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

Deliverable D3.2.8
Task 3

ELRC Workshop Report for Spain

Author(s):	Maite Melero
Dissemination Level:	Public
Version No.:	Vfinal
Date:	2021-07-06

Contents

Executive Summary 3

1 Workshop Agenda 4

2 Summary of Content of Sessions 5

2.1 Welcome and introduction 5

2.2 National Strategy for Artificial Intelligence in Spain, framework of the National Plan for Language Technologies 5

2.3 The European translation platform CEF AT, free to use for Public Administrations and SMEs 6

2.4 Language data management, exchange and sharing (Panel session) 6

2.5 Language technologies by/for the public sector (Panel session) 8

2.6 Language technologies showcase 11

2.7 Conclusion 13

3 Country Profile: Language data creation, management and sharing 14

Executive Summary

This report contains a summary of the presentations and the discussions carried out during the Third ELRC workshop in Spain held online on May 27th 2021. All workshop materials are publicly accessible through The Third ELRC Workshop in Spain website.

It also contains information and materials about participants and the feedback received. Its contents are confidential.

1 Workshop Agenda

9:45-9:55	**Welcome and presentation** Maite Melero (T-NAP) and Alberto Merchante (P-NAP)
9:55-10:15	**National Strategy for Artificial Intelligence in Spain, framework of the National Plan for Language Technologies** Laura Flores - Deputy Director General for Artificial Intelligence and Digital Enabling Technologies - Secretary of State for Digitalisation and Artificial Intelligence
10:15-10:45	**The European translation platform CEF AT, free to use for Public Administrations and SMEs** François Thunus - European Commission
10:45-11:20	**Management, sharing and reuse of language data in Spain: practices and challenges (panel discussion)** Ona de Gibert - BSC (moderator) Marta Villegas - Text Mining Unit - BSC Javier Hernández Díez - Deputy Director General for the Development and Implementation of Digital Services at the Ministry of Justice Raquel Xalabarder - Professor of Intellectual Property - UOC Alícia Ollé - Director of Customer Operations - Parlem Telecom
11:20-11:30	**Coffee break**
11:30-12:05	**Language technologies for the public sector (panel discussion)** Maite Melero - BSC (moderator) David Pérez - Director of Business Development and New Technologies - Segittur David Griol - Full Professor - University of Granada Manuel Herranz - CEF NTEU project coordinator
12:05-12:55	**Language technologies showcase** Vicomtech Pangeanic Inbenta Seedtag 1million bot M47 labs Flaps Botfoundry by Inetum
12.55-13.00	**Conclusion** Maite Melero (BSC)

2 Summary of Content of Sessions

2.1 Welcome and introduction

The T-NAP and the P-NAP welcomed the participants to the 3rd ELRC workshop in Spain. Afterwards, Maite Melero presented the ELRC and the CEF initiatives:

ELRC, a European network with representatives in all EU countries, plus Iceland and Norway, was created with the objective of bringing translation and language technologies closer to the public administrations of European countries and facilitating the reuse of language data. ELRC is supported by the CEF programme, which provides funding for the vision of a digitally connected Europe across administrative, economic and language barriers. A fundamental part of this programme is the support to language technologies, with a specific focus on translation technologies, and their implementation in the public administration of European countries. The CEF programme started in 2014 and has been running for 7 years. Now, on the threshold of this second digital era, the Digital Europe Programme will take over CEF and with a dedicated €7.6 billion budget until 2027, will fund projects in five key areas: supercomputing, artificial intelligence, cybersecurity, advanced digital skills as well as ensuring the use of digital technologies in the economy and society.

In Spain, this programme has had great synergies with the Plan for the Promotion of Language Technologies led by the Secretary of State for Digitalisation and Artificial Intelligence, which began in 2017. On this very year, a novel deep learning architecture appeared[1] and revolutionized, first, the area of machine translation and, later, the whole field of language processing, through pre-trained language models, such as BERT[2] and GPT[3]. Since our last Workshop in 2018 in Madrid, language-based applications have taken a quantum leap forward. And for the first time we are seeing great progress not only for English but for many other languages as well.

2.2 National Strategy for Artificial Intelligence in Spain, framework of the National Plan for Language Technologies

Laura Flores Iglesias, Deputy Director General of Artificial Intelligence and Digital Enabling Technologies at SEDIA, presented the National Artificial Intelligence Strategy in Spain, and the renewed National Plan for Language Technologies:

Spain's National Artificial Intelligence Strategy has seven strategic objectives:

- Scientific excellence and innovation in Artificial Intelligence: to position Spain as a country committed to promoting scientific excellence and innovation in Artificial Intelligence.
- Projection of the Spanish language: to lead the world in the development of tools, technologies and applications for the projection and use of the Spanish language in the fields of application of AI.
- Creation of qualified employment: to promote the creation of qualified employment, boosting training and education, stimulating Spanish talent and attracting global talent.

[1] The Transformer architecture was first described in Vaswani et al., 2017 "Attention Is All You Need"

[2] Devlin et al, 2018 "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

[3] Radford et al, 2018 "Improving Language Understanding by Generative Pre-Training"

- Transformation of the productive fabric: To incorporate AI both as a factor for improving the productivity of Spanish companies, the efficiency of public administration and as a driver of sustainable and inclusive economic growth.
- Environment of trust in relation to Artificial Intelligence: to generate an environment of trust in relation to AI, both in terms of its technological development and its regulatory and social impact.
- Humanistic values in Artificial Intelligence: to promote global debate on the technological development of humanist values (Human-Centered AI), focused on ensuring the welfare of society when making technological advances and developments, creating and participating in forums and dissemination activities for the development of an ethical framework that guarantees the individual and collective rights of citizens.
- Inclusive and sustainable Artificial Intelligence: to promote inclusive and sustainable AI as a cross-cutting vector for tackling the major challenges facing our society, specifically to reduce the gender gap, the digital divide, support the ecological transition and regional cohesion.

2.3 The European translation platform CEF AT, free to use for Public Administrations and SMEs

The European translation platform CEF AT was presented by **François Thunus** from the European Commission's Directorate-General for Translation.

He provided information on the target audiences of eTranslation, the CEF AT machine translation system, its use scenarios, languages coverage, etc. The translation output quality issue was discussed with engaging examples in different text types. Finally practical information on the registration options were presented, putting emphasis on the fact that the use of the platform is now freely offered, not only to public administrations but also to small and medium-sized European companies.

2.4 Language data management, exchange and sharing (Panel session)

Ona de Gibert Bonet, the chair of the session, started with a brief introduction stressing the importance language data as an asset. Then, she opened the panel discussion involving 4 experts on the field:

- **Marta Villegas**, Principal Investigator at the Barcelona Supercomputing Center, as an language data policies expert
- **Raquel Xalabarder**, Professor of Intellectual Property at UOC, as a legal domain expert
- **Javier Hernández Díez**, Deputy Director General for the Development and Implementation of Digital Services at the Ministry of Justice, representing the public sector
- **Alícia Ollé**, Director of Customer Operations at Parlem Telecom, representing the private industry

Although it plays an essential role when building any Language Technology (LT) application, language data is most often not considered as a valuable asset, as stated in ELRC White Paper *"Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe - Why Language Data Matters"*[4]. In fact, Machine Learning systems are trained with large sets of language data through language models such as XLM-R or GPT-3. Also, language data must be of high quality in order to achieve representativeness and avoid possible bias that could eventually affect a LT system. Therefore, it can be difficult to obtain this data which explains the needs for protocols for data sharing

[4] https://www.lr-coordination.eu/sites/default/files/ELRC_Conference/ELRCWhitePaper.pdf

and reusing, an essential step in the development of NLP across domains. Our experts shed light on the current trends regarding this issue.

Marta Villegas started by sharing her point of view on the political aspect of data sharing. There exists an awareness of how important data sharing is, that's why there are many initiatives such as European directives, as well as open data portals at different levels. However, these mostly contain structured data (statistics, lists, numbers…) and we find very few examples of language data, due to lack of awareness of the value of language data. This brings up the issue that we need to teach and spread knowledge with events such as this one to raise awareness about the value of language data. And besides the general misconception that is already an obstacle when sharing data, there exist all the technical problems: the format of the language data, the metadata, the traceability, the protection of personal data, etc. These are only a few of the many challenges we deal with when trying to share and reuse this type of data.

Next, **Raquel Xalabarder** discussed the legal aspects in relation with reusing data. According to her, the regulations both at European and national levels regarding the reuse of information have evolved since 2003. The European Union has made a huge effort to legislate the field of Artificial Intelligence, issuing recommendations, strategies as well as funding projects. However, while the legal aspect is covered, the actual implementation of the directives could be improved. It is also worth noting that the directives don't put particular emphasis on the reuse and sharing of language data. They just mention data, so the legal coverage of text mining could be improved. Furthermore, there exists a big difference on how these laws apply in the public and in the private sectors. The public sector is making an effort to implement the directive to be able to share data, while the decision on sharing data in the private sector depends solely on the owners of the data. There's no regulation regarding the private sector and that's probably the new milestone that should be tackled by the EU.

To further clarify the situation in the public sector, **Javier Hernández** shared his view in this regard. In his opinion, data exchange is not encouraged in the public sector. It is carried out only if mandatory, rather than willingly. And it is regarded as a difficult process that needs many preprocessing steps such as anonymization and decontextualization before being able to share anything in order to legally protect the people involved. Nonetheless, this is something that should be worked out and encouraged, as the data does not belong to the administration, but rather it is a common good that belongs to the people. The public sector can really benefit from collaboration between different actors to work on protocols and tools for data exchange.

As a representative of the private sector, **Alicia Ollé** also took part in the debate and shared her specific use case. Parlem Telecom is a telecommunications company that offers a virtual assistant in Catalan. However, to do so it makes use of real-time machine translation into Spanish as there are no resources available in Catalan to build such an application. Currently, they are willing to share their data to potential developers of technology in Catalan, as their aim is to contribute to the presence of Catalan in the digital world. Collaboration with researchers in the field seems to be the answer to this issue. Yet, sharing their data is proving trickier than expected, as the GDPR protects the data of their users. This shows, in relation to what Raquel said earlier, that even when the owners of private data are willing to share it, the process is not straightforward and still poses unsolved issues.

After the individual presentations, a lively chat ensued about how data sharing could be improved and further fostered. Those were the main points that were raised:

- More financing is required. Even if data exchange is encouraged by the legal framework, there needs to be resources to comply with it.
- Define protocols and release open source tools so that data exchange can be carried out in a systematic, ordered and efficient manner.
- Improve laws regarding text mining, such as the already existing European Directive Exception for Text and Data Mining. This Directive should be expanded to include exploitation of data, not only for research but also commercial purposes. Currently the owner of the Intellectual Property has to explicitly give their permission to use their data.
- Find ways to share non anonymized data based on trust. For that we need to change the cultural perspective regarding data, and put the focus on the common good it brings to society.

These are just a few of the ideas that came up during the panel discussion. The general conclusion of the panel was that there is a willingness to share data but there is still some work to do and it will only be achieved through collaboration between the different actors. Where there's a will there's a way.

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

This panel discussion was moderated by Maite Melero and brought together the following panelists:

- **David Pérez**, current Director of Business Development and New Technologies at Segittur, was the main promoter of the Language Technologies Plan during his time at SEDIA, and was its coordinator from the beginning until a few months ago. Due to his professional experience, he has a great knowledge of the Spanish Administration, as well as of the language technology sector.
- **Manuel Herranz** is founder and CEO of Pangeanic, a company with which he has participated in and coordinated several projects financed by the CEF programme with the objective of creating language technology infrastructures for use by the public administration.
- **David Griol** is Associate Professor in the Department of Computer Languages and Systems at the University of Granada. He has been conducting his research for more than 15 years in the development of conversational interfaces, statistical modelling of dialogue and user adaptation. As part of his collaboration in SEDIA's Plan of LT, last year he developed the Hispabot-Covid 19 intended to answer frequently asked questions related to the pandemic.

After the presentations, and to start the debate, **David Pérez** was asked about the **degree of integration of language and translation technologies in the digital services of the Spanish Administration** and what were the **main obstacles to a more widespread adoption** of these technologies in public administration, if any. He was also asked to provide ideas about possible applications in the tourism sector.

David Pérez started by acknowledging the importance of the LTs, which, in his opinion, are currently one of the most promising technology sectors. He then pointed out that the degree of integration of the LTs in the public services depends on the specific technology. Machine translation, for example, is the LT with a longer history in the public services, and conversational systems are starting to become popular, but most of the other NLP technologies are practically unknown in the Administration, with the possible exception of the occasional NER system or anonymizer. He also made the point that while MT or other LT systems may be used, most often the technology behind is obsolete, and not based on the current neuronal paradigms.

In summary, there is room for improvement in the integration of said technologies in the digital services of the Spanish Administration.

As for the obstacles to a more widespread adoption of LTs, he mentioned the lack of specific training of the civil servants in these novel technologies. Another obstacle has to do with the vertical nature of the Administration since Napoleonic times, with very little transversality throughout departments. For example, there aren't any organisational structures centered around AI and LTs (as an example, he cited the National Statistics Institute which plays a similar role with statistics). The Secretary of State of Digital Administration is emerging as a body that could take up this function. Such a dedicated structure would be able, for example, to validate or certify vital LT tools, such as the ones used for processing medical records, with a role similar to the FDA's the USA.

As for applications that he would like to see being deployed in the tourism sector, he mentioned high quality, multilingual machine translation, conversational systems for a range of domains, as well as active monitoring of the overall sector.

He then made a case for the need of data sovereignty by reminding the importance of the data-based economy, which has already surpassed the oil-based economy, but that so far has been only capitalized by the big corporations. He said that the value of the data held by the public administration is huge. The whole administrative life of the citizens is in their hands, and it needs to be exploited only in the interests of those same citizens, also by reusing it to improve the LTs.

He added that workshops such as this one, able to put together people from different public sector bodies so that they could explain and share their experiences and use cases, are really useful.

After David Pérez's intervention, the moderator asked **Manuel Herranz** about his **experience participating in European projects** where machine translation tools and other language technologies have been developed, aimed at public administrations.

After stating that his experience had been very positive, Manuel described Pangeanic's participation in a number of CEF-funded projects, as a coordinator for most of them. The first project dates back to 2017, it was IADAATPA, which produced MT-HUB, an open-source platform for routing MT engines, and which they have evolved into their own corporate tool. Then came NEC-TM, which published surveys all across Europe on public expenses related to translation services. In Spain, the expenditure amounts to around 25M € a year. Such large public spending has little return on investment since the huge amount of parallel data generated (translation memories) are not exploited for the public benefit. To remedy this, the project produced an open-source TM server and CAT tool.

The next MT-related project coordinated by Pangeanic is Neural MT for EU (or NTEU) which is a farm of translation engines for all language-pair combinations of EU's official languages. The project has produced 552 engines, which are dockerized and are freely available, though ELG. The engines feature direct translation from and to all EU languages, not using any pivot languages. In the administrative language domain for which they have been trained they outperform the equivalent Google translate engine, in all cases. Importantly, NTEU has also produced large amounts of bilingual data, part human and part synthetic.

The last project coordinated by Pangeanic is MAPA which is building a multilingual anonymization toolkit, and which is already working very well in several use cases.

Lastly, Manuel indicated that although their overall experience with these projects is very good, they have mixed feelings towards the receptiveness of the Administration to new technologies. In workshops similar to this one, they have met people working in different ministries expressing interest in the tools but admitting that their own equipment was inefficient and obsolete.

Manuel highlighted the importance of ELRC workshops wishing that there would be more cross-fertilization between them, so that people from different countries could benefit from each other's experiences.

Finally, the third panelist, David Griol was asked to elaborate on his **experience developing language technology solutions for the Administration**, and also to provide his opinion on the **quality of the technology for Spanish and the co-official languages**, and whether this quality is sufficient to meet the needs of public services.

David Griol said that his collaboration with SEDIA had been very positive because of the good performance level of the technical team. In fact, he considers that the academic and industrial level of NLP in Spain is very good. He went on explaining that the National LT Plan has three main lines: Text, MT and Conversational systems, and that he was involved in the latter, notably with the rapid deployment in April 2020 of a chatbot, intended to provide the citizen with information about COVID. The purpose of such a device was twofold, on the one hand it should liberate health emergency phones, and on the other hand it would contribute to provide reliable information to the public. The deployment was carried out in particularly difficult circumstances, known to all. It involved tight coordination between academia, public administration and industry, which provided their help almost altruistically, with a very low cost for the administration. The result was very satisfactory and with a major impact on the Spanish public, receiving more than 350000 queries between April and June 2020.

In David Griol's opinion, LT-based solutions are needed to enhance the quality of the communication between the citizens and the administration, and they will become increasingly common, not only to provide information, but for all kinds of administrative procedures.

As to the issue of the quality of the LT solutions in Spanish and regional languages, he pointed out that although large corporations do provide ASR, NLU and STT solutions for Spanish, the quality is much inferior than the one provided for English. And they do not provide any solution for Catalan, Basque or Galician in these technologies. For this reason, and also for privacy reasons, he is of the opinion that the Public Administration has to develop their own solutions.

After the presentations of the three panelists, **Khalid Choukri**, from the audience, completed Manuel's intervention about European projects, and shared the information about the forthcoming Horizon programme's budget of 95 M€ for research projects, out of which 75 dedicated to Language Technologies.

Francisco Javier González Castaño from the University of Vigo indicated that one thing hindering progress of LT in Spain, and Europe in general, was that the SMEs tended to be considered only as users of the technology and not as providers, when fostering the European LT industry is key for achieving global competitiveness.

2.6 Language technologies showcase

As pointed out by one participant in the previous panel, the full perspective of the LT sector in Spain would be incomplete if we did not include the small and medium-sized companies involved in Language technologies in our workshop. We wanted SMEs to give their particular vision on the subject, and we also wanted to give the audience the possibility of attending a showcase of sorts in which to find out what these leading companies in Spain were doing.

If the workshop had been face-to-face, maybe we would have asked the companies to set up a small stand with a poster and a demo and to explain their products to the attendees. The virtual format forced us to come up with a creative solution. On the one hand, we put together some videos that give us a complete vision of what these companies are doing, at https://linktr.ee/elrc_spain, and on the other hand, we invited representatives of 8 SMEs belonging to the Language technology sector in Spain to express their opinions in a panel discussion moderated by Maite Melero. The panelist were:

- **Amando Estela** (AE) from Pangeanic
- **Belén Alemán** (BA) from 1millionbot
- **Thierry Etchegoyhen** (TE) from Vicomtech
- **Enric Plana** (EP) from M47labs
- **Eva Martínez** (EM) from Seedtag
- **Julio Prada** (JP) from Inbenta
- **Eudald Camprubí** (EC) from Flaps
- **Nuria Sanchez-Almodovar** (NS) from BotFoundry by Inetum

After introducing themselves and their companies, the panelists were asked to give their opinion on a series of questions. Here we try to summarize the issues that came up in form of bulleted lists:

1. **What is the level of quality of applications in Spanish (and in the co-official languages), compared to applications in English?**
   - The quality of the technology in Spanish is quite good and continuously improving, even if still inferior to English. (all)
   - Technology for regional languages is far from optimal, but it has recently made major breakthrough, particularly in the field of MT (TE)
   - Speech technology lags behind English, for Spanish and even more for regional languages (TE)
   - Better access to data (raw, annotated) is needed for the technology to improve (EP)
   - Recent breakthroughs in technology (Transformers, pre-trained models, multilingual models) are helping progress for all languages (EM, EC)
   - The law of the market predicts that bigger markets, such as English now and predictably Spanish in the near future, will have more technology, better quality. (JP)
   - All languages matter. Multilingualism should be promoted. Neglecting the languages with smaller markets means impacting negatively on a group of citizens. (EP, EC)

2. **Does technology transfer between research and business function adequately in Spain?**
   - Technology transfer works reasonably well (all except JP)
   - Already established workflows in industry are a challenge to overcome (TE)
   - SMEs are interested in using mechanisms that allow for technology transfer, such as industrial PhDs, programmes like INNOVATEC, Marie Curie, etc. (NS, EP)
   - Technology transfer almost never succeeds because industry and research speak very different languages. (JP)
   - It is necessary to look for synergies between academia, SMEs and the big companies (NS)

3. **How can SMEs in the sector (specifically yours) compete with the big technology companies? Is there a capacity for internationalisation in Spanish companies? Towards Europe, towards Latin America?**
    - The keys that bring added value to SMEs over the big techs are (all):
        - Protection of privacy (GDPR)
        - Good service due to proximity with the client
        - Adaptation and integration of the solution to the client's needs and workflows.
    - SMEs strengths compared to big companies are: flexibility, capacity of innovation, closeness to their clients, ability to "land & expand" in micro-niches (JP, EC)
    - Very often a solution by a local SMEs outperforms Google's. An example is the Basque-English translator, and many MT systems customized on specific domains. Behind these successful systems there is a lot of time spent on cleaning the data, fine-tuning the models, etc.(TE)
    - Local LT SMEs can perform tighter quality control, e.g. control over bias by influencing the sampling of data, demography, annotation, etc. (EP)
    - Open-source solutions are crucial to SMEs success (EC)
    - SMEs must look for strategic alliances with the big techs, which tend to externalize many aspects of their business, such as generation of resources (BA)
    - Outsourcing of services by the big techs often takes place in developing countries (e.g. India) where cheap labor cost often entails quality loss. (EP)
    - The better the technology, the easiest is to internationalise it.(TE)
    - To effectively compete internationally better financing mechanisms are needed.(EP, BA)
4. **How can public administrations help the sector?**
    - The cooperation of the public administrations is necessary (all)
    - They can help through financing (all)
    - Support of Public Administrations is needed to compensate for technology gaps among languages. The project AINA, promoted by Catalan administration, is a good example of that. AINA aims at creating resources (data, pre-trained models) needed to lower the initial investment barrier (EC, EP)
    - Public Administrations have the largest volume of non-confidential data, and data is at the basis of LTs. They should make an effort to make this data available. (TE)
    - Public Administration's big asset is their own internal consumption, which should boost the demand. Countries such as France or the Netherlands always buy national technology. Spanish PAs should do the same, although they rarely do.(JP)
    - For the PA to consume technology they need first to augment their capacity for innovation. They need to understand that this is the best way to enhance the service they provide to the citizens.(EC)
    - Spanish administration has a record of favoring contracts with the big companies and consulting firms, such as Indra, which are easier for them. They should be promoting SMEs instead. (EC)
    - Events promoted by the Public Administration such as this workshop are very good because they give a voice to the SMEs. (BA)

2.7 Conclusion

At the end of the Workshop, **Khalid Choukri**, as a representative of ELRC, took the floor to thank the organisers and the participants. He commented on the importance of European funding for LTs and the usefulness of tools such as eTranslation in order to connect SMEs across Europe. All these efforts should lead to a stronger Europe, able to compete with the US and China in the field of LTs. He concluded by recalling that multilinguality is a major European asset.

To conclude, **Maite Melero** thanked the speakers, panelists and participants and called them to participate in a next, hopefully in-person, workshop.

3 Country Profile: Language data creation, management and sharing

In practical terms, the situation has not changed much in the Spanish Administration, with respect to the description provided in the Country Profile. What is slowly starting to change is the overall awareness of:

- language technologies, which are acknowledged as an important aspect of Artificial Intelligence
- linguistic data, which is starting to be valued as a technological asset that needs to be reused and shared
- the need to go further in the initial wave of digitalisation that has taken place along the last decade
- the crucial importance of technology to enhance the services to the citizens, and that, far from posing a threat to humans, it is able to optimize their work.

With the new AI strategy starting to be put in place in the coming months and the targeted funding from the EU Next Generation funds, we expect to see important developments at all levels in the Spanish Administration soon.

Most of the panelists and many participants expressed their satisfaction at having had the opportunity to participate in the Workshop and highlighted the importance of running such events regularly in order to put together the different actors (real-world use cases from the Administration, SMEs providing IT services, etc.) and give them a voice.