

10:30 – 10:45 **Language Technologies and European cooperation in the fields of Culture, Education and Research**
Antoine Cao, Director of the Digital Accessibility Program, Inter-ministerial Directorate for Digital and the State Information and Communication System
Pascal Brunet, Director, Relais Culture Europe

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

10:45 – 11:05 **Session 1.1 Connecting public services across Europe: ambitions and results so far**
Philippe Gelin, Head of multilingualism Sector, DGCONNECT, European Commission

11:05 –11:20 **Session 1.2 Public services in the era of Artificial Intelligence: vision and perspectives**
Victor Kahn, Policy Officer for Open Government, ETALAB

11:20 – 12:05 **Session 1.4 CEF in France: current needs, future challenges and good practices in French Public Administration – Round Table**
Modérateur: Thibault Grouas, Delegation for the French language and the languages of France

- Pascal Brunet, Culture Europe relay Director
- Téophile Fournier-Outters, Banque de France
- Nicolas Beckers, Language Department Manager, Arte
- Marion Dupuy, Communication Officer, OFPRA (French Office for the Protection of Refugees and Stateless Persons)

12:05 – 12:25 **Session 1.3 The CEF eTranslation platform @ work**
Michael Jellinghaus, Machine Translation Expert, DGT

Session 2. Engage: hands-on data

12:25 – 12:45 **Session 2.1 The European Language Resource Coordination (ELRC) action**
Khalid Choukri, ELDA CEO

12:45 – 14:00 *Lunch at the Ministry of Culture Cafeteria*

14:00 – 14:15 **Session 2.2 ELRC in France**
François Yvon, LIMSI-CNRS Director

14:15 – 14:45 **Session 2.3 Preparing and sharing data with the ELRC repository – and what happens next**
Khalid Choukri, ELDA CEO

14:45 – 15:05 **Session 2.4 Can Public Administration resources be shared and how?**
Mickaël Rigault, Legal Expert, ELDA

15:05 – 15:35 **Session 2.5 Identifying and managing your data: Questions & Answers**
Khalid Choukri, ELDA CEO and Victoria Arranz, R&D Head of Department

15:25 – 16:00 **Discussion and Conclusions**
16:00 – 16:45 *Coffee Break & Networking*

3. Summary of Content of Sessions

3.1. Welcome and Workshop Objectives

Welcome speeches and presentation of the workshop objectives are the object of this first session, introduced by Khalid Choukri.

The welcome speeches are given by the following participants.

Paul de Sinety, General Delegate for the French language and the languages of France, Ministry of Culture

Paul de Sinety opens the workshop by welcoming all the participants in the name of Mr Franck Riester, Minister of Culture. He then recalls the missions of the DGLF-LF, namely the coordination of the French Government's language policy, and plays a prominent role in the implementation of the plan of the President Macron: "An ambition for the French language and multilingualism" presented on 20 March 2018 at the Académie française, which encompasses measures including the teaching of two European languages in addition to the mother tongue, language training in European and international institutions.
Both digital as a tool for multilingualism and translation are strong focus in the Government's strategy. On the DGLF side, the topics Language & Digital and French Language Use & Dissemination are prevalent and involve DGLF-LF executives such as Paul Petit, Thibault Grouas and Claire-Lyse Chambon.
In 2019, a call for projects on Language & Digital will be issued, and the project of a Francophone Dictionary, Laboratory will be launched, as part the creation of an international residence for francophony located in Villers Cotterêts in 2022.
Finally, the French EU Presidency in 2022 will highlight digital technology for multilingualism and translation. After reasserting the main topics of the 2nd ELRC workshop in France, i.e. the sharing and collection of LRs for eTranslation and the dissemination of information towards the administrations in France, Paul de Sinety gives the floor to Alban de Nervaux.

Alban de Nervaux, Head of the Legal and International Affairs Department

Alban de Nervaux welcomes all the participants and states that he is impressed by the expertise of the participants who are the actors of multilingualism in Europe. The department he is heading relays the priorities of the French Government in all European and international instances. Multilingualism and translation are strong topic in France with the Work plan Culture 2018-2022 focusing on multilingualism and translation, as well as in Europe with Europe Creative (and ARTE Europe) that provides support to all types of translation.
Alban de Nervaux is grateful to the European Commission and the EU member States for these initiatives in favour of multilingualism.
As a conclusion, he reminds the audience that, from a legal point of view, the Legal and International Affairs Department has been expressing France's position during the authors' right debate. It was a very difficult negotiation compounded by power games, which however found its conclusion in an important victory for Europe, for Culture in Europe and the creators that the Ministry of Culture was defending. During the negotiation, the role of translation was crucial. All discussions took place in English and despite the high technicality of the discussions, no interim text was provided in other languages than English, which was a challenge for all the negotiating teams. In the future, Machine Translation could help getting interim texts in all the European languages to support better the negotiation work.
At the end of the 3-year negotiation period, it took an additional 3 months to the lawyer-linguists to define the right terminology with the services from the European Parliament. As an example, we can quote the creation/adoption of two emblematic terms (1) « téléversement » as the

translation of upload (as opposed to « téléchargement » for download) and (2) « proportionnel » as the translation of «proportionate » which solved a deliberate ambiguity. Both terms will be integrated in the French law.

Welcome by the European Commission

Philippe Gelin, Head of multilingualism Sector, DGCONNECT, European Commission

Philippe Gelin thanks the participants for coming, including those working in the Ministry of Culture. He praises the synergy between culture and economics, which he calls cross-fertilization of language and digital. According to him, regarding technology, we have reached a turning point. Machine translation quality has improved, even with post-editing.
Since he is giving a presentation during the next session, he keeps his welcome speech short and gives the floor to Antoine Cao.

Antoine Cao, Director of the Digital Accessibility Program, Inter-ministerial Directorate for Digital and the State Information and Communication System

The mission of Antoine Cao is to represent the French Authorities for many digital issues in a number of committees including CEF, but also ISA2 for interconnecting tools, eGovernment Action Plan steering Board, (Digital Europe). He is also the national coordinator for a large European programme: the Single Digital Gateway. This portal makes information available to every European citizen on the rights, procedures, they need to complete before they can move to another EU country to study, settle a business, work. The portal will make use of eTranslation so that the information is available to all in their own language.
Digital Europe is also looking into integrating the eTranslation tool as well as other DSIs and adopt an AI approach.

Pascal Brunet, Director, Relais Culture Europe

What triggers the interest in multilingualism? Multiple challenges! There are technical challenges attached to new tools, but the use and more specifically the change in access to cultural production is a matter of concern. 30% per year is the figure representing the growth in social media use on mobile phones. Ethics is also challenging: we can think of the changes in relation between the public and private spheres (including intellectual property), the evolution of civil society (unlock data or being « unlocked » by the data). From an economic perspective, new ideas need to arise and cooperative ideas must be invented since we will never have Netflix's capital. Culture is also at stake, in particular the preservation of diversity. What will the Europeans do? Finally, the evolution of artistic production (digital creativity) and accessibility will challenge the artists.
Multilingualism issues arise differently in France, which is a monolingual country, than in Spain. However, there is a strong creative ability in France which is the first country to mobilize Europe Creative. When it comes to education-culture ratio, France is less efficient. But we can expect that the Erasmus programme will evolve and become more massive.
To conclude, we can consider that research programmes are the places where we must look at the evolution of uses more closely but also at possible cooperation.

3.2. Session 1.1: Connecting public services across Europe: ambition and results so far

Philippe Gelin, Head of multilingualism Sector, DGCONNECT, European Commission:

Philippe Gelin gives a presentation of eTranslation at glance, comparing the interface with that of Google Translate: the user copies the text on the left and the output is displayed on the right. eTranslation is available for all European Member States languages, plus Norwegian and Icelandic which take part in CEF. The machine-machine access is also available through an API.
About 8M€ were invested to modify eTranslation in a 2-year time. Some portals used the system automatically. Since about 6 months, we have witnessed an explosion of the use of the

eTranslation tool going from 41M pages in 2018 to 18M pages as of the first quarter of 2019. It is interesting the mention that 1M pages are translated by external users from the Member States (not belonging to Europe institutions).
Currently, over 800K pages (875,773) are translated per day.
The other important factor for a successful MT platform is Language Resources. Good quality data are needed. Beyond translation, some fields are fast evolving and according to a WIPO report, Language Analysis and Speech are considered to be the 2 key technologies for Artificial Intelligence. Europe focuses a lot on AI, but Asia is progressing fast.
Language Technology is at the crossroad of the technology development involving Artificial Intelligence, Language Resource and High Performance Computing. SMT to NMT eTranslation transition increased quickly between 2015 and 2019.
Philippe Gelin continues his presentation and gives an overview of the new programs for research & innovation from 2021:

- Horizon Europe: 2 actions: 1) Cultural Heritage: digital preservation of languages and 2) Next generation internet:
- Digital Europe Programme: access to data, computing power, algorithms, localisation and uptake of language technologies

In addition, the new project ELG which has already started will provide an easier access to Language Technologies.

Philippe Gelin then calls for questions or comments.

Marie-Louise Desfray from the SFT, worries about the authors' rights on the translation, especially for freelance translators and wonders where the data come from. Philippe Gelin answers that the authors' rights are respected, but of course it should be clarified with the administrations (translation contracts). He specifies that for Europe, the Commission already holds a large amount of data.

Daniel Henkel from Paris 8 asks whether there is difference in the output if the user submits a text sentence by sentence or the full document. What control is implemented? Is there a mention added when a text is machine translated as like when a photo is "photoshopped".

Philipe Gelin answers that the tool is providing a first draft version which cannot be considered as final. The fluidity has been improved with NMT engines, but errors are still there.

Most translation services specify that the production is automatic so that everyone knows that it is automatic translation.

Nathalie Kübler from the University de Paris asks whether the system's objective is to cover all fields, if so there is a need for specialised data.

Philippe Gelin answers that eTranslation is based on the legalese corpus. Performance are good on terminology. If you want a specific domain, you need to have training data on this specific domain. The new trend is that NMT engines need less data but more quality data.

Leonor Agüera-Jacquemet, JIGGSAW, wants to know how the system manages variants of languages (US English vs UK English), because it may cause problems from a legal perspective.

Philippe Gelin answers that the tool is a support to translation only and does not replace human translators. He adds that language is alive and evolves and gives an example of a complaint from a Bulgarian citizen complaining that the Iphone he bought for his father did not carry a Bulgarian interface.

Joseph Mariani, from the LIMSI CNRS asks clarification on the participation of Norway and Iceland that are not EC members, and he wonders if such partnership can be extended to regional languages?

Philippe Gelin answers that Norway and Iceland take part in the ELRC programme and see their interest, provided many databases which enabled to provide adapted translation system. For other languages, there is a need to have bilateral agreement with EC with the corresponding budget allocated to this. Regarding the intra-extra community question, there are languages that have European impact whether industrial or economic The European Commission is looking into languages with a strong economic impact such as Chinese, Russian, Japanese or Korean to build translation systems. Non EU indigenous languages with a strong social impact such as Arabic, Turkish or Ukrainian, as well as European indigenous languages such as Corsican or Bavarian could also be considered for an extension of eTranslation.

This will depend on the on the technology, finances and languages.

3.3. Session 1.2: ELRC and Open language data in France

Victor Kahn, in charge of Open Government at Etalab, gives an overview of Etalab. Under the authority of the Prime Minister, Etalab is the government's task force for open data and data policy. The agency brings support to all French public administrations to facilitate the publication and re-use of public information. It is also responsible for the open data platform data.gouv.fr. Etalab's mission is threefold:

1. coordination: accompany administration to open access to their data
2. openness: organize events to encourage administrations and co-build
3. expertise: promote data science

Victor Kahn reminds the audience of the legal framework (Digital Republic Bill 2016) voted to foster data openness, then goes into further details on how the data.gouv.fr platform operates. The platform is open to all (administration, firms, citizens...) to deposit data sets and resources open by default. Since its launch, there have been over 5 million visitors. Research by (culture, agriculture, international...), by keyword, by production date is available on the 30K+ datasets uploaded on the platform. He provides details on the datasets, explaining how they are different from resources: several datasets can be published on a same one page and different resources can be associated to one dataset. A stamp certifies the producers and/or providers of the dataset (it can be a public administration). Dataset reuse can be referenced in the page (who reused the data, what they did with it...). Also, an API was developed to allow easier access to users (SIREN, RNA, BAN), eg geotracking.

There is also a specific service called Data Public Service which covers 9 datasets forecast by law (higher quality and regular updates) which needs to be always available.

Khalid Choukri asks whether Etalab works with other international open data platforms. Victor Kahn answers that there are collaborations with international platforms to have common datasets and that Etalab takes the international mission into account.

3.4. Session 1.3: The CEF eTranslation platform @ work

Michael Jellinghaus, Machine Translation expert at the DGT and MT@EC, provides some background on automated translation at the EC. MT@EC, the EC statistical Machine Translation system launched in 2013, was mostly trained on EU legislation texts. In 2017, the system evolved into what is known today as eTranslation, a neural-based MT system with an improved web interface and possible integration to public services.

As shown by Philippe Gelin earlier, eTranslation is available either for individual use or machine-machine use. Users are translators and officials from EU institutions, Digital Services

Infrastructures, information systems services, civil servants from Members States public administrations.

A quick demo is given on how the platform operates, how to upload the source text and then receive the translated output by e-mail, with the same formatting as the input document. One source text or document can be translated into all 24 EU languages and several documents can be sent simultaneously. Translated information remains confidential, and all documents processed are deleted from the server (after 24 hours or after upload when option cancellation selected), meaning that the intellectual property rights are not transferred to a third party over the translation.

eTranslation can be integrated in online public services through an API. For example, RAPEX, the Rapid Alert System for dangerous non-food products, can translate its pages automatically in the various European languages. An icon indicates that the pages were machine-translated and that the EU does not take responsibility for translations' bad quality. The API is also deployed in the European Data Portal allowing cross-lingual information search: the interface is in English but results can be displayed in other languages. All institutions interested should send a mail to Helpdesk CEF-AT@ec.europa.eu to obtain API integrated in administration tool.

Next, Michael Jellinghaus emphasizes the importance of Language Resources in determining 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.

He then provides a comparative analysis of some news translated using SMT (such as MT@EC, based on MOSES) and NMT highlighting the differences. Even with out-of-domain data, the NMT system gives better results. When using in-domain data such as legislation, the translation document can be used almost with no further corrections. Michael Jellinghaus concludes his presentation by emphasizing that key to success is to gather more data. By data he means any electronic text, both monolingual and parallel; more training data and more domain-specific training data. He encourages all those present to share any data that they might have. He then presents 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.

He concludes his presentation by giving insights of the ongoing work at the DGT to improve the quality, integrate more formats and languages (Russian, Chinese, Arabic, Turkish, Japanese), open to SMEs, scientific publications and foster the integration in other tools.

There are some questions and remarks. Philippe Gelin points out that the email address should be corrected. Ms Desfray from the SFT asks whether (1) the integration for SMEs will mean that SMEs sell automatic translations and (2) the users will be aware of difference between automatic and human translation.
Mickael Jellinghaus acknowledges that the risk exists but the interest of SME is to see that quality is needed by the users. Philippe Gelin highlights that improved automatic translation has contributed to the evolution of Alibaba.

All interested institutions should send a mail to Helpdesk CEF-AT@ec.europa.eu to obtain API integrated in administration tool and ask for the access credentials.

3.5. Session 1.4: CEF in France: current needs, future challenges and good practices in French PA - Panel

This session is chaired by Thibault Grouas (Public Services National Anchor Point) who frames the discussions around the challenges posed by translating in the public services and using Machine translation systems. The uses, practices, needs and perspectives should also be covered.

Panellists address these topics in the following way:

Pascal Brunet, director of the Relais Culture Europe, introduces the panel with the Kitty AI video, by artist and researcher Pinar Yoldas. This video, featuring a cat which becomes a governor, serves as illustration of the usage of AI and LRs in the art creation field. Mr. Brunet emphasizes the importance of multilingualism in the framework of this artistic work. The artists behind this work have had to carry out a thorough analysis of data and then apply AI techniques.

Théophile Fournier-Outters, Digital Transformation Department, Banque de France, is responsible for the implementation of MT at the ACPE. He describes MT as a "crutch" for the team, that is, as an element in support of the team with the English language. However, if the texts have a legal component, the agency ACPE needs to resort to translators as no risk can be taken in this matter.

When referring to eTranslation, Théophile Fournier-Outters raises a "confidentiality problem", as the location of the files to translate is a major issue for them. The data (TMXs) need to be made reliable.

Last but not least, given that the content of the COMOFI, the French Monetary and Financial Code, is outdated with regard to the Brexit, insurance code, etc., Machine Translation cannot be up to date.

Nicolas Beckers, responsible of the Language Department, Arte

Translation is in ARTE's DNA: they have an in-house translation and interpreting team, working with the support of a network of experiences freelancers. Their linguistic team has tested MT in their workflow: while some domains benefit from MT, the output is not ideal for others. Nicolas Beckers explains that ARTE uses MT but the type of content (and objective) determines its use. Their teams must be very cautious when using external MT services for confidentiality reasons.

ARTE also makes use of Speech recognition (SR) with some post-editing. This is done with a European system that can guarantee confidentiality: there is no data sharing.

Marion Dupuy, communication officer, French Office for the Protection of Refugees and Stateless Persons (OFPRA)

The OFPRA also makes use of interpreting services. They are confronted with asylum requests from refugees and they require such services during the interviews, which are individual and confidential. They work with 127 languages, some of them extremely rare, forcing them to go through a pivot language to communicate. A crucial point for this Office is that of ensuring civil protection (for instance, in the documents that are produced).

They have a research center with speakers carrying out in-depth research. They collaborate with the OFPRA's counterparts in Europe.

The language and education level of the refugee asking for asylum is critical, that's why it is crucial to have a translator with them. Speech Recognition had been suggested, but their technical means is limited and do not allow this technology.

Simon Karleskind, Inter-ministerial Delegation for refugees' reception and integration (DIAIR)

Created in January 2018, this Delegation is placed under the authority of the Ministry of Internal Affairs. Their objective is to support refugees, centralizing information for them, in order to help them with all the administrative paperwork they need to go through (Employment agency, social security, welfare, education, etc.).

They have an information platform (AGI'R) with the list of mechanisms in the territory to help refugees.

Their relationship with language technologies can be described as follows:

- Regarding MT services, they make use of Google Translate and they provide its output to translators so as to either finish up, post-edit, translate what is missing. Then, they have the texts validated with regard to confidentiality.
- Regarding AI: they use Facebook's free algorithms.
  - They use speech synthesis.
  - They make data available to associations.

Following questions were asked at the end of the panel:

- What about getting help from voluntary translators?

DIAIR is collaborating with INALCO, the University for Languages and Civilizations, and their students whose work is validated and verified by professors supervising their work. At OFPRA, the level of complexity is too high for them to implement such a mechanism

- What about Facebook's App?

For DIAIR, it is useful as it provides figures on each pair of sentences probability, allowing prediction and pointing items to be checked.

A key point for the French administration is that of confidentiality.

3.6. Session 2.1 The European Language Resource Coordination (ELRC) action

Khalid Choukri, ELDA, presents the action of the European Language Resource Coordination and introduces the organizations that form the consortium: Tilde, ELDA, DFKI and ILSP. Then, Choukri describes the role of the National Anchor points (NAPs); in the case of France: Thibault Grouas represents the French public administration and François Yvon as Technical NAP.

Khalid Choukri then continues with a description of the goals of the ELRC actions: gathering Language Resources, identifying needs of the public sector and fostering the engagement of the public sector in the identification of language resources that the EC translation system can use to improve its engines. ELRC, he explains, is providing potential Language Resource providers with technical and legal support and is in fact an observatory of Language Resources, gathering information from workshops such as this one.

Khalid Choukri emphasizes that the action's core activity is to identify data in to the domain of public services to improve the EC translation engines. He also reports that more than 90 resources have already been released, accounting for more than two million translation units. These figures remain however below the expectations, in particular in the case of the number of resources gathered for French.

Khalid Choukri ends his talk by underlying again the importance of eTranslation to manage the actual needs for multilingual communication and exchange of information in Europe on the one

hand, and the important role of sharing Language Resources on the other hand. He then shows the ELRC web site, and the ELRC-Share repository to facilitate the access, sharing and contribution of Language Resources, as well the contact forms and help-desk for requesting technical and legal assistance for those interested in contributing with resources.

3.7. Session 2.2: ELRC in French

François Yvon, Director of the LIMSI-CNRS and ELRC P-NAP, gives an overview of the situation in France. He starts by reminding the participants that language technologies, including machine translation, dialogs, information retrieval, have evolved tremendously over the past few years. The current state-of-the-art approach for machine translation is now based on deep learning algorithms and the use large annotated corpora.

HPC (High Performance Computing) and powerful machines are required. Language resources are valued and their discovery is easier. Within ELRC, all Language Resources collected and their associated metadata (usage conditions, data description, etc.) are stored in the repository ELRC-Share.

At the moment, the number of Language Resources (10) submitted by institutions in France and stored on the ELRC-Share repository is very low compared to other countries (up to 200). However, for French, about 80 corpora and 95 lexical resources, whether monolingual or multilingual, are available in the ELRC-Share repository. This can explained by the fact that French appears a lot in multilingual resources collected by other Member states (Germany, Spain, Sweden, etc.). There are few monolingual resources although they are very useful in Machine Translation development. All resources collected during the initiative are stored in the ELRC-Share repository and documented with metadata which allows search by corpus type, language pair, IPR, etc.

François Yvon continues his presentation by summarizing the contributions of the main providers in France: ANR (National Research Agency), Ministry of Finances, other foreign institutions. Other sources are expected among the participants to the workshop.

He then concludes by recalling the main obstacles to Language Resources collection: LRs preparation and upload, legal issues, access to translations (no access for technical reasons), lack of awareness or interest, reluctance to use MT systems or contribute to their development. The main outcome is definitely some very voluminous contributions, from all over Europe, and the upload/access issues that are now solved.

There are some questions:

Joseph Mariani asks for clarifications on the country contributions by Latvia and Greece, which are very high. Khalid Choukri replies that Tilde, the Latvian consortium partner, has a large team involved in Microsoft Office+ translation in addition to having developed the Presidency Translator (MT system) and that they brought many resources. For Greece, contributions were strongly encouraged by the Deputy Minister of Telecommunication which resulted in the collection of many LRs.

Victoria Arranz underlines that each contributing country does not necessarily reflect language as the resources are distributed per region over 4 partners. Some countries, Ireland, for instance, are more enthusiastic than others.

Pascale Elbaz, from ISIT, suggests to solicit translations from all the translation schools and universities labelled EMT. She wonders what domains and type of texts are targeted by ELRC. Collecting for the public administrations seems very wide and vague to her. Then she adds that a lot of linguistic data are available in Switzerland.

François Yvon replies that many parallel corpora are available, but we need texts that pertain to the activities of specific public administrations. Financial reports for example are relevant for the tax administration. He adds that Switzerland is not part of the CEF programme.

Leonor Agüera Jaquemet, from JIGGSAW, worries about the data quality and requests more details on the quality assessment, arguing that legal translations are not necessarily checked by legal experts.

François Yvon goes through the process again and explains that the translation corpora that are collected are not produced for the project. They are produced for a purpose by professional translators with the support of experts for specific domains. For instance, a report from the Agence France Trésor (tax agency) has been validated by tax and financial experts.

Khalid Choukri insists on what François Yvon said, as this is an important issue. The collaboration is made with translation services and the assumption is that documents have been produced in a professional way and that the quality is approved, as they come from professional translators. We need to work with experts, but the collector of data does not have to be an expert of the language. We assume that the work provided within administrations is good. ELRC just checks the technical part (e.g., alignment check).

3.8. Session 2.3 Preparing and sharing data with the ELRC repository – and what happens next

Dr. Khalid Choukri follows with a presentation on how to prepare and share language resources through the ELRC-SHARE repository. His presentation starts by explaining what linguistic data are and what they are in the context of the EC's eTranslation platform. Dr. Choukri shows several examples of the aligned bilingual data that the eTranslation engine uses to learn and improve its translation capabilities. ELRC-SHARE stores these data, duly described with the corresponding metadata for easy location and retrieval.
Following this introduction, Dr. Choukri goes in detail into the type of language data that the EC needs. Large amounts of translation memories (TMs) are already used by eTranslation but many more data are still needed, data held by public institutions. Some examples of the data required are reports, communication documents, news and website content, among others.
However, for data to be useful to train translation systems they need to follow certain specifications in terms of format. For instance, parallel data will be aligned and in .tmx format. Dr. Choukri provides an overview of the formats which are particularly suitable for parallel/multilingual, and monolingual corpora, as well as for terminologies. Furthermore, he goes through all the recommendations for data preparation, given that sometimes very rich data can be stored in inadequate formats that make their use very complicated. In that regard he mentions the ELRC On-site-assistance (OSA) service which can assist data holders prepare their data. One last thought on data preparation refers to file naming and data grouping, which are key points when handling large numbers of files covering data in several languages and domains. At that point, Dr. Choukri lists the most relevant domains with reference to the Digital Service Infrastructures (DSIs).

The presentation closes with a step-by-step demonstration of the use of ELRC-SHARE (www.elrc-share.eu) to contribute language date and the explanation of what happens to the contributed data. Collaborations with other initiatives such as ELRI are also mentioned as a guarantee of data collection and preparation in partnership with the EC. Having said that, the OSA service is mentioned as closing line to remind potential data donors that this service is there to assist them and make their contributions simpler.

3.9. Session 2.4 Can language data from Public Administrations be shared and how?

Mickaël Rigault, Legal Expert at ELDA, gives a very clear presentation about the legal considerations concerning data sharing and re-use. He gives an overview of the PSI Directive and its transcription in the French law. The PSI Directive allows the sharing and re-use of administrative and information documents. It is a simple framework, covering a large range of documents (texts, databases, audio-visual). The PSI Directive transposed in France also applies to education. In France, the *Code des Relations entre le Public et l'Administration* or CRPA (Code of Relations between the Public and the Administration) is the document that contains the guidelines and the rules for publishing and re-using administrative data, along with GDPR. The last modification brought in December 2018 addresses the lists published without being anonymised which include the administration organization charts or the list of approved translators. Using an on-screen image of a cake, he describes making public sector data usable as "a piece of cake". All documents held in a collection can be considered as a cake that could be split up into several parts. These parts include confidential information that is excluded from PSI, information under a 3rd party copyright that is excluded from PSI (it includes the work by translators covered by IPR), and personal data that is also excluded from PSI. The part of the cake that remains, hopefully more than 50%, is governed by PSI.

The data publication process is a 4-step process:

1. exclude non communicable information (secrets related to internal security: bank, medical...)
2. process data in conformity with GDPR
3. check that information in database is not submitted to 3rd party rights
4. check the compliance with CRPA (types of licences and publication procedures)

Users of ELRC-Share can have two roles: a) provider or b) user of resources.

a) The requirements for the provider are:

- apply CRPA dispositions
- follow GDPR dispositions
- avoid collection of personal data, unless within accepted exceptions
- do not infringe IPRs

It is advisable to integrate rights to use translation in your translation contracts.

b) The recommendations for the users of language resources are:

- Check existing repositories: ELRC and other repositories (ELRA, Metashare...).
- Avoid using web crawlers to build up databases as rights related to translated or original data are unknown (even with terms of services...)
- Legal experts available through ELRC project to clarify legal issues for the use/sharing of LRs.

If needed, potential users and providers can contact the ELRC Helpdesk to get support on the legal issues they may face, including clearing IPR or choosing the right license. Some of the licenses used in ELRC include:

- CC-BY 4.0: re-use open to all but need for attribution to author
- CC-BY-ND 4.0: re-use open to all, attribution required, no modification/derivative allowed unless permission from right holder.

Marie-Louise Desfray, from the SFT, is grateful to Mickaël Rigault for mentioning the translators' rights in his presentation and points out that there is a SFT brochure on translator's rights. Leonor Agüera-Jaquemet agrees and states that most translators do not have any contract to protect themselves.

Are there any rights attached to post-edition on machine translation output? Mickaël Rigault answers that the overall translation work is a creative work even if part of it is post-editing on the output of the machine translation. Post-editing is a new task and is not yet part of the IPR legal texts yet, so the translator should negotiate with his/her customer.

Regarding web crawlers, one should keep in mind that copy is illegal and the rights belong to the author. This should be taken into account when sharing language data stemmed from web-crawled data with ELRC. Many researchers do not apply this limitation. Daniel Henkel states that for researchers, the exceptions are very important. European framework is more restrictive than in the US, where the Fair Use allows limited use of copyrighted material without requiring permission from the rights holders, for education or research purposes. Mickaël Rigault answers that there is a new European directive on author's right recently adopted which forecasts (once transcribed in national legislations) exception for data mining and text mining that will enable researchers to mine and collect public and private data, for these research activities.

3.10. Sessions 2.5 and 2.6 - Identifying and managing your data: Questions & Answers + conclusion and discussion

Victoria Arranz, responsible for R&D and Language Resources projects at ELDA. She gives a quick overview of the projects she is currently in charge of including ELRC Lot 3 (LRs collection and processing) and ELRI (LRs / TMX sharing).

Before she starts the Q&A session, she introduces an important concept in data management (identification, collection, etc.): the Data Management Plan or DMP. Users are encouraged to implement a DMP to follow LRs throughout the whole LR lifecycle during and after the production cycle: data acquisition, data description (metadata) and ISLRN, legal issues, data preservation (sustainability and format interoperability), storing and sharing (e.g., catalogues/repositories, citation). The DMP also allows to anticipate on potential legal issues (personal data, anonymization, authors' rights, licenses to use, etc.) by defining the tasks. What are you creating data for? The DMP is elaborated with this question in mind. Reassigning data is also part of the DMP, when, for instance, data are provided in a given format and processed into a TMX.

The Q&A session opens with a concern raised by Marie-Louise Desfray about the commercial use of eTranslation. Could eTranslation be used by other translation services and private companies, threatening the work of human translators?

Philippe Gelin gives a legal answer reasserting that eTranslation is limited to public administrations and general interests. The next steps is to open it to non-competitive institutions, i.e, SMEs of specific sectors, employing (1 to 2 persons), whose needs in terms of translation is very low (about 5-10 pages per year). In any case, the objective is not to target the private sector.

Thibault Grouas then raises the question of confidentiality and asks whether ELRC can reassure providers on the confidentiality issues. How does it work? Philippe Gelin replies that with eTranslation, unlike online commercial engines, the document uploaded on the platform is deleted. Regarding the LRs uploaded on ELRC-Share, it depends on the metadata description. If the data is given only to the EC, only the EC will access it. For Philippe Gelin, it is the provider's responsibility to provide data cleaned of personal data or data that could lead to identifying people. For what concerns what can be retrieved from training data on NMT engines, the research is on-going and at this stage it is difficult to give a definite answer. Victoria Arranz reminds the potential providers that in ELRC-Share, metadata includes personal data or confidential data which is checked internally. ELRC-Share staff can provide support to clear rights or help in the anonymization process.

Daniel Henkel, from the University Paris 8,thinks that since some concepts do not exist in some languages, DGT should consider an ISO standard that would take into consideration the different steps of translation (human, MT, post-edition). Marie-Louise Defray informs the audience that the SFT participated into an ISO standard proposal on post-editing which has been published recently.

Nathalie Kübler, from the University of Paris, finds the eTranslation initiative along with the collection of data through ELRC very interesting and she thanks the organisers for the workshop. She would like to see this initiative extended to other domains and will bring this up with the new President at Paris University to see how universities can cooperate.

Can EMT Label schools and universities use eTranslation? DGT has agreed to extend the use of eTranslation to all the network of schools and universities labelled EMT. Marie-Louise Desfray, from the SFT, asks if SFT, through the training services on post-editing they offer, could access eTranslation? Philippe Gelin answers that it should be dealt with DGT directly.

Both Khalid Choukri and Thibault Grouas conclude the workshop by thanking the participants from the French scene, the EC participants, the teams from the DGLF-LF and ELDA.

Khalid Choukri reasserts that technology is evolving rapidly, in particular Machine Translation. We should be very careful to inform the user of what she/he gets: is she/he reading a text that has been translated by a human translator or processed through a NMT engine? Second issue is public data. Our common interest is to have a translation platform that works. eTranslation can be used by public services for free until 2021.

Thibault Grouas thanks everyone for attending, and commit to resuming the data collection to improve the score for France.

Finally, all participants are reminded to fill in the feedback and ELRC engagement forms that were handed out at the beginning of the day

4. Synthesis of Workshop Discussions

The most important topics that have emerged from the presentations and discussions can be summarised as follows.

The importance of digital in fostering multilingualism and translation for the French Government which has been strongly asserted by the Ministry of Culture representatives during the opening session is a positive sign for the ELRC initiative, and is expected to boost the collection of LRs in France.

Confidentiality and translators rights have also been important topics of the workshop with frequent questions/interventions from the University Translation Departments (well represented) and the French Translators Association (SFT).

Possible extension to other languages than those addressed within the current ELRC initiative have been questioned. Two of the public services represented at the panel are working with refugees and the number of languages they use exceed those spoken in Europe, including a number of indigenous languages.

5. Workshop Presentation Material

The presentations are published on the French (http://lr-coordination.eu/l2france_agenda) agenda page of the ELRC website.