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

European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.22
Task 3

ELRC Workshop Report for France

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

Dissemination Level: Public

Version No.: <V1.0>

Date: 2022-09-29

Contents

1 Executive Summary

This document reports on the 3rd ELRC Workshop in France. The 3rd French ELRC workshop was held online on Zoom for the first time and thanks to the support of DFKI, the presentations went smoothly, and the non-French-speaking participants could benefit from interpretation into English (by Interactio[1]), directly available from the Zoom interface. The event was attended by 52 participants.

The section 2 provides details on the workshop agenda listing all the presentations and presenters, to the exception of Nicholas Asher, from IRIT, who could not make it for personal reasons. François Yvon tool over his presentation. Section 3 provides a report of each session, including the panel discussion. Section 4 contains the update of the Country Profile. The results of the polls are available in Section 5 and Section 6 reports on participation.

The dedicated event webpage can be found at https://lr-coordination.eu/fr/france3rd. All presentations are published.

[1] https://www.interactio.io

2 Workshop Agenda

14:00 – 14:15	**Introduction** **Khalid Choukri,** ELRA/ELRC
14:15 – 14:30	**Welcome** **Paul de Sinety**, Delegate DGLF-LF
14:30 – 14:50	**A Language Data Space for Europe: from vision to implementation.** **Philippe Gelin**, DG CONNECT - Communications Network, Content and Technology - Unit G3 – Accessibility, Multilingualism and Safer Internet, European Commission
14:50 – 15:20	**Language Technologies and Artificial Intelligence – Where does France stand?** **Nicholas Asher**, IRIT[2]
15:20 – 15:50	**Language Technologies in France – Overview** **François Yvon,** LISN-CNRS, T-NAP ELRC **Victoria Arranz**, ELDA
15:55 – 16:00	*Pause café*
16:00 – 16:30	**Presentation/Demo of the eTranslation Platform** **Markus Foti**, DGT, European Commission
16:30 – 16:45	**National Programme for Artificial Intelligence** **Renaud Vedel**, French National Coordinator for AI
16:45 – 17:45	**Panel Session:** **Language Technologies and Artificial Intelligence by and for the public sector / Data** • M. Chevalier, ETALAB • M. Betting, Ministère des Finances • M. Conraux, AMD délégué, Ministère de la Culture • M. Bonnissent, DGLFLF Moderator: **Thibault Grouas**, DGLFLF, P-NAP ELRC
17:45 – 18:00	**Conclusions**

[2] Nicholas Ascher could not give his presentation. François Yvon kindly agreed to take over his presentation on LT and AI.

3 Summary of Content of Sessions

3.1 Introduction

Khalid Choukri, ELRA Secretary General, ELDA CEO and ELRC Consortium representative, welcomed the participants. He began with a few practical information on the Zoom platform: including the recording of the workshop and the interpretation from French to English, available for to English-speaking participants.

Then, he presented the agenda, informing the participants that because of Nicholas Ascher's absence, the slot on Artificial Intelligence would be taken over by François Yvon and Victoria Arranz.

He went on and introduced the ELRC consortium and its four European partners: DFKI (Germany), ELDA (France), ISLP (Greece) and Tilde (Latvia), reminding the audience of ELRC's mission. Started in 2015, the objective of ELRC was to collect Language Resources, also called Language Data, to allow the development of the language technologies (translation, speech, text, etc.) that would be discussed during this workshop and to identify the needs of public administrations for specific language technologies.

As part of ELRC, a team of technicians and legal experts supports all requests and questions related to the availability of use of data. Finally, ELRC-Share, an LR observatory and a repository, has been set up to store all the data that has been identified and collected.

The work done by ELRC is supported by the Language Resource Board members, a network of 60 experts, one duo per country of National Anchor Points: one technical representing the scientific community and one public representing the public administrations. For France, the NAPs are Thibault Grouas from the Ministry of Culture and François Yvon from the LISN-CNRS, respectively Public NAP and Technical NAP. ELRC also benefits from the expertise and involvement of the DGT field officers located in each Member State. Ms Sadji, the DGT Field officer in France, attended the meeting.

Khalid Choukri concluded the introduction by inviting the audience to take part in several polls that will be conducted throughout the workshop to collect information and feedback. He gave the floor to Paul de Sinety, General Delegate for the French language and the languages of France.

3.2 Welcome by DGLFLF

To begin with, Paul de Sinety thanked Khalid Choukri for contributing to the coordination of the workshop. He apologized for not being able to take part in the event this time. The last edition of the French ERLC workshop took place as an in-person event in the premises of Ministry of Culture on June 26, 2019. Given the sanitary context, the present workshop, after being postponed, was held as a virtual event, hopefully allowing very fruitful exchanges between the experts and the participants. Paul de Sinety appreciated that interpretation into English was provided for the European colleagues and wished that in the future, with the support of artificial intelligence, interpretation would be available in all the EU languages.

He recalled the missions of the DGLFLF, the unit of the Ministry of Culture in charge of coordinating the French Government's language policy. As such the DGLF-LF plays a prominent role in the implementation of the plan of the President Macron "An ambition for the French language and multilingualism". The measures proposed by the presidential strategy are reflected in France as well as in Europe: they include two European languages learning in addition to the mother tongue, and language training in European and international institutions. Both digital technologies, as a tool for multilingualism, and translation are strong focus in the Government's strategy. These are also part of the DGLFLF's major strands through its missions "Languages and digital" and "Employment and French Language dissemination". Paul de Sinety took the opportunity to thank Thibault Grouas and Claire-Lyse Chambon who were both attending the workshop.

Among the major projects undertaken by the Delegation, Paul de Sinety quoted a few. The creation of a Reference Center for the Language Technologies which will be attached to the International Center for French Language to be opened in 2022 in the castle of Villers-Cotterêts was the first. This International center was designed as a laboratory for the Francophonie aiming at being the forefront of prospective issues related to the evolution of the French language (learning and practice). With the increasing digitization of uses, dealing with

research and industrialization issues around exploiting linguistic data exploitation as well as building linguistic technologies is a strong requirement.

The second project he shed light on was the organization, within the framework of the next French presidency of the European Union, of a major event on February 7, 8 and 9, 2022, related to Technologies, Innovation and Multilingualism, with the focus on "multilingualism in the digital environment". He stated that this event would be open to all citizens, as well as European actors, public and private, and would be followed by the conclusions of the Council aimed at calling for a European strategy of innovation for multilingualism, a strategy wanted by President Emmanuel Macron. More information would be provided in the weeks following the workshop.

The third project was the Dictionary of the Francophones launched in March 2021 by the International Institute of the Francophonie (University of Lyon 3), with the support of the DGLFLF. This collaborative and contributive dictionary gives access to all the diversity in the French-speaking world through more than 600,000 definitions. As a reminder, the most important dictionary corpus published in France in hard copy is the Trésor de la langue française (French Language Treasury) that includes 90,000 entries.

The observatory of languages was another project planned to begin early 2022 and dealing with the numerous languages present in France. The report by Bernard Cerquiglini published in 1999 listed 75 languages in France, making France one of the countries in Europe with the greatest linguistic diversity. These languages are French, regional languages, some non-territorial languages, and the French sign language. This language observatory based on web data collection will offer the user a coherent and exhaustive set of information for each language, such as sound archives or statistical studies.

The discoverability of French-speaking scientific and cultural content, i.e., the capacity of such content to be discovered among other content in connection with the open science movement was the last project he described.

Regarding this workshop, Paul de Sinety reminded the participants of the two previous ELRC workshops: in 2015, the first workshop was organized by the ministry of Economy and Finances, the then P-NAP, and in 2019 by the DGLFLF which took over. The real challenge of this workshop was probably to investigate, beyond eTranslation, the place of French in the digital landscape, by asking ourselves whether (1) we could use French in all our digital interactions, (2) French language had enough resources to face the technological challenges posed by the artificial intelligence. Our objective is to collect feedback from the users, from the public or private sectors on the needs, the obstacles or the difficulties that prevent them from moving forward. Also, we will emphasize the importance of aligned corpora of French language with other languages to train the MT systems and improve their performances. In this challenging context, there is a risk that French will lose ground for lack of sufficient data, whether in terms of quality or quantity. Therefore, it is important to consider data access and sharing good practices.

As a conclusion, Paul de Sinety expressed his gratitude to the European Commission for this essential initiative to foster multilingualism, and thanked the ELRA association, and its agency ELDA, as an ELRC member and a DGLFLF regular and valued partner, for their support to the organization of this workshop, as well as LISN-CNRS, and all the workshop participants.

Khalid Choukri underlined how interesting this roadmap is and suggested initiating a joint reflection on the workshop follow-up, in the framework of the EU French Presidency starting in January 2022. Paul de Sinety committed to invite all participants to the February 2022 event in Lille.

Before giving the floor to Philippe Gelin, from the DG-CNCT, for his presentation on Language Data Space, Khalid Choukri reminded the mission of ELRC, set up in 2015 to identify, collect and make language data available for the EU languages plus Icelandic and Norwegian. Then, he highlighted the following points:

1. So far, 3000+ Language resources have been uploaded on the ELRC-Share repository (https://elrc-share.eu/). All LRs, identified, technically and legally cleared, encompass a wide variety of data such as audio, video, translation memories, language models, tools. 500 Language resources are French or contain information in French. One of the objectives of the workshop was to increase this number.
2. Legal and technical assistance can be obtained from: https://lr-coordination.eu/index.php/helpdesk
3. General Information on ELRC, including the organised events and workshops, is available at: https://lr-coordintation.eu

3.3 A Language Data Space for Europe: from vision to implementation

Philippe Gelin began his presentation by raising what he sees as a potential translation issue of "European Language Data Space for Europe", which according to him can be translated into French either by "Un espace de données linguistiques pour l'Europe" or by "Un espace de données pour le langage européen" Philippe Gelin admitted a preference for the second translation referring to a "European language" that goes beyond the language and could look like a unique European style of communication.

He continued and shared his personal view on the European Landscape.

Looking at the graph from the European statistics institute (2016) on the language skills (other than native tongues), the average value is 60% which means that 60% of the people can express themselves in a second language. On the other hand, when asked which second language they would prefer to study, most of the students answered "English". If we look at the European map, we can come to the easy conclusion that 60% speak English and 40% don't speak English. This means that from the unique market point of view, or even for transfer of information or social media, those mastering English can easily reach half the European market. On the other hand, for those who don't master English, then the market is limited to their own Member State. Of course, the situation varies from one Member State to another when, for instance, the same language is used across several countries. The difference however is that English, as an international language, facilitates the international exchanges with other countries (the USA or China), leaving the countries communicating in their own language aside. What if technology allowed us to speak in our language and be understood in all the other languages? The example of MT applies to written communication, but if the "Babel fish" existed, then everyone could communicate and disseminate information in their own language throughout the European Union. The consequences would be that (1) European citizens would improve their exchanges in Europe and (2) the Americans and the Chinese would not be privileged and would have to accommodate to using English and other languages. Culture is also an important part of this problematic: while speaking their language, Europeans favor their culture, through the cultural dissemination and exchanges throughout Europe. The "Babel fish" you could slide into your ear to translate all languages does not exist yet, but progress is real. For instance, the performances of MT or Speech recognition systems are improving and could help us reach this European dream.

Next, as you know, Europe is moving forward, and in a few days, a brand-new programme for Digital Europe will be launched. It will cover the 2021-2027 period with a budget of 7.6 billion €. This programme introduces an interesting concept: the European Data Space. The idea is to share and valorize data from one or several sectors to another in Europe. All sectors such as economics, finance, politics, culture, health, agriculture, etc. will be covered. A current example of data that could be shared is the information on how to travel in Europe during Covid and the kind of restrictions that apply. All sectors hold specific data that can be shared and valorized. The data sharing will be organized through a data warehouse, with restricted access, unlike open science or research. In the Digital Europe programme, deployment and competitiveness are the keywords. To enable a free flow of data between the various sectors and countries, a fully GDPR-compliant system will be set up as a horizontal framework ruling data access and governance. The EC will provide support for setting up standards fostering technical and semantic interoperability, which means that the effort done for data in a given sector could be replicated in another sector. The governance space by sector is essential: experts in the various fields can bring up specificities in the management of their data space. The EC will also try to support the digital infrastructure, the data storage, including cloud storage, services (HPC) and processes. Stepping aside his presentation, Philippe Gelin highlighted that statistically, half of the data are textual or spoken data, therefore with a language component which is very important.

Philippe Gelin's working hypothesis was: What if we created a language data space? Many already exist: EURAMIS, ELRC-Share, ELRA-ELDA, TAUS, MetaNet, ELG, CLARIN, etc., and propose language resources and tools. Unlike the previous initiatives on data over the past 40 years, this new programme will differ in setting up the governance body with a bottom-up approach. To achieve this, panels will be organized with industries, Member States, SMEs, research institutions, cultural bodies, and European institutions to collect information, identify the interests of all the parties and their will and capacity to contribute data. Some technical aspects are also important and that will change from the usual LR management: continuous media stream. This is especially true for production and broadcasting companies in the audiovisual sector that will contribute multimodal data (voice, text, image), that are language data. Language models are also data that can be stored in the Data Space.

As a conclusion, Philippe Gelin showed the slides describing the draft ecosystem, including data, HPC and Artificial intelligence, that will help achieve technological autonomy, digital transformation, and economic recovery in Europe through the provision of customized services, business intelligence and access to internal and international markets. He also introduced the tools developed by the European Commission: eTranslation (and additional tools) to be presented later by Markus Foti, and ELG, the European Language Grid.

There was only one comment by Thomas Ruiz (Ministry of Economy and Finances) in the chat who pointed out that if there is only one word ("language") in English for two concepts, in French, there are two translated terms: "langue" and "langage", which are not synonyms, except in the computer science domain.

Khalid Choukri gave the floor to François Yvon and Victoria Arranz who took over the presentation by Nicholas Asher, Director of IRIT in Toulouse, who could not give his presentation.

3.4 Language Technologies in France – Overview

3.4.1 Language Technologies and Artificial Intelligence

As a researcher at the LISN-CNRS, in Orsay, François Yvon's main research topic is Machine Translation. Within ELRC, he acts as the Technology National Anchor Point (T-NAP) for France, representing the academic community for the issues of linguistic data acquisition, pre-processing, and processing.

The purpose of his presentation was to raise awareness on the issues of natural language processing (NLP) and emphasize the importance of data, within the context of a fast-developing domain in relation with the progress of Artificial Intelligence and from the research perspective.

Language technologies is a generic term that covers a range of techniques, methods, and models whose evolution has been tremendous over the last 50 years. These models are intended to propose algorithmic tools to process linguistic objects that can take all possible forms: texts, writings of any form, oral documents, etc. Processing these objects mean analyzing them from the linguistic point of view, but also as part of a collection of documents. Language also takes a vocal form; then processing resorts to speech technologies including signal processing, as was as voice analysis, that provides information and meaning which are carried by the voice irrespective of the expressed words. There are other modalities of expression of the languages such as the sign languages. For sign language processing, a branch of the Language technologies, a different process requires the use of image/video processing techniques. Language technologies also covers the processing of images and videos aiming at analyzing writings embedded in images.

On top of that, NLP must consider the great variability of languages. As recalled by Paul de Sinety previously, in France, we must deal with all the diversity of a large number of languages, many accents, various styles of talking and of writing.

Through four main blocks, François Yvon provided some of Language Technologies coming in many applications.

Automatic Speech Processing: historically stemming from the field of Signal Processing, below are some examples of large applications that are now available to everyone, including on our mobile phones:

- Automatic transcription: indexing, subtitling, voice dictation
- Speech synthesis
- Speech translation: speech samples collection, translation
- Speaker characterization: identification, recognition, emotions ("Who is the speaker?", "What is their mental state, their emotions?")

Semantic analysis, "understanding"

- Dialogue systems: systems in which one exchanges freely with a machine (booking a plane ticket, a hotel room). There are also systems allowing free dialogue with an avatar.
- Automatic summarization (simplification of the access to information)
- Answer to specific questions (Q&A): provides specific answers to questions such as "Who is the President of the United States?" or "What is the status of the French debt?" which goes beyond the search engines outputs that returns documents or text excerpts in response to a query.

- Automatic text generation: generating text from a table with numbers, for instance, to synthesize/present the information.

Text mining

- Opinion and sentiment analysis: a popular application that associates scores with statements reflecting the positive or negative nature of the text or the position of the speaker or writer in relation to what they are saying. Online review (shopping, movie, etc.) where customers can associate text with stars is the best example of the application of this technology.
- Stylometry and author attribution: an application that can identify a person's style, or even detect the author from their text.
- Structured information extraction (terminology, ontologies, knowledge bases)
- Media/social media monitoring: a very important component of LTs that detects automatically hate speech or, fake news, as early as possible to keep them from spreading.

Translation technologies

- Machine translation (general, specialized)
- Computer Aided Translation and Post-Editing (CAT) applications, mostly for professional translators
- Translation memory management, including automatic alignment (sentences, words)

This list is non-exhaustive and does not cover for instance all the LTs for sign language processing.

All the above technologies can be considered as bricks that can be assembled in a single system. If we consider a multilingual meeting processing system, a typical example of complex systems that the industry is increasingly looking for, the assembled bricks are listed below and the application will allow the detection of the speaker, the production/synthesis/indexing of meeting minutes from audio-video recordings, their translation if the meeting is multilingual, etc.

- Speech turn detection (speech recognition, speaker characterization)
- Automatic generation of verbatims (summarization)
- Automatic translation of text and/or speech (automatic translation)
- Recognition of handwritten notes
- Search (multilingual) in past meeting minutes (automatic indexing, question answering)

The history of NLP is not new and has always collided with the complexity of processing languages automatically. Ambiguity in real statements is one of the issues that keeps generating work and effort. To illustrate this, let's look at a very simple statement in French: "*J'ai commandé une glace au serveu*r" (I ordered an ice cream from the waiter). For a human being, this simple statement in French is easy to understand. For a machine, because of each word polysemy (*commander*, *glace*, *serveur*), deep understanding and subsequent processing are challenging. For a very long time, expert systems, based on rules or algorithms to try to determine which term to use in what situation, were considered the best option to tackle this ambiguity issue. At the turn of the 1990s, machine learning (ML) methods became widespread in the NLP field. Over the last 30 years, ML has revolutionized the field to the point that today, all language processing techniques that work rely on the exploitation of annotated data. The annotated data are examples of the associated couple "System input" / "Desired output". In translation, the source language sentence is the input, and the translated sentence in the target language is the output. In sentiment analysis, a text is the input and a few stars the output. When annotated data is available, it is possible to circumvent ambiguity problems and produce powerful applications. Until the years 2005-2010, the progress was real but relatively slow, but then, the emergence of artificial intelligence-based technologies became prevalent, allowing substantial breakthrough in terms of quality and integration in applications for the general public.

The main qualitative leap that explains the power of Artificial Intelligence lies in the ability to learn to represent. This can be illustrated by the two following examples on Text classification and Translation (see the slides).

When we do text classification, as shown in the use case (slides), we have to take an email as input and decide whether this email is wanted or unwanted (spam). One can imagine representing the task according to the diagram (omitting the component of making the text suitable for processing by an automatic classification tool

– statistical decision rule). To make a text suitable for statistical decision, a representation operation is needed. The representation of linguistic statements can take several forms: sequences of symbols, bags of words or lists of the words that this text contains, irrespective of their order and structure. Once we go from a text to a word dictionary, we can use the dictionary to do classification (with a model). This is an extremely useful and efficient method for spam classification (the first public application with spelling correction). The last few years have seen decisive advances thanks to the use of representation learning methods. AI, or AI-based techniques, have made it possible to win on two fronts: to improve the classifier component (blue) but also (most importantly) on the representation component, by changing the way texts are represented in the machine, no longer as dictionaries but as hollow vectors (dictionaries) used in the models. There are many families of models, including BERT, which takes linguistic statements and transforms them into vectors in an optimized way to perform learning tasks.

In translation, the learning system used nowadays is called transformers. This extremely powerful system follows the logic of encoding (transforms a text or a series of sentences into digital representations) / decoding (generates a text from a digital representation). MT systems are fed with a set of texts in a source language (French in the example). The encoder transforms the set of texts into a series of digital vectors, which are decoded to produce a text in the target language (English). This output text is then be compared with examples of real translation into English. Based on the comparison between the output and the expected output, the set of parameters regulating the encoders/decoders operation is be updated. By repeating this learning loop, large scale systems are trained to obtain staggering performances. Increasingly, these systems are capable to generate texts that can hardly be distinguished from texts generated by humans.

Other systems also show staggering performances such as: the live automatic voice transcription of a TV news programme, the participation of a machine in the Jeopardy Q&A game (IBM Watson) or in a debate during which the machine can engage in an argument with human beings[3] (IBM Debater). The last example shown by François Yvon is the translation of a text from English to French using Deepl, with an output almost "perfect".

The systems are amazing yet misleading because they make errors that are largely unpredictable. They are highly dependent on learning data, and they can only reproduce what they have already come across. In translation systems, learning biases lead to gender stereotypes i.e., when switching from English to French the term nurse is systematically translated by *infirmière* (feminine). The systems are very data-dependent, they lack robustness to the unexpected. These systems are opaque, and their decision process is unintelligible when it comes to knowing why they succeed, or why they fail. These systems have no understanding, there are just extremely powerful machines into associating inputs and outputs but without understanding the manipulation of the inputs nor the data processing. Therefore, much work remains to be done to advance these systems. The state of the art is progressing in various directions. In machine translation, algorithmic efficiency can be impressive when the system learns from multilingual models to build a single machine translation system that can translate from many languages into many languages. Google and Facebook's search systems already work this way, from 100 languages to 100 languages with a single machine translation system.

One of the significant benefits of modern artificial intelligence is improved learning of representations. Although there is still significant work to be done on improving this type of unsupervised learning, we are also working on improving the complementarity of modalities, as Philippe Gelin pointed out, in order to build systems that process images, speech, and possibly texts embedded in the images. Finally, artificial intelligence has made possible to carrying out end-to-end learning. For example, automatic translation systems can now pass directly from the input speech signal to the output speech signal without going through an intermediate textual representation.

To develop these systems, data is needed, and data is scarce. A huge effort must be made to improve the processing of annotated data for low-resourced languages, but also to enhance the transfer between tasks and modalities: we are able to build systems that recognize speech for one application domain (e.g., news). However, systems performing well for newspapers may not be as efficient for TV series. The goal is to develop more general systems, like MT systems that can translate all languages.

For a few languages with a lot of resources (French, English, German, Spanish, Chinese, etc.) for which we have a lot of data, the systems are very efficient. On the other hand, for languages with few resources, (most of the

[3] The videos are available from the Slides.

world's languages), due to a lack of data, there is little hope that automatic processing systems can be produced. Making progress so that automatic processing systems become as efficient for less-endowed languages is one of the current challenges.

For the French language, François Yvon gave a brief overview. There are multiple actors that cover the whole spectrum of Language technologies for French: dialogue, synthesis and transcription, translation, generation, sentiment mining, Q&A. There is a wealth of general resources (dictionaries, corpora, models, including very large models). We are lucky enough to have a solid academic fabric with the 3IA Artificial Intelligence institutes, including the institute in Toulouse headed by Nicholas Ascher, but also CNRS, INRIA and universities. Despite this, we lack specialized data. Domain and regional languages coverage is still weak. And we don't have enough data very finely annotated. There are also, systemic obstacles, including data anonymisation, that impede data sharing. We have managed to overcome them for health data but are still struggling with language data. Victoria Arranz will be discussing this aspect in the next presentation.
As a conclusion of his presentation, François Yvon highlighted the quality measure and assessment as an important challenge to tackle. As a matter of fact, failing to assess the quality of the systems will hinder their development. More effort must be put on measuring the progress based on real usage to be able to orient the system development.

3.4.2 Managing Sensitive Data: MAPA Project

Victoria Arranz, who is responsible for LRs-related and R&D projects at ELDA, took the participants through the process of anonymisation. This has been the subject of the MAPA project. She first set the legal framework of data sharing by reminding the audience of (1) the PSI directive (2003-2013) that encourages the sharing and reuse of public data and (2) the GDPR (2016) on personal data protection imposing that personal information is removed from shared data. For the public bodies, it is very complex to remove the sensitive information from the texts. Not all the public administrations can afford to perform this task which is technically challenging and costly. Therefore, there is a need to develop anonymisation and de-identification tools. There are several levels and approaches that Victoria Arranz went through. Techniques to manage personal data include Anonymisation and Pseudo-anonymisation or de-identification.

During the anonymisation process, sensitive information is replaced by empty spaces or a chain of characters (e.g., XXX) and the identification of the person becomes impossible. Sensitive information includes all the personal information that should remain hidden: last name, address, phone number, social security number, credit card number, etc. Once data have been anonymised, the source cannot be traced back, but data's value and reuse are very limited for NLP processing. With the pseudo-anonymisation, or de-identification, the sensitive information is replaced by IDs (categories, sub-categories, etc.) or pseudonyms (last name "Dupont" can be substituted by "Dubois"). De-identified data is GDPR compliant and reusable in applications or for processing.

In this context, ELDA has been involved in MAPA (Multilingual Anonymisation for Public Administrations). The MAPA project was a 2-year INEA-funded Action for the European Commission under the Connecting Europe Facility (CEF) – Telecommunications Sector. The main objective of MAPA was to develop a toolkit for multilingual de-identification of personal data, easy to install and to implement through an independent docker and an API connection. The requirements were to address the legal and medical domains, that are critical when processing personal data, to comply to GDPR, to cover the 24 EU languages and to be available as a service on the eTranslation platform. For machine learning purposes, corpora have been created: raw corpora (1M sentences per language) and annotated corpora (a fraction of raw data).

The approach was based on named-entity recognition (using BERT model). Mono- and multilingual models have been created for this purpose. Then, entities (personal information) to be de-identified have been defined, in relation to medical and legal domains. Language and country specificities were considered. Finally, a manual of annotation for identified data was produced, with an ontology to be shown later. For training purposes, manual annotation has been processed for all languages and automatic pre-annotation has been run on Spanish. Several obstacles have been encountered, including identifying and collecting data in some of the languages that was not easy. Another problem was to obtain medical data such as clinical reports that are by nature confidential because they contain sensitive data. For French, as court data is already anonymised, we had to recourse to data from the Cour de Cassation (the highest court in the French judiciary system). Working with the Cour de Cassation was very fruitful because they have already implemented anonymisation process. We could

"reconstruct" data automatically using our lists of entities. An example of the "reconstructed" data is shown in the slides. For French, the EUR-LEX corpus was also used. For medical data in French, clinical cases were automatically processed to enrich their content with entities (e.g., the patient's name was replaced by "the patient").

Victoria Arranz took the participants through the process of annotation with the Inception tool, showing how this is instrumental in the machine learning process. Then, she made a quick demo from the MAPA interface to show the participants how the tool detects, obfuscates, and replaces the entities.

Victoria Arranz concluded her presentation by giving an overview of the preliminary results that are rather good in French and Spanish. She emphasized that with the BERT approach, the lack of data in some languages (there is no legal data in Maltese, everything is in English) is compensated by combined mono- and multilanguage models. Some use cases have been set up for both domains: Complaints Watch (Europe) and Spanish Ministry of Justice (legal), and a hospital in Paris for the system evaluation (French). Finally, the integration on the eTranslation platform has started and is on-going.
There was one question from Thomas Ruiz (Ministry of Economy and Finances) who was wondering whether the MAPA mechanism could be applied on fiscal data. Victoria Arranz replied that this is absolutely feasible: using MAPA would just require to adapt the entities.

More details on the project and on the partners can be found at https://mapa-project.eu/

3.5 Presentation/Demo of the eTranslation Platform

This talk was delivered by Marcus Foti, who is responsible for eTranslation at Directorate General of Translation (DGT). He went through a basic presentation of the tool, the objectives, and potential issues.

Why is the European Commission interested in Machine Translation? The European Commission has always been interested by MT. As early as 1975, the European Commission has worked with Systran on a rule-based system. At the end of the 23 years of collaboration, the system could operate on 28 language pairs. In 2010, a new project on statistical MT, MT@EC, was launched. At that time, the revolutionary technology has allowed the coverage of all the EU languages simultaneously, but statistical translation did not last too long. An ambitious collaboration with DG-CONNCT was set up to develop a neuronal MT system. If we go back to 2009, 2.5M pages were machine translated which is the equivalent of the number of pages translated each year by the 1500 translators working at the European Commission. This translation work of the EC translators was very useful to develop statistical MT engines because data (the European jargon) was already in the right format.

All 24 languages, apart from Irish, and Croatian who joined the EC at a later stage, were covered and available as parallel corpora. The lesser resourced languages are now catching up. All EC translation data are available online on DGT-Translation memory or Europarl websites, among others. After 2017, MT@EC was producing 18M pages per year which outpaces human translation by DGT. However, the quality and legal value of human translation does not compare with MT output that remains approximate. However, in some cases, MT is very useful.

Moving to eTranslation as part of the CEF, the user base was enlarged to include all the EU public administrations DSIs and lately the SMEs. At the beginning, MT@EC was meant for the EU translators and the EU institutions staff, the EU DSIs, and public services. eTranslation can be used two ways: through a web interface or an API that integrates the engine to information systems. When the user submits a document (.docx, .pptx, xlsx), the system returns the document in the same format. Lately, DGT has been working hard to integrate eTranslation to all the EC pages (the process is still on-going). European MPs claim that many European institutions sites are only available in English, French or German, but not in the other EU languages, deploying eTranslation is a way to address criticisms.

Through the CEF, the DGT has moved to languages that are socially and economically interesting: Arabic, Japanese, Mandarin Chinese, Russian and Turkish. Regarding the quality, feedbacks differ some find that eTranslation is good and others that Google is better. As discussed by François Yvon earlier, the MT output depends very much on the data used to build the engine. For the translation of EC documents, the results are very good, because eTranslation was built and trained using this EC data and documents.

The domain is also very important. For the financial domain, which is new for Translation, the DGT has worked with the French Ministry of Finances that contributed 120k new sentences. Also, the European Central Bank has asked all the national banks in Member States to collect data. National banks have provided data (from 60k to 200k sentences). Domain-specific data are instrumental in improving the quality output of MT engines. That is why the EC, with the support of ELRC, keeps collecting data for various domains and languages.

Now if we consider the performances of eTranslation and how the European institutions are using it, there is the Re-Open EU interesting experience. Re-Open EU is a website (and a mobile app) that was created to provide multilingual information on travelling to/in Europe during the COVID-19 pandemic. Initially launched using Microsoft MT engine, the translation output did not meet the expected quality and the DGT complained about this. Although very busy, the 1500 EU translators offered to translate the site that required very specific translations. Given the delay and the volume to be translated, it was impossible to meet the deadline, so it was decided to use eTranslation (with the General Text domain and a disclaimer). Overall, the general meaning is restituted, even if there were cases where incorrect translation was provided (for some specific language pairs).

Another example is the Conference for the Future of Europe, a very important event for the European Parliament, the Council, and the European Commission whose objective was to collect feedback from the European citizens. All languages were covered, even if only few were used for the actual exchanges. eTranslation allowed to "talk" to citizens directly. As in the previous example, DGT received some negative feedback for a couple of countries.

As a conclusion, Markus Foti reminded the participants that they can register, if they are eligible, to use eTranslation but also the other Language Tools such as NER, Speech-to-Text and soon Anonymisation as mentioned by Philippe Gelin earlier.

There were no questions, however Khalid Choukri asked Markus Foti to detail the access to eTranslation overtime. Markus Foti underlined that DGT is willing to sustain the access that had been set up in the framework of the CEF programme (until 2021), throughout the Digital Europe Programme that will take over and grant the access under the same conditions until 2027.

3.6 National Programme for Artificial Intelligence

Renaud Vedel, the coordinator for France's National Strategy for AI, gave an overview of the national strategy in connection with Language Technologies.

Further to the Villani Report on AI, presented by the President Macron in March 2018 (and updated in April 2021), the National Strategy is paving the way to structuring the French ecosystem in a long-term perspective. It also aims to position France as a world leader in this set of scientific disciplines and information processing technologies. The topic coverage is wide and addresses Research and Higher education, Economy, Health, Transport and Environment, Security and Defense. France is eager to push for the development of a "democratic model" of AI, so ethics is a topic that will be given strong attention.

At the European level, the Coordinated Plan on Artificial Intelligence (launched in 2018 and revised in 2021) is a joint commitment to maximise Europe's potential to compete at a global level and encourage Member States to develop national strategies. As part of this plan, both Horizon Europe and Digital Europe programmes are financing numerous actions (2021-2027) in various fields including Language Technologies. Renaud Vedel encouraged the institutions working in the LT field to apply and obtain financing within these programmes.

At the national level, there are the PIAs, the Investments for the Future programmes, whose 4th generation is covering 2021-2025 and France 2030, an additional investment plan announced by the French President earlier in October 2021 that will include a large part for education, including education to AI. The PIAs are financing both Research structures and innovation through call for projects. Two calls are specifically addressing language data: Digital platforms and data sharing for the sectors finances data hubs and IA Challenge for data enhancement. Then Acceleration Strategies, also called DeepTech strategies, include one dedicated to AI and more specifically Embedded AI/Trustworthy AI with related digital strategies that may incorporate AI (digital education, digital heath, culture, etc.) and in which the textual data is important.

The National Strategy, following a strong proposal of the Villani Report, states that the inequalities in access to intensive computing should be reduced. With the very large language models used in machine learning,

researchers and startups working on LTs should also have access to intensive computing. A major achievement of the National AI Research Programme has been the creation of a supercomputing facility for the ecosystem. The Jean Zay supercomputer devotes a section of its computing power to AI. In 2022, its capacity will be doubled. For now, an increasing number of NLP projects resort to the supercomputer. The best example is the NLP project BigScience, led by Hugging Face, whose purpose is to train Bloom, the largest multilingual open-source language model.

To sum up the National Research Programme on AI and the strategy, here is a list of the main policy instruments that are available:

- 190 teaching and research chairs
- Substantial extension of doctoral research programmes, with a total of 500 new PhDs in AI
- At least 7 interdisciplinary institutes including the 4 3IA institutes and 3 other independent institutes of excellence (Data-IA Institute in Paris-Saclay, SCAI in Paris-Sorbonne Sorbonne University and Hi Paris! (IPP and HEC groups)
- The National Research Organisations (ONR): CEA, CNRS and INRIA
- Research partnerships and transfers with the Carnot Institutes and the IRTs, and other competitive clusters
- Future European poles for digital innovation funded at 50% by the European Commission

In the French ecosystem, there are startups working on NLP and it may be very interesting to develop partnerships between the academia and the startups. In addition, the European commission is promoting Data Spaces in 9 domains. Language is involved in many of these domains. This model could also be set up in France provided that the governance allows interoperability (objects, metadata, etc.) and that an infrastructure is setup.

In terms of market, France lags behind other countries of comparable economic size (Germany and the United Kingdom). Therefore, one of the objectives is to consolidate the language technology market in France. The impact of LTs can be substantial in many sectors, including tourism with the provision of multilingual services, customer relations, etc. When browsing Hugging Face global library, French ranking is not bad compared to other languages, but still far behind English. What can we do to preserve linguistic diversity and counter the hegemonic position of English and Chinese? In Europe, some countries have already developed national strategies on language technologies as part of their AI strategy. This is the case of Sweden with AI Sweden, with a resource center, the animation of an ecosystem, a work agenda for the administrative authorities, etc. The provision of linguistic resources is available in Denmark at the Center for Sprogteknologi (University of Copenhagen), but also at ELDA in France, at ELRC in Europe. Finally, France can also draw inspiration from the German and Spanish examples of NLP.

Renaud Vedel concluded his presentation by asking how to better structure the Language Technologies sector in France suggesting avenues for reflection on the governance, the implementation of a French programme on language technologies (in the context of the opening of the Cité de la langue française at the Château de Villers-Cotterêts), language resources, evaluation campaigns, legal issues, etc. He announced that he will take part in the Forum on Plurilingualism organized in February 2022 by the DGLFLF and offered to relay the proposals, if any, to the French government.

Khalid Choukri thanked Renaud Vedel for his presentation and recalled the Technolangue[4] programmeme that was held a few years ago. The environment is different now, more competitive, more global, but Khalid Choukri hoped that by the end of the French Presidency, a roadmap could be finalized.

Philippe Gelin informed the participants that the Digital Europe Programme has been approved by the Member States and he confirmed that there will be a European digital space dedicated to language, one of the applications of which will be set up to support the development of language models.

[4] Cf http://www.technolangue.net/rubrique.php3?id_rubrique=22 . The main objective of the national programme was to set up a sustainable infrastructure for producing and disseminating language resources and evaluating LTs.

3.7 Panel Session: Language Technologies and Artificial Intelligence by and for the public sector / Data

As the chair of the sessions, Thibault Grouas, Digital Project Manager at the DGLFLF and ELRC P-NAP, set the framework of the panel session for feedbacks from the panelists on data and data sharing.

Pascal Betting, Head of Tools at the Translation Center of the Ministry of Economy and Finances, was the first speaker of this panel.

Pascal Betting indicated that he had followed the entire workshop with great interest and stated that the translation center is not involved in the AI field. As part of the collaboration with the DGT, the Bercy center provided 100k FR-EN segments in . tmx format. He added that injecting this data into eTranslation had improved the quality of the translation of full documents, confirming that training the MT engine with in-domain delivers results. When asked about the type of data shared, he answered that only public documents (published administrative reports, Trésor-Eco magazine, etc.) were shared with the DGT. He found the presentation on anonymisation very convincing and if the process works well, it is not impossible that, in the future, the center could provide translation data containing personal information (to be anonymised). So far, this could not be done, due to lack of time to devote to this task.

Thomas Ruiz, Translator, did not have much to add. He just said that the translation center is using CAT tools for their translation activities. However, he admitted that for certain types of documents of a technical nature, the DGT's MT translation solution could work well, especially for documents of a technical nature.

Christine Coubard, Head of the translation center, explained that eTranslation is used to translate EU texts, for volume and workload reasons, and to provide references for in-house translators. The center does not offer, and does not wish to offer, post-editing services. Post-editing is a very specific task that requires university training. She considers that tax-related documents are not currently suitable for public MT processing.

Aurélien Conraux, Delegated Data Administrator, Ministry of Culture, presented the Ministry's data strategy. The goal is to develop the uses of data. The Ministry has large and multidimensional collections, whether text (BNF), image/video (INA), national archives or digital legal deposit. The Ministry's roadmap on cultural data and content defines the Ministry of Culture's policy for data, algorithms, and source codes. This includes opening, sharing and valorizing data, notably through digital innovation projects, to promote the dissemination and visibility of cultural content. In this context, it is possible to share linguistic data with ELRC.

Jean-Christophe Bonnissent, in charge of the mission on the Presence of French in society, DGLFLF, addressed the subject of discoverability.

The discoverability of a content in the digital environment refers to its availability online and its ability to be found among a large set of other contents, especially by a person who was not specifically looking for it. The Ministry of Culture supports and subsidizes the development of digital initiatives and research projects in favor of the digital discoverability of cultural content.

The interest of the DGLFLF does not lie so much in the languages as in the speakers, and its action aims to produce actions that protect the different French-speaking communities. In the field of science, one must consider the irreducible dimension of multilingualism, even if it’s tempting to think that science is converted to English. This is not always true, especially in the social sciences.

The concern for users leads the Ministry of Culture to want to re-establish the continuum between research and transmission/teaching. Discoverability must be multilingual within scientific content. The production of terminology and repositories will allow scientific production to be equipped with multilingual metadata so that users can find and discover content using their own language. We can enable the discoverability of scientific content in French for non-French speaking users.

Alternatively, more French content can be circulated: with the support of machine translation, it will be possible to produce versions in other languages (in particular French) of English content. Philippe Gelin confirmed that both academic support and open science topics were supported by the European Parliament. One of the objectives of the next framework programme is to open the services to academics to cover the languages of eTranslation.

Boris Bayard, Operations Manager of the Entrepreneurs of General Interest programme, Etalab, could not attend and provided a written contribution (below).

The Entrepreneurs of General Interest (EIG) programme, a programme of Etalab at the Inter-ministerial Digital Directorate, is issuing their next call for projects mid-January 2022. The call targets public administrations, throughout France, that wish to conduct innovative digital projects.

The programme will select skilled digital specialists (developers, data experts, designers, and digital lawyers) to test and experiment new possibilities with public agents.

Judging criteria for the selected projects must include:

1. A clear, high-impact need, brought to a high hierarchical level by the host administration
2. A well-defined problem
3. A motivation to participate in the programme, to open the field of experimentation to digital talents, and to grant them autonomy.
4. Strong capabilities for digital transformation of the administration.
5. Opportunities to open, share and reuse the resulting tools and data.

Successful administrations receive support to pilot these talents (group seminars, team coaching, peer-to-peer learning) and are eligible for co-funding from Inter-ministerial Digital Directorate (under certain conditions).

For this new call for projects 2022, the EIG programme wishes to support more projects carried out within local authorities and decentralized government services.

To discover the ISG programme, examples of projects conducted, and discuss this call for projects, any agent of a local authority or a deconcentrated State service can register for the webinar organized as part of the "Innovation Month" on the DTIP web site.

At the end of the panel sessions, Binta Sagi, DGT Field Officer for France, took the floor to announce the preparation of a workshop on eTranslation, a European value-added translation tool. Khalid Choukri thanked Ms. Sagi and moved on to the conclusion.

Khalid Choukri emphasized the richness and interest of this half-day workshop. The presentations on technology and AI applied to language processing by François Yvon and Victoria Arranz were excellent. Philippe Gelin's presentation on (linguistic) data spaces was timely, especially in articulation with Renaud Vedel's presentation of the French national AI programme and with the work done by the DGLFLF as presented by Paul de Sinety and through the panel session run by Thibault Grouas. Finally, Khalid Choukri welcomed the confirmation that access to the eTranslation platform will be continued for EU public administrations and SMEs and extended to academics.

Khalid Choukri concluded the workshop by thanking the speakers, the DGLFLF team, François Yvon and Thibault Grouas, the 2 NAPs, the colleagues from the European Commission, Philippe Gelin, Markus Foti and Binta Sagy and the ELDA team for the organization.

4 Country Profile: Language data creation, management and sharing

A couple of changes have been made to the Country profile.

Setup of a data policy at the ministerial level with the appointment of Data administrators.

The French government has acknowledged that sharing and enhancement of data and algorithms supports innovation and research and contributes to create and boosts the development of new uses, such as artificial intelligence. In April 2021, as part of the implementation of the data policy, data administrators (AMD) have been appointed in each ministry to develop the strategy for data, algorithms, and source code. A roadmap detailing the objectives related to the opening and sharing of data has been produced by each ministry (roadmap of the Ministry of Culture). Following quality and interoperability standards, the open data of each ministry must be referenced on the data.gouv.fr, the portal managed by Etalab.

The role of LT and language data in France's AI regulations:

Make France a global leader in AI is the objective of the of French National Strategy for AI. To achieve this objective, investments will focus on research and education, data, IT infrastructures, etc. Numerous initiatives, programmes, projects, partnerships, etc. will be set up at the national and European levels.

Natural Language Processing, as a branch of AI, will also benefit from these mechanisms boosting research programmes, technology developments and partnerships in Language technology start-ups and companies. AI progress in the field of machine learning has had a strong impact on numerous fields including speech processing and NLP. Within the framework of the National AI Research Programme, HPC has been setup and is increasingly being used by NLP projects. The largest of these projects is BigScience, led by Hugging Face, whose purpose was to train Bloom, the largest multilingual open-source language model.

Updates in the Action plan:

- Continue to promote the in-domain data as means to improve the performance of MT engines.
- Promote anonymisation features to convince the translation centers to anonymise their data before submitting them to eTranslation
- Work with the data administrator in the Ministry of Culture to identify data that can be shared on ELRC-Share

5 Workshop Participants

	Number of participants	Percentage
Public sector	18	36%
Industry LT providers	0	0%
Research / Academia	8	16%
SMEs	0	0
EC and ELRC consortium	20	40%
Local organiser staff	2	4%
Other	2	4%
Total:	50	