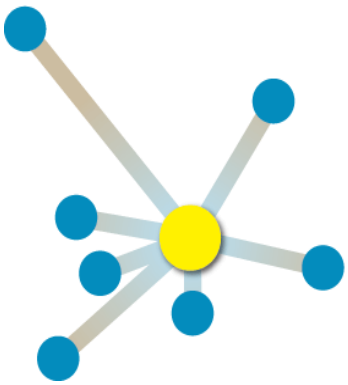
Deliverable D11.1

Report on the 5th ELRC Conference



European Language Resource Coordination

Connecting Europe Facility

Author(s):	Andrea Lösch (DFKI) Eileen Schnur (DFKI)
Dissemination Level:	Public
Date:	2021-04-12
Copyright:	yes

For copies of reports, updates on project activities, and other project-related information, please contact:

Prof. Stephan Busemann
Stuhlsatzenhausweg 3
Campus D3_2
D-66123 Saarbrücken, Germany

stephan.busemann@dfki.de
Phone: +49 (681) 85775 5286
Fax: +49 (681) 85775 5338

Contents

1 Introduction

This deliverable provides the report on the 5th ELRC Conference. The conference took place as a virtual event via Zoom on March 10, 2021. It was also streamed live on YouTube.

Figure 1: Visual Youtube Streaming

The deliverable is structured as follows: First, we describe the aims and objectives of the conference including also the target audience, followed by an overview of the thematic structure and organisation of the conference (see Chapter 2, Focus and Contents of the Conference). We then present the digest of the presentations and discussions of the conference. Last but not least, the Annex provides the conference programme, the attendance list of the conference participants as well as an analysis of their geographic distribution and the sectors covered.

All presentations are available online on the ELRC website: https://www.lr-coordination.eu/node/304. The full recording of the conference can be found on YouTube: https://www.youtube.com/watch?v=DRZpbmV6SfE.

2 Focus and Contents of the Conference

2.1 Context

For public services and administrations all over Europe, information exchange across borders is not only vital, but also increasingly difficult because of language barriers. Language technologies present a meaningful way to overcome these barriers. With CEF eTranslation, the European Commission has created a corresponding machine translation tool which is not only available to EU institutions, but also to public administrations, public services and small and medium-sized Enterprises (SMEs) in all EU Member States, Iceland and Norway. Moreover, additional language services such as speech-to-text, anonymisation and multilingual tweets (see https://language-tools.ec.europa.eu/) were made available to public services and SMEs across Europe.

However, in order to successfully adjust the CEF eTranslation platform to the requirements of public administrations, public services and SMEs (different domains, language pairs), corresponding language resources (LR) are needed. In accordance with the Tender Specifications (p. 29), the focus of the ELRC Conference was hence "to raise awareness of the relevant stakeholders about issues related to the CEF AT, to collect information, views and expectations on the CEF AT platform at Member State level, and to promote collaboration, networking and best practices in view of providing language resources, tools and other useful contributions to the CEF AT, in order to improve the quality of the multilingual services provided by CEF AT."

2.2 Target Audience

The conference targeted representatives of different public administrations and public services across Europe that are involved in the creation and sharing of language resources (data holders and/or potential data donors). Moreover, it explicitly also included representatives of SME that may potentially use and/or benefit from CEF eTranslation. In order to reach both target groups, a first save the date with follow-up email was sent to the participants of the last ELRC Conferences and country-specific workshops. The event was promoted on the ELRC website (www.lr-coordination.eu), and through the ELRC Social Media Accounts on Twitter (https://twitter.com/LR_Coordination), LinkedIn (linkedin.com/in/lrcoordination) and Facebook (https://www.facebook.com/EuropeanLanguageResourceCoordination).

Overall, there were 548 registered participants and 58 participants on the waiting list[1], adding up to 606 people who were interested in joining the event. In total, more than 365 attendants joined the Zoom meeting[2] during the conference day. In addition, approximately 80 people followed the conference on Youtube, adding up to a total participant number of 445. The drop-out rate hence was 23 %. As of 20 March, the full conference recording, which was uploaded on the day after the conference (11 March) has been watched more than 170 times already.

[1] In order to ensure maximum capacity for the actual ELRC target groups, interested parties from outside the EU were added to the waiting list and provided with the link to the Zoom livestream.
[2] 20 participants could not be clearly identified because of ambiguities in their usernames.

2.3 Focus and Contents

The focus of the 5th ELRC Conference was on the following key topics and issues:

- Status quo of the LT industry and new funding opportunities within the upcoming Digital Europe Programme;
- Judging the quality of translations and language data and effects of data quality on MT development;
- Anonymisation of language data: technical possibilities and legal constraints;
- Approaches to LR creation and collection: Resource projects within CEF.

The detailed conference programme is provided in Figure 2 below and was also published on the ELRC website (https://www.lr-coordination.eu/node/304).

Figure 2: Programme of the 5th ELRC Conference

3 Synthesis of Discussion Points

3.1 Welcome Address by June Lowery-Kingston

The opening speech was delivered by Ms. June Lowery-Kingston, Head of Unit G.3 “Accessibility, Multilingualism & Safer Internet” at the European Commission Directorate-General for Communications Networks, Content and Technology (DG CONNECT). Starting from the COVID-19 crisis, she underlined the importance of the COVID-19 Multilingual Information Access (MLIA) initiative. MLIA is a collective effort by the Language Technology (LT) community to improve information exchange about the virus, across all EU languages and beyond, by supporting the development of applications and services in relation to the COVID-19 pandemic. Here, ELRC has been and will be contributing a significant and large number of urgently needed language resources (please visit the ELRC-SHARE for a full overview of COVID-19 related language resources that were made available by ELRC https://www.elrc-share.eu/repository/search/?q=&selected_facets=projectFilter_exact%3ACOVID-19).

Figure 3: Welcome address by June-Lowery Kingston (Head of Unit G.3 "Accessibility, Multilingualism and Safer Internet", European Commission)

Ms. Lowery-Kingston also stressed that it is the goal of the new Digital Europe Programme to make Europe fit for the digital age. She particularly stressed that a single market for data is needed to support and serve Europe as an open, democratic and sustainable society. European LT would play a key role in supporting the Digital Single Market, by enabling people to work together, exchange and share information without language and/or speech barriers. Especially low-resourced languages and minority languages should be supported to build the necessary bridges. She emphasised that “Europe needs multilingualism, and Europe needs powerful language technologies made in Europe for Europe” – and in order to achieve this, language data as being collected by ELRC is the key.

Discussion points in response to the presentation:

Overall, one question was raised by participants who wanted to know which projects were mentioned on minority and low-resourced languages. ELITR, GOURMET and EMBEDDIA were then named as examples for such projects.

3.2 Welcome and Introduction by Andrea Lösch

In her welcome presentation, the ELRC Project Manager Andrea Lösch (DFKI) focussed on the overall frame of ELRC. She provided evidence on the importance of the Language Industry market in and for Europe (drawing on a corresponding Slator Industry Market Report and the CEF Market Study on the European LT-Market which illustrate the size and considerable growth rate of this market). Since LT is a key market

for Europe and since language data is the key for LT development, the collection of language resources is of utmost importance in Europe.

Following a brief explanation of how the language resources collected within ELRC can help improve the output of eTranslation (and other MT systems), she provided an overview of the ELRC network of National Anchor points and the main objectives of the ELRC service. Last but not least, she introduced the agenda of the 5th ELRC Conference and the key question this conference would address. Participants were informed about the code of conduct for this virtual conference, and they were also encouraged to participate in the corresponding evaluation survey.

Figure 4: Welcome and Introduction by ELRC Project Manager Andrea Loesch

Discussion points in response to the presentation:

One question from the audience concerned the process of feeding the machine with lexical resources to train it (in comparison with bilingual texts). Reference was made to the presentation on using crawled data for MT development where Marcis Pinnis explained the use of different types of LR (parallel corpora, monolingual resources and lexical resources) for MT development.

3.3 Inside ELRC and Digital Europe

3.3.1 ELRC Update

In her sub-sequent presentation on the status of data collection within ELRC, Andrea Lösch (DFKI) illustrated major achievements of the initiative: In the course of ELRC, almost 2.500 unique language resources could be made available through the ELRC-SHARE Repository. The vast majority of these resources (i.e. nearly 1.800 LR) are bi- or multi-lingual corpora which are most valuable for the training of MT systems. Moreover, more than 80% of all language resources are actually re-usable by anyone, as they have open licenses. Overall, ELRC resources are available for a vast majority of domains as shown in Figure 5 below:

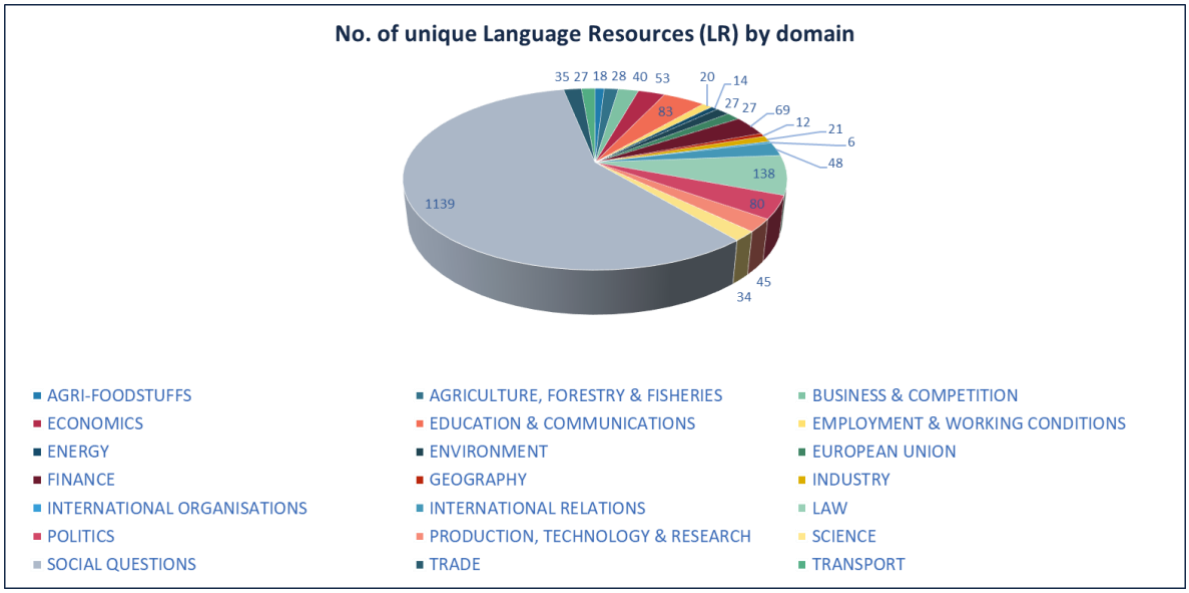


Figure 5: No. of unique Language Resources (LR) by domain

In addition to ELRC's supporting service (in particular the Helpdesk), attention was drawn to the CEF AT Catalogue of Service which to date provides access to more than 670 language tools and services in Europe (from more than 530 different European providers), including for instance more than 100 tools for speech recognition and more than 100 tools for text and data analytics. This catalogue presents an important source of information for any public service and/or SME seeking language technologies to support their language and needs.

Last but not least, Andrea Lösch gave an overview of ELRC's social media activities, showing that almost 1 million users could be reached by the campaign in the last 9 months, which also resulted in a considerable interest to check out the CEF eTranslation tool (more than 76.500 link clicks to access/apply for access). Figure 6 summarises the different social media channels operated by ELRC.

Figure 6: ELRC Social Media Channels

There were no further questions associated with this presentation.

3.3.2 European Language Industry Survey 2021

John Anthony O'Shea, Legal Translation Practitioner and Board Member of FIT Europe, provided insights into initial findings of the European Language Industry Survey (ELIS 2021), which covers market trends, expectations and concerns, challenges and obstacles, as well as changes in business practices. His presentation focussed particularly on language technology tools. The initial findings showed that machine translation (MT) remains the strongest technology trend followed by workflow technology, automated question answering, computer-aided translation tools (CAT) and automated interpreting. The greatest operational change hence was found to be due to machine translation and post-editing. Even for independent professionals, it was found that the use of CAT tools continues to dominate, and MT usage is on the rise. At the same time, 25% of independent professionals perceived the pace of technology as a stress factor in 2021 – which is very similar to the findings in 2020. Last but not least, the survey showed that the global industry's willingness to invest in language technology has decreased compared to previous years. The detailed survey findings will be available from 15 April 2021.

Discussion points in response to the presentation:

Several questions emerged from the audience in response to John Anthony O'Shea's presentation, which are summarised below:

- After the presentation, the speaker was asked to explain what the abbreviation "LSC" stands for. It was clarified that he referred to "Language Service Companies"- sometimes also called Language Service Providers (LSP). Subsequent to this clarification, there was a discussion on the terminology and whether LSCs and LSPs were really synonyms. It was explained that in most cases LSPs were companies, even though they also included governmental agencies and their language services.
- Following the question about auto interpretation being behind MT in terms of popularity and whether on-the-fly automatic subtitling was counted as being automatic interpretation or as MT, Mr O'Shea replied that most probably automatic subtitling was counted as a stand-alone technology.
- Regarding potential reasons for professional translators' resistance towards MT, Mr O'Shea explained that according to the initial results, the main concern was the impact on their payment, resulting in a drop of rates and income. He further clarified that there was no evidence of major concerns about the technology itself and that CAT-Tools were already used without any problems. In addition, he stated that MT would be adopted gradually.
- Following the question about when MT would be mature enough to translate e.g. legal texts, Mr O'Shea reported about his own experiences with translating legal documents into smaller languages, e.g. Greek – English. He explained that for lower resourced languages like Greek, it would probably take much longer than for French or German, for example. He explained that he was conducting some research on this issue and that a corresponding academic paper would be published in June. Mr O'Shea further elaborated that in the legal domain, even with highly trained engines focusing on specific data, it was difficult to achieve satisfying results for Greek. In response to that, one participant pointed out that for some legal documents (very standard texts), MT could already achieve satisfactory results, while for more complex legal texts, MT might not work well. He also stated that at

DGT, translators had to deal with legal documents on a regular basis, but as legal acts were often compiled from other legal acts, translation memories (TMs) containing the relevant legal reference acts were being used instead of MT. He added that there might be several linguistically correct translations, but often only one translation that was correct from a legal point of view.

- Following this, the question came up how the study's findings on MT adoption married with the fact that large content producers were already requesting 90% of human parity without machine translation post-editing within two years. The participant, a developer, who had already received such requests, further elaborated that to such companies, it was more important to have a cost-efficient and manageable solution providing an acceptable MT quality which suffices to understand the text, than to make use of post-editing and provide high-quality translations. He added that to his knowledge, there was hardly any translator who did not use MT for his/her work, as this would result in a financial loss. Mr O'Shea followed up on this comment by explaining that the aim of the survey he presented was to measure trends, expectations and outlooks related to technologies. Highlighting the differences in domains and language combinations, he stated that the number of freelance translators who were not using MT was still substantial and that some of his clients were not even aware of the existence of MT. According to the speaker, MT was a useful tool to translate short and clearly written sentences, but legal texts with poor punctuation and long sentences would still be a challenge. On the other hand, patents might be several pages long and there were MT solutions that could handle them. Mr O'Shea concluded that MT was another useful tool in the translators' toolkit and that in order to increase their income, translators needed to find their own way of using MT with its current limitations. Last but not least, one of the participants pointed to an interesting paper on "When Will AI Exceed Human Performance? Evidence from AI Experts" which was published in May 2017 (https://arxiv.org/abs/1705.08807).

3.3.3 Inside Digital Europe

Philippe Gelin started his presentation looking back at the 4th ELRC Conference and the draft of the Digital Europe programme he presented back then. He pointed out that a major development since the last ELRC Conference was the opening of eTranslation to European SMEs. In March 2021, more than 8.700 SMEs were registered (starting from March 2020). The promotional campaign also increased the number of public administrations with more than 2.100 additional public administrations having registered for using the eTranslation service. Moreover, eTranslation also significantly expanded its coverage with new languages (Russian, Turkish, Chinese, Japanese, Arabic, …) and additional tools were made available on https://language-tools.ec.europa.eu such as anonymisation for English and German, speech recognition, etc.

Philippe Gelin also demonstrated that language technologies and multilingualism will be relevant topics for both future framework programmes (Horizon Europe and Digital Europe). In the Digital Europe programme, LT plays an important role in the capacity building actions (in particular cloud-to-edge, data spaces support centre, AI on demand platform, AI testing and experimentation facilities) and in accelerating best use of technologies (see digital innovation hubs, EDGES Common Services Platform). Also in Horizon Europe, LT will continue to be present in different actions (safeguarding

endangered languages in Europe, strengthening Europe's data analytics capacity, trustworthy open search and discovery, AI for human empowerment, strengthening Europe's data analytics capacity, extended reality modelling). As Philippe Gelin noted, it is important to understand that unlike Horizon Europe, Digital Europe would not fund research. Figure 7 below illustrates how Digital Europe and the deployment of LT will support Europe's public administrations and SMEs.

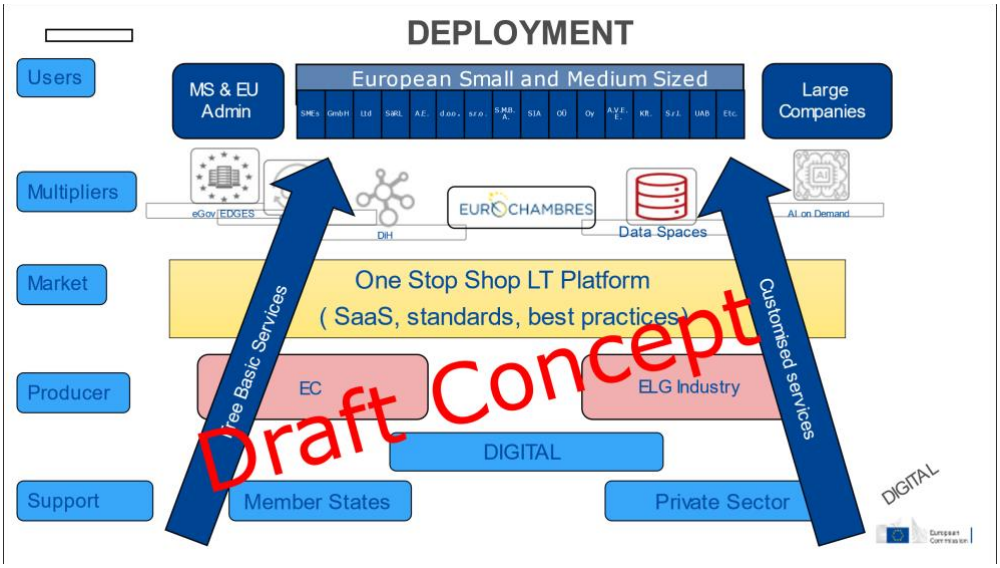


Figure 7: Draft Digital Europe at a glance

Discussion points in response to the presentation:

Several questions emerged from the audience during and after Philippe Gelin's presentation and were addressed by the DG CONNECT representatives (Philippe Gelin and June Lowery-Kingston):

- Following the question about extending the eTranslation language coverage to include Albanian, June Lowery-Kingston explained that Albania was formally a candidate country and that in March 2020, the members of the European Council endorsed the General Affairs Council's decision to open accession negotiations with Albania. In July 2020, the draft negotiating framework was presented to the Member States. She affirmed that under the Digital programme, adding further languages to eTranslation would be supported and that most probably, Albanian would be added. A clear indication of when this would happen could, however, not be given.
- When asked to elaborate on the funding conditions, especially the funding rate for companies and research organisations, Ms Lowery-Kingston referred to the first work programmes for Digital Europe and Horizon Europe which would provide all the details. It will probably be published in April 2021.
- In response to the question about the European Language Grid's place in the Digital Europe Programme, June Lowery-Kingston referred to Philippe Gelin's presentation of the "one stop shop", explaining that this will be supported by ELG (see above).
- As translation of texts containing named entities can be an issue, one participant asked if there were plans to better support this in eTranslation. In addition, it was

asked whether real-time translation was foreseen. It was answered that Anonymisation / Named Entity Recognition (NER) was already included in the CEF AT catalogue of services (see Catalogue of Service for all available solutions in Europe: https://cef-at-service-catalogue.eu/, but also see the European Commission's CEF AT platform for NER German and English: https://language-tools.ec.europa.eu). However, this was only a first step towards data anonymisation. The first priority was given to the data that requires anonymisation.

- On the question of funding possibilities to provide basic descriptions and resources such as gold standards for languages threatened by digital extinction, e.g. lexical descriptions, grammatical markup, semantics, discourse and dialogue structure, Philippe Gelin explained that this was part of the Horizon Europe Programme - Culture, Creativity and Inclusive Society - "Safeguarding endangered languages in Europe". In this respect, one of the research aspects were multilingual models and how the language coverage could be increased more easily. Mr Gelin also explained that languages which are socially and economically relevant have priority. He concluded by saying that overall, the European Commission's goal is that everyone in Europe will be able to communicate with each other.

3.4 Spotlight: Quality of language data

3.4.1 Assessing the quality of translations

Renate Müller is Quality and Language Coordinator at DGT. In her presentations, she gave valuable insights into the process of judging the quality of translations which needs to respect the quality of translations and at the same time be cost-efficient. Given the fact that quality always depends on the purpose and context, she explained that DGT employs both general quality standards as well as specific requirements in order to define the quality of a translation (see Figure 8 below).

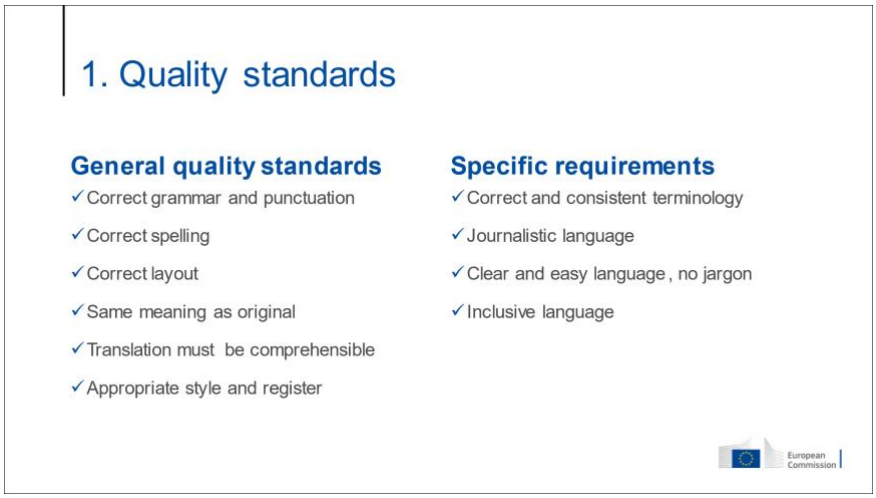


Figure 8: Quality standards employed in DGT

It is important to note that the investigation of the quality of translations typically only applies to outsourced translations. The corresponding error typology covers accuracy, terminological errors, conformance with linguistic norms, design and style. The process of evaluation is summarised in Figure 9 below. In addition, minor and major errors are differentiated with the help of error weights.

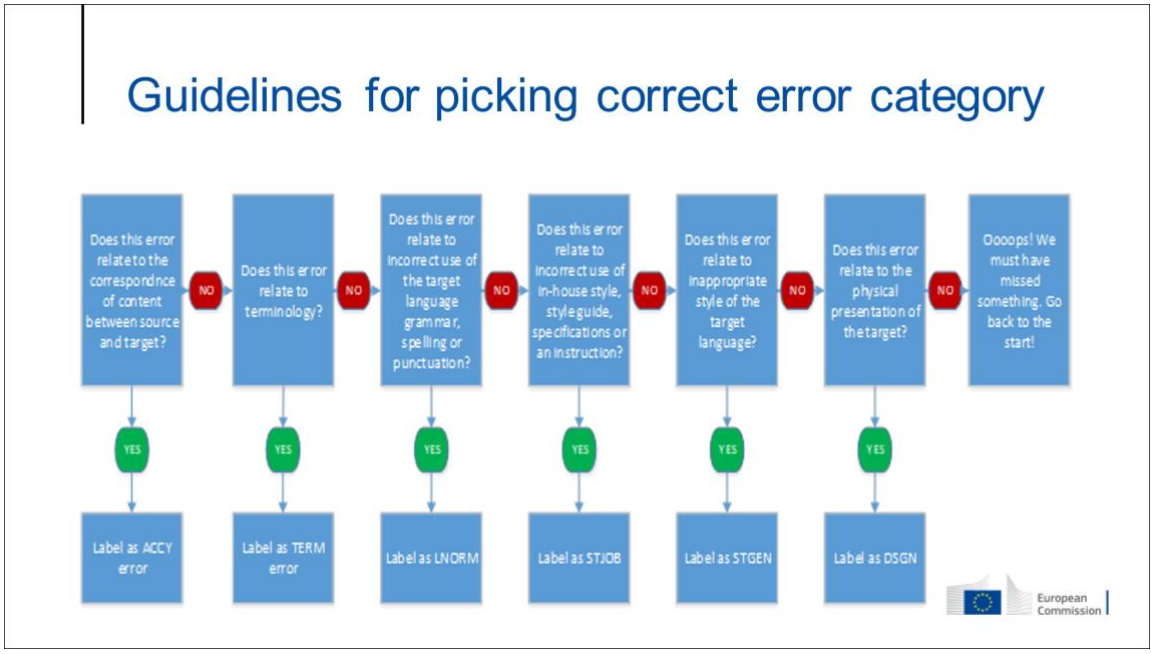


Figure 9: Process of evaluation employed

Typical errors associated with neural MT are terminology errors (i.e. inconsistencies, as well as invented words) and missing elements in a text.

Discussion points in response to the presentation:

Following Ms. Müller's presentation, several questions were asked by the the audience:

- In response to the question if and how the quality of 2.3 million pages of translation could be revised according to the presented error typology, Ms. Müller confirmed that it was definitely a challenge, at the same time clarifying that the evaluation was only carried out on outsourced translations, adding up to one third of all translations. She further explained that even in the case of outsourced translations, not all texts were being revised, as this highly depended on their relevance. An example of texts that were usually revised were legal texts unless they were standard texts or elements which had already been translated.
- When asked about the percentage of automatic post editing, Ms. Müller explained that every MT segment needed to be post-edited. She added that as the texts were fairly repetitive, as much as possible was recycled to increase efficiency. Text sections originating from other texts would not require post-editing though.
- Following a question about the standards that could be used to assess translation quality, Ms. Müller pointed out that this very much depended on the purpose and the context of the translation, clarifying that standards to be used in a legal context (e.g. the translation of regulations, like in our case) might be quite different from those used for the translation of films, poetry or literature.
- One participant wanted to know which CAT Tool(s) Ms Müller and her team are using. She explained that they were using SDL Studio at the moment.
- Another question concerned the policy on using external MT tools like Google Translate or DeepL and whether this was explicitly forbidden. Ms Müller explained

that inside the EC, eTranslation was the only translation service in use, because it worked best for this environment and context. In addition, she stressed the importance of privacy and security and that sensitive and personal information should not be shared with externals.

- Following up on the calculation of error numbers, one participant wanted to know if e.g. mis-translated words and fully wrong sentences would both be counted as one unit. According to Ms Müller, if there was more than one error in a sentence, it would be calculated as more than one error. This, however, would also depend on the type of errors and potential relations between them, which is why it was the evaluator's responsibility to decide on that. She stressed that in such cases, it was important to assign different error weights.

3.4.2 Using crawled data for MT development – promises and pitfalls

Marcis Pinnis, Chief AI Officer at Tilde, started his presentation with an illustration of different types of language resources (parallel corpora, monolingual corpora and lexical resources) and their role for MT development. He explained that monolingual data was typically used for back-translation – a method that allows the acquisition of synthetic parallel training data. However, back-translation cannot work wonders – it will not improve the translation quality of out-of-vocabulary words or phrases as Mr. Pinnis pointed out. As such, the domain of monolingual data needs to match the domain of the text that will be translated, and corresponding lexical resources can support the right terminology. However, the data needs to be of sufficient quantity to achieve the maximum quality increase.

With regard to crawling parallel data from the web, there are two main possibilities: (i) focussed parallel data mining where mining is performed for known, relevant web sites that contain parallel data (advantage: higher quality with smaller efforts) and (ii) large-scale parallel data mining where the whole web is considered a potential source of parallel data (disadvantage: quality heavily depends on the tools applied in the process). Most importantly, parallel data from the broad data mining processes often contain noise which then negatively impacts the quality of the MT systems trained with such data. Figure 10 below illustrates the different BLEU scores for training data obtained from large scale crawling (in this case: ParaCrawl) and for training data without large scale crawling as was found in the WMT 2020 experiments.

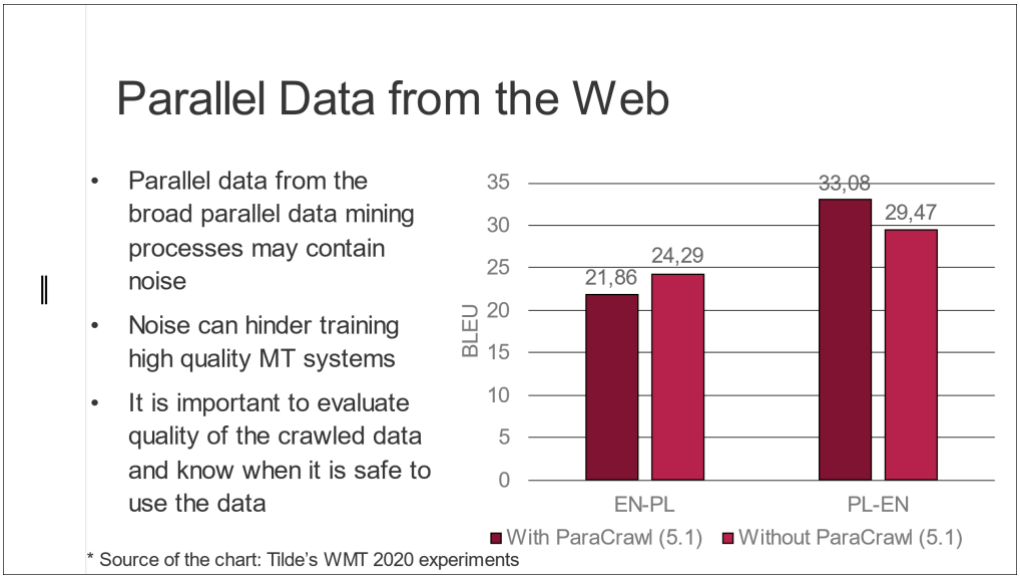


Figure 10: Training MT systems with crawled data – differences in BLEU scores

Discussion points in response to the presentation:

- Following the question about the costs of mining and crawling language data, Marcis Pinnis confirmed that data collection could be a costly process, but that the actual costs depended on many different factors, e.g. if you use a focused crawl or rather a broad crawling approach, requiring more processing resources and computing power.
- In response to the question about service providers who actually provide crawling and data mining services, Mr Pinnis pointed out that many research institutions were engaged in crawling activities. He added that such services were also offered by commercial providers like "Dolphio Technologies" (www.dolphio.hu) which is listed in the CEF AT Catalogue of Services (https://cef-at-service-catalogue.eu/). Further entries can be found when searching for "Data and Data Management" in the Catalogue.
- One participant wanted to know how many sentence pairs were required to get reasonable MT results. According to Mr Pinnis, the number of required sentences was domain-dependent. Speaking from his experience, he explained that systems being trained with 1 m sentences were very good for broad translations and that it was possible to achieve reasonable quality with 300-400.000 if the domain was very narrow. Mr Pinnis added that the usual amount was 20+ m sentence pairs, sometimes even more than 100 m. He concluded by saying that the more data you have, the more it can represent the language used – and the better the results will be.
- When asked to suggest a type of model suitable for the described translations (e.g. which type of seq2seq), Mr Pinnis answered that the current state of the art was based on transformer networks. He added that these networks were also used within Mr Pinnis' organisation with a number of tricks and adaptations.
- In response to the question about how machine-translated contents can be identified when crawling data, Marcis Pinnis clarified that he and his team did not try to identify such data. Referring to the example of ParaCrawl, Mr Pinnis explained

that it was possible to train a classifier to differentiate human content from MT content. However, he stressed that this was a challenging task and that current classifiers were not yet able to reliably remove MT content. He suggested to opt for a more careful crawling approach instead, as this would ensure high quality and increase efficiency.

- Following up on the topic of back translation, one participant raised a question about its usefulness for legal translation where concepts and underlying legal terms were different according to the different legal systems. Mr Pinnis confirmed that back translation could indeed be useful in such cases and that the considerations he presented were also applicable to the legal domain (translations of terms need to be present in the parallel data, the domain of the monolingual corpus has to match with the intended target domain, and the monolingual data have to be of sufficient quantity). However, when talking about terms specifically, the back translation process (i.e., the reverse NMT system trained on the original parallel data) would need to be able to translate a target term into a correct source term or at least a synonym. Only then, when training a system on the parallel and back-translated data, the new system would be able to learn to translate the domain-specific terms correctly. This is why the parallel corpora needed to feature correct term translation examples. He further explained that back translation could be useful if the style of the monolingual data to be used for back translation was different from the style of the original parallel corpus. In this case, back translation allowed for the alignment of the style of the MT system output towards the required style (e.g. British English vs. American English, sentence construction choices, formal vs. informal language, etc.). This could be particularly useful for legal translations when systems were trained on a broad variety of corpora. (Additional note on back-translated data: Jörg Tiedemann recently released 500+ m. translated sentences in 188 languages. If you want to have a look, simply search for 'Tatoeba Challenge' in your favourite browser. The corresponding paper can be found here: https://www.aclweb.org/anthology/2020.wmt-1.139.pdf).
- One participant stated that by including machine-translated texts, one machine would essentially teach another machine and a potentially poor quality might be transferred. Therefore, it was asked how using machine-translated content could be avoided when harvesting parallel texts from the web. Marcis Pinnis clarified that the quality was checked when using the approach of focused crawling, as it was almost impossible to reliably separate MT contents from human translated contents (see above).
- When asked for sentence alignment methods used in broad parallel data mining and their accuracy, the speaker explained that state-of-the-art methods compare neural representations (neural sentence embeddings) of sentences from a model trained from data in multiple languages.

3.4.3 LT development for low-resourced languages

In his presentation, Andrew Bredenkamp (Chairman of Translators without Borders) pointed out that over half of the world's population suffers from a lack of access to information in their language. While 3.5 billion people speak a "major" language (like English, French, Spanish or Chinese) which is well supported by language technologies, it is much more difficult for the remaining 4 billion who are speaking smaller languages. As such, Translators without Borders (TwB) aims to deliver a

combination of language technology and innovative language service platforms that support under-resourced languages. It draws on the experience of more than 120 experienced professionals (staff members) and a community of more than 60.000 linguists in 148 countries. Especially the latter are of utmost importance with regard to the collection and generation of relevant language data and hence provide the basis for the LT development. He stressed that digital engagement needs to be on an equal footing if it is to be successful.

With regard to the actual system development, he proposes an incremental approach to low-resource language support which he refers to as “fattening the tail”. Figure 11 below illustrates this approach which starts with a relatively limited size of language data to build a conversational system.

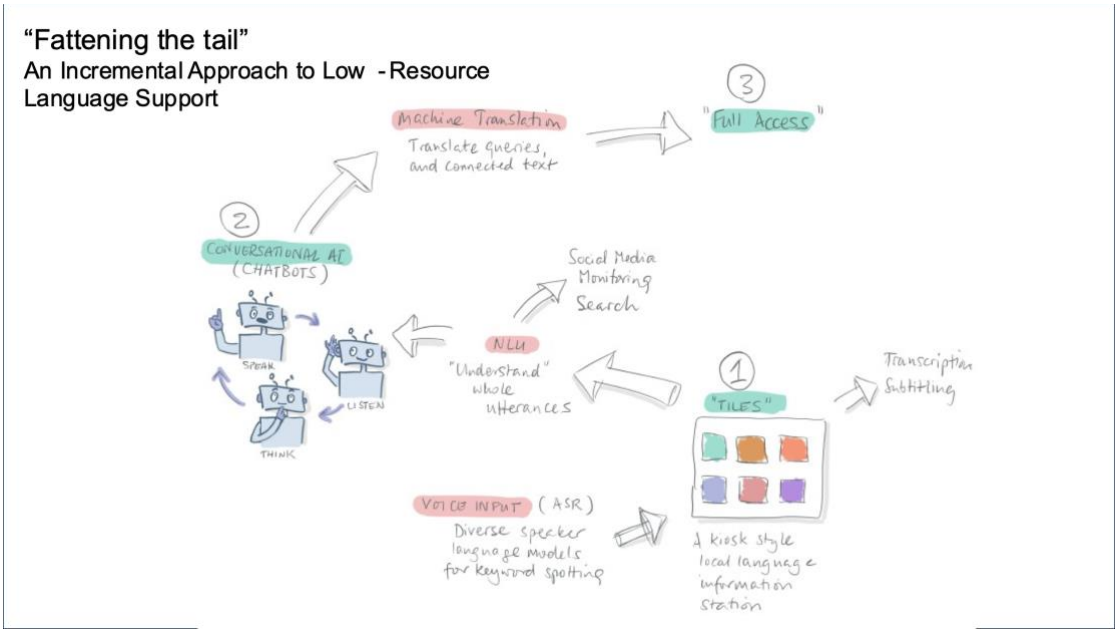


Figure 11: Incremental LT development for low-resource languages

Discussion points in response to the presentation:

Following his talk, two questions were asked to Andrew Bredenkamp that concerned details about the LT development:

- Following the question about examples where the “fattening the tail” approach was used, Andrew Bredenkamp mentioned that TwB developed an MT system for Tigrinya - English, a language spoken in the North of Ethiopia. Despite the lack of available data, they were able to start training a MT system with the help of so-called “mini kits”, initial sets of language data translated by volunteers. Even though the initial MT system had a specific focus, it was later also deployed in other contexts due to the Tigray Crisis to support e.g. the Red Cross Organisations, the UN, etc. As such, it became useful in a much broader sense in the context of crisis response. According to Mr Bredenkamp, other languages were Levantine Arabic (for the Syrian crisis) and Congolese Swahili (for DRC). Overall, approximately 50 to 100 languages were identified as being part of the “piece of the tail” TwB wanted to “fatten up”. The corresponding mini kits were currently being created and deployed in certain cases to get more traffic. Last but not least, Mr Bredenkamp highlighted that MT systems could only improve if they were actually in use.

- With regard to the amount of data required to get a useful MT system one could start with, Mr Bredenkamp indicated that at the beginning, more than 100.000 segments were necessary, at the same time stressing that the system would be barely usable at this stage. At this point, willing collaborators were required to further improve the system.

3.5 Re-using language data: Technical possibilities and legal constraints

3.5.1 Language data validation and curation in ELRC

This presentation was held by Victoria Arranz (Head of Language Resource Projects R&D at ELDA) and Mickaël Rigault (Junior Project Manager and Legal Counsel at ELDA). ELRC collects different types of language data (in particular corpora, language or translation models and lexical/conceptual resources) which come from different sources: (i) external donors (e.g., public organisations, other EC-funded projects) and (ii) webcrawling (whenever legally feasible[3]). As a consequence, data validation (i.e. the quality control of a language resource against a list of relevant criteria) may be conducted in two different ways:

- **Quick Content Check (QCC)**: for high-quality data (e.g., human translations)
- **Extended Content Validation**: e.g., for data derived from automatic processing or for high-quality data which requires further processing

In both cases, corresponding validation and processing reports are prepared and attached to the particular resource to make the clearance process transparent and retrievable if questions should arise.

While the Quick Content Check typically comprises the evaluation of a language resource with regard to its scope, format, correctness of metadata and legal status, the Extended Content Validation comprises both automatic and manual procedures as you can see in this figure. For instance, for crawled data, the list of crawled URLs is manually checked to find out if the websites are under the scope of the Public Sector Information Directive and can hence be re-used and shared. Content from websites that do not fall under the PSI Directive, or content that is not explicitly marked as open with a permissive license, is hence excluded. Also, errors in Translation Units (TU) are reported and Translation Units marked as containing errors are automatically removed; the remaining TUs are then annotated with an indication on the probability of finding the same errors. Among the tools that we use for the automatic data processing part are for instance DictMetric for document alignment, Microsoft Bilingual Sentence Aligner, language detection tool PYCLD2, and many others. The TMX files are validated using TMXValidator. Figure 12 below summarises the process for Extended Content Validation as employed in ELRC.

[3] See Webcrawling Report at http://www.elra.info/en/dissemination/legal-issues-webcrawling-report/

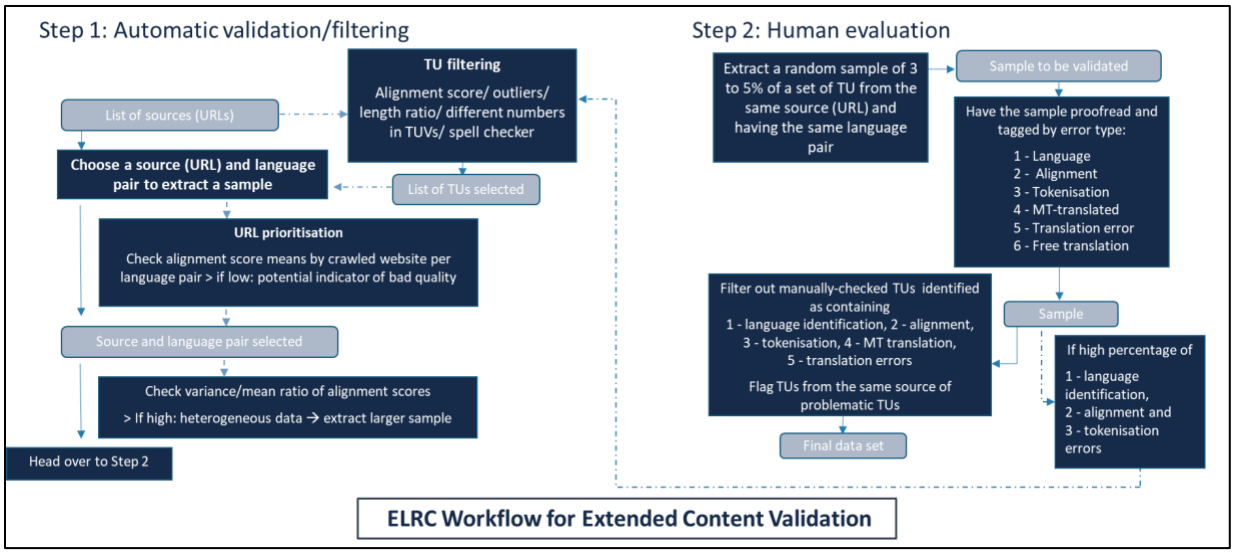


Figure 12: Extended Content Validation process in ELRC

Most importantly, every resource in ELRC is also checked for its legal status. The objective of this so-called **legal validation** process is to allow reuse and redistribution of Language Resources by identifying and assigning the correct legal status. Key aspects to be addressed and assessed as part of the legal validation include in particular relations to Public Sector Information Directive (PSI), Copyright. Public licenses and Terms of Use. Figure 13 below provides an overview of the legal validation process employed in ELRC.

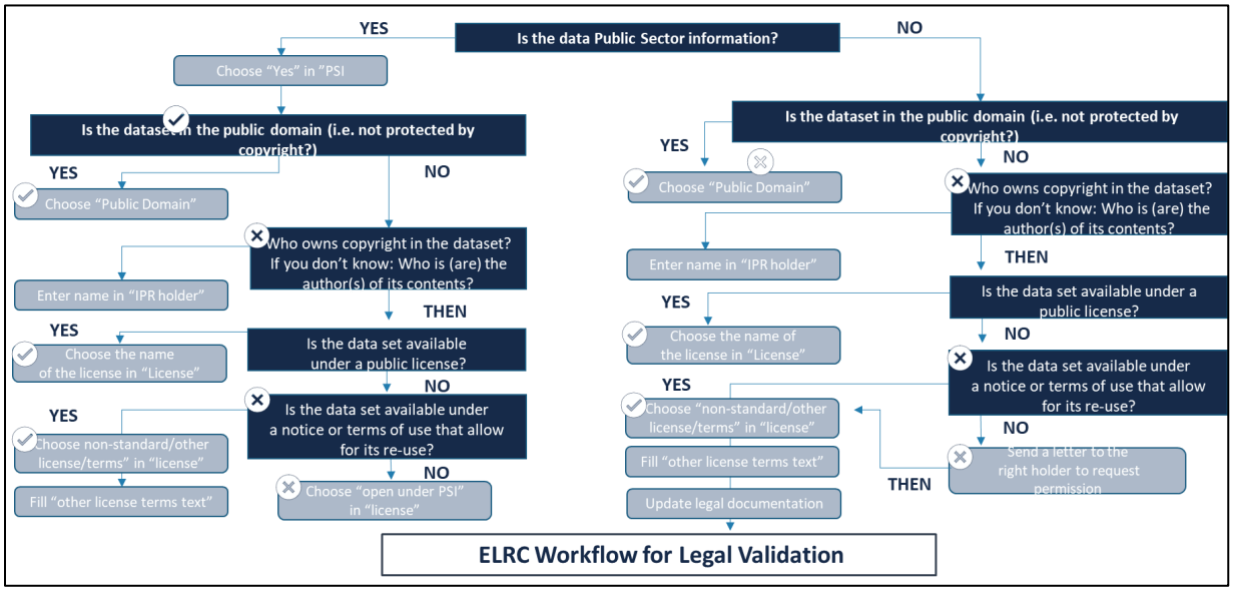


Figure 13: Legal Validation process in ELRC

An important aspect in the legal validation process is the handling of personal data (in accordance with the GDPR) which can present a challenge. Anonymisation (i.e., detecting and removing personal data) may help. There are several tools available, including MAPA (Multilingual Anonymisation Toolkit for Public Administrations) that uses a de-identification of personal data in order to keep the data set usable for training MT systems. The tool is available through a secured docker that can be installed by

Public Administrations (use of Domibus security and will also be connected to CEF eTranslation).

Discussion points in response to the presentation:

In response to the presentation, several questions emerged:

- When asked for an example of a poor-quality TU to be removed automatically, Victoria Arranz explained that if problematic TUs were detected during the automatic processing, they had to be located, labelled and extracted. The corresponding guidelines on the ELRC website provide useful information in this respect: https://www.lr-coordination.eu/sites/default/files/Documents/Resource%20Collection%20Guidelines_20160506.pdf
- In addition, there was a question about whether the anonymisation tool MAPA was open source or not. It was stated that the tool was going to be open source and that it would soon be available through a panel and API.
- Following the question about further information about the anonymisation tests foreseen within ELRC, Ms Arranz mentioned MAPA and CEF Marketplace as solutions that were currently under investigation. She explained that for the evaluation, a large number of features were taken into account, e.g., multilinguality, performance, coverage, open source, domains, etc. Ms Arranz concluded by saying that overall, the solutions needed to be suitable in the context of MT training / language resources as it was important to keep the context. If that was not the case, the data would no longer be usable for MT training.

3.5.2 Anonymising language data within the CEF Data Marketplace

This talk was presented by Amir Kamran, Head of NLP at TAUS who started with an introduction to the CEF Data Marketplace, which is an **easy-to-use**, **easy-to-explore**, **easy-to-trade**, and **easy-to-trust** commercial platform for language data with features that add value for data sellers and buyers. Figure 14 below illustrates the general workflow from publishing to actually purchasing/selling language data.

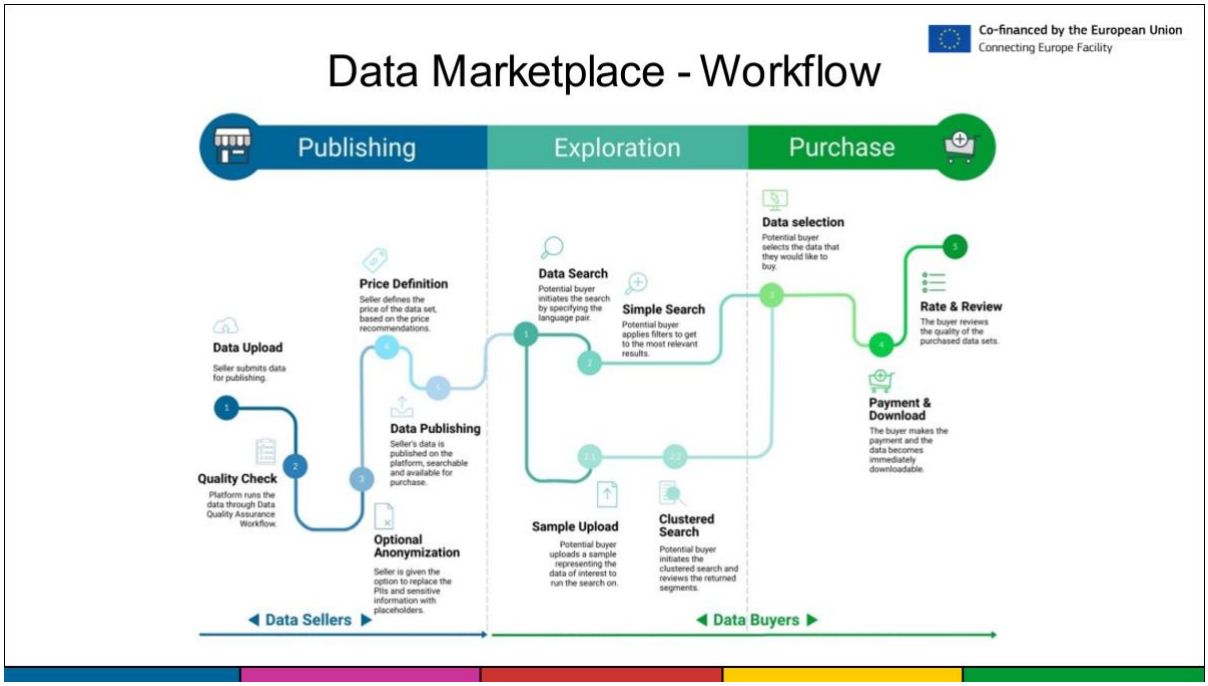


Figure 14: Workflow within the CEF Data Marketplace

As Mr. Kamran pointed out, Personally Identifiable Information (PIIs) requires anonymisation – and the need to anonymise such personal data depends on the context of the text. Typical information that needs to be removed includes emails, URLs, addresses, long integers (e.g., phone numbers, credit card numbers) and alphanumeric codes (e.g. driver's license numbers, identity card numbers, passport numbers, social security numbers, license plate numbers).

With regard to the anonymisation tools employed within the CEF Data Marketplace, evaluations showed that MBERT clearly outperformed Polyglot in all languages, finding significantly more named entities than Polyglot. However, Polyglot proved to be much faster than MBERT (with MBERT processing on average only 12 TUs per second).

Uploaders will then be provided with the recognised PIIs in different categories along with example sentences, allowing them to easily decide if they want to allow or remove certain segments. It is important to note that this is still work in progress and that the final integration and release of the anonymisation tool will only be available in June 2021.

Discussion points in response to the presentation:

The following discussion points emerged from the audience during and after the presentation:

- One participant wanted to know how the presented model took the GDPR differentiations of regular personal data (emails, names, etc.) and sensitive data (gender, synodical affiliation) into account. According to Amir Kamran, the judgement of whether data was sensitive or not could be annotated in the training data alongside the fact that it was personal. The rule system or ML algorithm would then need to generalise what had been annotated and to apply it to unseen data.
- When asked about typical vendors and buyers of data sets CEF Data Marketplace, Mr Kamran answered that anyone who was trying to train or customise MT models

could be interested in buying data, e.g. big language technology providers, companies, etc. Mr Karman further explained that it was possible to create a specific data set out of all available data sets if someone was looking for specific data on the marketplace. On the seller side, it could be anyone having and/or owning data, including translators.

- Following the question about the language coverage of CEF Data Marketplace, Amir Kamran explained that it was the goal to cover all possible languages; there were no limitations in this respect.
- When it comes to the error rates for PER and LOC types, the speaker explained that detailed analyses were not available yet, but that for MBERT, a score between 85 to 95 could be reached most of the time.

3.5.3 How much anonymisation is needed – a legal perspective

In their presentation, Dr. Andreas Sesing (Manager of the Institute of Legal Informatics at Saarland University) and Jonas Baumann (Research Associate at the University of Johannesburg) explained the scope of the data protection laws and also investigated to what extent it was possible to bypass the data protection laws by means of anonymisation. Figure 15 below illustrates the legal framework for data processing in the EU.

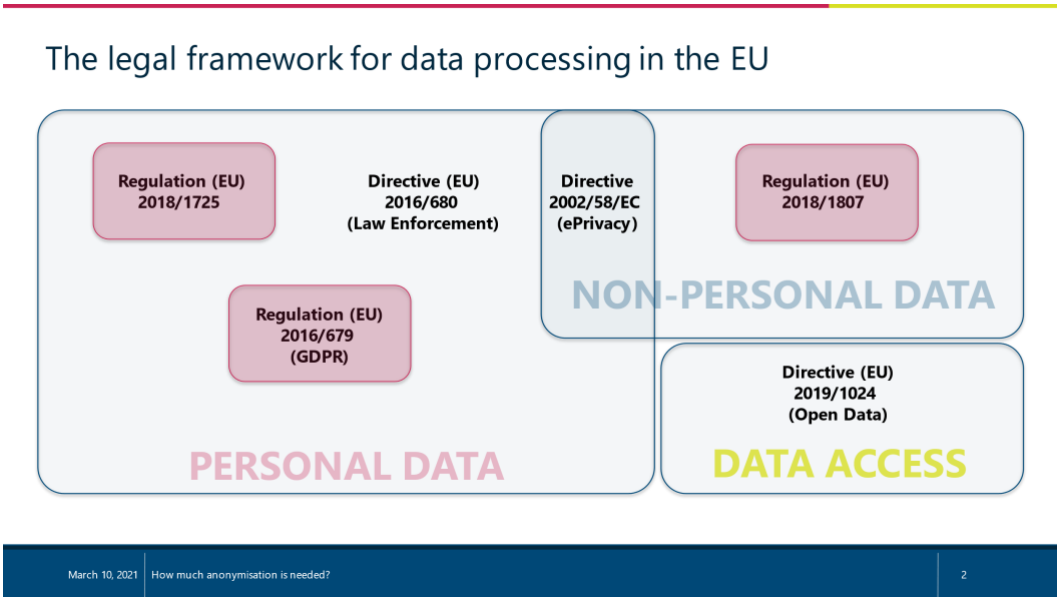


Figure 15: Legal framework for data processing in the EU

According to Article 4 (1) GDPR, “personal data” means any information relating to an identified or identifiable natural person (“data subject”), e.g., names, addresses, social security numbers, but also any other information that may be used to identify a natural person directly or indirectly. As such, even IP addresses may count as personal data. As the judgement of the European Court of Justice, C-582/14 (Breyer vs. Germany) from 19.10.2016 showed, even fully anonymised court decisions may allow for the identification of a natural person (in this case: the plaintiff). The judgment in the case of Breyer vs. Germany hence determined the following with regard to the identifiability of a natural person:

- "[…] to determine whether a person is identifiable, account should be taken of all the **means likely reasonably to be used either by the controller or by any other person** to identify the said person". (para. 42)
- "[…], that would **not be the case** if the identification of the data subject was **prohibited by law or practically impossible** on account of the fact that it requires a disproportionate effort in terms of time, cost, and manpower, so that **the risk** of identification appears in reality **to be insignificant**." (para. 46)

This judgement acknowledges the fact that anonymisation may never be perfect. The publication of texts that may allow for the identification of the data subject appears to be lawful in case it requires unreasonable efforts in terms of time, cost, and manpower to actually identify the person, so that in reality, the risk of identification appears to be insignificant. In terms of the means used for identification, the judgement refers to "means likely reasonably to be used either by the controller or any other person" to identify the said person.

The presenters also pointed out that personal data "infects" the whole data set, i.e., if the data set contains only one piece of personal data, it already counts as "personal data set" as stated in Art. 2 (2) Regulation (EU) 2018/1807: "In the case of a data set composed of both personal and non-personal data, this Regulation applies to the non-personal data part of the data set. Where personal and non-personal data in a data set are inextricably linked, this Regulation shall not prejudice the application of Regulation (EU) 2016/679." As such, bypassing data protection laws require the anonymisation of all data sets in a repository.

Nonetheless, the use and sharing of personal data may be legitimate anyways for data donors from the private sector, if so-called "legitimate interests" of the controller and 3rd parties apply. These could be economic interests, efficient AI training, freedom of information, or building a public repository for LT training data. In such cases, anonymisation mitigates the risk of the data subject.

The speakers also pointed out that Art. 32 (1) GDPR explicitly calls for anonymisation efforts to be undertaken: "Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller […] shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, **including inter alia** as appropriate: **(a) the pseudonymisation and encryption of personal data (…)**"

Sesing and Baumann hence concluded that anonymisation indeed appears to be an effective way to protect the interests of the data donors. With regard to data donors from the public sector, Sesing and Baumann consider the provisions of the Regulation 1/1958 (consolidated version from 1.7.2013) and also the Open Data Directive (if data is necessary to fulfil transparency requirements) as possible legal basis for the upload and sharing even of personal data. However, an explicit regulation of the processing of personal data relating to the training of AI systems could provide better legal certainty in this respect.

Discussion points in response to the presentation:

Two major questions were discussed as part of the presentation, namely:

- When asked about the limits of the degree of data anonymisation, Mr Baumann confirmed that there were clear limits. He referred to the presented judgement by the ECJ on Breyer vs. Germany, which stated that data did not have to be considered personal data if the risk for the data subject to be reidentified was insignificant in relation to the manpower and other means required to identify the person. More precisely, the judgement determined that it should require "a disproportionate effort in terms of time, cost and man-power, so that **the risk** of identification appears in reality **to be insignificant**". However, Mr Baumann pointed out that this was only an ECJ judgement and that there was no grey area with regard to personal data, because either data was considered to contain personal data or not. Therefore, Mr Baumann was convinced that there would be continuous discussions about when data could be considered sufficiently anonymised.
- When asked about the difference between anonymisation and pseudoanymisation, Mr Baumann pointed out that for pseudoanonymisation (which is subject of the data protection law framework), there was a key to retrieve the identity of the data subject, meaning that either the processing or the key to retrieve the original data was known. Contrary to that, there was no way to identify the data subject in a legal sense in the case of anonymisation.

3.6 Data creation and sharing in CEF Generic Services Projects

3.6.1 Massive collection and curation of monolingual and bilingual data focussing on under-resourced languages

The MaCoCu project (Massive Collection and Curation of language data) that focusses on under-resourced languages in Europe was presented by Miquel Esplà-Gomis, Research Assistant and Project Leader at the University of Alicante. The project will start in June and draws on the extensive expertise of all partners in previous activities such as AbuMATRan, ParaCrawl, and Gourmet. The resources collected in MaCoCu shall be relevant to 10 DSIs, namely e-Health, e-Justice, Online Dispute Resolution, Europeana, Open Data Portal, Business Registers Interconnection System, e-Procurement, Safer Internet, Cybersecurity, and Electronic Exchange of Social Security Information. The objective is a minimum size of 5m tokens per parallel and 10m tokens per monolingual corpus. TLD crawling is employed, avoiding the re-use of previously crawled data.

Miquel Esplà-Gormis pointed out that MaCoCu data will be enriched with:

1. **Quality scores** and other indicators from **ELRC guidelines** for a cleaner corpus
2. Language variety identification
3. Information for **anonymisation**
4. For parallel data: **Source language** identified (*translationese*)

The final outcome shall be 10 monolingual corpora and 10 parallel corpora with which one can generate MT training data that is anonymised, for specific language variants and for the DSI domains mentioned above with optimal compromise in terms of size vs. cleanness.

Discussion points in response to the presentation:

Two major questions were discussed as part of this presentation, namely:

- Referring to the optimal compromise between size and cleanness of the data to be provided, it was asked if the speaker could roughly estimate the data size he would end up with. Mr Esplà-Gomis explained that from his experience in past projects, setting a threshold was rather difficult, as the final outcome very much depended on the needs of the final user of the data set. He added that when the data was provided to the user, it was up to him/her to decide on an appropriate threshold. This also depended on the intended purpose, e.g. if the data would be used for MT training, a lower threshold could be used. If, however, it should be used for a translation memory, a higher quality would be required.
- Subsequent to that, it was asked how eHealth-related data could be collected even though it typically contained personal information and was therefore not shared by health authorities. Mr Esplà-Gomis explained that the initial goal was to collect as much data as possible with the help of general crawling on top-level domains; data classification would follow in the second step. The speaker confirmed that for some DSIs, this might result in less data. If that was the case, ways to obtain more relevant data would need to be explored.

3.6.2 Federated Termbank

The Federated eTranslation Termbank Network was presented by Gabriele Sauberer (Director of the International Network for Terminology) and Arturs Vasilevskis (Head of MT Solutions at Tilde). The speakers pointed out that there are still several challenges with regard to terminology:

- Many organisations still do not manage their corporate language.
- Language Service Providers use translation memories or spreadsheets for glossaries, professional terminology management and termbases are rare.
- Language professionals are looking up terms and definitions by using online search engines.
- Quality of Machine Translation output currently depends on the machine, not on high-quality termbases. Artificial Intelligence is rather ignorant when it comes to terminology.
- Terminology data is still fragmented in many separated silos, complicating access and use.

That is why the Federated Termbank project aims to establish a terminology data infrastructure for the creation, management and sharing of terminology resources. A corresponding open FedTerm software toolkit is developed to enable organisations and institutions to locally deploy interlinked and synchronised FedTerm nodes. Figure 16 below illustrates the FedTerm toolkit components.

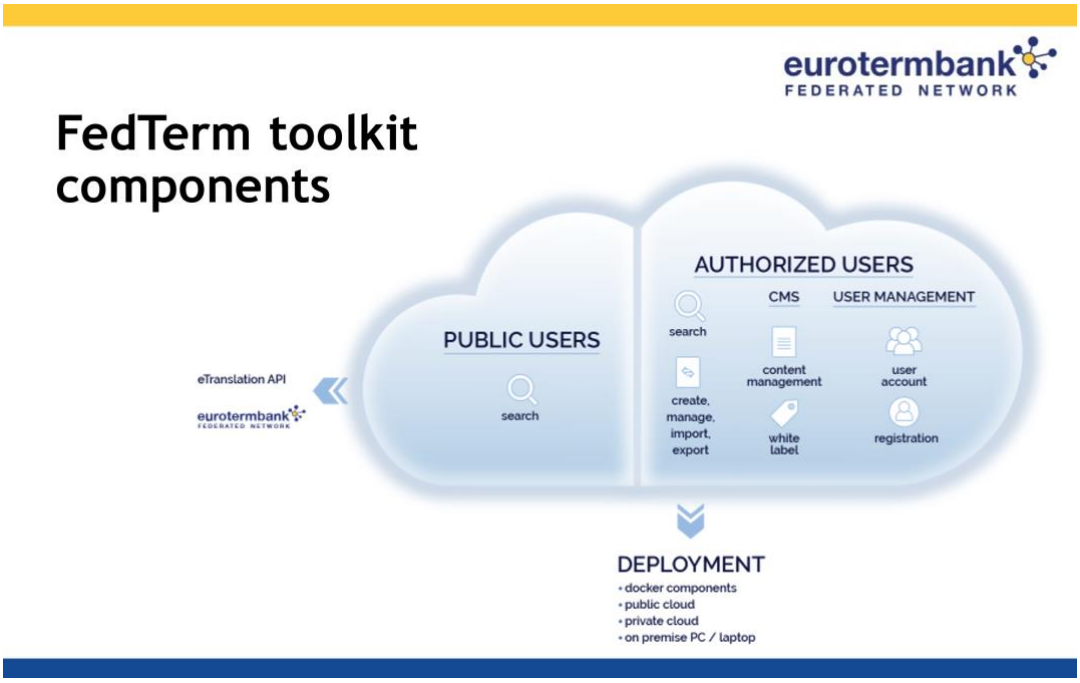


Figure 16: FedTerm toolkit components

While public users can use the term search (prefix search, exact match, full text search, language filtering, domain filtering, search in specific collection) and see term-related information, authorised users may create, manage, import and export term collections. The user management provides a secure authentication and authorisation, also including several self-services for users (e.g., registration, editing of the account, recovering password etc.).

Most importantly, FedTerm is easy to deploy as all components will be provided as Docker images. They will also provide a sample configuration for Docker compose and Kubernetes, and it will be possible to run on cloud services providing Kubernetes (or other Docker based deployments), on local servers (by installing Kubernetes services or Docker Swarm), and on PC or laptops (by pre-installing Docker Desktop).

Discussion points in response to the presentation:

Several questions were raised by the audience during and after the talk, including in particular:

- Following the question on the possibility of hierarchical relationships of terms (narrower, broader terms), Arturs Vasilevskis stated that the tool would be used in tbx format, allowing for connections in the terminology.
- Closely related to the above-mentioned question, the participant added that it would be good to have the possibility of hierarchical relationships etc., as this would make the TermBank much more useful for future AI applications. Mr Vasilveskis explained that this was initially not foreseen, but that potential options to include this in the FedTerm toolkit were currently under investigation.
- Referring to the presentation of Mr O'Shea, which showed that limited resources were being invested in terminology management, although terminology tools have the second highest priority in terms of the actual use by the survey respondents, several questions were raised: The audience wanted to know 1) which conclusions

they should draw from these outcomes, 2) if the available tools were good enough for translators' needs and 3) if there should be more solutions, more awareness and more networking. Ms Sauberer explained that there was always a more efficient way to manage and maintain terminology and that there was a gap in the terminology management and synchronisation. She confirmed that there was much more to do, especially when it comes to raising awareness about the available networks and tools. Ms Sauberer added that even though the solutions could of course always be improved, good solutions already existed and that a lot of resources were available, too. According to Ms Sauberer, it was important to spread the word and connect the world with the help of databases like the Eurotermbank and the FedTerm project.

- Following up on the previous questions about terminology, Gabriele Sauberer was asked about the role of the European Association for Terminology (EAFT). She answered that it was a valuable association for individuals, as it was easy to join and facilitated discussions about terminology issues, challenges, solutions, etc. However, referring to the International Terminology Network TermNet, she stated that associations where membership was granted to organisations instead of individuals typically had more impact and a wider reach.

3.6.3 PRINCIPLE

The PRINCIPLE Project which focusses on collection of high quality language resources for Croatian, Icelandic, Irish and Norwegian (Bokmal and Nynorsk) was presented by Jane Dunne (EU Research Project Coordinator at the ADAPT Centre of Dublin City University). In return for their language resources, early adopters (such as National University of Ireland Galway, CIKLOPPEA D.O.O, Icelandic Ministry of Foreign Affairs, Standards Norway, Norwegian Ministry of Foreign Affairs) were offered MT systems which were built with the help of these resources and which were meant to demonstrate the benefits of language data sharing. Data contributors in all four countries completed a questionnaire to receive further information about the translation process (needs, demands, workflows) and the type of LRs available (formats, quality, quantity). Figure 17 below provides an overview of the ELRC-SHARE data used for the 1st baseline engines.

ELRC-SHARE Data used for 1st Baseline Engines*

Language	No. of TUs
Irish	588,663
Croatian	3,337,608
Icelandic	702,139
Norwegian (Bokmål)	1,140,351
Norwegian (Nynorsk)†	-

[*After Iconic cleaning/filtering of bi-lingual corpora]
[† a lack of public data for Nynorsk meant that it was not possible to train a Nynorsk engine]

Figure 17: ELRC-SHARE data used for first baseline engines

The internal evaluation comprised the following steps:

- By using a 3.000 sentence test set, automatic metrics (BLEU, METEOR, TER, chrF) were computed.
- The same 3.000 sentence test set was then translated through eTranslation and the public Microsoft and Google interfaces.
- The automatic metrics for eTranslation, Microsoft and Google where then compared to the iconic baselines.

The project will now start confirmation of early adopters for phase two and the development of corresponding MT systems for phase two. Corresponding evaluation will then follow.

Discussion points in response to the presentation:

Several questions were raised by the audience during and after the talk, including in particular:

- When asked if a pivot language or linguists/corpora with rare combinations such as Icelandic-Croatian were used within PRINCIPLE, Jane Dunne explained that all corpora collected involved English: Norwegian bokmål-English, Irish-English, Icelandic-English etc.
- Following up on this, one participant invited the audience to visit the website of the NTEU project (https://nteu.eu/), in case they were interested in rare EU language combinations.
- One participant wanted to know if any reports about the project's activities, especially on the use-case analysis and the evaluation of MT systems, were available online. Jane Dunne explained that there were no reports available, at the same time offering the participant to contact her in case of any specific questions. Ms Dunne also invited the audience to visit the project's website https://principleproject.eu/.

- In addition, it was asked how the project ensured that no personal data was included in the sentences and whether anonymisation was being used for that. Ms Dunne answered that from the beginning of the project, all potential data holders were informed that there was no anonymisation facility available in PRINCIPLE. Therefore, the data holders had to provide data that did not contain any personal information.
- At the end of the discussion, one participant pointed to an example of research-related vocabularies, which might be useful: https://vocabularies.cessda.eu/.

3.7 Summary and conclusions

Andrea Lösch (ELRC Project Manager at DFKI) concluded that the 5th ELRC Conference was a day packed full of interesting presentations and insights starting from new insights into the quality of language data, how to judge it, how to use crawled data for machine translations and its consequences. In the afternoon, the focus was on technical possibilities of re-using language data, in particular using anonymisation, also referring to the legal constraints of doing this. As a consequence, it became clear that it is indeed possible to make language data GDPR compliant – and she stressed that we all should consider this important message in our efforts for data collection.

The different CEF project that were presented at the end of the conference underlined once more the importance and feasibility of successfully collecting language data for the development of machine translation systems, even for under-resourced languages. As was shown, there were different approaches starting from massive crawling to personal data collection, all successfully pursuing the goal of providing much-needed language data.

As such, the participants can look back very positively on this conference day, because there were many important insights and lessons that we can take "home" with us for our future work and activities. Andrea Lösch also expressed an official and fourfold thank you, stating:

“That is why my first thank you of the day goes to all our presenters. Thank you for your availability, thank you for sharing your knowledge and expertise with us today and thank you for answering our questions! At the same time, I need to thank each and every single participant today. Thank you for your interest, thank you for your comments and questions. Without this, we would have learnt a lot less today. Also, I want to say a big “Thank you” to all representatives of CEF Digital, starting with June Lowery-Kingston and Philippe Gelin. It is DG CONNECT who made this conference and the collection of language data within ELRC for language tools and services like eTranslation possible. And last but not least, I would like to thank some people who unfortunately have been invisible today, but who brought this conference alive and provided the necessary technical and organisational support that it takes. Thank you, Eileen and Stefania, for all your support today. So, when closing this conference now, I will not just say goodbye. Instead, I would like to wish you a wonderful afternoon and evening – and say: Let’s stay in touch, even virtually, and see you soon at one of our events. And we look very much forward to your feedback in our conference evaluation survey.”

Philippe Gelin, Head of Sector Multilingualism at DG CONNECT, confirmed the thanks to all participants, contributors, organisers and supporters of this event, wishing everyone a nice evening.

Figure 18: Screenshot of "Thank you" Visual

4 Annex

4.1 Annex 1: Conference Participants

The following sub-sections provide an analysis of the conference participation based on their location (geographical coverage) and the sector they operate in.

4.1.1 Geographical coverage

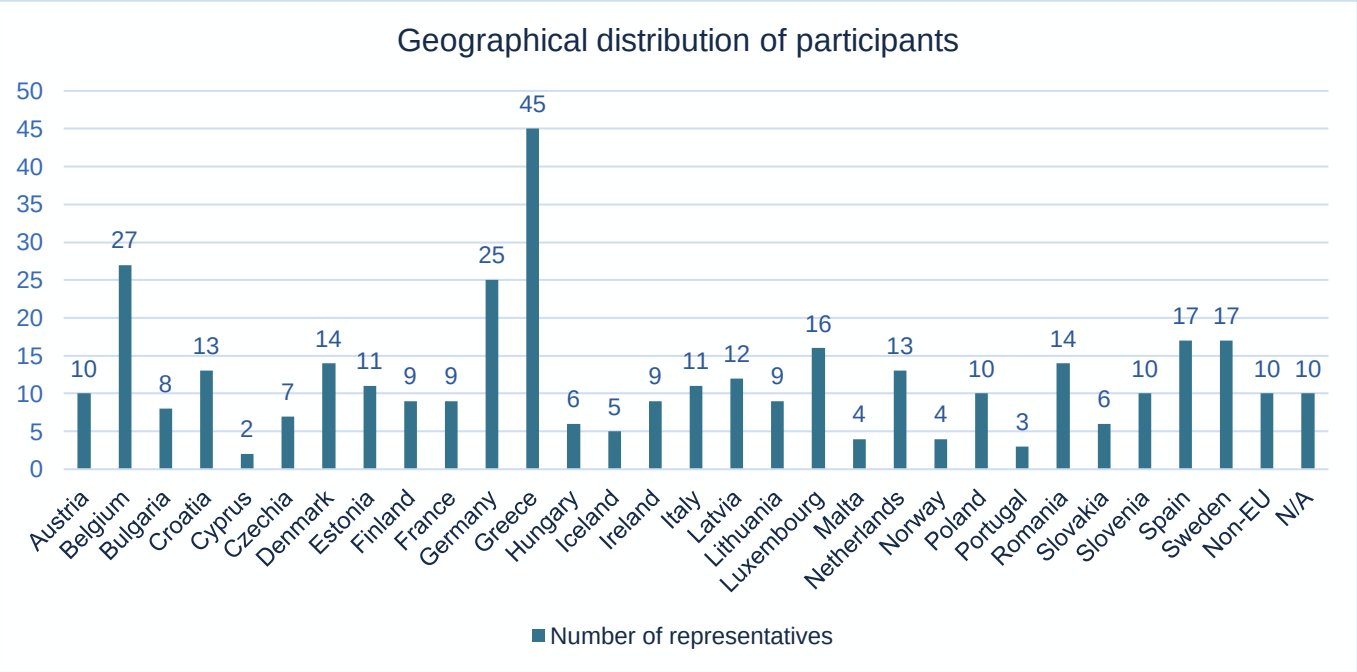


Figure 19: Participant distribution by country

4.1.2 Sectors covered by conference participants

In the registration form, participants were also asked to indicate the sector they are representing, leading to the following distribution[4]:

[4] Some participants assigned themselves to more than one sector, which is why they were counted twice/several times.

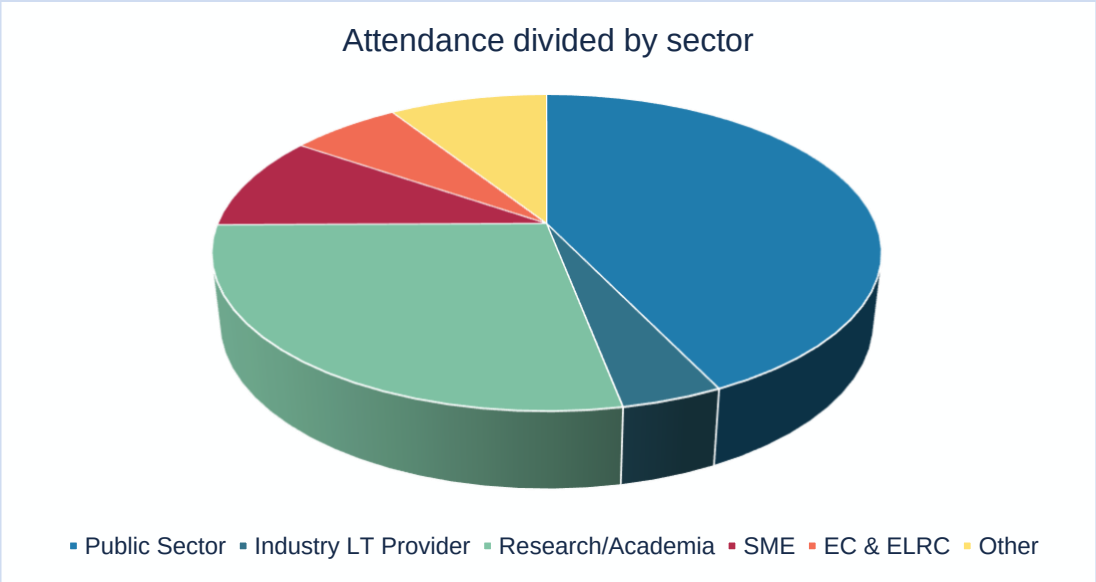


Figure 20: Participants divided by sector

4.2 Annex 2: Conference Presentations

All presentations are available online through the ELRC website: https://www.lr-coordination.eu/node/304

Moreover, the full recording of the 5th ELRC Conference can be found on YouTube: https://www.youtube.com/watch?v=DRZpbmV6SfE