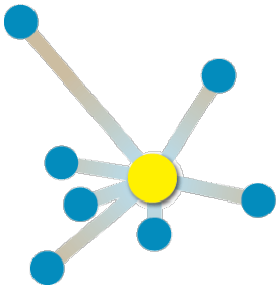

European Language Resource Coordination
Connecting Europe Facility

## Deliverable D3.2.1 Task 8

# ELRC Workshop Report for Ireland

Author(s): Meghan Dowling, Teresa Lynn, Andy Way

Dissemination Level: Public

Version No.: 1.0

Date: 22-12-2017



# Contents

# 1 Executive Summary

This document reports on the ELRC+ Seminar in Ireland, which took place in Dublin on the 13th of October 2017 at the Marker Hotel. It includes the agenda of the event (section 2) and briefly provides details on the content of each individual, interactive and panel workshop session (sections 3 & 4). The event was attended by 53 participants spanning a wide range of government departments and public organisations.

The dedicated event webpage can be found at http://www.lr-coordination.eu/l2ireland.

# 2 Workshop Agenda

| | |
|---|---|
| 08:00 – 09:00 | **Registration** |
| 09:00 – 09:10 | **Welcome and introduction**<br>*Prof. Andy Way, Dublin City University* |
| 09:10 – 09:15 | **Welcome by the EC**<br>*Gerry Kiely, EC Representation Ireland* |
| | **Session 1. Connecting a multilingual Europe: European context and local needs** |
| 09:15 – 09:35 | **Connecting public services across Europe: ambition and results so far**<br>*Aleksandra Wesolowska, DG CONNECT (Video presentation)* |
| 09:35 – 09:55 | **The CEF eTranslation platform @ work**<br>*Colmcille Ó Monacháin,*<br>*Irish Language Translation Unit, Directorate-General for Translation, European Commission* |
| 09:55 – 10:40 | **CEF in Ireland: an outlook into current and future challenges – Panel session**<br>Moderator: *Micheál Ó Conaire , Department of Culture, Heritage and the Gaeltacht*<br>Panelists:<br>• *Fiona Pennolar, Department of Foreign Affairs*<br>• *Fiona Morley Clarke, Department of Public Expenditure*<br>• *Deirdre Lee, Derilinx* |
| 10:40 – 11:00 | **National initiatives for digital public services and (open) data**<br>*Owen Harrison, Department of Public Expenditure and Reform* |
| 11.00 – 11:30 | *Coffee Break* |
| | **Session 2. Engage: hands-on data** |
| 11:30 – 11:50 | **The European Language Resource Coordination (ELRC) action**<br>*Khalid Choukri, ELDA, ELRC* |
| 11:50 – 12:05 | **ELRC in Ireland**<br>*Micheál Ó Conaire, Department of Culture, Heritage and the Gaeltacht* |
| 12:05 – 12:35 | **Preparing and sharing data with the ELRC repository**<br>*Dr. Teresa Lynn, Dublin City University* |
| 12:35 – 13:35 | *Lunch Break* |
| 13:35 – 13:55 | **Can language data be shared and how?**<br>*Pawel Kamocki, ELDA* |
| 13:55 – 14.10 | **New services provided by the ELRC consortium**<br>*Khalid Choukri, ELDA, ELRC* |
| 14:10 – 14:40 | **Identifying and managing your data: Questions & Answers**<br>*Moderator: Prof. Dave Lewis, Trinity College Dublin* |
| 14:40 – 15:30 | **Discussion and Conclusions**<br>*Prof. Andy Way, Dublin City University* |
| 15:30 – 16:00 | *Coffee Break and networking* |

# 3 Summary of Content of Sessions

## 3.1 Welcome and Introduction

Dr. John Judge opened the ELRC+ seminar by thanking everyone for their attendance, and notifying the audience that live interpretation and bilingual slides would be available throughout the ELRC+ seminar. He introduced Prof. Andy Way - deputy director for the ADAPT Centre and former president of EAMT.

Prof. Way echoed Dr. Judge's gratitude to those in attendance, and mentioned his delight in seeing familiar faces from the first ELRC Workshop in Ireland, located in Europa House (EC Representation in Dublin) in January 2016. He made reference to the co-located event which had taken place the day before: '*Irish as a full official and working language of the EU – building a network of experts in Ireland*'. That conference had focused on the increase in translation needs as Irish becomes an official EU working language. This ELRC+ seminar, he said, would deal with how to support that process using technology.

Prof. Way also suggested the use of a Twitter hashtag, #ELRC_Dublin, for all those attending to interact with.

Prof. Way highlighted the goal of this ELRC+ seminar as persuading public institutions to offer up data to help build better machine translation (MT) systems which those stakeholders would in turn benefit from, especially with the Irish language derogation coming to an end in the near future. He noted the high demand in the Dept. of Culture, Heritage and the Gaeltacht (DCHG), a demand that is rising constantly, with over 40,000 words translated so far this year. Prof. Way noted that DCHG already receives support from MT; the 'Tapadóir' MT system has helped translate more than 1.5 million words within DCHG to date. He stressed that MT should be seen as an additional tool in the translator's armoury; it exists to support translators, never replace them.

Prof. Way expressed his thanks to Gerry Kiely, the European Commission representative in Ireland, for his involvement in the ELRC+ seminar.

## 3.2 Welcome by the EC

Gerry Kiely, the European Commission (EC) representative in Ireland, welcomed the audience to the ELRC+ Seminar on behalf of the EC. He thanked Prof. Way for an interesting summary, and expressed his full support for this initiative. Mr. Kiely referenced a speech made by Mr. Juncker, President of the European Commission, in which he called for a more united Europe. He highlighted the critical importance of interaction between Europeans, citing miscommunication as a major issue.

Mr. Kiely then drew the audience's attention to the Connecting Europe Facility (CEF), which aims to support communication in Europe. He noted that the European Union (EU) operates in 24 languages, and that many CEF services will not be accessible if not made available in each of these languages. He noted the increase in translation needs since the 2004 enlargement, an increase that required support from MT.

Mr. Kiely spoke of the free availability of MT@EC for national administrations, and that while it was very useful in translating legal texts it quickly became clear that it was unsuitable for translating out of domain content. He stressed the need for high quality data from public administrations to aid the ELRC in their efforts. Building on this, he said that public administrations are not only in a suitable

position to provide language data, but are obliged to provide bilingual documents. He notes the increased demand for Irish language translation, and the need to improve MT resources for Irish.

He finished by strongly urging the audience to help improve Irish resources that will benefit all EU languages, and thanked them for their support.

## 3.3 Connecting public services across Europe: ambition and results so far

A video presentation from Aleksandra Wesolowska (DG CONNECT) was played to the audience with live interpretation into Irish.

## 3.4 The CEF eTranslation @ work

Colmcille Ó Monacháin, Head of Unit for Irish Language at the Directorate-General for Translation (DGT), began his presentation by thanking the organisation team for preparing bilingual slides for his presentation, but as they were quite technical he would prefer to explain in his own words.[1]

He first explained that statistical machine translation (SMT) works by recognising patterns within a large bilingual corpus. As well as a large parallel corpus, for SMT to be effective it also requires that corpus to be of a high quality, and for it to be in the same domain and style of the text it intends to translate. He then moved on to discussing neural machine translation (NMT). He explained that NMT is a new method of machine translation, which recognises patterns in a manner intended to mimic neural processes in the brain.

Mr. Ó Monacháin revealed that MT@EC works using SMT, and is tailored to the translation of European law. He stressed that a human always post-edits machine-translated text, and makes sure that the final draft does not appear machine translated. He explained that in July of this year, a new MT service was launched: eTranslation. Unlike MT@EC, it uses NMT and can be used in translating texts from a wider domain. Among the advantages of eTranslation, Mr. Ó Monacháin explained, is that the user can translate entire documents without affecting formatting and can be sure that the information translated remains private.

Mr. Ó Monacháin then described the data gathering efforts in the Irish context, led by the ELRC. He explains that in this situation, 'data' refers to electronic text, specifically monolingual (Irish) and parallel (Irish and English) text. He stressed the need for Irish, as an official EU language, to have the same resources as every other EU language. He said that every person in the room with text in a digital form should contact the ELRC team, and that it would be quite an easy process. He referenced the derogation on the production of Irish language text within the EU that would be lifted in 2021, and the huge increase in Irish language translation that comes with that. He ended by encouraging all those present to make a good effort to share any data that they might have.

## 3.5 CEF in Ireland: an outlook into current and future challenges - Panel session

The panel session began with a presentation from Fiona Pennolar, Department of Foreign Affairs. She gave an insight into the many interesting details on the Irish passport. She highlighted the multilingual nature of the passport design; it contains multiple examples of English, Irish, Ulster Scots and Ogham. She described the new online portal for renewing passports. As Ireland has two official languages, it was necessary for the website to be completely accessible in both Irish and English. She described the

[1] Note that for all talks, slides were presented simultaneously in both English and Irish.

bilingual nature of the website as 'seamless' - the website was always going to be bilingual so it was built thus, neither language was an afterthought or an alteration.

The panel session was then opened by the moderator, Micheál Ó Conaire from the Department of Culture, Heritage and the Gaeltacht. He introduced the three panelists: Fiona Pennolar, Department of Foreign Affairs, Fiona Morley-Clarke, Department of Public Expenditure and Deirdre Lee, CEO of Derilinx.

His first question 'what is open data in Ireland?' was opened up to the panel. Fiona Morley-Clarke, Head of Open Data Unit in the Department of Public Expenditure, responded that open data has huge benefits for Ireland, in particular social and economic benefits. Data held by public bodies should be available online, and the open data initiative dealing with the reuse of public sector information (PSI) is a step in the right direction. She described how an open data strategy was launched in July, a 5 year framework which would shape how open data would develop in the future. The key outcome of this strategy would be the development of a national data portal. In achieving this, Morley-Clarke explained, a close relationship with the European data portal team would be maintained.

Mr. Ó Conaire then posed a question on the availability of open data in Irish. The panel explained that the EU data portal offers the choice to machine translate the data portal, but in the long-term they are more interested in a higher standard of translation. Interoperability is the key component to make things easier, e.g. ease of translation, ease of linkage. There has been some outreach and engagement with public bodies in Ireland, including seeking Irish language datasets.

Mr. Ó Conaire asked who can access open data, and whether or not members of the audience had access. The panel clarified that fully open data means that it is open to everyone: public bodies, businesses, researchers, those in academia, etc.

Fiona Pennolar was asked a question regarding the use of the passport portal abroad, and whether or not there are any issues or obstacles with the use of the portal abroad. Ms. Pennolar replied that the aim of the Department of Foreign Affairs is to provide the portal in English, Irish and the language of the country the person is based in. However, Pennolar explains that in application only English and Irish are needed and no difficulty has been encountered with speakers of other languages. Every Irish citizen should be fluent in either Irish or English and therefore it should be sufficient to provide the portal in English and Irish only.

The next answer posed to the panel was 'In terms of local needs - is there any push to improve systems in terms of open data?' Fiona Pennolar responded that from a passport point of view, 80,000 citizens have already successfully applied for a passport renewal using the new online portal since March 2017, and that this number is steadily increasing. By 2019 the portal will be open for all Irish citizens, not just adults renewing their passports, and by this stage it is envisaged that approximately 900,000 citizens annually would use this service. The plan for the improvement of the open data portal is to upgrade to a new version, improve the user-friendliness, include a search function, increase datasets available and also include an online survey to gather feedback on user experience.

Deirdre Lee, CEO of Derilinx, described working with data publishers to advise how best to publish data, and to make it as easy as possible to access existing open data. She stressed the importance of raising awareness of open data and working directly with public bodies to increase the quantity of open data available. Interoperability and standardisation of data is of high importance for it to be as reusable as possible.

Having opened up questions to the audience, Síne Nic an Ailí from Conradh began by thanking Ms. Pennolar for her interesting discussion about the Irish passport, then went on to ask the panel their views on bilingual English-Irish websites. In Ms. Nic an Ailí's experience, often the same information isn't available in Irish as it is in English. Regarding Irish character accents, they are often not recognised and it seems that they pose a problem in terms of the programming of a website. She gave the example of the Eircode website, in which names are translated to English. She asked whether any of the panel had encountered these issues, and if there are any solutions to this issue.

Ms. Morley-Clarke responded that she had encountered a lot of problems regarding the translation of names and that the language support provided in the open portal was not of a sufficient standard regarding Irish.

Ms. Lee enclosed that within Derilinx no data was translated to Irish, although metadata contains vocabulary that can be easily translated, e.g. categories or themes.

Ms. Pennolar responded that from a passport point of view, your name is your identity and a fundamental principle of identity is that it cannot be translated. However, a fundamental design principle of the passport portal is bilinguality. It was built to be bilingual, as opposed to being altered as an afterthought. In this way it was much easier to provide the site bilingually, Ms. Pennolar explained, and it would appear to be much more difficult to provide bilingual content on a previously monolingual website.

Mr. Ó Conaire concluded the panel session by thanking the panelists for their time and expertise.

## 3.6 National initiatives for digital public services and (open) data

Owen Harrison from the Department of Public Expenditure and Reform began by revealing the focus of his presentation: to give the audience a feeling for where the government is at in terms of digitisation and how important data is to that. The connection to this topic and today's event, he explained, is the importance of data, the management of data and the opportunities that data can bring.

In terms of digital performance, Ireland is doing well - it ranks 8th among EU member states. In other areas (connectivity, human capital, business, etc.) Ireland ranks between average and well above average. However, Mr. Harrison explained that the rate of change in other states is astounding - Ireland will need to double-down in order to remain above average. In terms of open data, Ireland now places 3rd in Europe. However, in terms of online service completion, Ireland ranks 10th, a position that is likely to worsen, according to Harrison. This is due in part to a weakness in how public services use data online. On pre-filled forms within the websites of public bodies, users must provide the same information again and again. To combat this, Mr. Harrison explained, the Government have launched an eGovernment strategy on the coordination of various online public services - to increase efficiency for citizens and businesses.

Mr. Harrison revealed that the new version of the online site gov.ie will be launched within the next couple of months, with the goal of standardising the homepage for online government services. Harrison views this website as the first step of a 20-30 year programme that will unify and simplify how people interact with the Irish government online.

A challenge they are currently facing is the presentation of online services consistently and with a high editorial standard in both English and Irish. Mr. Harrison sees this as a difficult issue, given the finite amount of resources allocated. However, as the aim is to provide centralised portals under one portal,

the localisation to Irish should be relatively easy: it will be easier to maintain a higher level of consistency and the coding of accents etc. only needs to be carried out once to be applied to all aspects of the platform.

A similar platform from which this project draws inspiration from is the government portal in the UK. All services are online under one portal; citizens don't care which department is associated with the service, they just want to access the service.

Mr. Harrison then described the governance bill: a bill public bodies must legally follow to share data, making clear exactly what is legal and illegal. Relating to this, the new governance board oversees data sharing and advises on the correct procedures.

He concluded his presentation by again stressing the importance of data and data management: good data management gives rise to better and more reliable data.

## 3.7 The European Language Resource Coordination (ELRC) action

Khalid Choukri, ELDA, presented the session describing the ELRC action.

Dr. Choukri began by providing the audience with an explanation of the ELRC. He described it as a coordination that was founded in 2015, and that is headed by 4 organisations: Tilde, ELDA, DFKI and ILSP.

It is supported by National Anchor Points (NAPs): in Ireland's case the NAPs are Micheál Ó Conaire, who is the public administration NAP, and Prof. Andy Way, who is the technical NAP. Prof. Way is also supported by a technical team within Dublin City University.

Dr. Choukri then moved onto the question of 'What does the ELRC do?' He explained that the aim of the ELRC is to try to set up a pipeline between EC services across EU member states, as well as Norway and Iceland. To achieve this, the ELRC collects datasets suitable for developing MT systems: parallel corpora, terminology databases - any digital text expressed in words by human experts.

Dr. Choukri also described the need to identify the various requirements across member states. He said that this is a critical issue, and that it is necessary to engage with each member state to locate and collect existing language resources in a suitable manner.

When the ELRC was first set up, Choukri revealed that they came across some issues, for example: technical and legal difficulties. Following this, a helpdesk was set up intended to deal with all related queries.

The next question posed by Choukri was "Why ELRC?" The answer: to facilitate cross-border interaction. Ireland, along with many EU countries, is a bilingual country and it is important to support communication using tools. As Prof. Way mentioned in his welcome address, we can't ask translators to do absolutely everything, there is simply too much to be done. Translators need support.

Dr. Choukri then revealed the next question: how to make it (MT) work? In-domain text that has been translated by experts is key. Data that is out of domain can be collected, although it may introduce noise to the MT system. In terms of resources already collected, Choukri suggests that the amount of Irish language data that has been collected has only 'scratched the surface' of the amount of data that exists in Irish.

Dr. Choukri concluded his presentation by repeating that help is available for any data holder who needs it, and can be accessed via the online helpdesk.

## 3.8 ELRC in Ireland

Micheál Ó Conaire, Department of Culture Heritage and the Gaeltacht was welcomed to the stage to present the audience with information about what the ELRC has achieved in Ireland so far.

He explained that the National Anchor Points (NAPs) usually involve a representative from public services, and a representative from a technological background. In the case of Ireland, Mr. Ó Conaire himself is the public services NAP and Andy Way is the technological NAP.

He then provided details of the amount of Irish language data that has been gathered by the ELRC so far. Organisations such as UCD *Bord na Gaeilge*, *Foras na Gaeilge* and *An Coimisinéir Teanga* contributed to a collection of resources containing well over 3 million Irish words. Mr. Ó Conaire thanked every person and organisation who had been involved so far for their support, and explained that the reason for his presentation was to convince more data holders to contribute to the initiative. He explained that for a suitable machine translation system to be built for CEF, both a large quantity of language data, and the right kind of data would be required.

In terms of the obstacles that ELRC in Ireland have faced, Mr. Ó Conaire began by mentioning the fact that data collection can be a slow process, partly due to permission from a high level being required before an organisation can share their data. In the case of some organisations, although they were willing to help with the project and share their data, there were difficulties in sharing the right data due to a lack of (mainly human) resources. In terms of legal issues, ELRC worked hard to provide licences that were suitable for the organisations and their data. Some organisations were happy to share data, with certain conditions put in place, and ELRC helped facilitate that.

Another challenge ELRC faced in Ireland, Mr. Ó Conaire explained, was the discrepancy in file names and formats. Although all data is valuable, data in TMX, TXT, Word, or XML format is much more useful and computer-friendly than data in PDF format. In many cases, data-processing was required before being delivered to the ELRC.

Despite these obstacles, ELRC in Ireland succeeded in meeting the data requirements for Ireland for that stage of the project, and have learned some valuable lessons to bring forward to future stages. Moreover, Mr. Ó Conaire explained that with some organisations, data collection worked very well. He gave the example of Foras na Gaeilge, who were very knowledgable about their data, proactive in terms of defining a licence that suited their needs, and willing and capable to share their data.

Mr. Ó Conaire went on to stress the importance of individual relationships and rapport in this type of project. He told of how DCU had hired a summer intern for this purpose, who was able to ring and call in to organisations and discuss their specific situations.

He concluded by urging any Irish language data holders to engage with the ELRC, and impressed on the audience that any issues or worries they might have would be addressed by the Irish ELRC team.

## 3.9 Preparing and sharing data with the ELRC repository

Dr. Teresa Lynn began her presentation by acknowledging that there has been some repetition in the presentations of the day, but that this repetition was done on purpose. The reason for the repetition has been to reinforce vital concepts and ideas, and it is important to make sure that these are clear.

Dr. Lynn explained that her presentation deals with how to prepare and share data, assuming that the relevant data has been identified. She accepted that the audience may have experienced 'information

overload' during the day, but that the ELRC and ADAPT Centre teams are always available to answer technical questions or provide on-site assistance.

Dr. Lynn then moved on to clarify some technical terms. She described data as a very broad term meaning electronically-stored content. In the context of the ELRC, data refers to language data, i.e. text. An example of metadata in this context could be the language of the data, the format of the document, etc.

Within the DGT, MT@EC has been the backbone of eTranslation. The difference between the two, Dr. Lynn explained, is that MT@EC is tailored to legal texts of a certain tone and style, and would not be suitable for translating other domains. MT works by identifying patterns in a statistical way from a large bilingual text, and using an algorithm to predict translations. Therefore if you train a MT system with text from the domain you wish to translate you are going to achieve much better results.

In terms of the domains eTranslation is most interested in, Dr. Lynn names reports, communications, news, web content, policies, terminologies, archives, forms and FAQs. All text formats that are electronically-readable are useful e.g. .TXT, .TMX, .DOC, .XLS. However, this does not include scanned PDFs, as it is very difficult for the computer to understand it as text. Often if a PDF file exists, then the same file in a different format (e.g. Word doc) also exists. If possible to locate, the original document would be more appropriate.

Lynn added that monolingual texts are also very useful - they can be used to build the language model of the MT system, the part that models how fluent Irish should look. She suggested that a lot of organisations may have an internal terminological resource, to maintain consistency in translations. This would also be gladly accepted by the ELRC.

Dr. Lynn concluded her presentation by reminding the audience that the online helpdesk is available for them all, and help will be given to all those who ask for it.

## 3.10 Can language data be shared and how?

Pawel Kamocki, ELRA, took to the podium to discuss the legal aspects of data sharing.

He began by giving an overview of European legislation on reuse of public sector information (PSI). He explained that the transposition of the PSI Directive in Ireland is as follows:

- S. I. No 279 of 2005: European Communities (Re-Use of Public Sector Information) Regulations 2005 (amended by S.I. No. 525 of 2015)
- S. I. No. 30 of 2014: Freedom of Information Act 2014
- Chapter 19 of the Copyright and Related Rights Act, 2000 (on Government Copyright)
- Circular 12/2016, Department of Public Expenditure and Reform

Accompanied by an on-screen image of a cake, he described making public sector usable as 'a piece of cake'. He described all documents held in a collection as a cake that could be split up into several parts. These parts include confidential information that is excluded from PSI, information under a 3rd party copyright that is excluded from PSI, and personal data that is also excluded from PSI. The part of the cake that remains, hopefully more than 50%, is governed by PSI.

Mr. Kamocki then explained the steps one must go through in order to make data usable under PSI:

1. Exclude confidential information
2. Obtain prior informed consent, find a legal basis, anonymize or exclude personal data

3. Ensure there is no 3rd party copyrights, that the material is in the Public Domain or that the necessary licences have been obtained
4. Follow the national PSI transposition rules (e.g. use the national Open Government Licences or the standard procedure for releasing PSI)

If a public body organisation is not sure whether their data is PSI compliant, Mr. Kamocki recommends contacting the national open data portal (data.gov.ie) or the ELRC. He also recommends that public service organisations apply the PSI regulation, follow the General Data Protection Regulation (GDPR), avoid personal data and check with the appeal Commissioner.

Mr. Kamocki moved on to provide examples of licenced ELRC resources in other countries and concluded his presentation by urging the audience to make use of the online helpdesk, and to approach the legal team during the tea break for any detailed questions.

## 3.11 New services provided by the ELRC consortium

Dr. Choukri took to the stage again to discuss the new services that are provided by the ELRC consortium. He explained to the audience that it was his intention to convince them to donate data not only to improve systems within the EC, but also to make such data available to researchers at a later stage, e.g. in the ADAPT Centre.

Dr. Choukri explained that the ELRC can provide support in dealing with IP, technical and legal issues in a suitable and convenient manner. He urged the audience to get in touch with any question they might have about the ELRC and data sharing.

He cited the types of language processing services available from the ELRC as including data conversion, tag removal, reformatting, cleaning, alignment, metadata validation and anonymisation. If a language provider wished to make use of any of these services, Choukri explained, they should make a request for onsite assistance. Following the request, the ELRC would provide the expertise to make the data usable and shareable. When the data undergoes any processing that the ELRC deems necessary, the data-holder would also receive a copy of their cleaned data.

Dr. Choukri then provided the audience with a brief and concise description of how to request services. He explained that all the request involves is the filling in of an online form. He urged the audience to feel comfortable requesting assistance when necessary and stressed the importance of working together. He concluded his presentation by reiterating that the ELRC are prepared to discuss anything data providers feel that they need.

## 3.12 Identifying and managing your data: Questions & Answers

Prof. Dave Lewis from the ADAPT Centre greeted the audience and began his presentation on the identification and management of data.

He explained that his presentation would involve looking at the practical processes in trying to address some of the issues discussed already. Those in research, such as Prof. Lewis himself, are required to undertake data management planning. He explained that this is very good practice when undertaking any activity involving the creation, collection and sharing of data, and all employees should be aware of what their data policy and plan is. He recommended taking the whole lifecycle of data into account, and documenting everything; why is the data being collected? How was it required? What is the associated metadata? Any legal issues?

In terms of curating data, data holders should be aware of actions they can take to make the data more useful and ready to share. Data holders should also be aware of where their data is stored, and how much of their data is suitable for sharing.

Prof. Lewis assured the audience that good data management is a tool that can be used over and over, saving time and resources.

Although the audience were not forthcoming with questions, Prof. Lewis described and answered some frequently asked questions surrounding data management. A selection of these questions and answers are as follows:

**Q:** If a public agency outsources a translation of a text of which it owns the rights, who owns the copyright of the translated version? Can the translation be shared?

**A:** It depends on what the outsourcing contract establishes with regard to IPR. Public agencies should make sure that the outsourcing contract grants them the right to freely reuse and share translation memories.

**Q:** I have created a corpus of literature texts for my research. Can I donate it to ELRC?

**A:** All the texts included in the corpus must be IPR cleared. Some of them, especially old works, may be in the public domain (e.g. if copyright has expired). For the rest, a licence must be obtained from the copyright holders authorising redistribution to third parties.

**Q:** I am the owner of the translation, but not the owner of the source text (or vice versa). Can I share the parallel dataset? What are the necessary steps to take?

**A:** In order to be able to distribute a parallel dataset, IPR must be cleared both for the source text and for the translation. If the source text (or the translation) is copyrighted a licence must be obtained from the owner of the source text (or the translation) to be able to share it with third parties. The first step is thus to contact the owner of the text to find out whether the text is available under an open licence or if a different licensing agreement has to be negotiated.

Prof. Lewis concluded his presentation by urging any members of the audience to approach him or any of the ELRC if they had any further questions.

## 3.13 Discussion and Conclusions

Prof. Andy Way from Dublin City University returned to the podium to round up the ELRC+ seminar.

He urged the attendees not to feel intimidated or uneasy after having heard the in-depth discussion surrounding the legal aspects of data-sharing. The ELRC would much rather that data holders contact the National Anchor Points to deal with the more complicated cases than the data never being shared/identified. He confirmed that the ELRC is available to help with any issue locally. He stressed that the main job of the data holders is to identify possible resources. After that the data processing would pass to the ELRC.

Prof. Way then went on to applaud the three interpreters, and said that he felt grateful and privileged to make use of their services throughout the day.

He reiterated that the main goal of the seminar was to encourage data holders to provide data to make engines that will better serve them. If data is provided, the development team will be in a better position to improve technology that supports translation. In this way, we will all have better access to pan-European services.

Prof. Way noted that translation technology can always be improved - especially so in the context of the Irish language. With the progression of SMT to NMT, he underlined Dr. Lynn's earlier claim that as much data as possible was needed to achieve good results.

He also announced that the ELRC were seeking interested parties to attend the ELRC conference, and to contact Micheál Ó Conaire (Public Admin NAP) for more details.

The audience were reminded to fill out the feedback and engagement forms to benefit future ELRC seminars for other countries, and that certificates of attendance were available to all those who required one.

Prof. Way concluded his presentation, and the ELRC+ seminar, by thanking Gerry Kiely (EC), Micheál Ó Conaire (DCHG), Dr. Aodhán Mac Cormac (DCHG), Dr. Colmcille Ó Monacháin (DGT) as well as Dr. Khalid Choukri and Lilli Smal and their team. He also thanked the local organisation team of the ELRC+ seminar: Dr. Teresa Lynn, Meghan Dowling and Abigail Walsh for their hard work in ensuring the seminar was a smooth and bilingual success.

Dr. Choukri then in turn thanked Prof. Way himself. The meeting was then brought to a close, with the strong feeling that this event -- like the previous event in 2016 -- was a successful one and brought together researchers and potential data donors under one roof; it is only with joint work between these two communities that improvements to MT will be made, to the benefit of the stakeholders supplying the data to train the MT engines.

# 4 Synthesis of Workshop Discussions

## 4.1 ELRC and Open language Data in Ireland

The Open Data subject was given a central focus during the workshop. In July 2017, Ireland launched the Open Data Strategy for 2017-2022. The Department of Public Expenditure and Reform, represented at the workshop by both Ms Morley-Clarke, Head of Open Data Unit, and Mr Harrison, Principal Officer, will lead this 5-year framework shaping the development of Open Data in Ireland in the future, including the key outcome of developing a national data portal, in close relationship with the European data portal team, with an improved language support of a sufficient standard regarding Irish.

Currently, Ireland is doing well both in terms of digital performance, ranking 8th among EU member states, and even better as far as open data is concerned with a 3rd rank in Europe. Both speakers acknowledge the huge benefits of open data for Ireland, in particular social and economic benefits. Data held by open bodies should be available online, and the open data initiative dealing with the reuse of public sector information (PSI) is a step in the right direction. However, special attention should be given to data sharing. The Irish Government has approved the Data-sharing governance bill, a bill public bodies must legally follow to share data and which clarifies what is legal and illegal. Relating to this, the new governance board oversees data sharing and advises on the correct procedures. Data interoperability and standardisation are also of high importance for reusability purposes.

**ELRC Workshop Report for Ireland**

In Ireland, the PSI Directive has been transposed by the European Communities (Re-use of Public Sector Information) Regulations (2005), amended in 2008 and in 2015. The applicable copyright framework in Ireland seems relatively favorable to the re-use of PSI: the chapter 19 of the Irish Copyright and Related Rights Act 2000 foresees that copyright in works created by public servants in the course of their duties belongs to the Government (the exception for works communicated to the Government is provided in section 75 of the same Act).

Moreover, Ireland (unlike many other EU Member States) has adopted a clear recommendation regarding the use of licenses for re-use of PSI (Department of Public Expenditure and Reform, Circular 12/2016 of 26 April 2016, which recommends the use of CC BY 4.0). Despite this overall favorable context, some doubts persist among stakeholders which affect the production of language resources from PSI. These doubts seem to be mostly related to copyright ownership in translations made by freelance translators (it is not clear to some public sector bodies whether they hold copyright in such translations, and consequently whether they can be released and re-used without infringing the rights of the translators). This seems, however, a matter that can be solved by adopting good practices for the future, rather than one that would necessitate a revision of the legal framework. A more burning issue is the one of time wasted for seeking appropriate permissions from further up the hierarchy; in other words, it is difficult for stakeholders to pin the responsibility for making PSI re-usable down to a specific organ of the public sector body. This may necessitate a more substantial intervention (e.g. appointment of a 'PSI manager' in every public sector body).

#### 4.1.1 Success stories and lessons learnt

**2nd Workshop Effect**

The overall content of the 2017 workshop led to much more engagement and understanding from the attendees, compared to the 2016 Workshop content. It seems that keeping it more high level is particularly effective, as is the avoidance of too much technical explanation of how MT works. In general, the audience benefitted more from the workshop this time around. They felt more confident in understanding what was required from them - and as such will feel more confident communicating it to respective management or decision makers in their respective organisations.

### 4.2 Session Questions

#### 4.2.1 Questions from the session The CEF eTranslation @ work

Síne Nic an Ailí from Conradh na Gaeilge asked whether it was necessary for data providers to have technical knowledge. Mr. Ó Monacháin responded that technology plays a central role in the work of a translator, and translators should take advantage of new facilities, for example translation environments such as Trados.

#### 4.2.2 Questions from the session National initiatives for digital public services and (open) data

Mr. Harrison was asked whether everything on the new portal is to be made available in Irish. He replied that yes, it would be - not just menus but every single aspect.

Mr. Harrison was also asked, by Mr. Ó Monacháin, about data protection, and what to do if there is doubt surrounding it. He replied that in the general sense of data protection, one must approach the sharing/processing of data in the following manner: every public body should be seen as the data controller, with a series of legal obligations to honour, all aiming to achieve the highest standards possible. In terms of the reuse of data, it is important to have good governance, with standards that

govern how data is shared, e.g by having governance board. He highlighted the importance of transparency in data management.

Kevin Boushel, Irish officer for Maynooth University, asked when Irish citizens can begin using gov.ie. Mr. Harrison replied that the first launch would be at the end of this year, and would just consist of a portal linking to the right government department. He expects this to hold some value as a first step, incrementally adding more functionality and combining content over hundreds of government-related websites as appropriate.

### 4.2.3 Questions from the session The European Language Resource Coordination (ELRC) action

Síne Nic an Ailí from Conradh na Gaeilge asked whether the ELRC is interested in data from the public, in other domains. Dr. Choukri replied that the ultimate goal of the ELRC is to develop MT for public administrations, and therefore the best data is reliable data from public bodies. However, the algorithms are very robust, so quality is not an issue as long as it has been translated by professional translators.

### 4.2.4 Questions from the session *ELRC in Ireland*

Mr. Ó Conaire was asked which language direction the machine translation system is capable of translating in (English to Irish or Irish to English). Dr. Judge, ADAPT Centre, answered that with parallel data translation can work in both directions, and theoretically doesn't make a difference. At the moment the demand is for English to Irish translation but it isn't an issue for developers to reverse that process.

### 4.2.5 Questions from the session Preparing and sharing data with the ELRC repository

Seán Hayde, from DGT, enquired whether there has been any work done regarding artificial intelligence, with which the computer could translate text that it had not seen before.

Dr. Lynn answered that translation memory (TM) is often used to provide the translator with similar phrases that have been previously translated. MT, however, can be used to translate text it hasn't seen before using statistics and pattern-matching. A helpful environment for translating would be a dual-system, combining both TM and MT. TM would provide the previously-seen text, and MT could fill in the gaps for the unseen text. This is currently the set-up for the Tapadóir system, which is integrated inside the Trados TM tool.

Aislinn McCrory (DGT), said that while the MT systems Dr. Lynn described have been built using corpora, she has heard that there is another method for building a MT system based on rules. She wondered whether there was any work with this type of system in relation to Irish. Dr. Lynn replied that there is a group in Trinity College Dublin developing a rule-based machine translation (RBMT) system, but the problem is that it takes a long time to build the complex rules needed for a successful RBMT system. In the future, a hybrid system is envisaged. At the moment, the Tapadóir system contains some rules necessary for automated post processing but a fully hybrid (RBMT plus corpus-based MT) system is the intention going forward. This will take some time, and may be superseded by the significant impact that NMT has already had.

John Toolan, Rannóg an Aistrucháin, noted that in his opinion, translation standards are declining in some State organisations in Ireland. He was interested in the Tapadóir system, and whether NMT is being considered, that works not just on statistics but also on themes. Dr. Lynn replied that she had two answers. One is that the EC plan is to implement NMT for all languages. The other is that in the

point of view of the Tapadóir development team, there is not yet enough data for NMT to be considered for Irish. Initial experiments revealed very poor results for Irish-English NMT due to the small amount of data available. NMT is the trajectory, but in Lynn's opinion much more data will have to be gathered before it is viable. Of course, engaging more stakeholders as part of the ELRC programme will make NMT more viable in the future for Irish.

#### 4.2.6 Questions from the session Can language data be shared and how?

Mr. Kamocki was asked if the data holder wanted to submit existing translations to the ELRC but didn't have copyright clearance from the translator, would that be OK, and would they be copyrighted to the translator? He replied that if they are translated by a 3rd party they are covered by a 3rd party copyright so the data provider should ensure that the copyright be transferred back to them.

Dave Lewis, ADAPT Centre, mentioned that Mr. Kamocki discussed that when getting clearance on TM, permission is required for both the source and translation. He enquired whether Mr. Kamocki had seen any examples where someone has claimed ownership following alignment procedures. Mr. Kamocki replied that alignment is not protected by copyright. Mr. Lewis followed up by asking what amount of effort put into structuring databases would deem them suitable to be copyrighted. Mr. Kamocki replied that it should be taken on a case-by-case basis, but in his opinion aligning two documents is not enough to claim an exclusive copyright.

#### 4.2.7 Questions from the session New services provided by the ELRC Consortium

Colmcille Ó Monacháin, DGT, asked whether the form is available in Irish. Dr. Choukri replied that the website is multilingual so he would presume so, but confessed to not knowing for sure. He said that he would strive to make sure it was, as his 'homework'.

Dr. Choukri was then asked a question regarding the anonymisation service that the ELRC could provide. In terms of the EC, the penalties for a breach of anonymisation are much more serious. How would the risk for non-anonymisation break down between the ELRC and the source organisation? Dr. Choukri replied that the potential user is the EC, therefore they could not take any risk regarding anonymisation. If it could be done easily, e.g. within tax forms, then identity could be hidden but wherever there is a risk the data could not be accepted.

#### 4.2.8 Questions from Section the session *Identifying and managing your data*

Prof. Lewis was asked a question regarding extracting terminology from a large amount of text - could copyright be gained? He advised that if it was done manually there may not be issue with copyright, but it is a 'grey area', that should be taken on a case-by-case basis.

### 4.3 ELRC and Ireland's Open Data Portal

Please see Section 3.6

## 5 Workshop Presentation Material

The presentations are published in both Irish and English on the respective agenda pages (http://lr-coordination.eu/ga/l2ireland_agenda and http://lr-coordination.eu/l2ireland_agenda).