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

European Language Resource Coordination
Connecting Europe Facility

Deliverable Task 6

ELRC Workshop Report for Spain

Author(s):	Núria Bel (UPF)
Dissemination Level:	Public
Version No.:	V1.1
Date:	09.02.2016

Contents

1. Executive Summary

This document reports on the ELRC Workshop in Spain, which took place in Madrid, on the 26th of January 2016 at the Representation of European Commission in Spain. It includes the agenda of the event (section 2) and briefly informs about the content of each individual, interactive and panel workshop session (section 3).

The ELRC-Madrid Workshop was attended by 81 people. The participants distribute as follows: 48 people come from the public administration, 2 people from local administrations, 19 people from the University, 6 people from industry and 6 from other organizations.

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

2. Workshop Agenda

Agenda for CEF.AT - ELRC Training Workshop
(European Language Resource Coordination)

SALA EUROPA

Representación de la Comisión Europea en España, Paseo de la Castellana 46

Madrid, 26.01.2016

08:30 – 09:00	Registration
09:00 – 09:15	Opening by **Carmen Zamorano**, EC Representation in Spain **Víctor Calvo-Sotelo**, State Secretary of Telecommunications and for the Information Society
09:15 – 09:25	Welcome by **Carmen Zamorano**, EC Representation in Spain **Khalid Choukri**, ELRA/ELDA **Núria Bel**, UPF
09:25 – 09:35	Workshop objectives **Khalid Choukri,** ELRA/ELDA
09:35 – 09:50	Europe and multilingualism **Carmen Zamorano**, EC Representation in Spain
09:50 – 10:20	Language and related language technologies in Spain **Núria Bel**, UPF
10:20 – 11:00	Panel: Public multilingual services in Spain Moderator: **Núria Bel**, UPF Panelist: **Leandro Valencia**, Head of the Office of Language Interpretation of the Ministry of Foreign Affairs and Cooperation. **Pablo de Amil**, Head of the Information Technologies and Communications of the State General Administration. **Marta Xirinachs**, Deputy Director of Linguistic Policy of the Generalitat of Catalunya. **Josu Uribel**, Basque Institute of Public Administration
11:00 – 11:30	**Coffee Break**
11:30 – 12:00	Machine translation: how does it work? **Marta Ruíz Costa-jussà**, UPC
12:00 – 12:30	How can Public Institutions benefit from the CEAF.AT Platform? **Spyros Pilos**, EG, DG Translation, Head of Language Applications
12.30 – 13.00	The data for CEF-AT

	Khalid Choukri, ELRA/ELDA
13:00 – 14:00	**Lunch Beak**
14:00 – 14:30	Legal framework for contributing data. **Prodromos Tsiavos**, ELDA **Salvador Soriano,** Area coordinator in open data and intellectual property of the Ministry of Industry, Energy and Tourism
14:30 – 15:30	Panel: Data and language resources in Spain. Moderator: **Asunción Gómez Pérez**, UPM Panelists: **Lucía Escapa Castro**, Deputy Director General of the General Secretary of Technology and Information Services of the Ministry of the Presidency. **María Dolores Ortigosa Lorenzo**, Translating and Interpreting Service of the Directorate General of Police of the Home Office. **Carlos Romero Dexeus,** Director of R&D of SEGITTUR of the Ministry of Industry, Energy and Tourism.
15:30 – 16:00	**Coffee Break**
16:00 – 16:30	How can we engage? **David Pérez**, Secretary of State for Telecommunications and the Information Society of the Ministry of Industry, Energy and Tourism.
16:30 – 17:00	Data and language resources: Technical and practical aspects. **Khalid Choukri,** ELRA/ELDA
17:00 – 17:15	Wrap-up and Conclusions. **Khalid Choukri,** ELRA/ELDA

3. Summary of Content of Sessions

3.1. Session 1: Opening

Ms. Carmen Zamorano expressed her welcome message to the audience and thanked Mr. Víctor Calvo-Sotelo (State Secretary of Telecommunications and for the Information Society) for attending the workshop, on behalf of Ms. Aránzazu Beristain (Head of the EC Representation in Spain). Ms. Zamorano set the CEF scene in supporting the Digital Single Market (DSM) strategy. She introduced ELRC as the consortium in charge of CEF-AT and she insisted on the need for the Member States to participate by providing data.

Mr. Víctor Calvo-Sotelo, State Secretary of Telecommunications and for the Information Society, thanked Ms. Carmen Zamorano for hosting the Workshop and for having invited him. He thanked Spyros Pilos and Khalid Choukri for being there. He pointed out the key role that Machine Translation (MT) and language resources are increasingly playing in Europe and the importance of the DSM strategy to generate opportunities in Europe. He mentioned the complexity of managing language and culture diversity in Europe and he expressed his belief in that language technologies will help achieve the goals. He presented the aspects of multilingual Spain, which has an official language and 3 co-official languages, and the platform for MT (PLATA) used by the public administration. Finally, he mentioned the new program to support language technologies of the Ministry of Industry, Energy and Tourism in Spain.

3.2. Session 2: Welcome

Núria Bel (UPF) thanked the audience for participating in the workshop and pointed out the success of the event, which showed that the topic is of great interest and may help identify the needs and share interests. She asked the participants to contribute with suggestions, needs and questions. She thanked the EC Representation in Spain for their support in the workshop organization.

Khalid Choukri (ELDA/ELRA) thanked Mr. Víctor Calvo-Sotelo, State Secretary of Telecommunications and for the Information Society, and Ms. Carmen Zamorano (DGT local officer) for their support to the event, and thanked the assistance. He also thanked the the organizers of the workshop who made it possible to have such opportunity to discuss how to share linguistics resources, a main topic for ELRA, which he presented. He also invited everybody to participate and to share their hopes and questions about the importance of Language Resources for Language Technologies.

3.3. Session 3: Workshop objectives

Khalid Choukri introduced the context in Europe where language diversity is the basis of the European culture and identity and he introduced CEF-AT. Then, he went through the objectives of the workshop, which mainly stand in working with public services, administrations, ministries, etc. to find and share appropriate data to get the best results from the proposed Machine Translation Technology for their needs and language.

3.4. Session 4: Europe and multilingualism

Carmen Zamorano showed the European scene with its 24 official languages and 60 major regional/minority languages pointing out that all languages are equal in Europe and that Europe is committed to multilingualism. She described the EU's Digital Single Marker strategy and how language barriers affect private and public services. Pan-European public services face a Multilingual challenge: 90% of EU web users prefer to use their own language in online services. But human translation is inappropriate in most of the use scenarios of public services and available language coverage by online translators is partial and not secure. The solution is to build CEF-AT, aiming at making public digital services equally usable by all EU citizens and facilitating cross-border information exchange in public administration. The role of Member States, then, is to take ownership of their own language and to make sure their language is adequately supported in CEF-AT. Their involvement is essential.

3.5. Session 5: Language and related language technologies in Spain

Núria Bel introduced language technologies and machine translation and she showed the state of language technologies for the languages discussed in the META-NET White Paper Series, providing some cases of machine translation uses in Spain. She presented the aspects of multilingual Spain, which has one official language and 3 co-official languages. In addition, in daily life, multilingualism may be extended by tourism, trade, cooperation between regions of Europe, cross-border legal issues, immigration and International Relations. She presented the language technology companies and centers spread all across Spain. She concluded by briefly talking about the forthcoming action plan to boost language technologies in Spain.

3.6. Session 7: Machine translation: how does it work?

Marta Ruíz Costa-jussà, from the Universitat Politècnica de Catalunya, gave a brief overview of the technical basics of automated translation. First, she discussed the motivations to use, depending on the context, automated translation in addition to human translation and she gave an overview of the history of machine translation. Then, she presented the statistical machine translation approach, through some basic illustrations, the data needed and the learning process.. She concluded by highlighting the importance of abundant data.

3.7. Session 8: How can Public Institutions benefit from the CEAF.AT Platform?

Spyridon Pilos, Head of Language Applications of DGT, started his presentation by elaborating the reasons why machine translation is essential in a multilingual Europe. Then, he described in detail the MT@EC system, i.e. the supported languages, the technologies upon which it is based, its user interface and other technical features, including input data format, delivery of results and security in document transfer. He concluded by describing the CEF-AT platform, emphasizing that the focus will be the coverage of more domains (not just typical EU texts) and increasingly in-domain texts are needed to implement this move at sufficiently high quality.

One of the participants asked what would happen with those organizations working with rule-based MT systems e.g. Lucy and Apertium, whether they would have to create resources for both rule-based and statistical systems.

Another participant raised the issue of terminological resources and asked for help to share terminological databases between administrations. Terminology, however, has not been addressed yet within the project/platform.

Finally, one participant from a private company asked whether companies will be able to use the resources.

Khalid Choukri pointed out that the project will not ask to produce more data, but to share what is already there. He also explained that they hope that public data, in addition to its delivery to the EC/DGT, will be shared between public administrations in different countries and between public administration and private companies (and t*his will depend on the owners of the data sets)*.

3.8. Session 9: The data for CEF-AT

Khalid Choukri focussed on the data that need to be collected for MT: texts (translations, aligned translations, collections of comparable texts), glossaries, terminological databases, dictionaries and lists of words in one or more languages. He discussed how different formats are useful to varying degrees and stressed the importance of metadata. He showed several examples of how language resources are produced from data. Finally, he encouraged the participants to contribute to the project by providing or identifying the required textual content.

3.9. Session 10: Legal framework for contributing data

Prodomos Tsiavos expressed the need for a clear and easy to follow regime for data re-use across the EU. He stressed the existence of legal and technical interoperability and simple redress mechanisms aiming to develop a single European market for innovative apps based on public data. He briefly presented the Public Sector Information (PSI) Directive and explained the structure of the data rights. He listed the stages for releasing the public data governed by the PSI Directive, and showed several case studies.

Salvador Soriano, area coordinator in Open Data and Intellectual Property of the Ministry of Industry, Energy and Tourism, said that, barring a few exceptions, all data that the public administration in Spain produces is public. He also mentioned that they have a lot of data without personal data. As for personal data he said it should be anonymized, which he considered a complex task. He talked about the efforts they are doing to promote the reuse of data.

3.10. Session 12: How can we engage?

David Pérez, from the Secretary of State for Telecommunications and for the Information Society of the Ministry of Industry, Energy and Tourism, presented a new program to support language technologies launched by the Ministry of Industry, Energy and Tourism in Spain. This program has four main lines of actions: (1) the development of language infrastructures, (2) the promotion of language technology industry to improve its visibility and transfer and to support its internalization, (3) to promote Public Administration as a promoter of the language industry and (4) the development of several applications.

One of the participants asked whether local administration will be able to participate in the platform and whom they should contact for that. David Pérez answered that it was also for local administration and that they should contact his office.

Another participant asked whether there will be an open call for projects to be funded by the new program. David answered positively.

Finally, another participant asked about the role and contribution of the universities in this program. David Pérez said that they hoped universities contribute by developing applications.

3.11. Session 13: Data and language resources: Technical and practical aspects

Khalid Choukri presented in detail the workflow for the collection, processing and sharing of language resources. Most of the stages of this process, e.g. the identification of the data sources and datasets, the basic metadata documentation, data cleaning, privacy and ethics management are tasks in which the public sector providers will collaborate with ELRC. He then encouraged the audience to participate in these activities and work together with ELRC, and he showcased the mechanisms with which ELRC will fully support the providers throughout the whole process, i.e. the helpdesk and user forum mechanism, the ELRC repository and the ELRC website.

Spyros Pilos pointed out that the data will be accessible only to public administrations.

Núria Bel thanked the participants for having attended the workshop.

4. Synthesis of Workshop Discussions

4.1. Panel 1: Public multilingual services in Spain

The panel was moderated by Núria Bel. The panelists were representatives from the Ministry of Foreign Affairs and Cooperation, the Information Technologies and Communications of the State General Administration, the Generalitat of Catalunya and the Basque Institute of Public Administration.

Leandro Valencia, Head of the Office of Language Interpretation (OLI) of the Ministry of Foreign Affairs and Cooperation, explained that the 19 translators of the OLI work exclusively for the Ministry of Foreign Affairs and Cooperation. He described the type of documents they translate from English, French, German, Russian, Catalan, Arabic and Greek, and sometimes they contract with external translators. They work with a CAT system that only works with MS-Word.

Pablo de Amil, Head of the Information Technologies and Communications of the State General Administration, presented the MT platform PLATA. This platform uses a hybrid approach to translate from Spanish into English, Basque, Catalan and Galician, which uses Apertium and Moses.

Marta Xirinachs, Deputy Director of Linguistic Policy of the Generalitat of Catalunya, presented the MT system of the Catalan Government. This system uses the system developed by Lucy LT to translate to/from Catalan into Spanish, English, French and German, and it uses Apertium to translate from Catalan and Spanish into Occitan and Aranese. From 2006 to 2010, the system was free and received and processed 13,962,921 requests. From2011 on, the system could be accessed via an intranet service for government departments and in 2015 553,455 requests were received. The Generalitat de Catalunya also offers a service to solve linguistic queries.

Josu Uribel, from the Basque Institute of Public Administration, presented IDABA: the tool for multilingualism in the Basque Administration. IDABA is a multilingual database based on Oracle that includes segmented and indexed texts, accompanied by certain metadata: date, source entity, translator and document type. All documents are in Spanish-Basque. Currently, it has 7,833 documents, with over 2.2 million translation units. IDABA provides access to the database sequences of characters, words or text, it also allows retrieving source texts in a specific language and exporting translation memories, defined by some parameters or metadata in any of these formats: TMW for Trados, TMX for Wordfast and tabulated text TXT. The query can be normal, introducing the character sequence and specifying the language, choosing one or several parameters or metadata or choosing certain documents or document types.

4.2. Panel 2: Data and language resources in Spain

The panel was moderated by Asunción Gómez Pérez. The panelists were representatives of the Ministry of the Presidency, the Home Office and the Ministry of Industry, Energy and Tourism.

Carlos Romero, from the Ministry of Industry, Energy and Tourism, explained how the tourism industry in Spain undergoes a rapid change and then he focussed on the current challenges which include: the need to offer relevant and personalized content at a low cost; the adaption of tourist services to different linguistic and cultural segments; the customization of ads, the online positioning strategy, the brand monitoring, the web design, in the respective languages; the increasing need for market intelligence about the preferences and opinions of our customers; the growing role of the recommendations and local knowledge, the new business models (sharing economy). Translation is a constant concern in tourism (cost, quality, speed, local, adapted, multiplatform, labelling). There is a real need for translation solutions which are cost-efficient, high-quality, real-time, locally adapted and mobility platform.

Lucia Escapa described the functions and the structure of the Ministry of the Presidency. She explained that 6 translators are responsible for translating about 2000 documents, mostly in English, but also in French and German.

María Dolores Ortigosa Lorenzo, from the translating and interpreting service of the Directorate General of Police of the Home Office, explained that the 260 translators that work in the offices around Spain do not share their translations, they do not even use translations memories, they don't have access to internet, due to security issues.

Spyros Pilos asked who had the ownership of the translations and Lucia Escapa said that they belong to the Ministry which is stated in the contract with the translators. The same goes with SEGITTUR (Sociedad Estatal para la Gestión de la Innovación y las Tecnologías Turísticas, Ministry of Industry, Energy and Tourism).

5. Workshop Presentation Materials

The workshop presentations can be accessed at the event webpage, at http://lr-coordination.eu/spain_agenda