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 Ireland

Author(s): Meghan Dowling, John Judge, Andy Way

Dissemination Level: **Public**

Version No.: V1.0

Date: 26/02/2016

Contents

1. Executive Summary

This document reports on the ELRC Workshop in Ireland, which took place in Dublin, on the 28th of January 2016 at Europe House. It includes the agenda of the event (section 2) and briefly informs about the content of each individual, interactive and panel workshop session (sections 3 & 4). The event was attended by 37 participants spanning a wide range of ministries and public organisations.

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

2. Workshop Agenda

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

Dublin, January 28th, 2016

Europe House

08:00 – 09:00 Registration

09:00 – 09:05 Opening and welcome

09:05 – 09:15 Local Welcome - Prof. Andy Way (ADAPT Centre)

09:15 – 09:25 Welcome by the European Commission - Tim Hayes (European Commission Representation in Ireland)

09:25 – 09:35 Aims and Objectives - Dr. Khalid Choukri (ELRC)

09:35 – 09:50 Europe and Multilingualism - (Markus Foti DG Translation and MT@EC)

09:50 – 10:20 Languages and Language Technologies in Ireland - Dr. Teresa Lynn (Adapt Centre)

10:20 – 11:00 Panel: Multilingual Public Services in Ireland - Moderator: Séamus Mac Giolla Chomhaill (Dept. Arts, Heritage & Gaeltacht)

11:00 – 11:30 Coffee Break and Networking

11:30 – 12:00 Automated Translation: How does it work? - Prof. Andy Way (ADAPT Centre)

12:00 – 12:30 How can Public Institutions benefit from the CEF.AT Platform? - Colmcille Ó Monacháin (DG Translation)

12:30 – 13:30 Lunch Break

13.30 – 14.00 What Data is needed? - Yvette Graham (ADAPT Centre)

14:00 – 14:30 Legal framework for Contributing Data - Khalid Choukri (ELRC)

14:30 – 15:00 Interactive Session: Data and Language Resources in Ireland (Facilitators: Dr. Briain Ó Raghallaigh (Fiontar, DCU) and Séamus Mac Giolla Chomhaill (Dept. Arts, Heritage & Gaeltacht)

15:00 – 15:30 Coffee Break and Networking

15:30 – 16:00 Data and Language Resources: Technical and Practical Aspects - Dr. Joss Moorkens (ADAPT Centre)

16:00 – 16:30 Interactive Session: How can we engage? - Facilitator: Dr. Aodhán Mac Cormaic (Dept. Arts, Heritage & Gaeltacht)

16:30 – 17:00 Best Practice for the Future - Capitalizing on your Valuable Data - Prof. Dave Lewis (ADAPT Centre)

17:00 – 17:15 Wrap-up, On site Conclusions and Commitments - Prof. Andy Way (ADAPT Centre)

3. Summary of Content of Sessions

3.1. Opening

Dr. John Judge, opens the event by welcoming the audience and introducing the key persons in conceiving and organizing the event, namely the ELRC consortium and the EC/DGT representatives.

3.2. Session 1: Local Welcome

Prof. Andy Way, the national anchor point in Ireland, welcomes the participants to Europa house in Dublin.

He extends a special welcome to Dr. Khalid Choukri, who has travelled from the ELRC Workshop in Madrid to attend.

He states that Horizon 2020 funding will crack the language barrier.

He notes that there are similar consortia in other European countries, with the aim of inviting people to share data. The ultimate aim of the ELRC workshops are to gather as much data as possible, and to raise awareness of the benefits of data sharing. Data, especially data of a certain quality, is crucial in machine learning and translation systems, and it is possible that there is a lot of data in Ireland which could be put to better use.

He expresses his thanks to Colmcille Ó Monacháin from DG Translation, Aodhán Mac Cormaic and Séamus Mac Giolla Chomhaill from the Dept. of Arts, Heritage & Gaeltacht, Dr. John Judge and Dr. Teresa Lynn from the ADAPT Centre, and also colleagues from the North of Ireland, whose presence ensured that this workshop was an all-Ireland event.

3.3. Session 2: Welcome by the EC

Tim Hayes, Deputy Head of the European Commission Representation in Dublin begins by quoting Jean Monnet: "We do not create federations of States, we are uniting people", and highlights the still present need for unity among European citizens.

He tells how the Connecting Europe Facility (CEF) was created to improve interaction between Europeans, and specialises in trans-European digital services. However, as many of these services cannot fully benefit European Citizens if they don't speak the language, this increases the need for translation. Legislative acts of the EU must be translated into 24 languages, and this creates a translation burden. Following the 2004 enlargement, the EU dealt with this increased burden by investing in translation technology, i.e. translation memories and machine translation.

He explains that the CEF.AT want to help the Irish public administration to perfect the EU machine translation tool for the Irish language, and in doing so save time and money and provide a better service to Irish citizens. He stresses the importance of ensuring our technological independence, before becoming obliged to pay for machine translation at a high price.

He mentions the derogation on Irish language translations in the EU, which is due to lift in 2020 and will bring both challenges and opportunities. Irish translation demand in the EU will increase tenfold, and for this to be feasible significant improvements in Irish machine translation will be necessary. As well as this, he highlights the additional benefits to Irish machine translation - it will further protect the Irish language and cement it as a living language, flourishing in the digital age.

3.4. Session 3: Aims and Objectives

Dr. Khalid Choukri (ELRC) begins by thanking the organisers.

A multicoloured map depicting the different languages tweeted in during a single evening in Europe is used to highlight the multilingualism of Europe, and Choukri poses that multilingualism is Europe's greatest asset. Languages are core to the rich tapestry of European cultures and identities and it is important that no language is discriminated against. He impresses that in a multicultural setting such as Europe it is crucial that language barriers do not restrict the mobility of people, information, ideas or commerce and therefore the true language of Europe is translation. However, he concedes that to translate all of Europe's content is a monumental task, with the example of 500 million tweets being sent in Europe every day. He then provides a solution - translators supported by automated translation.

He then describes the automated translation of the Connecting Europe Facility (CEF.AT). While automated translation has been carried out by the EC for many years, the CEF.AT is different. It focuses on the translation needs of citizens and public bodies and is provided freely.

To ensure that CEF.AT provides a service of the highest standard, high quality data is needed. The translation system will learn from high quality human translations, and in turn produce a high quality translation service. Therefore in order for this to work, public bodies need to share the right kind of data with the CEF.AT and in return they will receive better support for their own language and multi-linguality in Europe.

3.5. Session 4: Europe and Multilingualism

Markus Foti, from DG Translation and MT@EC begins his talk with an overview of European policies, at the core of which are multilingualism and linguistic diversity. He reminds us that while the EU now recognise 24 official working languages, this has grown from an original 4 languages and is continuing to grow today, e.g. Catalan and Basque. He explains that the initial approach to translating content into all official languages was to hire more translators, but this proved to be an insufficient solution. Language is a national matter, and there is no lingua franca in the EU. As well as this, human translation for this amount of content is too slow and too expensive. In the context of the Digital Single Market, Foti emphasises the need for help from EU member states in creating translation tooling to break down language barriers and ensure that public digital services are available for all EU citizens. Part of the solution is to gather good quality data from member states, and develop language resources with the CEF Automated Translation Platform. He highlights the advantages of member state collaboration with CEF.AT, among which are free automated translation, easier sharing of information, and more accessible public services for all European citizens.

3.6. Session 5: Language and Language Technologies in Ireland

Dr. Teresa Lynn, ADAPT Centre, begins by providing a background to the Irish language, noting the urban revival of Irish, as well as the increased usage of Irish online. Census data showing languages spoken other than English or Irish highlight Ireland as a multilingual European nation.

Moving on to Irish language technology, she mentions the strong historic link to translation and localisation in Ireland, and presents the Irish language technology resources which are already present or under development. She also stresses the lack of resources available for Irish technology, which severely inhibit the development of modern language processing technologies for Irish. While some resources exist, Dr. Lynn emphasises the need for more good quality open-source resources, which, in consultation of a graph depicting language resources for EU languages, are very sparse for Irish.

The presentation then describes the English-Irish statistical machine translation project 'Tapadóir', developed by the ADAPT Centre and funded by the Dept. Arts, Heritage & Gaeltacht. She explains that Tapadóir was developed as a tool to aid human translators in the DAHG, improve productivity and reduce translation costs. It has been integrated into the workflow of the professional translator, who has full control over the final translation. It is trained on translations from the DAHG, and is optimised to suit translations of this type. Dr. Lynn presents the key results of the Tapadóir project, which include better automated testing results than the online translation tool Google Translate.

She concludes by stressing that everyone has a role to play in Irish language technology.

3.7. Session 7: How does machine translation work?

Prof. Andy Way, the national anchor point in Ireland and the deputy director in the ADAPT Centre, begins by posing the question 'Why MT?'. Reinforcing previous presentations, he answers by highlighting the multilingualism of Europe and the cost of translating such a huge volume of data. He then explains the challenges associated with MT, among which are the elegant and complex nature of natural languages, cultural differences, and ambiguity. He readily concedes that MT is imperfect, an approximate solution that requires human professional translators to perform post-editing.

He describes the process of MT by using a simple bilingual example to first examine word to word translation, and then phrase-based translation. In an example of a bilingual menu, the need for data is stressed - "an MT system based on this data would know a great deal about Chinese soup translation, but very little else!".

In a section entitled 'ELRC: The Wider Context', the goals of the Multilingual Digital Single Market and the CEF.AT are laid out.

To conclude, Prof. Way stresses the importance of data for MT, and describes the type of data needed by the CEF.AT in order for everyone to benefit.

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

Colmcille Ó Monacháin, DG Translation, begins by using a diagram to illustrate the complicated interaction between administrations, citizens, and businesses in two EU member states and appeals to the audience to imagine the potential language barriers in these situations.

Ideally, within the EU, all citizens would be able to access all EU information in their own language.

The public services in Ireland are working hard to provide access to services to the Irish language community, and CEF.AT aim to help this process. MT@EC already provide an online translation service, with which users are ensured integrity and security. However, there is a discrepancy in corpus size among EU languages and Irish is very low in comparison.

CEF.AT will build on the existing MT@EC service, putting emphasis on secure, quality, customisable MT. It will provide public institutions with faster and more secure translation, better quality translation in general and include translation for specific domains. However, for this to become a reality, a great deal of linguistic data is necessary, especially for the Irish language.

A question regarding the cost to state bodies is posed to Colmcille, to which he replies there is no cost. Regarding the usability of CEF.AT, he replies that any state body can use it as long as they register. It includes a user-friendly interface, and the maximum time delay for a translation is 17 seconds.

3.9. Session 9: The data for CEF-AT

Dr. Joss Moorkens, ADAPT Centre, describes the data needed using the image of a bag of words. He explains that anything with words can be useful; both monolingual and parallel documents, dictionaries, brochures, etc. He explains that all document types can be useful, and that optical character recognition (OCR) could even be carried out on non-digital corpora. He describes the concept of metadata, and which metadata in particular would be useful for MT.

Dr. Moorkens also explains the process of creating a linguistic resource from data, from collecting bilingual data, to cleaning, tokenising and aligning it. He shows the positive correlation between more data and better results, and reaches out to public institutions to share the data which isn't easily accessible.

He concludes by asking the audience not to underestimate the value of their resources.

3.10. Session 10: Legal Framework for Contributing Data

Dr. Khalid Choukri, ELRC, describes how the Public Sector Information (PSI) rules address the need for a clear and easy-to-follow regime for data reuse across the EU, and provide legal and technical interoperability. They cover all public information, and regulate the exchange of public information between users. Dr. Choukri explains that PSI rules sit above copyright laws in terms of the structure of rights and ensure that, when

the right conditions are met, public sector information is available to third parties for reuse with the minimum possible frictions.

Dr Choukri explains that, regarding the ELRC, data will not be reused as such, rather data will be used to create language models, from which the original data will be impossible to extract. He mentions the very permissive government licence in Ireland, and concludes by inviting participants to speak to the legal experts for more information.

3.11. Session 12: Data and Language Resources: Technical and Practical Aspects

Dr. Joss Moorkens, ADAPT Centre, takes to the podium again to speak about the issues surrounding data collection and curation. To put this into context, he describes the collection and curation process, listing the possible challenges at every step. He explains that public sector collaboration during these steps could eliminate a lot of the issues; some reasons for this being that public institutions know their own data, and the licenses associated with it, and also that access to derived forms of data (e.g. PDF) is less efficient than access to internal source content repositories.

When asked if there is a limit to the amount of data that could be uploaded, Dr. Khalid Choukri answers that there is no limit, and that the ELRC would be grateful for as much data as possible.

3.12. Session 14: Best Practice for the Future – Capitalizing on your Valuable Data

Prof. Dave Lewis, Associate Director at the ADAPT Centre, gives advice regarding data management. He argues that public bodies have an obligation to make a data plan, to find out more about their licence terms and assert ownership of their data. He stresses that a data management plan is a reusable tool that will ensure the quality of data produced. Prof. Lewis shows the benefits of using the Web as an additional publication channel, and advises on the best practices for doing so.

3.13. Session 15: Wrap-up, On Site Conclusions and Commitments

Prof. Andy Way (ADAPT Centre), summarises the day by emphasising the value of data. He explains that the goal of the ELRC workshop was to emphasise to public bodies that their data is valuable, and that everyone can benefit from data sharing.

He also stresses the importance of maintaining this conversation, especially as at the moment only 5% of the information that should be translated, is translated.

Prof. Way brings up the recurring theme of quality, and how to define an acceptable level of quality. He defines some of the ELRC user groups; MT@EC, government departments, county boards, 3rd level institutions, etc. He explains that efforts would be made to match the user group to the appropriate use-case or service. He emphasises that Google are only interested in profit, and have no associated security.

Prof. Way concludes by thanking everyone for attending, in particular the host, the interpretations team and the audience.

4. Synthesis of Workshop Discussions

4.1. Panel 1: Public multilingual services in Ireland

Moderator: Séamus Mac Giolla Chomhaill, Dept. Arts, Heritage & Gaeltacht

Panellists:

Dr Aodhán Mac Cormaic, Dept. Arts, Heritage & Gaeltacht
Tony O' Dowd, Kantan MT
Markus Foti, DG Translation
Dr. Patrick Cadwell, SALIS DCU
Orla Ryan, Lionbridge

The moderator, Séamus Mac Giolla Chomhaill, Dept. Arts, Heritage & Gaeltacht, begins by laying out the objectives of the panel: to hear first-hand from multilingual public services and provide useful information to the ELRC.

He mentions that the temporary derogation placed on Irish language translation by the EU will be lifted in the year 2020, and also notes the positive impact of the Official Languages Act (2003) which is still present. He stresses the importance of providing Irish language services, which are dependent on good resources. In this regard, corpus development is extremely useful.

Dr. Aodhán Mac Cormaic, Dept. Arts, Heritage & Gaeltacht, speaks about how Ireland can learn from good language practices in other countries. He speaks of important Irish language legislation, including the Gaeltacht Act, and the 2005 inclusion of Irish as an official EU language. He mentions the importance of bilingual documents in the government, e.g. court rules and road signs. Dr. Mac Cormaic describes the government decision to end the derogation at this time as a 'pain point'. He mentions the increased demand on human translation this will bring, as well as the need for automatic translation as a translation aid.

Orla Ryan, Lionbridge, describes her experience with training translators to work with technology. She mentions that 42% of Foras na Gaeilge translators are not available for translation.

Tony, O' Dowd, from Kantan MT describes the work of Kantan as solving the problem of language for clients. He tells of a revolution within the EU, of communicating multilingually instantaneously. According to O'Dowd, language is the biggest barrier to international trade - a huge challenge but with huge benefits, and Ireland has a long way to go in order to be on the same scale as other EU member states. Ireland has a responsibility to uphold their national identity. The Irish language is almost digitally extinct, and MT is necessary to preserve a language or it will become extinct. He describes the productivity gains associated with MT, mentioning that in the time taken to translate 2000 words, MT-aided translation could produce from 4000-6000 words.

Dr. Patrick Cadwell, SALIS DCU, describes his work on a project which dealt with the human experience of using technologies such as those mentioned. He tells how he questioned human translators about how they feel about using translation memory or machine translation in their work, and which factors are taken into account when they

choose TM or MT. He describes the first stage of his project, which involved speaking to 70 translators in small focus groups and expresses that he wishes to speak on behalf of these translators. He emphasises the human aspect of MT, and expresses that the most important things to translators when using language technology are quality, freedom and control.

Markus Foti, DG Translation, discusses the value of working to preserve a language. He tells of his previous experience as a translator, and also mentions the divergent attitudes to TM - it being the default in the present day, compared to the strong resistance it first met. He predicts the same situation will unfold with MT. Translators view MT as the 'enemy', but Foti insists that it is there to help, and aims to produce something that is good enough to help human translators.

MT usage by human translators is questioned, with Patrick Cadwell answering that based on his focus groups, the majority of translators use MT, but it depends on the task, and Markus Foti answering that from his recent tracking he has found that more than 50% of the time MT is chosen to aid the translation.

The quality of translation is raised as a concern. This is addressed by explaining MT as a facility to be used by translators. The MT output isn't perfect, and post-editing always needs to be performed to ensure a high quality translation. The misuse of MT results in poor results, and negative attitudes towards MT. It is stressed that training and informing users is crucial and that a human must always be involved in the translation process.

Questions are raised about how MT achieves a certain quality, and why some language pairs achieve better translations. The quality of data and the amount of data are marked as two very important factors. The difference in languages is also noted - Irish and English are very different languages with divergent word orders, and the richness of Irish morphology can prove challenging when translating from English. However, it is also emphasised that English-Irish MT is still instantly deployable, with the example of the Tapadóir engine used by translators in the DAHG as evidence.

Overall, Dr. Cadwell's statement of "quality, control and freedom" defines the sentiment of the discussion. These are, he says, the 3 most important factors to the human translators he interviewed when choosing tools and mechanisms to support their work.

4.2. Interactive Session 1 S11: Data and Language Resources in Ireland

Facilitators:

Dr. Briain Ó Raghallaigh, Fiontar
Séamus Mac Giolla Comhaill, Dept. Arts Heritage & Gaeltacht

Dr. Briain Ó Raghallaigh begins by describing the language resources made available by Fiontar, e.g. logainm.ie, the national placenames database. He explains that Fiontar aim to make their data as accessible as possible - both user and machine friendly. Data is machine readable, open licence, available for download but also accessible through a user-friendly interface which is constantly updated.

The types of data which qualify as useful were questioned and discussed, and it was stressed than any data has the potential to be useful, whichever format it is in. For example, bilingual files, .TMX files, dictionaries, monolingual Irish files and even scanned versions of hand-written data were identified as having potential use.

Some questions arose concerning 'An Caighdeán', the official standardised form of Irish. Dr. Ó Raghallaigh explained that in Fiontar, An Caighdeán is always applied, and that there are computational methods of doing so. Regarding older data that may not utilise An Caighdeán, Dr. Andy Way explained that including a lot of this un-standardised data would harm the MT engine, but a suite of smaller systems could be built, with one of them specialising in translating data from before An Caighdeán was made official.

Some concern was raised about the loss of language niceties or a particular translator's personality if MT was used. In response to this, Dr. Way advised that it would be possible to build a MT engine based solely on one translator's work, and thus preserving his/her style but stressed that MT is always post-edited by a human translator, and that is where the opportunity for style is introduced. Máirín Ní Mharta, a translator from the DAHG who uses MT in her workflow (the previously mentioned Tapadóir engine) reinforced this point, saying that she would never submit something that she didn't write herself, and also feels that the MT engine she uses is learning from her style.

4.3. Interactive Session 2 S13: How can we engage?

Facilitators:

Dr. Aodhán Mac Cormaic, Dept. Arts, Heritage & Gaeltacht
Dr. Patrick Cadwell, SALIS DCU

Dr. Aodhán Mac Cormaic, DAHG, begins by expressing that he understands that the free sharing of data will be a difficult sell to employers, but implores the audience to try their best.

Dr. Patrick Cadwell describes the desired outcomes of this interactive session - to see what the opinions and feelings of the audience are regarding ELRC, and Dr. Mac Cormaic adds to this by asking if everyone sees the importance of their data.

The audience respond that they could see the importance, but are unsure if their employers would, and express the worry that other people may not see the advantages.

Who data would be shared with is also a talking point. Dr. Khalid Choukri explains that the aim of CEF is to gather data and give it to the public services, but would also be open to sharing for research purposes. Some audience members express that as their organisations have paid for professional translations it would be hard to persuade them to share the translations.

It is suggested that ELRC speak directly to the legal teams associated with the institutions, but this is rejected, citing that if someone from their own institution who has been informed of the benefits tries to convince the management it would be much more effective.

5. Workshop Presentation Materials

All presentations are available online on the ELRC website: http://www.lr-coordination.eu/ireland_agenda and also http://www.lr-coordination.eu/ga/ireland_agenda

Tim Hayes, Deputy Head of the EC Representation in Dublin welcomes Delegates

Panel session on Multilingual Public Services in Ireland

Dr. Teresa Lynn speaking on Language Techn