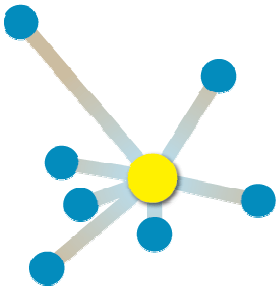

European Language Resource Coordination
Connecting Europe Facility

Deliverable Task 6

ELRC Workshop Report for Portugal

Author(s): Andreia Querido (UL)
Rita de Carvalho (UL)

Dissemination Level: Confidential

Version No.: <V1.1>

Date: 2016-03-16


European Language
Resource Coordination
Connecting Europe Facility

Contents

1


European Language
Resource Coordination
Connecting Europe Facility

1 Executive Summary

This document reports on the ELRC Workshop in Portugal, which took place in Lisbon, on the 1st of March 2016 at Representação da Comissão Europeia[1] (Largo Jean Monnet 1-10.º, 1269-068 Lisboa). It includes the agenda of the event (section 2) and briefly informs about the content of each individual, interactive and panel workshop session (sections 4 & 5).

The event was attended by 31 participants spanning a wide range of ministries and public organizations. From the 31 persons present, 23 were potential data providers.

After the workshop opening and welcoming of all participants, the workshop goals were exposed, and an overview of the linguistic context in Europe was given. Then, the language technologies and public services in Portugal were discussed.

The CEF.AT platform, which ELRC presents in the workshop as the successor of MT@EC, uses an automatic translation (AT) system based on statistical methods. This technology was explained to participants, as well as the kind of data that is needed to improve automatic translations. The legal framework concerning the release of data from public sector bodies was also addressed.

The relevance of this AT platform for the public services in Portugal was stressed out, and the audience was encouraged to participate in improving CEF.AT by providing the right kind of data needed by ELRC.

During the workshop, participants were informed about two ways of delivering their data: by sending them to DGT (through the email DGT-LISBON@ec.europa.eu) with a copy to the helpdesk or by uploading them to the ELRC repository (http://elrc-share.ilsp.gr/).

The dedicated event webpage, with the Agenda and the online registration form, can be found at http://lr-coordination.eu/pt/portugal.

[1] European Union Delegation in Portugal

2 Agenda

08:00 – 09:00 Registration

09:00 – 09:30 Opening of the workshop, welcome, and workshop goals (Paulo Batista representing João Tàtá dos Anjos, European Commission)

09:30 – 10:00 Europe and Multilingualism (Paulo Batista, European Commission)

10:00 – 10:30 Language Technologies in Portugal (António Branco, University of Lisbon)

10:30 – 11:00 Panel: Multilingualism in Portuguese Public Services (Paulo Batista, European Commission)

11:00 – 11:30 Coffee Break and Networking

11:30 – 12:00 How does Machine Translation work? (João Rodrigues, University of Lisbon)

12:00 – 12:30 How can Public Institutions benefit from the CEF.AT Platform? (Khalid Choukri, Evaluations and Language Resources Distribution Agency - ELDA)

12:30 – 13:30 Lunch Break

13.30 – 14.00 What Data is needed? Why? (João Silva, University of Lisbon)

14:00 – 14:30 Legal Aspects (Khalid Choukri, European Language Resources Association - ELRA)

14:30 – 15:00 Panel: Language Data in Portugal (Amália Mendes, University of Lisbon)

15:00 – 15:30 Coffee Break and Networking

15:30 – 16:00 Data and Language Resources: Technical and Practical Aspects (Khalid Choukri, European Language Resources Association - ELRA)

16:00 – 16:30 How can we participate? (António Branco, University of Lisbon)

16:30 – 17:00 Summary and next steps

3 Workshop Attendance

31 persons from 19 institutions, essentially public service translation departments, pre-registered to the workshop.

17 persons participated actively in the workshop either by presenting content or organizing the event.

5 pre-registered persons from 4 institutions did not attend the workshop.

In total there were 26 persons present during the workshop.

From these 26 persons, 23 persons from 16 institutions are potential data providers.

Again, from these 26 persons, 22 filled the feedback forms.

4 Summary of Content of Sessions

4.1 Session S1: "Opening of the workshop"

António Branco, the national anchor point (NAP) in Portugal, opened the event by welcoming the audience. After introducing the key persons in the organization of the event, Paulo Batista and Khalid Choukri, he invited each participant to introduce himself. Then, he gave orientations about the materials provided to each participant and the structure of the sessions (20 minutes for presentations and 10 minutes for discussion).

4.2 Session S2: "Welcome by EC"

Paulo Batista introduced himself as a European Commission translator in Portugal. He then welcomed the attendees on behalf of João Tàtá dos Anjos, Acting Head of Representation of the European Commission in Portugal, who could not be present, but sent a warm welcome to everyone.

He highlighted that this workshop is an initiative that has been carried out in multiple European Union Member States, 15 until now. In addition to discussing data sharing, or information exchanges, the workshop will provide the opportunity for participants to express their doubts and suggestions about translation work automation.

After a brief introduction of the CEF.AT platform, he reinforced the need to know what the participants' difficulties are regarding machine translation.

4.3 Session S4: "Goals of the workshop"

Khalid Choukri, from ELDA, made this presentation.

He started by providing some context: a multilingual Europe, with 24 official languages, many cultures and identities. As "multi-linguality is strongly supported by EU", communication becomes a challenge. Few people are proficient in more than one language; this situation can create restrictions or even discriminations. Translation is the way to make languages an asset rather than an issue.

In a technological world, where information is spread so quickly and so massively (for example, in Europe there are several thousands tweets per day), human translation has to be supported by technology (automated translation – AT).

Connecting Europe Facilities (CEF) is a program that supports the vision of a multilingual Digital Single Market (DSM). One way to accomplish this is through the new CEF.AT platform, a free platform for automated translation, provided by the DGT, and which focus on citizens, public services, ministries, translation services, etc.

If we can adapt this automated translation platform to our needs, it can be used for the benefit of all – national public administrations and EU citizens.

Since the learning process of this system is based on human translations, better data will result in better translations. CEF.AT can help us and we can help CEF.AT and our own language by providing specific domain data. It will improve the quality of the translations to be used in the day-to-day needs of public services in Portugal.

The main goals of this workshop are:

- To share data, so they can be used to improve CEF.AT;
- To raise awareness and engage the participants;
- To help with legal and technical issues associated with the collection of data.

4.4 Session S3: "Europe and multilingualism"

Presented by Paulo Batista, European Commission.

Thinking of Europe is thinking of *multilingualism*. In fact, this concept is so enshrined in UE legal basis, that all EU citizens have, by law, the right to address the European authorities in their own language.

With 24 official languages, the EU is committed to supporting multilingualism.

An example of that commitment in supporting multilingualism is the Digital Single Market (DSM) strategy. The European Digital Market is multilingual, and language can become a barrier to the economical growth and the creation of jobs, affecting private and public services.

Connecting Europe Facility (CEF) is a deployment program using mature technologies. Its multilingual e-services like eProcurement, eHealth, Europeana or Open Data Portal, help citizens, business and public administrations.

Statistics show that the majority of EU web users prefer to use their own language in online services and that nearly half of the EU citizens do not speak English. That is why we cannot say that there is a real *lingua franca*, and even if there was, citizens have the right to speak in their own language.

Translation by humans is too expensive and too slow for the needs of public services in Europe – we have to use machine translation, but the system must be efficient and secure. This is what CEF.AT can do for local public services of each EU Member State: make digital services equally usable for all users, whatever language they may speak; and facilitate cross-border information exchange in administration.

For the technology to improve, more data are needed - data from specific domains that are closer to the day-to-day needs of administration services. ELRC is a consortium created to facilitate this data collection.

With more linguistic resources, CEF.AT, an automated translation platform, free of charge, will provide better translations, thus helping all its users. The goal is to reduce the language barrier in pan-European public services.

After the presentation, questions were asked on how to use the system (how to get access and how the translations were done); how to transfer data to the platform and how to secure the copyright issues.

4.5 Session S5: "Language technologies in Portugal"

António Branco, University of Lisbon (UL), structured his presentation around three questions:

1) What is language technology?

It is a multidisciplinary technology that emerges from Natural Language Processing (NLP) and that is crucial for us to interact with artificial devices using natural language, and for people who do not share a common language to communicate among themselves.

Its impact is economic, social and cultural. That is why it is crucial to prepare our language, Portuguese, to be used within these technologies.

2) How prepared are the language and its speakers?

Portuguese is the fifth most spoken language in the world. About 127.8 million people use Portuguese to communicate on the Internet. The Internet reaches 47% of the population in Portuguese-speaking countries and 42% of the global population.

In the white book *The Portuguese Language in The Digital Age*[2], part of a series on the situation of language technologies for 30 European languages, Portuguese appear not to have a good technological coverage: references to Language Resources for Portuguese in scientific articles are way lower than the references to Language Resources for English. In fact, the number of research papers on English is 22 times higher than those addressing Portuguese.

From the perspective of academy researchers, this ELRC initiative can be important and a good contribution to change this scenario.

3) How to boost language technologies in Portugal?

We have PROPOR, a bi-annual conference specifically for the processing of Portuguese, and CLARIN, an initiative that provides resources and technologies to researchers.

Language technologies are important to ensure citizenship in the digital age, to promote innovation and economical internationalization. Portuguese researchers, companies and investors should work together with these purposes.

4.6 Session S7a: "Automated translation: How does it work?"

Presentation by João Rodrigues (UL).

Automatic translation started with rule-based systems, but the complexity of both language and translation emphasized the weakness of this approach.

Nowadays, another method proved to be more successful: statistical and probabilistic systems using machine learning algorithms. Statistical machine translation learns how to translate from data: existing translations (parallel corpora) and documents on the target language. That is why more data equals better results – and even better results if the right data are used.

After collecting them, the data will be aligned (sentence and word alignment). From this, translation probabilities can be extracted. These translation probabilities are then used to generate a set of possible candidate translations. The best translation from this set is picked by resorting to a language model of the target language.

---

2 António Branco, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, and Vera Lúcia Strube de Lima, 2012, *A Língua Portuguesa na Era Digital / The Portuguese Language in the Digital Age,* White Paper Series, Berlim, Springer, ISBN9783642295928 (impresso), ISBN9783-642295935 (ebook).

One significant improvement to the success of automatic translation was made when translation probabilities started to be extracted from phrase translations. When translating single words, the context is lost and the resulting translation is not so accurate. Phrase-based statistical machine translation is the standard technology, present in Moses, an existing open-source translation tool, used by the EC in CEF.AT.

With this technology, data are crucial, and CEF.AT is looking for the right kind of data to improve translations for public services.

After the presentation, questions were raised about the statistical extractions in phrases and the types of data needed. To the question "What do you need, translated documents or monolingual documents?", João Rodrigues, António Branco and Khalid Choukri reinforced that both types of documents are very useful to CEF.AT.

4.7 Session S11 a): "How can public institutions benefit from the CEF.AT Platform?"

Khalid Choukri (ELDA) gave the talk on behalf of the EC and started by sharing the current main goal of DGT: improving AT in a way that it could be possible to access information in any language, anywhere.

There are two types of machine translation users: the ones who do not understand the source language and the ones who understand both source and target language – AT is useful for both.

The existing system, MT@EC, a statistical MT system, was released on the June 26, 2013, and is able to translate between any of the 24 EU official languages. The system is available through a web-user interface as well as through web services. It has an highly secure protocol that guarantees confidentiality of data, and can be used by European institutions and bodies, online services supported by the EU and Member States national public administrations, as long as the staff members get an individual European identification through ECAS.

EURAMIS (European Advanced Multilingual Information System) stores all the EU data on which MT@EC is built: 940 million sentences (by the end of 2015), and it is growing at a rate of 2.6 million per month. CEF.AT will improve the existing MT@EC in terms of security, quality and adaptability. The digital European Library (Europeana), the pan-European Open Data Portal (ODP), the Electronic Exchange of Social Security Information (EESSI), the Online Dispute Resolution platform (ODR) and e-justice will be connected to CEF.AT platform.

For that improvement to happen, more data are needed, especially "domain" data.

Several questions were raised after the presentation. The discussion focused on two main topics: the format of texts that could be translated by MT@EC and the automation level of the system. The conclusions of the discussion were that the system accepts all formats (modulo the PDF format); that the system is automatic when providing a translation to a user; and that, before any translated documents enter EURAMIS (the database), they are validated by human translators.

4.8 Session S7 b): "What data is needed? Why?"

João Silva (UL) started by referring that his presentation would complement the session presented by João Rodrigues (UL).

MT uses a data-driven paradigm, which means that "the more data the better". Producing data from scratch is not viable, nor is the objective of projects involving machine translation. For ELRC, any document in a EU language is useful, especially multilingual data.

"Data" can be reports, speeches, web page content, brochures, etc. Everything is useful, in particular:

- Translations done by humans (because they will probably have less errors than automatic ones);
- Documents from specific domains (to improve the quality of translations in those domains);
- "Aligned" translations;
- Comparable collections;
- Dictionaries, terminologies, etc.

All these data can be found in several digital formats: *.doc, *.docx, *.txt, *.odt, *.xls, *.xlsx, *.csv, *.html, *.eml, *.pub, *.rtf, etc.

The subsequent processes of text extraction, text clean up, text alignment, etc., can be automated. The automation of the procedure can be, hence, turned into a Language Resource factory.

Nonetheless, only "visible" resources (the ones in the public web and indexed by search engines) are accessible. Most of the documents that are relevant to the CEF.AT improvement can be found in the deep web, within the organization-specific repositories protected by passwords. All those data already exist: what the audience can do is provide them.

During the subsequent discussion, the importance of having access to data that are not available in the surface web was reinforced.

4.9 Session S10 a): "Legal aspects" / S10 b) European Union data portal

Khalid Choukri (ELDA) spoke on behalf of legal experts, who could not be present at the workshop.

Years ago, the EU felt the need for a legal framework so that data could be re-used. The new Public Sector Information Directive (PSI) solves many issues by emphasizing the idea that what is public should be published and made available. It also makes it easier for Member States to provide data.

These PSI rules sit on top of other legal regimes and ensure that, when the conditions are met, public sector information is available to third parties. For example, public museums and libraries are now within the scope of this directive and have to share their data.

PSI rules are complemented by sector specific regimes, which assist the re-use of data in specific areas. Regarding copyright rules, it is important that the public sector information be owned by a public sector body (if so, data are available, if not, other licenses are needed).

Data sets that contain personal data can be re-used, as long as there is consent/ legal base or as long as those data are anonymized. If doubts remain about the possibility of re-using certain datasets (because they include trade secrets or confidential information, for example), it is better not to submit those data. Anyway, these cases are exceptions, rather than the rule.

It is important to keep in mind that collected data will not be used as such, nor will they be provided online. The AT will be trained by extracting translation or language models (e.g. occurrence counts) from the data.

PSI directives have different approaches, depending on the national legislation of each EU Member State. For example, in Portugal, all public sector data are available by default. In the UK, on the other hand, there is specific license regarding the Government data.

4.10 Session S9: "Data and language resources: technical and practical aspects"

This presentation was given by Khalid Choukri (ELDA).

From the data that can be provided by the organizations represented in the audience, until their usage as Language Resources, there are many steps: data identification, selection, documentation, cleaning, conversion, validation, processing, description and storage. During this process, as mentioned previously, the legal status determination of data also takes place. An important step is the data anonymization, when required, which determines whether data can be accepted or not.

As said throughout the workshop, ELRC is looking for existing data that can be provided by public institutions. This can be done through the ELRC portal (www.lr-coordination.eu), which has a very useful helpdesk. The person in charge in each public institution only has to register and create an account. Then, the data submission process is easy: they only have to fill out a data submission form or a source submission form. In the submission sources, there is an automatic check if the URL is already in the database.

Besides the upload option, data can also be sent on a physical medium or through a customized deposit account. Data can be provided as of today.

In the discussion following the presentation, there were questions about the waiting time when uploading files (which is a maximum of 20 minutes); about the registration in the portal (done by e-mail); about the accuracy of named entities translations (that will improve with a bigger data volume); and about the use of the new orthographic rules for Portuguese (every text is relevant, independently of the orthographic agreement that is followed).

António Branco intervened as an academic researcher, expressing the need for the scientific community to be granted access to the data collected by ELRC. Khalid Choukri guaranteed that these data are public will be made publicly available by the European Commission.

4.11 Session S14: "Summary and next steps"

The organizers thanked the audience and Paulo Batista (EC) presented some conclusions:

i) The national and European legislations declare that all Portuguese public administrations data are public;

ii) The more we feed machine translation system (MT@EC) with data, the more accurate results we can get;

iii) Changing the password of MT@EC account every 3 months is an important rule to ensure privacy and confidentiality of data;

iv) DGT has used the MT@EC system for many years. Hence, DGT translations will probably be the first to appear in a translation search in the platform. The long-term goal is to change this scenario progressively.

v) During the next week an email will be sent to the participants asking for the contact of the person who, in their institutions, can provide data to ELRC, even public domain data.

5 Synthesis of Workshop Discussions

5.1 Panel 1 - Session S6: "Multilingualism in Portuguese public services"

Paulo Batista, the representative of Directorate-General for Translation (DGT) in Portugal, European Commission, moderated the debate.

Paulo Batista wanted to hear the Portuguese public administration and, for this, he asked the participants some questions:

i) Which languages pairs are used in Portuguese public administrations?

ii) What kinds of documents need to be translated?

iii) What is the format of the documents?

iv) How do you translate?

v) What are the most difficult issues to deal with?

vi) Do you use an automatic translator?

The given testimonies were as follows.

The representative of the Portuguese Republic Assembly explained that they translate from and to Portuguese and English, Portuguese and French, Portuguese and Italian, Portuguese and Spanish, highlighting that the most difficult is to translate from Portuguese to foreign languages. She said that they use the CAT translation tool Trados, build their own terminological databases with Portuguese, English and French terms and use the Linguee Dictionary. It was reported that they tried to use MT@EC to translate from German to Portuguese but, since it was necessary to know German to correct the target language, they did not finish the translation. In general they use MT@EC instead of Google Translate.

The representative of the Portuguese Social Security said that they use Google Translate and the Linguee Dictionary.

The representative of the Portuguese Ministry of Foreign Affairs reported that they do not use Google Translate. They prefer to consult the Linguee Dictionary and MT@EC, especially for Portuguese-English and English-Portuguese pairs.

The representative of the Portuguese Institute of Registration and Notary Affairs (IRN) explained how the process of translation works in this office. Firstly the texts are automatically translated by a private system bought by IRN. Then, the translation is revised, which accounts for most of the hard work because there are many mistranslated words and expressions.

The representative of the Portuguese Directorate-General of Justice Administration (DGAJ) defended that MT@EC is a very relevant tool and suggested that it should be the responsibility of the Portuguese Agency for Public Services Modernization (AMA) to start recommending it for all public services, using a top-down approach.

The representative of the Portuguese Agency for Public Services Modernization asked if it was possible that the MT@EC interconnected with Portuguese systems used until now, since AMA manages a platform where several public administration services are shared.

Khalid Choukri said to the representative of the AMA that he thinks that it is possible, as CEF.AT would allow for interfacing with web service.

5.2 Panel 2 - Session S8: "Language data in Portugal"

Presented by Amália Mendes (CLUL). This presentation is about data and linguistic resources for the Portuguese language that can be used in language technologies.

In the white book *The Portuguese Language in The Digital Age,* data are provided on the situation of linguistic resources for the Portuguese language.

In the white book *The Portuguese Language in The Digital Age*[3], part of a series on the situation of language technologies for 30 European languages, Portuguese appear not to have a good technological coverage: references to Language Resources for Portuguese in scientific articles are way lower than the references to Language Resources for English. In fact, the number of research papers on English is 22 times higher than those addressing Portuguese.

It is necessary to put much more effort on the construction of treebanks.

There are few corpora with lexical semantic information.

The PAROLE, CORDIAL-SIN and CINTIL have part-of-speech, named entities and lemma annotation. Truly, while many corpora have part-of-speech annotation and other types of morphological information, syntactically annotated corpora are smaller and scarcer.

From the perspective of those who study the language, data are very important to develop the language technologies that support this study.

As a researcher, Amália Mendes has already contacted public institutions, asking them for data, because every kind of data is important for linguistic analysis. In most of the cases her team gave back the data with suggestions for improvement. Therefore, she considers the information sharing as very enriching.

5.3 Panel 3 - Session S11 b): "How can we participate?"

António Branco moderated the debate, where many issues were raised. It is possible to find below a list of questions or suggestions that were raised by several representatives of Portuguese public administrations and the respective solution or explanation as given by Khalid Choukri, António Branco and Paulo Batista.

i) Some representatives of Portuguese public administrations showed great concern with the fact that, on their own, they cannot decided whether data can be provided or not. The decision should be made by heads of departments or ministers or by the Portuguese Agency for Public Services Modernization (AMA). The representative of the AMA said that he will take the information to his institution, in order to facilitate this data sharing. It was explained to Portuguese public administrations representatives that there is a law that requires the sharing of public data, which is a reason why all institutions should provide their data. Therefore, it is expected that all the representatives present in the workshop collaborate with ELRC or point someone with whom to interact.

ii) Some of the most frequently asked questions were "If everyone send a great number of data, how can you manage them? How do we share these data? Where to send them?". After

[3] António Branco, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, and Vera Lúcia Strube de Lima, 2012, *A Língua Portuguesa na Era Digital / The Portuguese Language in the Digital Age,* White Paper Series, Berlim, Springer, ISBN9783642295928 (impresso), ISBN9783-642295935 (ebook).

these questions, the participants received a document with the addresses to where they can send the data. Khalid Choukri explained that there is a repository where data are saved and that there are expert people to process the documents, work on them and train the tool.

iii) It was suggested that the request for data should be sent to the middle managers of institutions. Khalid Choukri, António Branco and Paulo Batista said that the idea was welcome and they believe that it is important to disseminate the message at all hierarchical levels.

iv) It was said that most of the data are confidential or contain important information. The organizers of ELRC clarified that, for the purpose of AT development, they are not specifically interested in the content of the texts. Therefore, if there are documents about which the providers feel unsure (because, for example, they contain personal information), they should not provide them. Even so, the organizers are sure that there are still a lot of data that can be given.

v) Other question was: “What is or will be the feedback on provided data?”. The answer was that all data are welcome and it is important to note that the data has a bi-directional flow: the main objective of this workshop is to receive participants' data in order to improve the automatic translator (MT@EC). Then, if all collaborate with ELRC, its translation platform will be adapted and improved and could be really useful to national public services. It is a win-win situation.

6 Workshop Presentation Materials

All presentations are in PDF format on the ELRC website: http://lr-coordination.eu/pt/portugal_agenda