European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1091 programme.

European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.12
Task 3

ELRC Workshop Report for Ireland

Author(s):	Teresa Lynn
Dissemination Level:	Public
Version No.:	<V1.0>
Date:	2021-11-10

Contents

1 Executive Summary

This document reports on the ELRC3 Workshop in Ireland, which took place online on the 24th of June 2021. This report includes the agenda of the event and briefly provides details on the content of each presentation. We also report on updates in terms of activities in Irish LT, lessons learnt for future events and challenges that still exist in the Irish language technology landscape. Finally, polls were carried out during the workshop to gather information to update the Country Profile. The results are available in the accompanying document and a summary of the findings are outlined here. While the small number of participants and level of feedback is not sufficient to result in significant changes to the country profile, it was a useful exercise to see the continued trend in perspective as to how the Irish language is supported through technology.

The event was attended by 54 participants (including organisers), spanning a wide range of representatives from Irish government departments, universities, public organisations and Small to Medium Enterprises (SMEs).

This was the first time we held the ELRC Workshop online. One of the major challenges in organising this workshop related to the lack of availability of interpreters. The workshop incidentally coincided with an EU parliament event related to the Irish derogation and as such, many Irish-based freelance interpreters were unavailable. In the end we needed to enlist the services of a third-party interpreting service recommended by the ELRC (Interactio[1]).

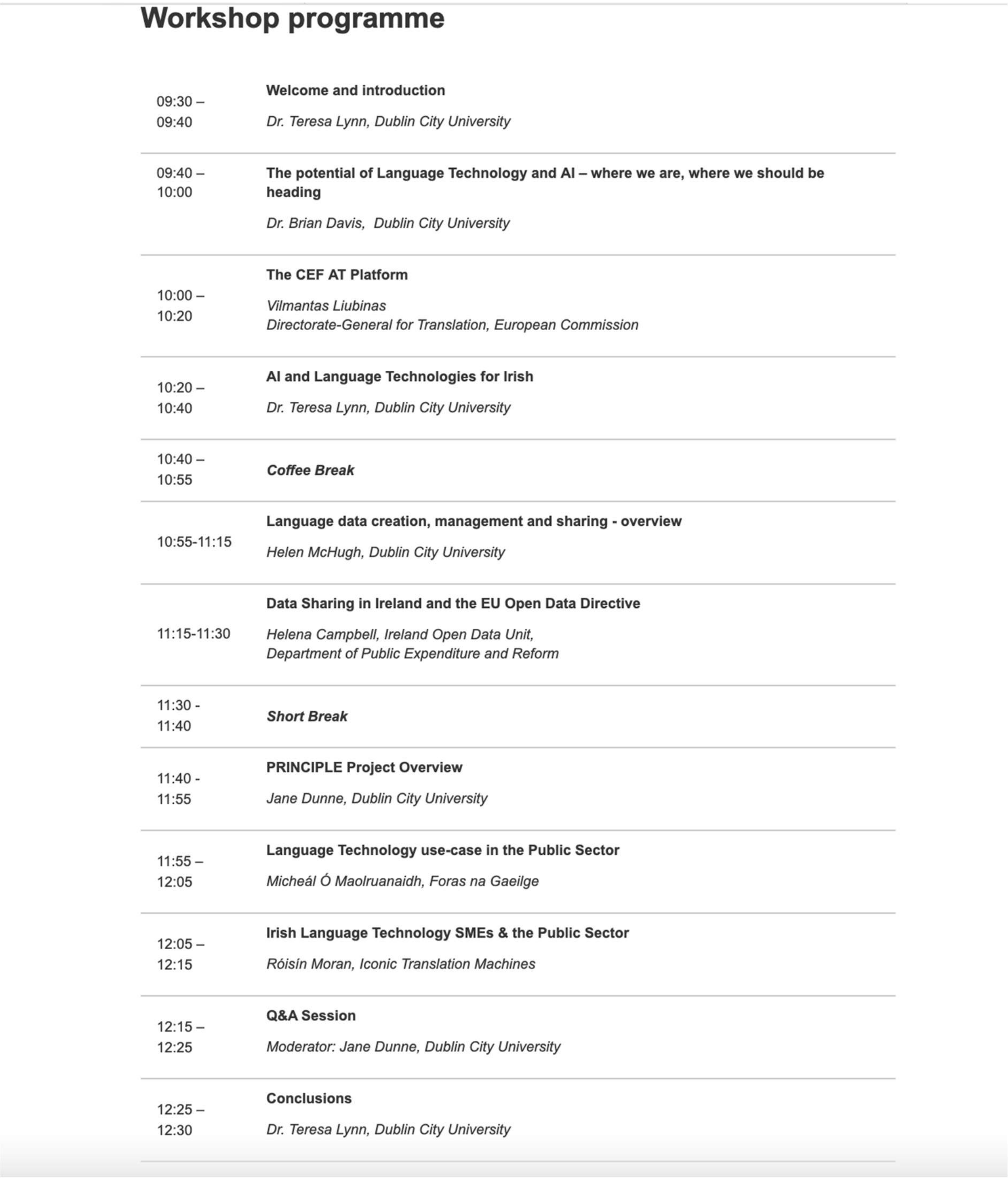
Point of note with respect to the deviation in some parts from the European Commission's suggested format of the ELRC3 workshops:

1. There is not enough LT activity for Irish or sufficient examples of exemplary data sharing in Ireland to have enabled a meaningful panel discussion at any stage during the workshop. Instead, presentations by experts were delivered on the relevant topics to help inform how we can improve in this respect.
2. The workshop was not used as an opportunity to co-host a European Language Grid (ELG) workshop, as the ELG community/ audience will be both English and Irish speaking (the ELRC workshop was aimed at Irish speakers and those working with Irish language content only).
3. The significant lack of Irish language technology in SMEs meant that the proposed focus on promoting eTranslation and language technologies to SMEs was not possible. In fact, the only SME present at the workshop develop their own machine translation (MT) engines.
4. The workshop evaluation poll was not carried out at the end of the workshop as a large number of attendees had left by the final presentation. A link to the online evaluation form was sent in a follow-up email to attendees that also included a link to the slides and recordings. A 3 week delay in the workshop webpage being updated with this content led to a delay with this follow-up email.

The dedicated event webpage can be found at https://www.lr-coordination.eu/ireland3rd

[1] https://www.interactio.io

2 Workshop Agenda



Workshop programme

09:30 – 09:40	**Welcome and introduction** *Dr. Teresa Lynn, Dublin City University*
09:40 – 10:00	**The potential of Language Technology and AI – where we are, where we should be heading** *Dr. Brian Davis, Dublin City University*
10:00 – 10:20	**The CEF AT Platform** *Vilmantas Liubinas* *Directorate-General for Translation, European Commission*
10:20 – 10:40	**AI and Language Technologies for Irish** *Dr. Teresa Lynn, Dublin City University*
10:40 – 10:55	***Coffee Break***
10:55-11:15	**Language data creation, management and sharing - overview** *Helen McHugh, Dublin City University*
11:15-11:30	**Data Sharing in Ireland and the EU Open Data Directive** *Helena Campbell, Ireland Open Data Unit,* *Department of Public Expenditure and Reform*
11:30 - 11:40	***Short Break***
11:40 - 11:55	**PRINCIPLE Project Overview** *Jane Dunne, Dublin City University*
11:55 – 12:05	**Language Technology use-case in the Public Sector** *Micheál Ó Maolruanaidh, Foras na Gaeilge*
12:05 – 12:15	**Irish Language Technology SMEs & the Public Sector** *Róisín Moran, Iconic Translation Machines*
12:15 – 12:25	**Q&A Session** *Moderator: Jane Dunne, Dublin City University*
12:25 – 12:30	**Conclusions** *Dr. Teresa Lynn, Dublin City University*

3 Summary of Content of Sessions

3.1 Welcome and introduction

Opening was made by Dr. Teresa Lynn, Technical National Anchor Point (NAP) for Ireland. Firstly, those who have been part of the ELRC workshop series since 2016 were welcomed back again, and a warm welcome was then given to newcomers. A quick run-through of technical advice was given (moderators' names, Zoom tips, *Interactio* interpreting app tips, alert that recording would take place, that polls would be conducted and that certificates of attendance were available on request).

The second part of the talk highlighted the motivation behind the workshop, how Europe is working towards a digital single market (DSM) and aiming to reduce language barriers. Digital Service Infrastructures (DSI) were explained and the benefits of the DSM. A brief introduction to the concept of speech and language technologies was given before explaining the role of CEF and the ELRC. The National Anchor Points were introduced (Dr. Aodhán MacCormaic named as stand-in as the previous Public NAP, Micheál Ó Conaire, has now moved on to a new role). The Digital Europe Programme was introduced and the presentation was finished off with a run-through of the agenda.

Session 1. Connecting a multilingual Europe through Language Technology

3.2 The potential of Language Technology and AI – where we are, where we should be heading

This presentation was delivered by Dr. Brian Davis, a lecturer at the School of Computing, DCU and an expert in Natural Language Processing/ Natural Language Generation. Much of this talk focused on explaining what AI was, how it relates to our lives, how it will impact our future lives, how the EU wants to harness its strengths to help ensure that we are self-sufficient in terms of language-centric Artificial Intelligence (AI). Language technology was explained in more detail, as was machine translation. With that, the concept of data-driven systems was explored providing context for the workshop and our aims in data collection. With respect to the Digital Economy and Society index it was pointed out that while Ireland is doing well compared to many other member states, this only relates to English language AI.

3.3 The CEF AT platform

This talk was delivered by Vilmantas Liubinas a computational linguist at Directorate General of Translation (DGT). The CEF platform language tools were introduced first with the acknowledgement that eTranslation is available in Irish for use by those in public administrations, universities, SMEs and CEF DSIs. A clear visual graphic was provided to demonstrate just how much text eTranslation has translated to date. The notion of domain was introduced and explained, and the eTranslation user-page was also shown to demonstrate the options available to users. Integration into translators' tools was explained along with a discussion of translation quality. Neural networks were explained briefly to show how these systems actually work, which led to the topic of the importance of the availability of parallel data and how it is required to build a reliable MT system.

3.4 AI and Language Technologies for Irish

There are so few AI and Language technologies for Irish (one SME builds Irish MT models and a few researchers are still working on building tools fit for purpose). As such, there was no scope for a fruitful discussion and instead a presentation was delivered to ask the question "Is Irish sufficiently supported in the AI era" and to then highlight just how much Irish lags behind in terms of being technologically supported. This was achieved through reference to the META-NET White paper series outcomes and also an overview of the little investment and work done to date in this area. The notion of data being needed to drive the state-of-the-art systems was introduced again and the Tapadóir MT system was highlighted as the first MT system used in the public sector. This was followed by a recommendation to read the 2019 ELRC White Paper and some suggestions as to how existing language resources could be integrated into public websites to support Irish speakers. In addition, it was discussed how easily some technologies like automatic subtitling could be created for the state broadcaster based on the archived availability of so much relevant data. From that point on, the focus was on CEF funding and how it is shaping the Irish LT landscape (ELRC, ELRI and PRINCIPLE). The European Language Equality (ELE) project was also highlighted, along with an encouragement for participants to contribute to the European Language Grid (ELG) where possible. Finally, the Digital Strategy for the Irish language, under review by Foras na Gaeilge, was mentioned and its value for the future of language technology was reinforced.

Session 2. Why Language Data Matters

3.5 Language data creation, management and sharing: existing practices and challenges

As we are still in the stages of promoting the use of the National Relay Station (NRS)[2] in Ireland and explaining how good data management can help data collection efforts, we decided to run this as a presentation instead of a discussion session. The visuals of this presentation made the concept of data management and sharing much clearer than a panel discussion could do.

This talk was delivered (in Irish) by Helen McHugh, a project officer at the ADAPT Centre, working on the NRS and the PRINCIPLE project. Helen presented the NRS portal, highlighted the bodies who have contributed data so far, the benefits of sharing on a national level, the benefits of having a shared location for language data, an insight into the types of files that can be shared, and the types of content that is useful. Translation Memory (TMX) files were also explained and those dealing with Language Service Providers were encouraged to request their TMX files to be returned with translation deliveries. Challenges often faced in collecting or sharing data were also highlighted, as were the benefits of the Open Data Directive and how this may address some of these challenges going forward.

3.6 Data Sharing in Ireland and the EU Open Data Directive

We invited Helena Campbell, a representative from the Open Data Unit in Ireland, to talk about the recent updates to the Open Data Directive and how it will impact the public sector from this year onwards. This talk followed on nicely from the previous talk and demonstrated how it will help to support the ELRC data collection efforts in the future, and to address some former challenges that data holders faced. The Open Data Strategy (2017-2022) lists the actions of public bodies with respect

[2] The NRS was developed as part of the CEF-funded ELRI project. Accessible at https://elri.dcu.ie/ga-ie/

to data sharing: data audits, requirements for Open Data publication data plan, facilitating requests for data, using Eircodes in addresses, collaborating with re-users and appointment of an Open Data Liaison Officer. The Irish Open Data Portal was discussed as was Ireland's Open Data progress to date with respect to other EU member states. Most crucially it was explained that data should be made open by default and design. Data management plans will be required, and the release of non-personal data will become obligatory as will Open formats and Open standards. There will also be new rules on free re-use of data. Specifications will also be made with respect to APIs and real-time data and bulk downloads. Additionally, metadata preparation will be a prerequisite for data sharing, which complements the ELRC collection efforts nicely. It was felt that this was a reassurance for those participating with respect to their abilities or authority to share data in the future.

Session 3. LT in Ireland: the PRINCIPLE Project use-case

3.7 PRINCIPLE Project Overview

Jane Dunne, project coordinator of the CEF-funded PRINCIPLE project presented an overview of the PRINCIPLE project as an example of how LT is used in the public sector in Ireland. The project, running since 2019, focuses on (1) the collection of data for 4 languages (Icelandic, Norwegian, Irish and Croatian) and (2) the development of MT models for use by data-providers for the duration of the project in order to assess the quality of the data collected. The project partners are Dublin City University and Iconic Translation Machines/RWS Language Weaver (Ireland), The National Library of Norway, University of Iceland and University of Zagreb. Data collection and pre-processing was explained, as was the automatic and human evaluation of the various MT systems that have been built for the early adopters (data contributors of the project). Data identified as high-quality will be uploaded to the ELRC-SHARE in order to improve eTranslation for the four low-resource languages involved.

3.8 Language Technology Use-case in the Public Sector

Mícheál Ó Maolruanaigh of Foras na Gaeilge presented on the organisation's experience of being an early adopter in the PRINCIPLE project (through Irish). Foras na Gaeilge is the public body responsible for the promotion of the Irish language in both the Republic and North of Ireland. They have an in-house translator and therefore do not outsource translations. Firstly their translation needs were discussed (e.g. application forms, annual reports, policies, social media, etc.). This was followed by an explanation of the translation memory tool in use by their in-house translator (memoQ and SDLTrados). Overall the feedback on the experience of being part of the project was positive and Mícheál explained how they learned a lot about translation technology and about human evaluation of the translation systems' output. He was particularly complimentary about the quality of translation output from their bespoke MT system and they are keen to continue using MT in their everyday work.

3.9 Irish Language Technology SMEs and the public sector

Iconic Translation Machines (now known as RWS Language Weaver) was the project partner responsible for building MT systems in the PRINCIPLE project. Iconic were one of only two SMEs in Ireland building Irish language machine translation systems (the other is Kantan.io). Róisín Moran, who worked on the PRINCIPLE project delivered the presentation in Irish. She discussed the company's role in the project, described how the systems were trained and evaluated and how they fared (in general better) when compared to Google Translate, Bing and eTranslation. Finally Róisín reported on the positive testimonials they received from the early adopters:

- “it did a good job at translating the text without much input from the translator”
- “It is easier to move clauses around and correct terms and grammar rather than starting from scratch”
- “Post-editing was by some distance faster than translating from scratch”
- “If the question to be answered in this testing procedure is whether the machine translation is helpful and saves time in this sort of translation, then the answer is "absolutely"”

3.10 Take-home message and conclusions

Topics raised were:

- The potential of LT to transform governance, administration, commerce and society
- The lack of technological readiness of Irish in the AI era
- Ireland’s position with regards to the development and adoption of LT for Irish
- Activities at the EU level towards a wider development and adoption of LT for Irish
- The EU Open Data Directive

Lessons learnt were:

- Speakers of less spoken languages are facing a digital time bomb
- While Ireland is a leader in AI technologies, this only applies to the English language
- Better funding and planning is required to ensure Irish is digitally prepared going forward
- LT and AI need more data!
- Share your language data nationally and with Europe

3.11 Demos session

There were no demo sessions.

4 Synthesis of Workshop Discussions

There was very little engagement with the audience in terms of questions. This is partly cultural (we've experienced the same wariness of participants to ask questions in previous workshops) and partly due to the online format. In our previous experiences, the lunch and coffee breaks (and follow up meetings) are often where most of the discussions take place. Questions both from the chat facility and orally are listed below.

We have learned that continual engagement with data holders is essential to ensure a sustainable model for data sharing. The recently announced funding for the National Relay Station (from the Department of the Gaeltacht) will allow for more resources to be allocated to this need. In addition, our advice to departments who outsource their translations to LSPs on how to define the need for the return of TMX files has proved beneficial over the past couple of years. Most importantly, the clarity and confirmation of the Open Data Directive, along with the support of the Open Data Unit in Ireland, will prove to be invaluable in terms of addressing some challenges faced by data-holders. We have found that including a presentation from the Open Data Unit is crucial in all of our previous workshops.

The invitation was sent out to over 470 contacts in government departments and public bodies, Irish language organisations, universities and SMEs. The workshop information was shared on social media (Twitter and LinkedIn), yet only 78 people registered and 54 people attended. Only 33 of those were non-organisers/speakers/ELRC related. It is possible that it might be more fitting in the future that the invitations are sent out by the Department of the Gaeltacht instead of from the ADAPT Centre email account. We feel that the workshop content might be taken more seriously if delivered by a government department who has a strong reputation amongst the Irish language public organisations. In addition, some departments' firewalls are so strict that the email invitation and reminder may have ended up in spam folders. This is less likely to happen if it is sent from a government account.

SMEs were largely underrepresented at the workshop, despite engaging with Údaras na Gaeltachta[3] to disseminate the event's details amongst their members. There are only a few privately owned businesses that operate through the medium of the Irish language across Ireland and one machine translation business who builds Irish language engines for public body use. Businesses with bilingual websites would have benefitted from the event in terms of learning how to keep their translation costs low.

While the online format allowed for the attendance of representatives from across the country who otherwise may not have attended, without a doubt the in-person event is much more beneficial for Ireland in terms of discussion, engagement and networking.

4.1 Questions/Discussion points

Question: Can anyone sign up for access to eTranslation?
Vilmantas put back up the slide that listed all organisations who had free access to eTranslation.

Comment: A note on the topic of Open Data, Welsh technologists noticed an uptick in understanding of speech to text (voice assistants) once Welsh Wikipedia reached over 100,000 articles (GA is currently 55,000). Welsh Wikipedia has benefited hugely from the Welsh government support and the publication of source material openly.
Response: Teresa advised that it's important to note the value of a Wikipedia size for minority languages as we know that larger tech companies (Google, Apple, etc.) refer to these as a sign of digital vitality and base their decisions to support a language on this amongst other factors.

[3] https://udaras.ie/en/

Comment: A suggestion was made to consider reaching out to new Irish language officers, Irish language development officers, or 26+ Irish language planning officers in the Gaeltacht, given that machine translation is probably not something that they discuss daily. The suggestion was to look at tools and resources for those people that are working "on the ground" and see how we can help make this more accessible to them. The attendee also noted that the Cumann Oifigeach Forbartha na Gaeilge is another organization to reach out to, as not all of those who work in that sector use Trados or machine translation, even though they might carry out translations in-house. These organisations might not be storing their data the way that would be useful for the ELRC.

The attendee also noted that it was interesting what Helen said in her presentation about difficulties encountered with getting permission to share data. She also highlighted that some are not completely and utterly comfortable sharing their data because they're not 100% sure of the quality (with over-cautiousness about the correct use of *fadas* (diacritics).

Response: Jane agreed and pointed out that eTranslation is available and is accessible. She also noted that DCU has received more funding to extend the running of the National Relay Station (NRS) for another 18 months, which will include outreach, so this will facilitate us to meet with grassroots initiatives and networks to understand what the issues might be with regard to sharing. Teresa also noted that those groups mentioned (language officers) were invited to the workshop but didn't show up or respond and encouraged those in attendance to spread the word to other relevant parties when possible.

Question- (WRT presentation on eTranslation) That first translation to Irish was very bad. Does Vilmantas know that?
Response: Vilmantas said he deliberately didn't check the quality of the output for the purpose of the presentation. Teresa explained that it's important to remember that MT needs to be used in a post-editing environment with professional translators as it's not a guaranteed perfect translation, but instead, a translation aid. Also explained that this example simply highlights the need for improving the Irish system through the collection of more data.

Question: Sula gcuirfear na téacsanna ar aghaidh, an ndéanfar eagarthóireacht orthu leis an chaighdeán a chinntiú? *'Before the texts are sent on, are they edited to ensure quality standards'*?
Response: No but because of the nature of the AI systems, any bad translations are drowned out by the weight of all the other good translations due to the nature of the statistical/probability approach.

Question: An attendee noted that in their experience eTranslation is very good for EN>GA formal texts, but observed that the performance declines on less formal the text and wondered if that is that the experience with other language pairs?
Response: Teresa explained that it's related to the domain/genre of the text that the system is trained on. In general, eTranslation for Irish is trained on legal and public administration text and will do better on these domains as a result.

5 Country Profile: Language data creation, management and sharing

There was no panel discussion during the Language Data Creation session as we needed to use this opportunity to provide detailed information to attendees on how to better manage data inhouse and across LSPs, and how to use the National Relay Station Portal.

We carried out Polls during the workshop to gather information related to the Country Profile. Some of the suggested ELRC poll questions/ answers were edited slightly for the Irish audience (see below). The poll feedback is available below, with a summary supplied here:

The Public Sector and Research/University were the most represented groups in attendance. Information Search and Retrieval is the most widely used AI tool amongst this cohort. But as there are no tailored or open-source IR systems for Irish, we can conclude that this is Google/Bing web-search engines. Speech recognition was named by one respondent but as there is no ASR system for Irish (not even proprietary software) we can conclude that this should have been a speech synthesis selection (there is a text-to-speech tool available for Irish). Overall, the level of satisfaction for Irish speech and language technologies is low.

Minimal changes are required in the Irish Country Profile. The Shared Translation Service still has not yet been established. The uptake on translation memory tool usage is still slow. Small numbers still avail of the eTranslation system. With respect to data sharing, the same challenges are still present: Licensing issues or not having the authority to share language data along with the lack of a data management plan. The updates mainly reflect recent activities in the CEF-funded PRINCIPLE project, changes to the Open Data Directive and recent CAT workshops.