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

Deliverable D3.2.9

Task 3

ELRC Workshop Report for Italy

Author(s):	Claudia Soria, Marina Omiccioli, Roberta Persia
Dissemination Level:	Public
Version No.:	V1.0
Date:	2021-10-05

Contents

1 Executive Summary

This document reports on the ELRC+ Workshop in Italy, which took place online via Zoom, on June 10th, 2021. It includes the agenda of the event (section 2) and briefly sums up the content of each presentation and of the panel workshop sessions (sections 3 & 4).

The 3rd ELRC Workshop was attended by 74 people, among which: 32 people came from public administration, 23 from academia and research institutions, 5 from EC, 6 from LT Industry, 3 from SMEs and 5 from other organizations.

The dedicated event page can be found at https://lr-coordination.eu/node/365

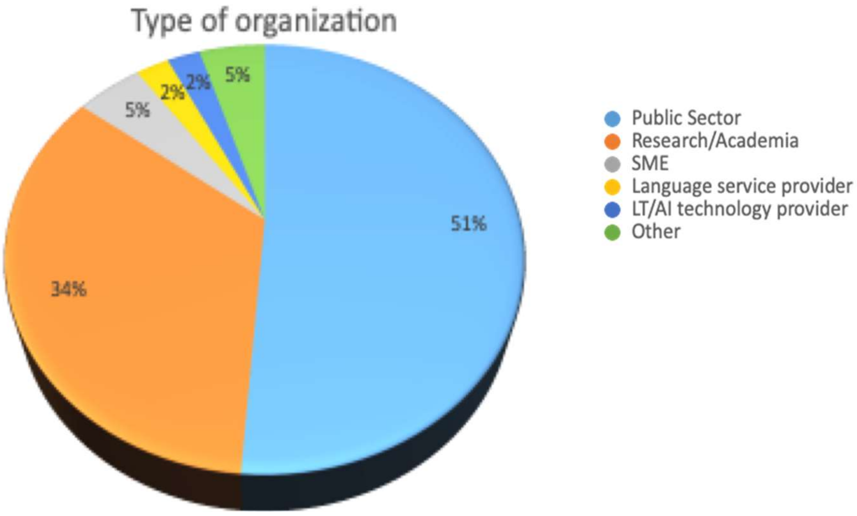
2. Workshop Agenda

09:00 – 09:10	**Welcome and introduction** *Claudia Soria, ELRC Technology NAP*
09:10 – 09:30	**The potential of Language Technology and AI – where we are, where we should be heading** *Malvina Nissim, Groningen University, Netherlands*
09:30 – 10:30	**Language Technologies for Italian - Panel session** *Bernardo Magnini, AILC and FBK* (Moderator) *Marco Turchi, FBK* *Andrea Bolioli, CELI - H-FARM Innovation* *Maria Palmerini, Cedat85* *Guido Vetere, G. Marconi University*
10:30 – 10:45	***Coffee Break***
10:45 – 11:15	**The CEF AT Platform** *Markus Foti, European Commission, DG Translation*
11:15 – 11:45	**Language technologies by/for the public sector and the SMEs - Panel session** *Marina Omiccioli, ELRC Public Services -NAP* (Moderator) *Mauro Draoli, AgID (Agency for Digital Italy)* *Barbara Altomonte, Presidency of the Council of Ministers* *Paolo Cappelli, Ministry of Defence* *Fiorenza Manfroi, PromoFalcade - Dolomiti*
11:45 – 12:15	**Language data creation, management and sharing: existing practices and challenges - Panel session** *Simonetta Montemagni, ILC-CNR* (Moderator) *Fabrizio Calabrese, Bank of Italy* *Gabriele Ciasullo, AgID (Agency for Digital Italy)* *Monica Monachini, ILC-CNR and CLARIN-IT*
12:15 – 12:25	**Q&A**
12:25 - 12:30	**Conclusions**
12.30 - 13.00	**Networking and demo**

3. Summary of Content of Sessions

3.1 Welcome and introduction

Claudia Soria, T-NAP, welcomed the participants to the virtual workshop. She presented the ELRC initiative and introduced the main themes that would be addressed during the various sessions. The workshop was conceived as an important opportunity to combine the demand and supply of language technologies for Italian and ELRC could be seen as a facilitator between, on the one hand, those who develop language technologies and, on the other, those who use these technologies in their daily work practice. She invited the attendees to make their voice heard, both during and after the workshop. After the announcement of some technical details, a first poll was launched to understand the composition of the audience.



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

This talk was given by Prof. Malvina Nissim, Associate Professor at the University of Groningen, the Netherlands.

The talk offered a brief introduction to Nissim's view on the current and future potential of language technology and artificial intelligence. Starting with some examples of language technologies used in everyday life, she showed how current tools are excellent and very advanced, and allow for truly impressive results, in many cases very similar to human behavior. However, the results we get are not without bias.

Indeed, technologies often introduce these biases, which in turn are determined by the data available. For example, Google Translate handles idiomatic expressions from English to Italian very well, but not so well in the opposite direction: the expression *to be over the moon* is correctly translated as *essere al settimo cielo*, while if we reverse the translation direction, the English correspondent we get is a literal expression *to be in seventh heaven* and not the corresponding idiomatic expression. This small example demonstrates the crucial importance of the amount of data available to language technology algorithms: it is the greater amount of data available in the English to Italian direction that determines the correctness of the translation.

It is therefore necessary to raise awareness among users about the existence of biases, in addition to research and development to try to eliminate this kind of deviation.

In short, the future perspective from the point of view of Prof. Nissim can be summarized as follows: our aspiration must be to have the best possible technological support to carry out a given task, without fear that technology may replace human work. To do this, it is necessary to find the best possible cooperation between humans and machines and achieve this cooperation by using LT critically in our daily work and life. It is also necessary to achieve a better knowledge and understanding of LT and the way they work to be able to use them critically. At the same time, investing in research and development of resources that can reduce biased results is equally crucial.

From this perspective, language professionals should never approach language technologies as competitors for their professional expertise but rather, as an important and helpful support.

3.3 Language Technologies for Italian (Panel session)

This session was chaired by Bernardo Magnini, Senior Researcher of Fondazione Bruno Kessler in Trento, and President of the Italian Computational Linguistics Association (AiLC). Bernardo Magnini introduced the topic of the panel.

Recent years have witnessed a sort of revolution in the research and development of applications based on language technologies, with a strong impact of artificial intelligence subfields such as machine learning and deep learning. The aim of this panel is to provide an understanding of the situation for the Italian language.

The round table introduced the perspective of those who develop and research these technologies, to have an opinion on what the state of the situation is and, on the limits, if any, that are encountered in creating applications.

The four experts invited were:

- Marco Turchi, Head of the Machine Translation Unit at Fondazione Bruno Kessler, a research institute.
- Alessio Bosca, Project Manager at CELI, a company with a long tradition in the development of technologies for the Italian language, in particular conversational agents and text analysis.
- Maria Palmerini of Cedat 85, Expert in speech recognition technologies.
- Professor Guido Vetere of Guglielmo Marconi University of Rome.

The panel was organized in two rounds. In the first, each of the invited experts presented their point of view on the current situation, also highlighting the limits of the technologies in their field of experience. After a first round of questions, the second series of interventions covered the future, in terms of next steps ahead and missing elements for advancing the development of technologies and availability as concrete applications.

Marco Turchi talked about his experience in the field of automatic translation to and from Italian, stressing how exciting the current period is: a new era that leads to continuous improvements.

Until five years ago, research in machine translation was at a standstill: systems based on statistical methods had reached a point where performances did not improve any longer.

The emergence of neural models has completely changed the sector, giving a strong impetus to automatic translation, especially for languages that have a completely different syntax.

Models based on deep learning and artificial intelligence have brought great advantages, but they require large amounts of data. This discriminates between those pairs of languages for which the system works excellently and those pairs of languages for which the system performs problematically. Fortunately, Italian is one of those languages for which we have large amounts of data.

Of course, everything changes for languages for which there is far less data. In these cases, the situation becomes quite complex and the quality drops dramatically. A similar dropping in performance applies for very specific and technical domains.

In the current paradigm, language skills are completely left out from technology. However, it is highly probable that the boom in data intensive exploitation will be over, and there will be a need to create synergies with people who both use these models and also have the language skills to be able to intervene and integrate the systems.

Alessio Bosca, Researcher and Project Manager of CELI, represented the point of view of an Italian company that has been working on the development of language technologies for many years now. The focus of Dr. Bosca was the interaction with customers in the use of resources for the development of commercial applications.

CELI uses a hybrid approach that combines models based on Artificial Intelligence with rule systems based on linguistic patterns. Models are pre-trained on large volumes of data (both in-house and open source) and then are specialized for a specific task, for example the classification of texts, using the same texts that are the subject of the analysis.

CELI therefore uses linguistic resources at two levels: generic resources that are used for the creation of the basic model, and domain resources that are used to specialize the general model. The latter are much smaller and much more specific resources that are provided by the customer.

This approach is used both for text analytics and text classification systems and for conversational agents, for example for systems in which the intention of the questions is to be recognized to provide adequate answers from a structured knowledge base. A recent case of this kind was that of a collaboration with a pharmaceutical company and a hospital to provide information to diabetic patients or to people who care for such patients. The system can provide information on topics of daily interest such as diet, or certain types of behaviors, the effects of certain types of drugs, the signals for which a person should be alarmed or not. All this was done by creating a dataset in which the answers to be provided to the customer were associated with questions classified according to the user's intent.

Interacting with customers and involving them in the development of applications is an approach that CELI finds very fruitful, because it allows to make the model not only more precise, but also more transparent and understandable to the client. A close collaboration between developer and client benefits both the product and the process and allows to raise the clients' awareness about the potential and limitations of the systems.

Automatic text analysis has also played an important role in the advent of neural models in the last ten years. Many of the traditional activities carried out on text, such as normalization, basic linguistic analysis, morphological analysis, recognition of headwords or analysis of a higher type, such as emotions and feelings recognition, identity recognition, and now classification, have benefited greatly from the introduction of these technologies.

Bernardo Magnini introduced Maria Palmerini, Speech Recognition expert at Cedat 85. About 10 years have passed since Apple launched Siri, one of the first virtual assistants on smartphones. There has been a long way since then and language technologies have played a very important role in this development. Speech was perhaps the first area where performance benefited from neural technologies, and this then paved the way for many developments.

Cedat85 is an Italian company that began working on the Italian language and then opened up to many other languages. Automatic speech recognition consists in transferring, decoding speech contained in an audio or a video into a written text. The fields of application of automatic speech recognition are numerous, some more famous than others.

The first field of application is that of producing a transcript in all cases where it is necessary to document the spoken word in written form, therefore in all the various sectors of reporting: political, judicial, trial, conference reporting, verbalization of various types, transcription of interviews, university lectures, interrogations and so on.

A lesser-known aspect is that speech recognition is itself an *enabling technology*. A whole range of technologies normally applied to writing can be applied to automatic transcripts of spoken speech, such as text processing, information retrieval, speech analytics (for example in the field of call centers), sentiment analysis, keyword spotting, and more generally information extraction.

Nowadays, the technologies are largely based on neural networks, however they are still using statistical technologies.

In order to make a correct use of technology, however, it is important to understand its limits.

When working with speech, the limits are those given by the nature of speech itself, which is the linguistic variety most exposed to external inputs. The challenges to be faced every day are, on the one hand, acoustic challenges (therefore managing audio collected in the most disparate environments and in very different acoustic conditions: microphone audio, telephone audio, band-restricted audio, ambient audio, etc.). On the other hand, there are difficulties that are related to the very nature of speech, first the variety of speech on the diatopic level and therefore the phonological diversity in the production of regional pronunciation. Italy is an extraordinary example of this linguistic variety. Speech also manifests a very marked changeability on the lexical level. It is the first variety that incorporates not only the different types of pronunciation and accent, but also the neologisms and the different linguistic creations. For those who work on spontaneous speech, this poses a challenge in terms of lexical coverage and management of vocabularies and requires continuous updating and maintenance of the models.

The last speech was delivered by Professor Guido Vetere from Guglielmo Marconi University of Rome on the topic of semantics.

Guido Vetere proposed a reflection on two aspects related to the techniques that have produced this advancement in language technologies to date. The first aspect is of a philosophical nature: neural networks and deep learning killed the linguistic sign, in its Saussurian dimension of signifier and signified. In fact, today's artificial intelligence takes up a distributional approach, typical of American linguistics of the early 1900s. Based on the distributional hypothesis, words occurring in the same contexts presumably have a similar meaning. The distributional hypothesis, with a strong emphasis on statistical methods, was functional to the linguistics of that period which studied Native American languages that lacked writing, grammars, or vocabularies.

However, the examples of bias introduced by neural models clearly demonstrate how it is not entirely possible to do without linguistic knowledge, which must also be reintroduced in relation to the cultural conventions of our communities.

The second aspect relates to the knowledge divide which is determined by the current computational techniques used in natural language processing. The resources and algorithms that make these performances possible are the prerogative of big techs that have enormous computing and energy capabilities. This favors the concentration of technologies in the hands of a few actors and a division between those who have knowledge and technologies and those who do not, as also reported by UNESCO in 2005.

Italy too runs the risk of falling off on the side of those who do not have these resources. It is important to raise the alarm on this concern, including towards political decision-makers.

Malvina Nissim commented that: in her experience, companies are investing heavily in trying to reduce the costs associated with the development and use of computational models currently in vogue.

Bernardo Magnini asked Marco Turchi whether, in the field of machine translation, the collaboration between the machine and the user of the translation is possible or if the complete replacement of the human by the machine is foreseen. Marco Turchi replied that at this stage, a lot of work remains to be carried out by translators to correct machine translation systems. Unfortunately, this type of information on a technical and scientific level is not exploited: in the future the interaction between man and machine should be made stronger, not to replace the professional translators but to make these systems more supportive of the work of the human. As long as these two worlds remain separate, we will lose many opportunities, from both market and research perspectives.

The session continued with a final round of speeches describing the future of technologies in the various fields.

According to Marco Turchi, the field of machine translation in Italy needs to create strong synergies between the excellences at university, industrial, and research levels: a real national infrastructure in which new researchers, future professors, and new professional translators are trained. This work is essential if we want to create the necessary conditions to confront ourselves with the big industrial players and at the same time to develop the collaborations and opportunities that will allow technology to progress further.

Alessio Bosca agreed with Marco Turchi and pointed out the development of audio resources as one of the sectors on which to intervene in the future. In fact, the neural transcription experiences highlight the need for vast and heterogeneous public datasets, close to real data in terms of quality. The Mozilla Common Voice dataset, for example, is too clean an audio type to be used to train systems that then process real speech. Beyond that, multi-shift annotated dialogues would also be needed. Bosca concludes by mentioning the publication of the EVALITA[1] campaign resources in ELG[2] (European Language Grid), a very useful initiative with a view to sharing and pooling the available resources.

Maria Palmerini underlined three aspects to keep in mind for the future. The first is the **availability and traceability of data**: many corpora collected and richly annotated in the scientific and university fields are very often not available for commercial purposes use. The data sets that are available for sale have very high costs and are often not in line with the needs of the companies that own them. This means that companies end up having to produce data in-house, with enormous difficulties and investment of time and resources.

The second aspect is related to new technologies literacy: in Italy there is still a conspicuous digital divide, both at the level of individual users and between producers and possible users of new technologies. It would be very useful if a new professional figure was developed to **help dialogue between users and producers**, both to train users in the use of new technologies and to bring user needs to producers. It is very important that new technologies are perceived as supporting production processes and as a means for developing new, more qualified forms of work. For their part, manufacturers have an acute need to know the needs of users to develop technologies optimized for market scenarios and workflows.

Finally, Guido Vetere recalled how there are recent results that show how embedding, that is the vector representations of words extracted from the English WordNet, are competitive with Bert-like ones. The advantage of extracting these embeddings from a WordNet is clear: they require modest computational resources, are based on knowledge rather than correlation, and are easy to modulate. Specialized technical modules can be inserted easily, and they eliminate the problem of bias at the origin: if there is a synonymy that we do not like, we remove it seamlessly. On the other hand, these techniques require WordNet - or with a more modern name - linguistic knowledge graphs - of excellent quality. These new approaches, which tend to integrate the empirical perspective of data with the lexicographic, rationalistic perspective of vocabulary knowledge are promising and can provide an even more balanced development perspective for this type of technology.

Question and Answer session

Q: In Italy, are the products in the field of machine translation competitive with respect to the offer of the major international players?

Marco Turchi replied: Absolutely. In Italy, companies are currently implementing techniques that make these systems not only competitive with large international players, but in many contexts

[1] http://www.evalita.it/

[2] https://www.european-language-grid.eu/

even superior to them It is important to remember that large translation systems are "general purpose", in the sense that they are excellent for translating generic content but don't perform so well in specific domains.

A comment from the public led Maria Palmerini to deepen the link between speech recognition technologies and the forensic field. A dialogue with the forensic world for technological innovation is desirable and would benefit to the entire justice system, by improving the processes, making them more fluid and also more reliable.

Moreover, CEDAT took part in 2010 in an experiment organized by the Ministry of Justice and the Consiglio Superiore della Magistratura (High Council for the Judiciary), for the introduction of speech recognition technologies in the courts. The experimentation was very successful but unfortunately there was no follow-up. It would be important to be able to develop this aspect.

3.4 The CEF AT platform

The CEF AT platform was presented by Markus Foti (Directorate-General for Translation, European Commission). He began his presentation by introducing the evolution of the EC's machine translation system from the statistical to the neural paradigm.

The target users of the CEF AT platform (in particular CEF eTranslation) are:

- Translators and staff of the EU Institutions
- Digital services of the EU Institutions
- CEF Digital Service Infrastructures
- Public administrations in EU Member States
- Universities
- European SMEs

The system that has been trained on a huge database of translated official EU texts performs very well for the translation of formal EU language. For non-standard or creative texts, the quality of the translations may not be so good. To reach an acceptable quality level, it is important to select the appropriate domain-adapted engine according to the text type to be translated.

Markus Foti stressed the importance of Automatic Translation (AT) when the number of translation requests has become so high that it cannot be fulfilled by human translators. Some EU websites use AT to convey their message to a wide public: for instance, Re-open EU, which provides information on the various anti-Covid measures in EU countries, and the Conference on the Future of Europe website, where people can say what kind of Europe they want to live in.

Other services are available on the CEF AT platform such as multilingual tweet, named-entity recognition, and speech to text. The latter is based on a customized Microsoft cloud, although an in-house solution would be preferable for the digital sovereignty.

Regarding the future development of the CEF eTranslation service, Markus Foti noted that the EC is working on developing more language technologies, such as topic modelling, anonymization, language recognition, summarization, and a new web interface.

Finally, he presented the links to self-register and use eTranslation:

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

As for the audience questions, Ms. Claudia Foti from the Ministry of Justice noted that two new specific domains were recently added to eTranslation (one for the Deutsche Bundesbank and one for the French Ministry of Finance). She asked how much data was required to create these domains and if Italian public administrations could also have specific domains on eTranslation.

Markus Foti replied that a finance domain is being developed on eTranslation at the request of the European Central Bank. National Central Banks were required to send as much data as possible and the Bank of Italy provided about 500,000 sentence pairs, so there should be a specific finance domain for Italian shortly.

3.5 Language technologies by/for the public sector and the SMEs (Panel session)

The panel session addressed the LT demand side. The panel was moderated by Marina Omiccioli, Coordinator at the Language Service Division of the Bank of Italy and ELRC Public Sector National Anchor Point. The panel discussion aimed to touch upon the demand of the Italian public sector for LT-enhanced digital services, the related need for specific procurement procedures and the demands of SMEs for automated translation services.

Marina Omiccioli introduced the panelists:

- **Barbara Altomonte**, Italian Presidency of the Council of Ministers,
- **Paolo Cappelli**, Lieutenant Colonel, Ministry of Defense
- **Fiorenza Manfroi**, Falcade Dolomiti, a tourism promotion company.
- **Mauro Draoli**, Digital Italy Agency (Agenzia per l'Italia Digitale - AGID).

The representative of the Italian Presidency of the Council of Ministers, Barbara Altomonte reported on the pilot experience in integrating eTranslation into an institutional website. With the support of Paola Rizzoto and Markus Foti, the Presidency of the Council of Ministers set up the institutional website of the Department for European Policies as a multilingual website that is available in five languages of the European Union (www.politicheeuropee.gov.it/en/). They created the website according to the Guidelines for Accessible Information promoted by the Digital Italy Agency. The selected platform for the new website turned out to be the same as the one used for deploying the website of the Italian Presidency of the Council of Ministers. The project lasted a few months and consisted of two steps: (1) the technical integration and (2) the editorial and linguistic phase. A multilingual editorial team was created with the support of the "Scuola Superiore of Modern Languages for Interpreters and Translators" (SSLMIT). SSLMIT has an extensive history in training translators and interpreters. While translating the website's contents, all recognized standards were applied.

The website of the Department for European Policies has been included in a list of model websites in the IV National Open Government Partnership’s Plan. Accordingly, the technical project and all the stages of the project are available to any Italian public administrative body. Any public administrative body can draw on those documents and on the resources that the European Commission made available and use the eTranslation machine translation platform for its own institutional website. In order to support the ELRC project, the Presidency of the Council of Ministers made all the Italian text and their multilingual translations available to CEF-AT for training the eTranslation system. The IV National Open Government Partnership’s Plan will be active for one more year.

In Italy, there is no coordination among editorial and translator experts working in different Ministries and Departments. Only a few administrative bodies formally recognize the role of *language officer*. There is no official network among people working in different Institution in the field of public communication and translation. This constitutes a major problem because the human element is a key element if you want to help people to better understand legal and administrative texts or drafting. To effectively communicate to citizens in different languages, we should reinforce a network among experts of institutional Italian to promote the production of texts in good and plain Italian, that can be easily understood by the public at large. Those texts can be easily translated and cross-checked. To move through a better institutional Italian, we need to identify actors to involve in an active form of coordination, devoted to enriching the systems of machine translation as eTranslation with natural Italian and other natural languages.

The representative of the Ministry of Defense, Paolo Cappelli, highlighted some of the limits machine translation systems. He pointed out the peculiar vocation of the Ministry of Defense, that is not so keen to address a public audience, although there are publications that are available to the public. Those publications represent 3% of the total documentary production and can be made available through the Ministry of Defence’s website in several languages to those who need them. The remaining documents are not intended to leave the perimeter of the Ministry of Defence for obvious security and confidentiality reasons. This poses limitations to the use of machine translation engines. That's because basically the word “cloud” in fact means someone else's computer that hosts data and applications that are made available to us, but we don't know how our documentary data travel before they get to those clouds, and we don't know who interacts with that computer. So, it is never possible for us to entrust the machine translation engine with a speech from the Chief of Staff because we would be providing information that should remain within the network of the Ministry of Defence. The expert of the Ministry of Defence explored the possibility of importing these translation engines, within the perimeter of Difenet, the Ministry’s internal network, but they registered an important shortcoming: those engines make external calls to the different sites and servers, attempting to access external resources outside of Difenet. That is obviously not acceptable for the Ministry of Defence.

Lieutenant Colonel Paolo Cappelli referred to the German *Bundessprachenamt* as a good example of networking among language experts. The German *Bundessprachenamt* is a large agency for language services, which provides language training for civil servants and carries on all language services translation and interpretation. Therefore, the interpreters of the German *Bundessprachenamt* can work for the different public institutions as needed, within a single working structure.

Italy can look at this and other experiences to create its own coordination scheme, probably at ministerial level. This would also allow translators to better harmonize different issues, such as terminology and the naming of offices and Ministries which is something that may seem trivial, but in fact is continuously changing.

The representative of the small and medium-sized enterprises (SMEs), Fiorenza Manfroi, highlighted the usefulness of machine translation for accelerating the release of translation, and appreciated the confidentiality of eTranslation. Ms Manfroi is marketing chief at PromoFalcade, a company promoting Falcade, its skiing and summer holidays resorts. Mastering foreign languages is essential for tourism. Touristic companies are constantly engaged in relations with foreign markets and need to communicate clearly, directly, and precisely in order to avoid misunderstandings. A correct translation of their contents reflects the quality, care, and professionalism of their efforts to deliver high quality products. The daily activity of a touristic promotion company results in translating various texts: hotel contracts, travel agency agreements, descriptions of accommodation facilities, social posts, websites.

The creation of the new Promofalcade website provided an opportunity to learn about the eTranslation machine translation system. The website has already been published in Italian, and currently Promofalcade is translating contents into other languages. Using eTranslation helped Ms Manfroi to overcome some skepticism about the machine translation tools. So far Ms Manfroi had used those systems mainly to translate from languages she did not master, simply to get an idea of the content of the message. The use of eTranslation proved to be simple: in a few minutes the output file was available, in the same format as the input file, carrying the Italian content that had been copied from the website. Manfroi evaluated positively the quality of the translation obtained, although post-editing was required to fix some expressions that belong to the technical touristic terminology. Ms Manfroi also appreciated the privacy and confidentiality granted by eTranslation. She then called for a broader promotion of eTranslation and committed to presenting the tool to her colleagues in the touristic sector. Finally, she called for continuous improvement of the tool over time.

Ms Manfroi also made a request: namely the possibility of directly translating entire sites. In fact, SMEs interested in translating their website could take further advantage of this tool if they could incorporate it into their website and obtain a machine translation without having to copy contents into a file and then reload them into the site, further reducing the time to allocate to this activity. For all SMEs and for SMEs working in the touristic sector in particular, it is important to have an easy access to reliable, high-quality and free support tools. Although outsourcing translation services can be expensive, Ms Manfoi expressed her personal concern that the use of these tools may shrink the workload of translation professionals.

The representative of the Italian Digital Agency Dr Mauro Draoli confirmed that public administrations need to use services based on NLP technology and artificial intelligence. This demand is probably broader than it might appear at first glance, as these technologies are sometimes embedded in many different systems. So, there is a widespread need to purchase services of this kind. Purchases of such innovative technologies and services are, of course, subject to the Italian Public Contract Code, and in most cases imply running a public procurement. All public administrations generally purchase pre-established products and services. If the characteristics of those services are well known they will be purchased through contracts, defined as ordinary contracts by AgID. These contracts are based on a specific list of requirements for a service whose characteristics are identified in the technical specifications, and which may also be purchased on the lowest price basis. The purchase of services involving NLP technologies and artificial intelligence, however, is different, because it involves new technologies and services, which have recently been introduced on the market and have rapidly evolving performances that are changing over time. Often the buyer is not in a position to provide a detailed description of what they intend to buy.

AgID pointed out that the Italian Public Contract Code defines a specific procurement category called ‘innovation partnership’, when the contracting authorities and the contracting entities need to develop innovative products, services or works and to subsequently purchase the ensuing supplies, services or works. This includes pre-commercial procurement, specifically designed to enable the public purchaser to make purchases, by negotiating with suppliers, possibly by recruiting multiple suppliers and promoting among them a call for competition, as well as organizing subsequent stages of screening tests while the partnership is active. These partnerships allow the identification of tools that deliver the best performance in specific a language context and in the document domains in which the application is intended to be used, since it is impossible to establish a priori the performance and the level of satisfaction that this new technology can offer.

Dr. Draoli also pointed out that the Italian Government has tried to help the Italian administrative bodies to create those original purchasing partnerships in recent years, including through pioneering activities. More information on this subject can be found at appaltiinnovativi.gov.it. On the website there is a collection of case studies. An additional initiative, called “Smart Italy'', is currently being developed. This is one of the major European programs to support public administrations, that propose innovative, original and, by their very nature, risky purchasing themes.

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

The panel session addressed the existing practices and challenges in language data creation, management and sharing. The panel was moderated by Simonetta Montemagni, Director of the Antonio Zampolli Institute of Computational Linguistics (National Research Council) and former ELRC Public Sector National Anchor Point.

The panel discussion aimed to touch upon creation, management and sharing of language resources, the three macro steps in the life cycle of language resources.

Ms Simonetta Montemagni stressed that artificial intelligence systems in the field of language technologies, in particular in the field of machine translation, are trained by our own language productions. The Italian language is one of the richest languages in terms of linguistic resources available for training machine translation systems. However, some translation results still remind us of the need to identify more relevant data for specific domains. Ms Montemagni also stressed the active role of the ELRC initiative in collecting linguistic resources (multilingual corpora, public administrative monolingual corpora enriched with language annotations and many multilingual lexical resources and translation memories) to train eTranslation in specific domains. A new eTranslation domain for finance is about to be released, trained with relevant resources provided by the Bank of Italy. During the panel discussion, Mr Fabrizio Calabrese from the Bank of Italy acknowledged Dr Marcus Foti for the effective cooperation in the deployment of this domain, stressing that the Bank of Italy values eTranslation as an important instrument and will keep contributing to its training process.

The translation practices currently adopted by the Italian public administrative bodies are described in the ELRC White Paper issued in December 2019, which details the situation in all countries. In Italy, translation is a process that can either be carried out internally, by a unit devoted to translation services, or outsourced to external translation agencies. **Computer Assisted Translation tools are rarely used by public translation units** and those instruments are still perceived as **expensive software**, even in terms of training. This evaluation does not consider their positive impact in terms of both quality and efficiency of the translation process. Sometimes CAT tools are also perceived as systems that could potentially replace translators, rather than tools supporting their activities.

The representative of the Bank of Italy, Dr Fabrizio Calabrese, Head of Language Service Division, stressed the importance of language and linguistic data for institutions in general and for the Bank of Italy in particular. Institutions act through language, issuing administrative acts, policy communications, scientific analyses, and various communications to citizens, which form the basis of their accountability. For this reason, there are dedicated language services units which also take care of managing linguistic data.

Dr Fabrizio Calabrese used a railway wagon as a metaphor to express the issues entailed by linguistic data management. Any railway wagon has codes identifying its characteristics. Rail transport requires the proper coverage of wagons according to the type of goods, similarly **linguistic data require a proper format to grant data integrity**, i.e., to ensure the availability of complete, reliable and up-to-date data. The use of CAT tools solves many problems, but databases can also be enriched through large amounts of data collected elsewhere. **Data sharing also depends on compatibility of formats**, in the same way as wagon size depends on track gauge. It is important to identify and **share common standards** in order to be able to share data across different institutions' databases and create a shared knowledge base. Data format conversion is key: in this respect, although proprietary systems can be useful, **it is always better to be able to reverse data in an open-source format**, for exploiting them at length. An additional internal management concerns the ability to deploy computer systems with the right physical capacity: both the allocation of space and the speed of the network affect the data refresh rate and response capacity during the translation process.

Finally, Dr Calabrese highlighted the importance of data confidentiality for central banks, echoing Dr Cappelli from the Italian Ministry of Defense. Translated texts embed valuable language knowledge that can be useful for forthcoming translations or for training automatic translation engines (such as eTranslation), but strict rules prohibit the use of data conveying confidential information. To **keep confidentiality, it is important to identify software solutions for anonymization or pseudo-anonymization**. This software is essential for granting full access to a whole knowledge base, both externally and internally, and to use data for training machine translation engines such as eTranslation.

The representative of the Italian Digital Agency Gabriele Ciasullo referred to Law Decree No. 76/2020, known as the “Simplification Decree” and containing “Urgent Measures for Simplification and Digital Innovation”, published in the Official Gazette in 2020. The Law emphasizes the importance of public administrative data and the need to share those data for institutional purposes. The decree imposes an obligation on the President of the Council of Ministers to adopt a national data strategy. AgID is involved in the drafting of this document, along with the Department for the Digital Transformation, the support structure for the Italian Minister for technological innovation and digitization. The European Strategy for data highlights some criticisms including: (i) the issue of confidentiality, arising from the increase in the amount of available data, allowing for more data correlations; (ii) the need to safeguard SMEs, that proved less performant than larger companies in exploiting the potential of data, probably because of their limited computational capacity.

At the European and national level, many analysts urged the public sector to establish a stable data governance. Data governance is intended to serve two different goals: (1) enhance data sharing among different public administrative bodies, to boost the efficiency of public administration; (2) facilitate the re-use of data by companies as envisaged by the open data policy. In fact, open data policy can have a positive impact on economic activity, increasing employment and providing new business opportunities. Dr. Ciasullo advocated for **stable data governance** in a Parliamentary hearing and in the drafting of the National data strategy. Data and services produced by different administrative bodies interact with each other: data provides services to businesses and citizens and vice versa, services provided produce further data.

Metadata management is based on semantics. In Italy, data description is widely provided according to European standards. Specific data management, for instance in the field of geographical data, widely respect European standards, as well. **Metadata shall converge in national catalogs** to reduce the effort for searching for existing data: the larger the amount of available data, the stronger the need for catalogs that allow stakeholders to search among data. In the last few years, a registry for geographical information, similar to the Metadata Registry of the Publications Office of the EU, has been in production. This tool provides controlled vocabularies and taxonomies. The Italian Registry participates in the European INSPIRE Federation. A new instance of this tool will be implemented. The tool is intended to manage in close cooperation with the Publications Officer of the EU any extensions to existing vocabularies. This new instrument is expected to help public administrative units to manage language issues, as well.

The representative of CLARIN-IT (Common Language Resource and Technology Infrastructure), Dr. Monica Monachini, presented the advantages that CLARIN offers to the ELRC initiative, providing a safe place to host language resources: digital linguistic data, both written and spoken, both monolingual and multilingual, as well as procedures for automatic processing and analysis of data. CLARIN collects data relating to different disciplines in the field of human and social sciences, whose subject of study is language. Additional data from public administrative bodies would close an existing gap and enrich the CLARIN community and data. CLARIN has a distributed infrastructure and is a federation of 68 data centers situated in 21 European countries. Many of these centers are also ELRC centers.

CLARIN centers provide digitized language data as well as NLP tools for automated analysis, enrichment, and data extraction, regardless of where data is retained. CLARIN provides access to researchers through a single sign on system. In Italy, CLARIN-IT is one of the three centers currently certified by an external body with the SIL co-trust. With the rise of FAIR (Findable, Accessible, Interoperable, Reusable) Data and open science paradigm, many Italian organizations are looking for open and secure institutional archives to be entrusted with their data. CLARIN-IT ensures that data are findable, accessible, interoperable, and reusable. The CLARIN-IT data center services provide data description and grant that data are findable by applying a harmonized set of metadata descriptors. CLARIN-IT collects this metadata and transfers them into a CLARIN central catalog, greatly increasing data visibility. NLP services provide the possibility of analyses, increase, integration, and enrichment of data. In addition, the CLARIN-IT Center identifies all conferred data through a permanent identification system, ensuring their persistence, reuse, and versioning. Finally, CLARIN-IT center uses a licensing system which ensures the enforcement of the desired access restrictions.

Dr Monachini also highlighted the concept of interoperability. CLARIN has a strong commitment to formats and standards preservation. CLARIN recommends that both researchers and organizations use a common set of standards for annotation and encourages the scientific community to comply with these standards, recognized by ISO and the Test Encoding Initiative. Standard formats grant that data and tools provided by different data centers are fully interoperable. Dr Monachini defined **CLARIN as a bank for data, where data is secured, and its value can be increased**. In fact, by publishing data on a CLARIN data center, data becomes part of an ecosystem that increases their value.

3.7 Conclusions

Claudia Soria briefly thanked all the attendants for their participation. Special thanks were conveyed to the speakers for the enthusiasm with which they accepted the invitation to participate and for the support they gave to the workshop. This workshop is a small step towards the integration between supply and demand, between academia and research on the one hand, industry, and users and the public and whoever uses these technologies. Years ago, when we were invited to take part in the ELRC initiative, we didn't even know where to turn. Thanks to ELRC, there is now an open channel and an active dialogue between LT providers and users. It is worth to invest time and energy on this open channel so that to start creating an infrastructure between LT supply and demand.

Special thanks go to Marina Omiccioli and Roberta Persia, without whom all this would not have been possible. Also, we are grateful for the support brought by Eileen Schnur, Stefania Racioppa and Andrea Lösch, the ELRC project manager. We also extend our thanks to our colleagues from ELDA, Khalid Choukri and Hélène Mazo for the help and support that they have given us all this time. Finally, we thank the interpreters very much for their precious work, and the technicians who also helped us to allay the panic in the hours preceding this event.

3.8 Break-out rooms

At the end of the workshop, the participants were offered the additional opportunity to interact directly with some of the speakers to ask them specific questions and build relationships. Seven virtual rooms have been activated, for each of the following institutions or initiatives:

- Fondazione Bruno Kessler (Marco Turchi)
- Cedat85 (Pierpaolo Barnaba)
- CELI (Andrea Bolioli/Alessio Bosca)
- CLARIN-IT (Monica Monachini)
- Banca d’Italia (Fabrizio Calabrese)
- AgID (Gabriele Ciasullo)
- CEF AT (Markus Foti)

4. Synthesis of Workshop Discussions

4.1 State of the art

- There are companies in Italy that are currently implementing techniques that make MT systems not only competitive with large international players, but in many contexts even superior to them.
- MT remains one of the most important technologies because of the high number of translation requests that cannot be satisfied by human translators.
- Public Administrations need to use services based on NLP technology and artificial intelligence. It may not be visible, however, these technologies are sometimes embedded in many different systems that are used in the public services.
- As far as SMEs are concerned, MT can be used safely and fruitfully. In a business context, it is important not only to convey information in different languages accurately, but also to showcase a trustworthy and professional image of the enterprise. A correct translation of their contents can reflect the quality, care and professionalism brought to deliver high quality products. The variety of texts at stake is considerable and include product descriptions, contracts, agreements, manuals, social media posts, all requiring appropriate genre and style adaptations.
- A national Data Strategy is currently under development.

Remaining Issues

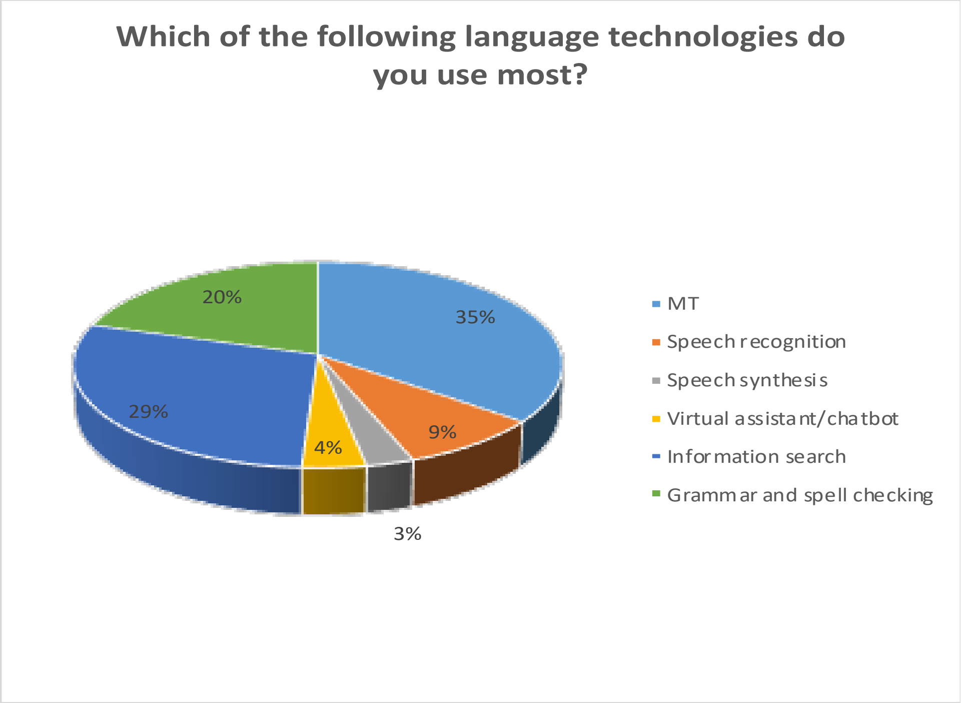
- One of the biggest issues lay in data availability.
- At present, the use of MT can be hindered in domains where both privacy and security is an issue.
- In Italy there is no coordination between editorial and translator experts working in different Ministries and Departments. Only a few administrative bodies formally recognize the role of *language officer*. Also, there is no official network among people working in different institutions in the field of public communication and translation. This constitutes a major obstacle to exchanges between people that could foster a better understanding of legal and administrative texts or drafting.
- Public Administrations need NLP tools and services, but these cannot be purchased according to the ordinary public procurement rules.
- **Anonymization or pseudo-anonymization is the bottleneck for any use and re-use of data by and for Public Administrations**. Translated texts embed valuable language knowledge that can be useful for forthcoming translations or for training automatic translation engines (such as eTranslation), but strict rules prohibit the use of data containing confidential information. Therefore, it is of utmost importance to identify software solutions for

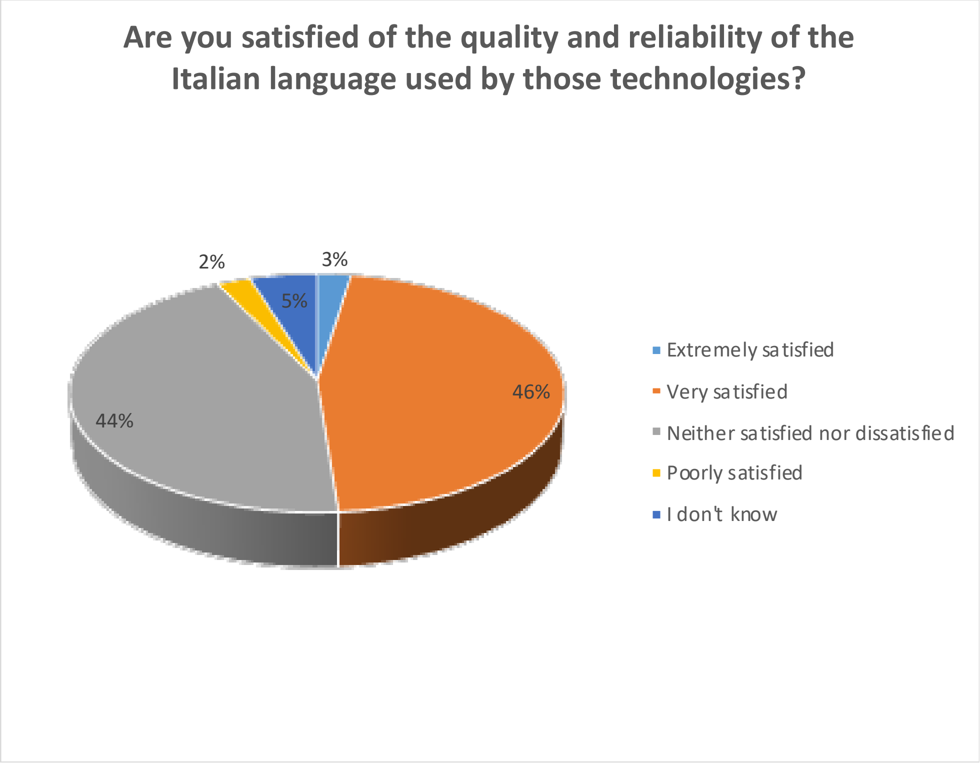
anonymization or pseudo-anonymization to guarantee the confidentiality and to be able to exploit the data produced by Public Administrations.

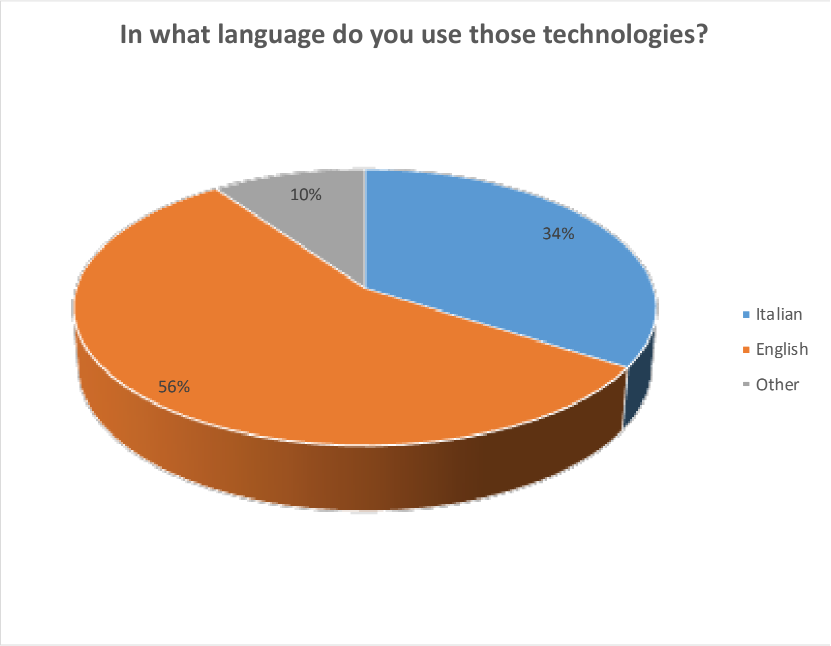
Expectations and futures needs

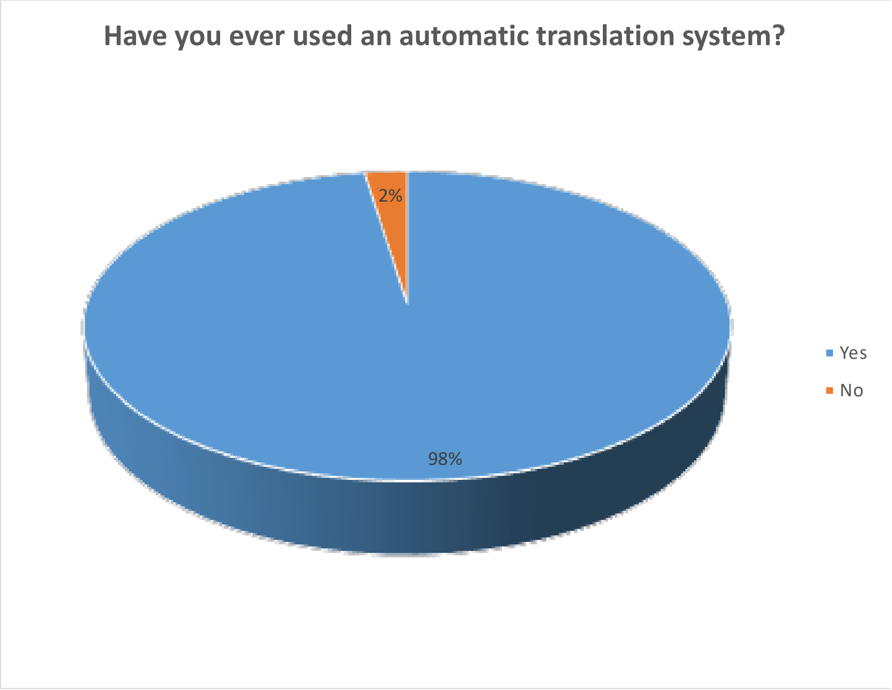
- A closer connection should be established between speech recognition technologies and the forensic field. There is the need for a dialogue about technological innovation within the forensic domain. Trials could benefit from such technology.
- In order to improve communication towards citizens in different languages, we should reinforce a network of institutional Italian experts to help promote the production of texts in good and plain Italian that can be easily understood by the public at large. Those texts can be easily translated and cross-checked.
- The presence of professional figures to help dialogue between users and producers would be very useful.
- A good example of networking among language experts is the German Bundessprachenamt, which is a large agency for language services, which also provides language training for civil servants and carries out translation and interpretation services for all languages. A similar coordination scheme could be put in place in Italy. This might be achieved at the ministerial level and improve the harmonization work for translators such as the terminology and the names of offices and Ministries.
- The Italian Public Contract Code defines a specific procurement category called ‘innovation partnership’, when the contracting authorities and the contracting entities need to develop innovative products, services or works and to subsequently purchase the ensuing supplies, services or works. This includes pre-commercial procurement, specifically designed to enable the public purchaser to make purchases, by negotiating with suppliers, possibly by recruiting multiple suppliers and promoting a call for competition among them, as well as organizing subsequent stages of screening tests while the partnership is active. These partnerships allow the identification of tools that deliver the best performance in a specific language context and in the document domains in which the application is intended to be used, since it is impossible to establish a priori the performance and the level of satisfaction that this new technology can offer.
- Linguistic data is the backbone for all NLP tools and services. Data sharing must be encouraged, and education must be provided regarding data formats and formatting standards.
- A stable data governance is required in the public sector.
- **Metadata shall converge in national catalogs** to reduce the effort for searching for existing data.
- Linguistic variation must be integrated and represented in the data used to train systems **to avoid biased results and enhance the performance of NLP tools.**
- Research and development in language technologies must be driven by a vision of LTs as supporting and integrating human work, not substituting it.

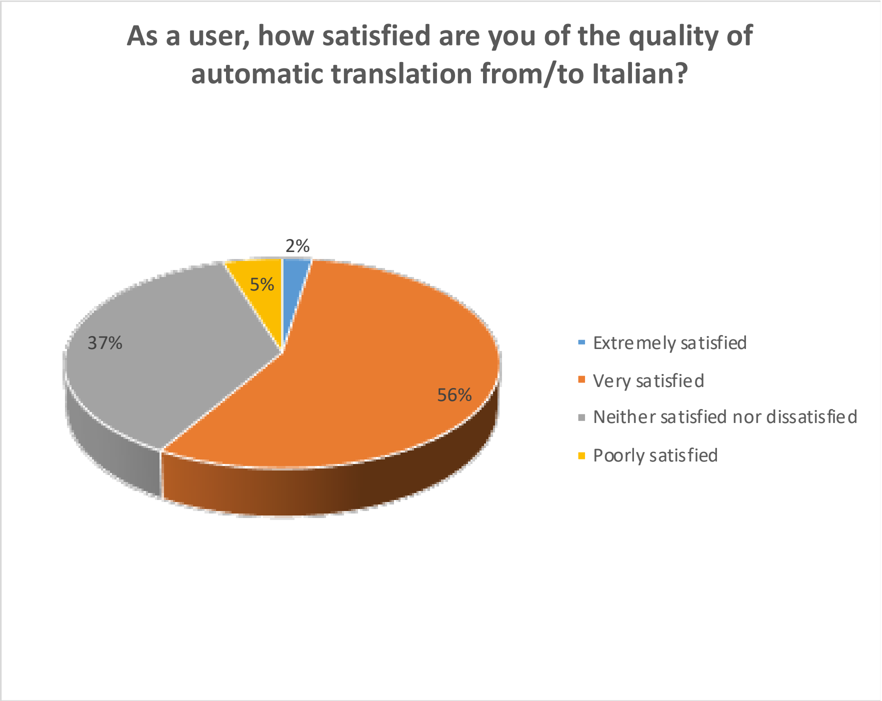
4.2 Results of polls



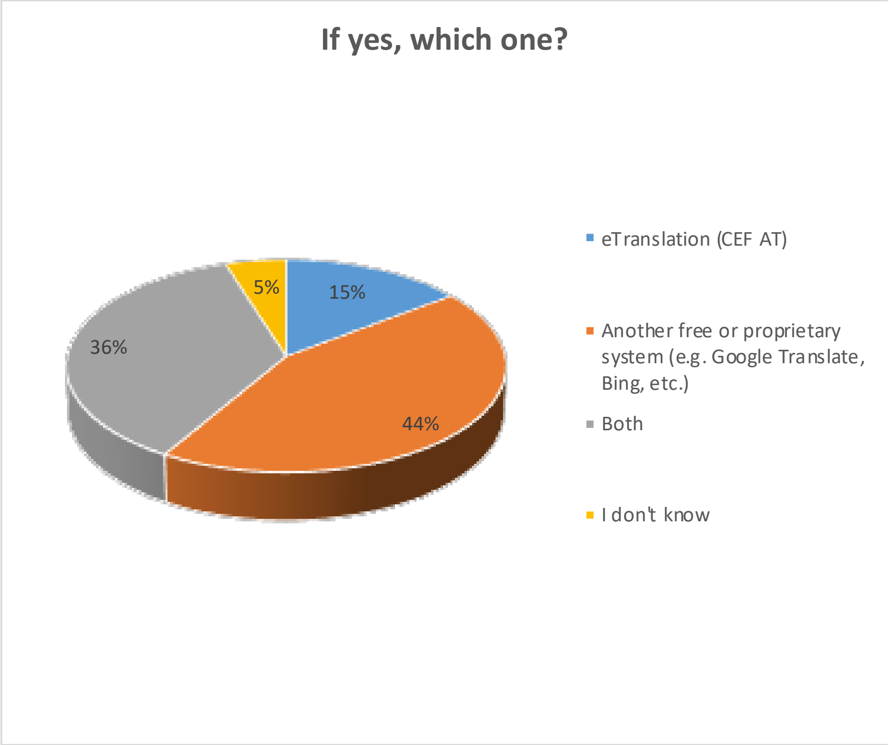



In what language do you use those technologies?
10%
34%
56%
Italian
English
Other

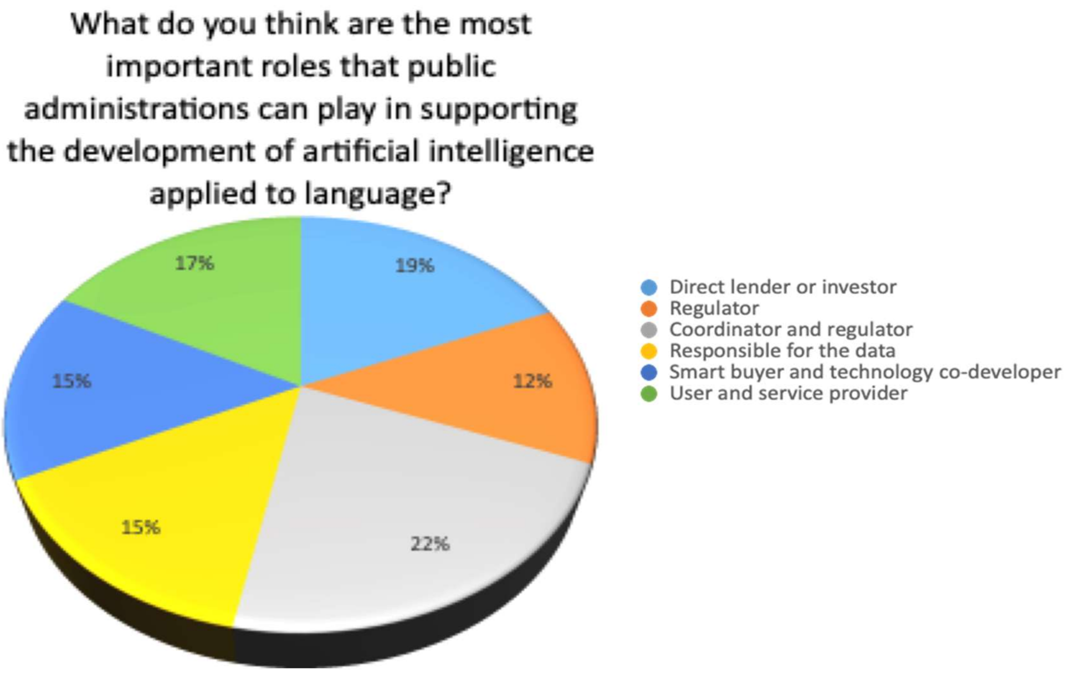

Have you ever used an automatic translation system?
2%
98%
Yes
No

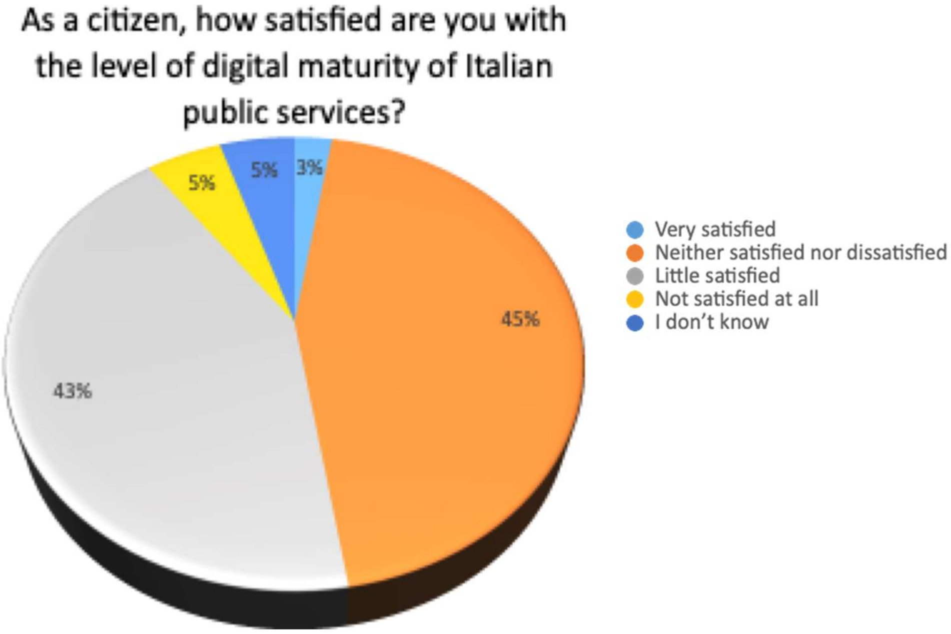


As a user, how satisfied are you of the quality of automatic translation from/to Italian?



If yes, which one?


What do you think are the most important roles that public administrations can play in supporting the development of artificial intelligence applied to language?
19%
12%
22%
15%
15%
17%
● Direct lender or investor
● Regulator
● Coordinator and regulator
● Responsible for the data
● Smart buyer and technology co-developer
● User and service provider


As a citizen, how satisfied are you with the level of digital maturity of Italian public services?
3%
45%
43%
5%
5%
● Very satisfied
● Neither satisfied nor dissatisfied
● Little satisfied
● Not satisfied at all
● I don't know

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

The main challenges Italy is facing in sharing data and restructuring the translation workflow to make it more efficient can be grouped in two categories:

- Public perception of language data
    - Awareness of the importance of language resources for machine translation and other applications of artificial intelligence is growing steadily.
    - Language data is increasingly considered a valuable resource and regulation of its management is starting. An appropriate language data management structure is still lacking at the institutional level.
    - Permanent education re. data sharing and open data is needed in order to increase willingness to share translations.

- Structural issues
    - Little knowledge about automatic anonymization tools results in valuable data remaining not shareable, in addition to manual anonymization process being very time consuming.
    - Privacy concerns regarding confidential and personal data (GDPR) are an obstacle for many Ministries (incl. Justice and Interior) to share language data.
    - Strict rules prohibit the use of data conveying confidential information. To keep confidentiality, it is important to identify software solutions for anonymization or pseudo-anonymization.
    - The representative of the Italian Digital Agency Gabriele Ciasullo referred to Law Decree No. 76/2020, known as the “Simplification Decree” and containing “Urgent Measures for Simplification and Digital Innovation”, published in the Official Gazette in 2020. The Law emphasizes the importance of public administrative data and the need to share those data for institutional purposes. The decree imposes an obligation on the President of the Council of Ministers to adopt a national data strategy.

Updated action plan (new items are in italics):

Italy needs awareness raising activities on the value of language data both with translators and decision makers in public administrations. This should be done through examples of how language data management practices can reduce costs and improve quality.

Legal, privacy and ownership concerns should be addressed and best practices in the use of CAT tools and language data management should be developed, preferably by a central body.

The public sector should establish a stable data governance. Data governance is intended to serve two different goals: (1) enhance data sharing among different public administrative bodies, to boost the efficiency of public administration; (2) facilitate the re-use of data by companies as envisaged by the open data policy.

The following specific objectives are suggested to address the challenges Italy is facing when it comes to sharing language data:

- To establish good data management practices in public services
    - Further investigation of data management practices
    - Definition of confidential and personal data that can be used to introduce the practice of clear separation between confidential and personal data from public sector information in the translation process
    - Establish shared language data management practices to reduce costs, improve quality and leverage on existing language assets
    - *Exploit the CLARIN-IT archival facilities as an introductory step to the benefits of centralized language data archiving and documentation.*
    - *Identify and share common standards in order to be able to share data across different institutions' databases so as to create a shared knowledge base.*
    - *Embed anonymization or pseudo-anonymization tools into the data management life cycle, so as to overcome the strict limitations of confidentiality issues.*
    - *Establish a stable data governance*.

- To raise awareness on the importance of language data as a valuable asset and as Open Data
    - Raise awareness on the value chain of language data and the importance of LR
    - Share benefits of sharing language data
    - Enhance the publishing of Open Data making Italy one of the trendsetters of Open Data in Europe
    - Integrate language data in the national Open Data policy
    - Emphasize the role of digital texts in the digital economy
    - Establish practical guidelines for language data as Open Data