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 D2.3.2
Task 3

ELRC Workshop Report for Luxembourg

Author(s):	Dimitra Anastasiou
Dissemination Level:	Public
Version No.:	V1.0
Date:	2021-01-18

Contents

1 Executive Summary 3

2 Workshop Agenda 4

3 Summary of Content of Sessions 5

3.1 Welcome and introduction 5

3.2 The Public Services Ecosystem in Luxembourg 6

3.3 The potential of Language Technology and AI – where we are, where we should be heading 6

3.4 The CEF AT platform 7

3.5 Language Technologies for the Luxemburgish public sector with sub-sequent live poll 9

3.6 Language data creation, management and sharing: existing practices and challenges 11

3.7 Summary and conclusions 12

4 Synthesis of Workshop Discussions 13

5 Country Profile: Language data creation, management and sharing 20

1 Executive Summary

This document reports on the *2nd Luxembourgish ELRC Workshop*, which has been held online on 11.12.2020 from 09:30-12:45 due to the COVID-19 pandemic situation. Most invited speakers spoke in the official languages of Luxembourg, mostly German and French, and there were interpretation services offered, so that the participants could follow the event either in English, German, or French. We particularly thank www.josten.lu for the interpretation services.

There were overall 81 participants who attended the online event, a significant number given the small population size of Luxembourg (current population of Luxembourg is 630,823 as of 24.12.2020). The participants were mostly from public institutions, public administration and ministries of Luxembourg, as well as European Institutions, Academia, industry partners/SMEs, and freelance translators.

The ELRC office, Andrea Lösch, Eileen Schnur and Stefania Racioppa, fully supported the local organiser, who was the Technology NAP, Dimitra Anastasiou, in the organisation of the event, including inviting speakers, participants, preparing a promotion video, advertising the workshop in social media, and providing the zoom license including the interpretation services. In the invitation email, the ELRC office prepared and sent a useful document including practical hints about participating in an online event. This facilitated a smooth conduction of the workshop without any technical difficulties. The ELRC office was very helpful and supportive during the whole preparation phase, but also during the virtual event, such as accepting participants who joined late, recording, moderating, etc.

The following section includes the agenda of the event (Section 2); Section 3 briefly informs about the content of each presentation of the workshop (Subsections 3.1-3.7). In Section 4, there is a summary of the discussion raised during the workshop. In the last section (Section 5), we focus particularly on Luxembourg and the current situation on language data creation, management, and sharing.

All presentations by the speakers are available at http://lr-coordination.eu/luxembourg.

2 Workshop Agenda

The workshop agenda was as follows:

Time	Title
09:30 – 09:40	**Welcome and introduction** ***Dimitra Anastasiou***, *Luxembourg Institute of Science and Technology*
09:40 – 10:00	**The Public Services Ecosystem in Luxembourg** ***Pit Schneider***, *Bibliothèque nationale du Luxembourg*
10:00 – 10:20	**The potential of Language Technology and AI** ***Christoph Schommer***, *University of Luxembourg*
10:20 – 10:50	**Demo Session: The CEF AT platform** ***Andreas Eisele***, *European Commission*
10:50 – 11:00	***Coffee Break***
11:00 – 11:45	**Language technologies for the Luxemburgish public sector with sub-sequent live poll** ***Thomas Vavra***, *IDC;* ***Alexandros Poulis***, *Multilingual AI Advisor and* ***Enrico Santus***, *MIT/Bayer*
11:45 – 12:30	**Language data creation, management and sharing: Existing practices and challenges** ***Anita Sempels*** *and* ***Andrea Benedetti***, *Wordbee and* ***Andrea Lösch***, *DFKI GmbH*
12:30 – 12:45	**Summary and conclusions** ***Dimitra Anastasiou***, *Luxembourg Institute of Science and Technology*

Apart from a slight delay towards the end, the agenda was followed as planned; there were no changes or any technical difficulties.

3 Summary of Content of Sessions

3.1 Welcome and introduction

Dimitra Anastasiou (Luxembourg Institute of Science and Technology)

Dimitra Anastasiou was the main moderator for the whole event. She first welcomed all participants in English and she introduced the housekeeping rules, e.g. to switch the camera and microphone off when the speakers are talking, and to ask questions either directly after each presentation or in the chat. After the welcome, she explained how participants can select their preferred language (English, German, or French) to listen to the talks. There were no technical difficulties, and participants got settled very quickly concerning this.

Thereafter and during the whole workshop, Dimitra Anastasiou spoke in German. In her introduction, she described the CEF programme as a key EU funding instrument that supports the development of high performing, sustainable and efficiently interconnected trans-European networks in the fields of transport, energy and telecommunications. The Telecom[1] strand, among other things, funds the deployment of pan-European digital services and one of these Digital Services is eTranslation, the Machine Translation system of the European Commission. ELRC is as an Action supporting eTranslation, by coordinating the collection of language resources that are necessary to enhance the system and by raising awareness of the role that not only translation, but also other language technologies have to play in Digital Europe as well.

Then she went through the objectives of ELRC which are the following five:

- *Identify* the multilingual needs of public services (for instance, what kind of contents need to be translated, in which languages, etc.);
- *Engage* with the public sector in order to identify language resources relevant to these needs;
- *Gather* these language resources in a central repository – the ELRC-SHARE Repository;
- *Help* public services with any technical or legal issues that may come up when sharing language resources;
- *Act* as an observatory for language resources across Europe.

She explained that in every ELRC member country, there is a Public and a Technology National Anchor Point (NAP). In Luxembourg, the Public NAP is Mr. Gérard Soisson from the Ministry for Digitalisation and Dimitra Anastasiou from Luxembourg Institute of Science and Technology has been appointed as Technology NAP. Later on, Dimitra briefly went through the official languages spoken in Luxembourg and their rankings depending on the context:

- **At home,** *Lëtzebuergesch* is the most widely spoken language (74 %), followed by French (32 %) and Portuguese (15%).
- **At work**, 98% of the Luxembourg population speaks French, 80% speaks English, and 78% speaks German. Luxembourgish is used by 77% of the population.
- **In a social context**, French (81%) narrowly outstrips *Lëtzebuergesch* (77%): the latter is the preferred language in particular among young people aged 15 to 24 (92%) and those aged 65 and above (80%) in the context of their free time.

Noteworthy is though, the language that the citizens use when they ask for a public service in Luxembourg. According to the 1984 law on the language regime, there is a possibility to use any of

[1] https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom/apply-funding/2020-cef-telecom-calls-proposals

the three official languages French, German or Luxembourgish in the field of justice and administration. The public servants are then entitled to answer in any of the three languages. However, laws are drafted in French with the important consequence that, legally speaking, French alone is authoritative at all levels of public administration. A brochure[2] published by the Information and Press Service of the Luxembourg Government describes multilingualism in different contexts in Luxembourg and is available in English, German, French.

Finally, Dimitra Anastasiou gave the floor to Andrea Lösch, the ELRC project manager who introduced the agenda of the workshop.

3.2 The Public Services Ecosystem in Luxembourg

Pit Schneider (Bibliothèque Nationale du Luxembourg/BnL)

Pit Schneider from BnL[3] presented the status of the Public Services Ecosystem in Luxembourg. He mentioned the newly founded Ministry for Digitalisation in Luxembourg and particularly one of the projects he and his team are currently working on. Pit Schneider is working on improving Optical Character Recognition (OCR) and Newspaper Exploration. He discussed the recognition of named entities, entity relations, timeline, maps and wikidata. As data, Pit and his team are using and training the Newspaper Corpus containing about 200 million text lines. They take as ground truth about 100 thousand text lines and based on this, they run binarization, segmentation, font, and OCR. The binarization includes cleaning, dilation, padding, inversion and white on back detection steps, so that the original text block becomes a binarized text block. Segmentation includes morphology, connected components, and horizontal histogram projection. The font recognition uses convolutional neural network, and a binary classifier. For OCR, they are using open-source software, such as *kraken*[4]*,* *Tesseract*[5], and *Calamari*[6] to do a custom model training. As for the named entity recognition, they use the parser *spacy*[7]. As a result of their work, the accuracy depicted an improvement for an estimated 30% of text blocks. As for improvement of newspaper exploration, they are currently observing first results on named entity recognition and are currently developing a new User Interface for a new BNL Labs platform.

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

Christoph Schommer (University of Luxembourg)

Prof. Christoph Schommer from the Department of Computer Science, University of Luxembourg, gave an overview of AI along the years and highlighted recent research projects focusing on Luxembourgish language. He began his talk presenting a few multifaceted Natural Language Processing (NLP) applications. He explained the crucial role Deep Learning plays for AI. He also illustrated that AI can be so advanced that it is even capable of writing a poem. Computer scientist professors explain how AI and Machine Learning (ML) make it possible for AI to affect art and write poems[8]. In 2021, Ranjit

[2] https://luxembourg.public.lu/en/publications/ap-langues.html
[3] https://bnl.public.lu/fr.html
[4] http://kraken.re/
[5] https://github.com/tesseract-ocr/tesseract
[6] https://github.com/Calamari-OCR/calamari
[7] https://spacy.io/usage/linguistic-features
[8] https://www.youtube.com/embed/WGt8MkeGpNA

Bhatnagar, an artist and programmer invented the *Pentametron*, an art project that mines the Twittersphere for tweets in iambic pentameter[9].

Christoph Schommer continued his presentation introducing two projects funded by the Fond de Recherche National (FNR): *Deep House* [10] and *STRIPS*. The internal address of *Deep House* is http://10.240.2.56/ and the collaborators are David Matos Freitas, Fabio Di Biase, Gabriela Vieira Pinto, Gilles Chen, Joshgun Sirajzade, and Christoph Schommer. Taking the Open Research Dataset Challenge (CORD-19) as a resource of almost 60000 scholarly articles, *Deep House* has two central goals: i) to consolidate available data in a Covid-19 data warehouse by applying appropriate data integration techniques, and ii) build a web-based platform being extendable, which should demonstrate a discovery of relevant information.

The second project *STRIPS*[11] with collaborators Joshgun Sirajzade, Christoph Schommer, Christoph Purschke, Peter Gilles, Daniela Gierschek, has as a goal to develop a toolbox of semantic search algorithms for Luxembourgish. The main focus lies in the linguistic processing of texts written in Luxembourgish (particularly stemming, use of phonetic dictionaries and tagged word list for Luxembourgish; Part-of-speech-tagged text corpus), in similarity learning aspects to allow fuzziness in search queries, and in the identification of temporal cross-dependencies inside the Luxembourgish text corpus. The collaborators have been given heterogeneous text sources (official news items and user-contributed comments) by RTL[12].

After presenting the two aforementioned projects, the speaker illustrated two errors of MT in the DeepL and Google Translate system. The two examples were “The astronomer marries a star” and “The old man the boat”, sentences which still pose a lot of difficulties in MT. Both DeepL and Google Translate translated the sentences wrong, both from EN-DE and EN-DE: “Der Astronom heiratet einen Stern” and “Der alte Mann das Boot”. *Star*, in this case, is not the “fixed luminous point in the night sky”, but rather “a principal performer of a show”. Similarly *man* is, in this case, a verb, meaning “to operate sth./serve in the force”, and not the noun referring to a “male adult”. As for future prospects, Christoph Schommer referred to BERT *Bidirectional Encoder Representations from Transformers* (Devlin et al., 2018[13]), GPT-3, OpenAI’s language generator, and wav2net (Unsupervised pre-training for Speech Recognition).

3.4 The CEF AT platform

Andreas Eisele (European Commission)

Dr. Andreas Eisele from the European Commission is a Chief Scientific Officer of DGT R3.4 working on the CEF Automated Translation Platform eTranslation. He began his presentation with a short history: the previous name was MT@EC and was based on statistical MT. eTranslation is built on neural MT, supports domain adaptation, and can be integrated with online public services, whereas CEF.AT besides eTranslation, includes also other LT tools and supports actions for data collection (such as ELRC) and other funded projects (see Figure 1).

[9] https://www.newyorker.com/culture/annals-of-inquiry/the-mechanical-muse

[10] https://wwwfr.uni.lu/fstm/actualites/article_series_the_experts_behind_luxembourg_s_covid_19_fight2

[11] http://sentilux.uni.lux/stripsannotation/sentences.php

[12] https://www.rtl.lu/

[13] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

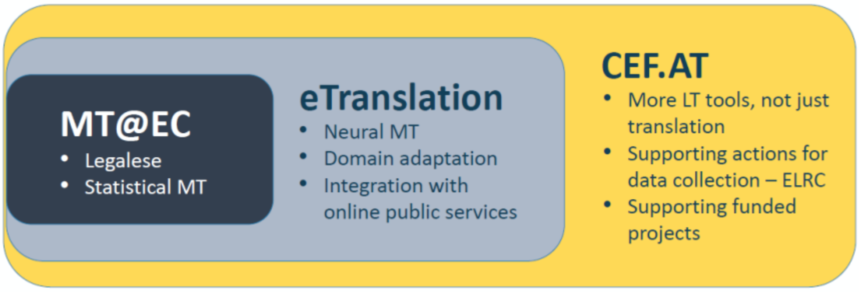

MT@EC
Legalese
Statistical MT
eTranslation
Neural MT
Domain adaptation
Integration with online public services
CEF.AT
More LT tools, not just translation
Supporting actions for data collection – ELRC
Supporting funded projects

Figure 1. History of CEF.AT

Andreas Eisele made clear that eTranslation is open to European SMEs since March 2020. He also mentioned the two use scenarios of eTranslation, as a web interface and as an API. The interface can be used by human users to automatically translate text and get rough translations which should then be revised by human translators. The API can be integrated in workflows, websites, digital services, etc.

eTranslation currently supports translation to and from any of the 24 EU official languages, Icelandic and Norwegian, and since recently also Russian and Chinese. The documents supported can be in various formats (doc, pdf, odt, xlsx, html, etc.). Some of the domains that eTranslation covers are:

- EU formal language
- General text
- Court of Justice Case Law
- Cultural
- Deutsche Bundesbank
- IP Case Law
- Ministère des Finances (France)
- Public Health
- Technical Regulation Information Systems
- Valtioneuvoston Kanslia

He mentioned that what works best are texts related to EU policies, whereas what works less well is the non-standard, new or creative text, and also single words or expressions that depend on context.

Andreas Eisele demonstrated many examples extracted by eTranslation and was very honest about the quality of the MT output. Just for instance, he provided the example of *The chair was broken for the duration of the meeting,* which is translated in French as *Le président est cassé pour la durée de la reunion*. As eTranslation is trained mostly for EU formal language, *chair* is not translated as it should, i.e. the piece of furniture, *la chaise*, but as the president, *le president.*

Regarding the future prospects of CEF AT, Andreas Eisele mentioned i) extending the domain coverage to more general text, but also scientific texts and social media, ii) extending language coverage (non-EU languages of social & economic importance), iii) adding further language technologies, such as speech-to-text, anonymization, named-entity recognition and also to be used as a basic CAT tool.

Before he finished his presentation, he provided some useful links about self-registration[14] of individual users as well as the web service integration[15].

3.5 Language Technologies for the Luxemburgish public sector with sub-sequent live poll

Alexandros Poulis (TransPerfect) and Enrico Santus (MIT/Bayer)
Thomas Vavra (IDC)

In this session, it was Alexandros Poulis, manager of Dataforce of TransPerfect and Enrico Santus, Data Scientist from MIT/Bayer, who had a presentation first, followed by Thomas Vavra from IDC.

Alexandros Poulis and Enrico Santus began their presentations with two very enlightening mottos: *With great data comes great responsibility* (Alexandros Poulis) and *Changing the world, one project at a time* (Enrico Santus). Alexandros Poulis first presented a very evolutionary example of AI in 2040: Dr. Monique Weber wakes up after 7.34 hours of sleep, has a healthy breakfast prepared before she catches the self-driving bus to Kirchberg hospital. Her kids get offered a personalized curriculum when the family has an emergency. There is a pandemic alert in Australia, but through forecasting technology and accelerated drug development, the virus is expected to be under control in 5 days. Moreover, AI-based financial systems analyse Luxembourg's budget and national dept for 2041. While all this sounds like science fiction today, thanks to advances in AI, the foundations are already being built. According to the speakers, the main goals of 2040 that AI can certainly enhance are:

- Healthy Population
- Inclusion and education for all
- Sustainable mobility
- Environment
- Sustainable finances

Enrico Santus continued with the advances of deep learning in a large number of tasks involving language, vision, motion. In the clinical domain, for example, machine learning (ML) can create risk prediction models, which help to understand if a person is likely to develop some diseases in the future. Systems are then able to recommend the most appropriate preventive drug or monitoring approach. Computer vision algorithms can instantly evaluate the patients' X-rays, providing accurate feedback with no waiting time and at no cost.

The presenters then illustrated why data labelling is so important. They presented examples of investment banking, waste management, and healthcare. They highlighted the human should be in the loop, when after running the machine learning (ML) model, there is still ambiguous data that should be sent to human annotators.

There were no questions asked to Alexandros Poulis or Enrico Santus.

Thomas Vavra from the market research company IDC, who works with the European Commission and the ELRC consortium, presented the results of a survey to see how eTranslation is being used, by whom, and for what, but also how the services can be improved. 58% of the respondents were from public administrations, whereas 35% were SMEs, followed by representatives from Academia (4%), and market players (3%). The satisfaction with the platform was depicted as follows: 58% were

[14] https://webgate.ec.europa.eu/etranslation/public/welcome.html

[15] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/How+to+submit+a+translation+request+via+the+CEF+eTranslation+webservice

satisfied or very satisfied, 24% were dissatisfied. LU users gave 4/5 points. What prevents users from greater use is the accuracy of translation (66%), followed by ease of use (22%), and speed of translation (13%). Luxembourgish users were more concerned about the ease of use.

The most requested domains, from the side of public administration, are:

- Economy, finance and tax
- Science and technology
- Business
- Judicial and law enforcement
- Medicine and healthcare

For SMEs, they are:

- Science and technology
- Judicial and law enforcement
- Business
- Intellectual property and trademarks

Thomas Vavra finished his presentation showing the Catalogue of CEF eTranslation services[16]. It includes almost 600 language tools/services from 444 different European providers, there is free text search which can be filtered by function, task, language coverage, and provider's country. In Luxembourg, there are seven providers listed: *Amplexor, SDL Muliterm, Lingua Custodia, Wordbee, Docbyte, everis knowler, EvidenSSE*.

After the presentation of Thomas Vavra, Andrea Lösch moderated a poll with the following questions:

- Does your organisation use/intend to use eTranslation?
- Does your organisation use/intend to use any other Language Technologies or services?
- Regarding the current functionalities of eTranslation, what would you want to improve or add? Please select your top 3 answers!
- eTranslation currently covers all EU official languages, Chinese, Icelandic and Norwegian. For which additional languages do you need translation support most? Please select the top three most important languages from your perspective!
- What are the domains / areas for / in which you mostly translate? Please select up to three options.
- From your perspective, which other functionalities should be added to eTranslation in the future? Please select up to three options!
- Are there any language resources / translations within your organisation?
- Does your organisation employ a data management plan, i.e. guidelines and/or standards for making the data created by the organisation findable, accessible, interoperable and reusable?
- What are, based on your experience, the main difficulties that may prevent the sharing language data? Please select up to three options below.

The detailed results of the poll are presented as part of Section 4 Synthesis of Workshop Discussions.

[16] https://cef-at-servicecatalogue.eu

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

Andrea Lösch (ELRC Project Manager)

Anita Sempels and Andrea Benedetti (Wordbee)

In this session, Andrea Lösch first gave a presentation, followed by Anita Sempels' and Andrea Benedetti's presentation from Wordbee.

Andrea Lösch, complementarily to the presentation of Andreas Eisele, provided some updated statistics with regards to eTranslation. She highlighted the importance of sharing language data, because CEF eTranslation helps the public service to operate multilingually.

She stated that today, there are more than 70 systems connected to eTranslation, more than 7.000 registered SMEs, and more than 40 million words were translated in 2020, a statistical result that would not have been reached by human translation alone.

After that, she introduced ELRC-SHARE, which currently hosts 2355 data sets and more than 200 billion words in all EU official languages. The majority of shared language data are bi/multilingual corpora, followed by lexical resources, and monolingual corpora. The licenses of the shared data are in most cases open licenses. She then mentioned the ELRC White Paper[17] and clarified why language resources are significantly needed to cover all the available languages and domains. Particularly, she spent some time on the case of Luxembourg highlighting that translation is often needed in public services in other languages than the official ones, such as English. For more information about Luxembourg, see Section 5.

In the second part of this session, as a best-case practice for translation management, firstly Anita Sempels introduced the company *Wordbee*[18] to the participants and then Andrea Benedetti made a demo of their translation management tool. *Wordbee* was founded in Luxembourg in 2008 and collaborates with Egypt, Greece, Spain, Poland, France, Germany, and USA. The data centers are located in Luxembourg, Zurich and Amsterdam and there are in total 455 languages that are being translated. *Wordbee* offers a combined project management solution plus translation editor in a single SaaS solution. In Luxembourg, *Wordbee* has several clients in the public sector, and specifically:

- Service Information Presse
- CTIE
- Ministère de L'Education nationale
- Ministère de l'Economie
- Agence pour le développement de l'emploi (ADEM)
- Ville de Luxembourg

The *Wordbee* solution offers the following characteristics:

- Multicolumn translation editor
- Project Management
- Terminology Manager
- Portal
- Management of requests and service providers
- Integration of neural machine translation engines
- Business management and financial reporting

[17] http://stg.lr-coordination.eu/sites/default/files/ELRC_Conference/ELRCWhitePaper.pdf
[18] https://www.wordbee.com/

- Cost management and invoicing
- API

Moreover, they offer Beebox, which is Middleware connecting any content source external to the translation platform.

3.7 Summary and conclusions

Dimitra Anastasiou (Luxembourg Institute of Science and Technology)

Due to a slight delay in the programme, the conclusion was kept short. Dimitra Anastasiou briefly described the project European Language Grid (ELG) and highlighted important information about funding opportunities:

- European Language Grid Calls for Pilot Projects[19]:
    - Development of missing services or solutions for underrepresented languages
    - Duration 9-12 Monate
    - Up to 200.000 EUR per project
- Horizon Europe Programme[20] :
    - Culture, Creativity and Inclusive Society (safeguarding endangered languages in Europe)
    - Digital Industry and Space (strengthening Europe's data analytics capacity, open search and discovery, leadership in AI based on trust).
- Digital Europe Programme[21] :
    - Capacity Building (cloud-to-edge, data spaces support centre, AI on demand platform)
    - Accelerating best use of technologies (Digital Innovation Hubs)
    - Common Services Platform

In the end, she thanked all participants for attending the workshop and run a poll about the evaluation of the workshop.

[19] https://www.european-language-grid.eu/open-calls/

[20] https://ec.europa.eu/info/horizon-europe_en

[21] https://ec.europa.eu/digital-single-market/en/europe-investing-digital-digital-europe-programme

4 Synthesis of Workshop Discussions

In this Section, we discuss the answers of the speakers as well as the poll results.

Regarding the presentation of Pit Schneider from BnL, several questions emerged from the audience, including:

- Is OCR completely language-independent?
- OCR application: How does this solution compares to state of the art, has this been evaluated on some common dataset?
- Once OCR is done, can you distinguish text passages from German and Luxembourgish?
- The font types used in your dataset probably changed or morphed throughout the time. Does that have any impact on your classification, which is binary? Or are these changes negligible?
- Is there a web page where we can read more about the pipeline?

The answer to the question whether OCR is language-independent, was that OCR is not necessarily language-dependent, the algorithms they use are mostly visual. The sequence of characters plays a role, but not that much, so language itself does not play a big role. Concerning Pit's comparison to state of the art, he has looked at "out of the box models" from modern OCR engines; this helped Pit and his team to use their own models and to train them on their data. They found better results compared to what is freely available. To the question about the distinction of text passages between German and Luxembourgish, he said that after the OCR output, they look at named entities and there, the language is indeed more important than for the OCR engine; they need to apply the correct language model to detect named entities; they are still working on it to improve the language detection accuracy. As far as the font types change throughout time, they started with a binary classification and mentioned that it is a good idea to add more font classes over time. This, however, would require more ground truth, in order to have sufficient data available for every class. For now, they see promising results using 2 classes, but the cursive fonts seem to be the biggest problem and probably would represent the next class to be added. He mentioned that there is currently no web page for the public to read about the pipeline, but they plan on making it open source next year.

Christoph Schommer has provided the web address 10.240.2.56 for the project Deep House which was one of the questions raised by the audience. However, he pointed out that the web address given on the slides was an internal address belonging to the University of Luxembourg. More information on Deep House, however, could be found here: https://wwwfr.uni.lu/fstm/actualites/article_series_the_experts_behind_luxembourg_s_covid_19_fight2.

Also with regard to Andreas Eisele's presentation, several questions emerged from the audience:

- Luxembourgish is not an official EU language. The problem I see with Luxembourgish is that it is a mix of German, French and English. The "Luxemburger Wort" could be a useful resource as it provides news articles in German, French and English.
- May I ask, as an interpreter, how does it look like for automatic interpretation?

The remark of a participant that Luxembourgish is a mix of German, French and English and that the "Luxemburger Wort" could be a useful resource opened the discussion about the current situation of eTranslation and Luxembourgish. Andreas Eisele said that Luxembourg probably joined the EU too early. Back then, making it an official EU Language was not on the table, but we should catch up on this. He mentioned the need for more data and to leave the difficult things to the machine, i.e. the neural nets, because they can identify the regularities.

Concerning automatic interpreting, Andreas Eisele cited the highly interesting work by Prof. Waibel (KIT in Karlsruhe and Carnegie Mellon University), who used to translate lectures in real time. He is rather sceptical, because the error sources of speech recognition and MT add up. As a tool for human interpreters, the challenge is the cognitive load, which would increase, if interpreters had to read the results of the MT and correct them. Their project is not yet ready for this, but starting to work on speech recognition, collaborating with the DG for Interpretation SCIC, working on meeting transcription etc. could at least lay the groundwork and may provide opportunities in the future.

An important comment by Dimitra Anastasiou was that the big issue for Luxembourgish is the lack of data, so she invited participants to help collect more data. In case someone has and/or likes to share any translations supporting the development of eTranslation, can upload them to https://elrc-share.eu/ for review. ELRC's Technical and Legal Helpdesk Team (https://www.lr-coordination.eu/helpdesk) will support them in identifying the right license.

As concerns the presentation of Tom Vavra, several questions emerged:

- Has security been included in the survey? This is a selling argument for eTranslation, as mentioned also by Andreas Eisele, as it is guaranteed by the European Commission
- My concern is about the possibility to participate in the improvement of the quality of translation by professionals and linguists… By the way, are there any trial/budget versions?

Regarding the first question, answer was that yes, IDC asked if users were concerned about that, and the overall response was no; they assumed that security measures are built in.

Regarding the presentation on Wordbee, there was a question whether IATE is integrated in their platform. The answer was yes, and Andrea Benedetti showed how this is possible: In the Editor view, you can open the lexicon and have access to IATE.

Coming to the polls´ results, we could gain some very useful insights. Regarding the first question, whether your organisation uses/intends to use eTranslation, out of 27 answers, there were 21 Yes and 6 No answers. To the question whether your organisation uses/intends to use any other Language Technologies or services, again out of 27 answers, there were 25 Yes and only 2 No. These overall very positive results show that indeed there is a stronger need for Language Technologies in general, and not only for MT, but this is why the CEF platform is available: to give access to free tools, support, and funding, and generally help building digital services.

As for the participants´ suggested improvements of MT, the results can be shown in Figure 2 below.

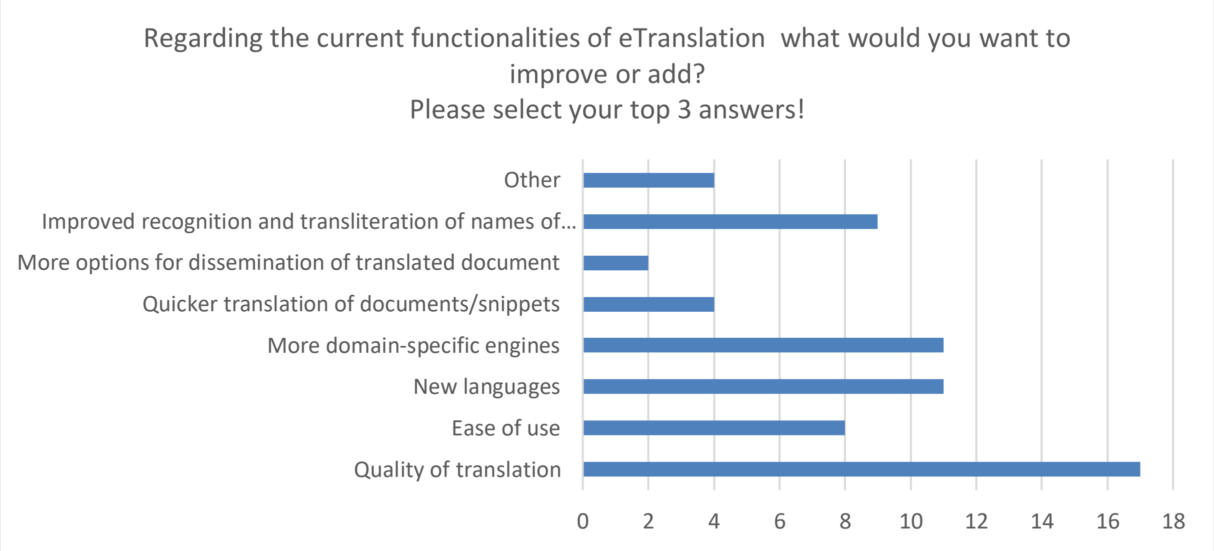


Figure 2. Participants' suggested improvements for eTranslation

As we can see, the quality of translation was the top answer, showing that indeed this is the most important topic that matters, in the end, for the end user. The second highest ranked answer was the addition of new languages (see Figure 3) and more domain-specific engines (equally ranked).

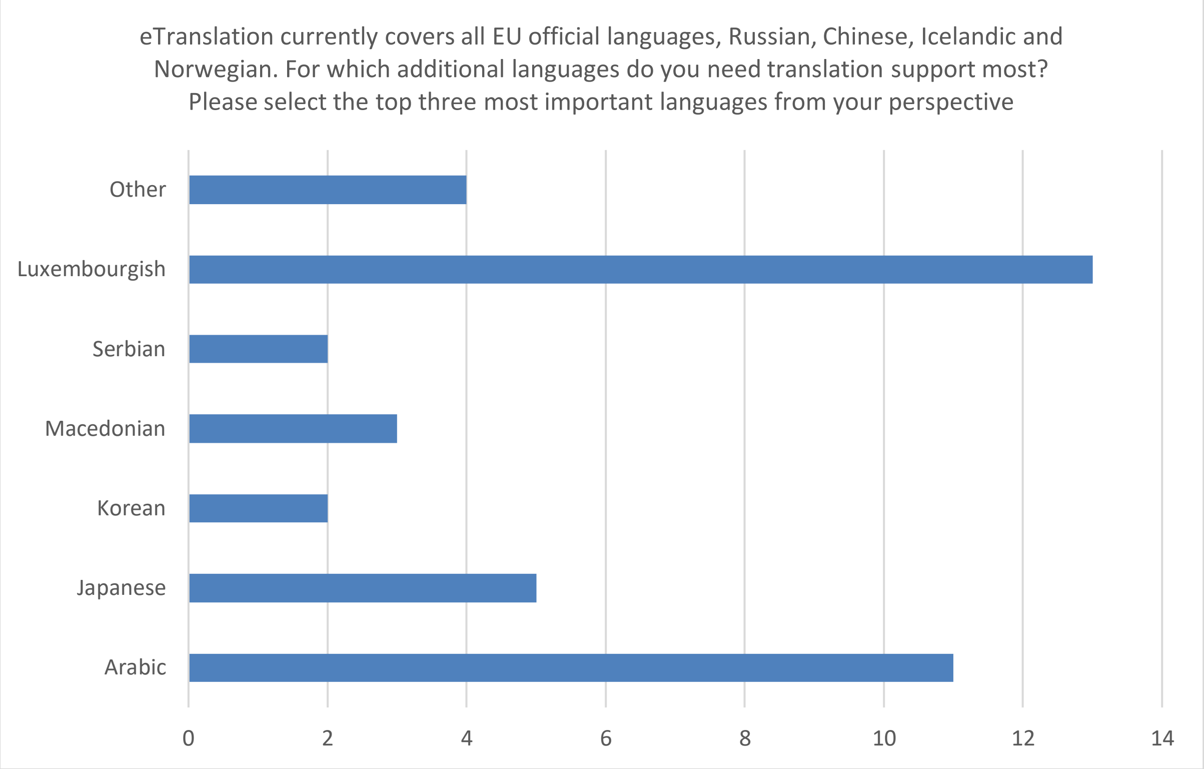


Figure 3. Languages that participants would like translation support

As for which languages the participants would need translation support, Luxembourgish outranked all other options, closely followed by Arabic and then Japanese. This shows that indeed Luxembourgish is needed for the stakeholders in Luxembourg. Although Luxembourgish is not an official EU language, it is one of the three official languages in Luxembourg and it is spoken and written in various contexts. Therefore, automatic translation from and into Luxembourgish would benefit many citizen and businesses equally.

The results of the next question about the domains that participants mostly translate in can be found in the following Figure 4.

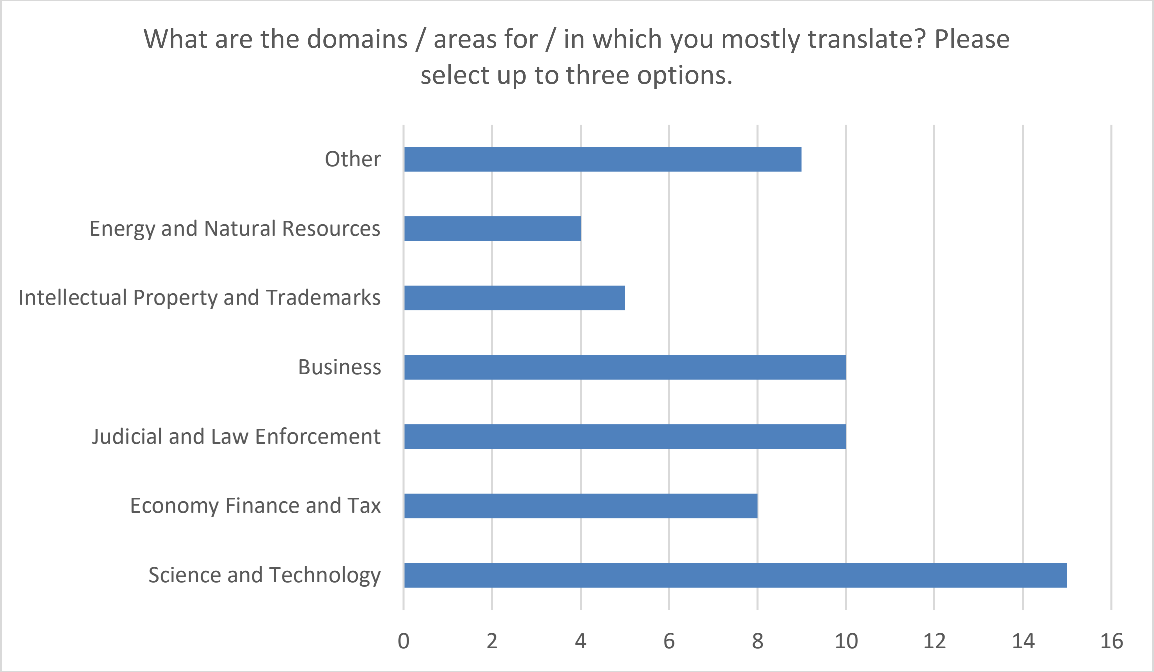


Figure 4. Participants' domains of translation

Science and Technology was the top ranked answer, followed by Business and Judicial & Law Enforcement. For the next question about which functionalities participants would like to have added to eTranslation in the future, the top answer was terminology support, including processing, database, integration, concepts, and definition). The second most selected option was automatic summarization and data anonymization, whereas the third option was automatic translation of text on images; see Figure 5 for the ranking of all options.

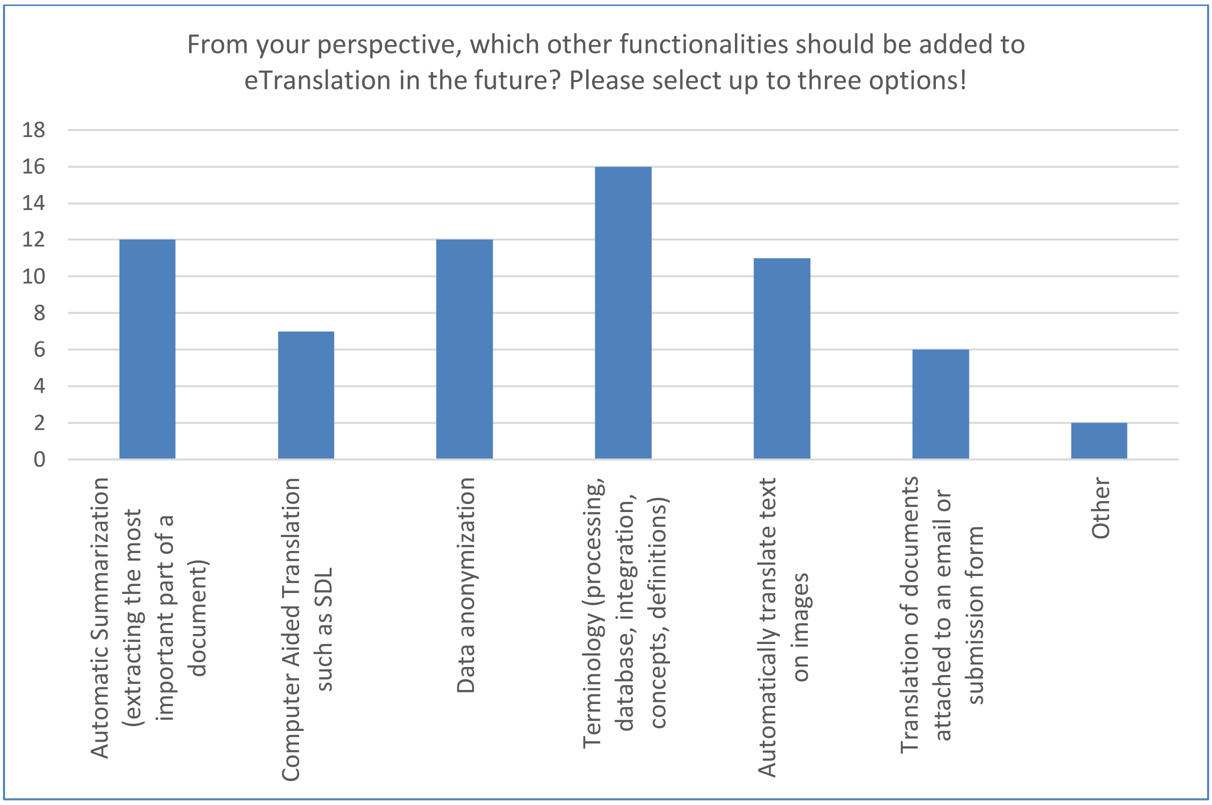


Figure 5. Participants' suggested functionalities of eTranslation

The general take-home message from the Luxembourgish workshop is that SMEs should register themselves in the CEF catalogue for more visibility and that public services should submit more data to ELRC-Share. Luxembourgish is not an official EU language; however, we should try to include Luxembourgish in eTranslation in the future, as this was also proved as a stakeholder requirement through the poll. For this, however, we do need a lot of resources to train the neural nets of CEF eTranslation.

As an answer to the question of the poll about whether there any language resources/translations within the participants' organisations, out of 20 answers, there were 16 Yes and 4 No. To the question "Does your organisation employ a data management plan, i.e. guidelines and/or standards for making the data created by the organisation findable, accessible, interoperable and reusable?", there were 12 Yes and 8 No answers. Regarding the difficulties that may prevent the sharing of language data, the participants answered as illustrated in Figure 6.

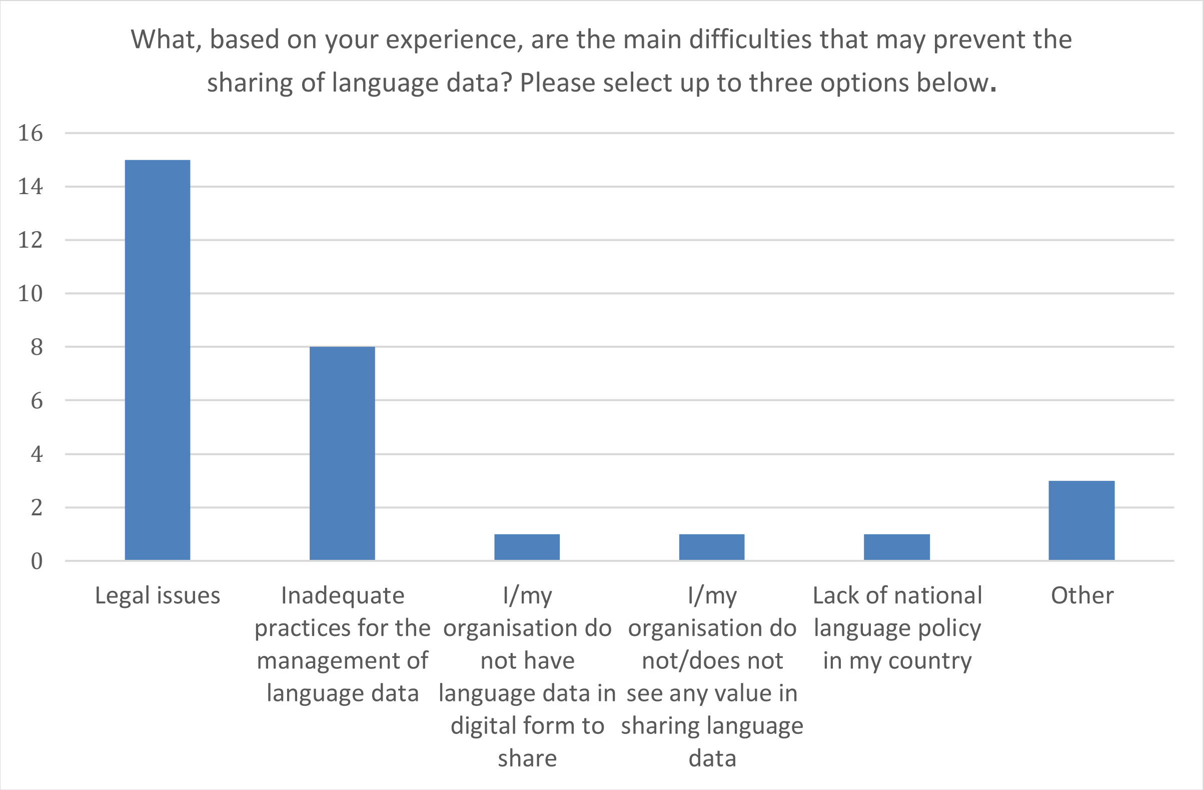


Figure 6. Participants' main difficulties of sharing language data

In fact, legal issues are the main difficulty for the participants of this workshop. To be noted here, the ELRC's Technical and Legal Helpdesk Team (https://www.lr-coordination.eu/helpdesk) will support stakeholders in identifying the right license. Another important issue were the inadequate practices for the management of language data.

Luxembourg is a strongly multilingual country, but this has often the drawback that translation is managed internally and not outsourced. Also, the need for MT is smaller, since many citizens living in Luxembourg are multilingual. That Luxembourgish is a mix of languages, is not a problem for training the language technologies; the stakeholders should share data and should leave the technical difficult things to the machine/neural nets, because they can identify the regularities. Dimitra Anastasiou highlighted once more that we can all benefit from sharing our resources, it is a win-win situation that drives our success.

Last but not least, when asking the participants for their feedback about the workshop, they stated that they were satisfied or even very satisfied with the event (see Figure 7).

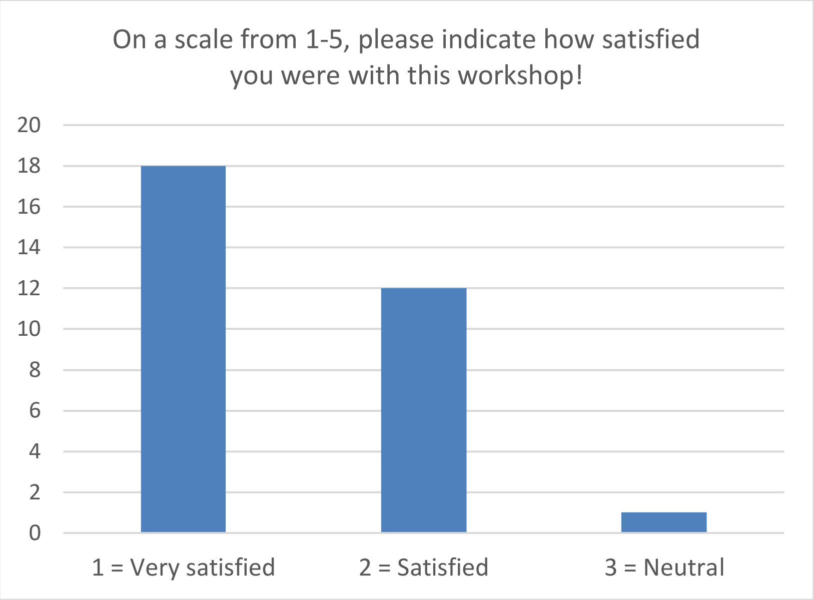


Figure 7. Participants' evaluation of workshop

5 Country Profile: Language data creation, management and sharing

Luxembourg has already provided data to the ELRC-SHARE on several occasions starting with the first language resource being uploaded by Jean-François Wioland from the Governmental Information and Press service (SIP) who submitted Luxembourg's very first language resource on 05.01.2018. This was a corpus and two lexical conceptual resources. The corpus is about SIP publications, while the lexical resources are SIP dictionaries. Two years later, on January 2020, Lynn Pundel submitted several multilingual corpora (German-French-English). These resources are all related to guichet.lu, one is generic, the second about business, and the third about Luxembourgish citizens. All data sets that have been provided so far are multilingual (German-French-English).

In the 2nd Luxembourgish workshop, we made it very clear how helpful it would be for all people to add Luxembourgish to the eTranslation platform. There was a lively discussion initiated by DGT representative Andreas Eisele, that more language data is needed to further improve and extend the coverage of CEF eTranslation. In Luxembourg, multilingualism is a reality for many citizens and this should result in a benefit for the development of language technologies. European Institutions, public services and governmental institutions should share monolingual and bilingual corpora and other relevant data. The sharing of such data will facilitate the development of language technologies, and why not the support of Luxembourgish in eTranslation in the future.

In the case of Luxembourg, it often is the case that public and private organisations do not outsource their documents to have them translated, simply because most employees speak already 3 or 4 languages. Consequently, translations tend to be made internally. This is certainly a drawback of multilingualism, since translation management lacks a specific systematic workflow.

Nonetheless, with regard to the infrastructures for sharing translations and language data, Luxembourg shows significant efforts into making public services digital and multilingual. The main web portal in this domain is Guichet.lu, which is managed by the Government IT Centre (CTIE). Guichet.lu has its own in-house translation service and regularly exchanges translation memories. In order to fully meet their translation needs, all public authorities outsource at least some translations to either freelance translators or language service providers. The government.lu website is the information portal of the governmental Information and Press Service, SIP. It federates all information and news concerning the Luxembourg government in three languages (German, French, and English) and sometimes also in Luxembourgish.

The current infrastructure of language data creation and exchange in Luxembourg is as follows:

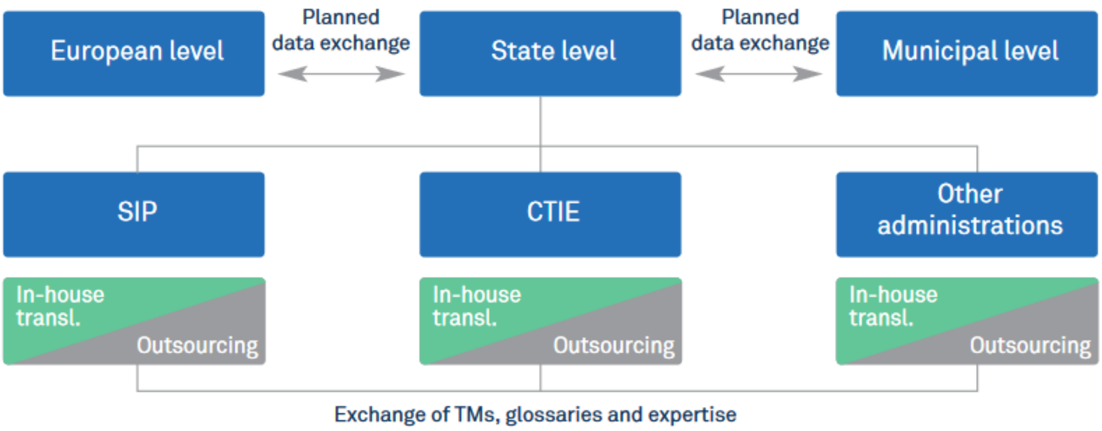
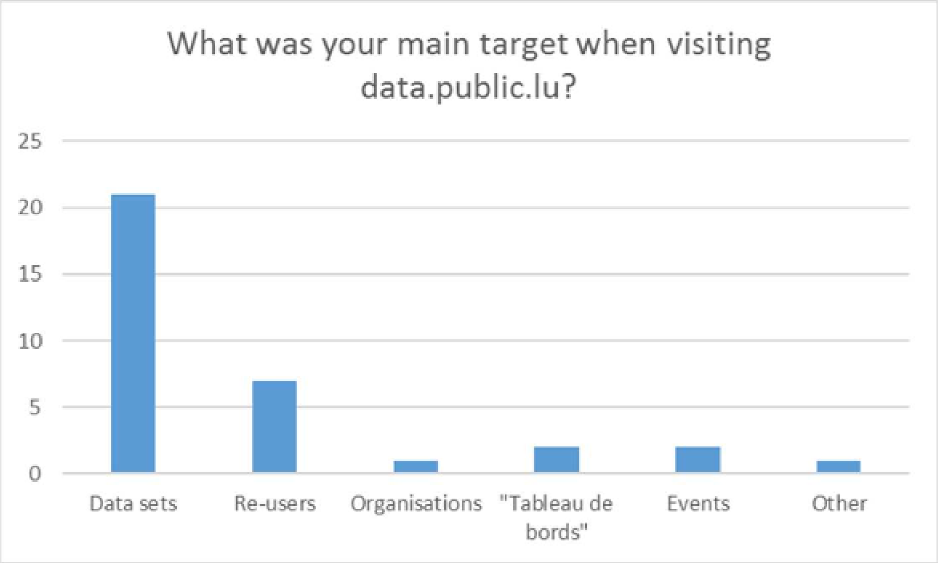


Figure 8. Current infrastructure of language data creation and exchange in Luxembourg

Talking about current advancements with regard to language data in the Luxembourgish public administration, there were no changes with regard to the preparation and processing of translations/language data in the Luxembourgish public administration. However, following Lynn Pundel from guichet.lu, this situation may change in the course of 2021 and corresponding updates will be shared with ELRC. Moreover, other Luxembourgish public administrative entities like the Official Journal of Luxembourg at the Ministry of State (http://legilux.public.lu) (SCL) have started investigating the opportunity of using machine translation tools like eTranslation in their workflows. However, in the case of SCL, the project was put on hold in 2020 due to the added difficulties within the COVID-19 pandemic but may be taken up again in the course of 2021.

As far as open data in Luxembourg is concerned, the Luxembourgish Open Data Portal[22] was launched in April 2016 and hosted more than 800 published datasets. However, the majority of the data sets remain numerical datasets, and not textual. In 2019, Luxembourg Institute of Science and Technology (LIST) and Digital Luxembourg published an evaluation[23] of the impact of Open Data in Luxembourg in order to better understand its users and their expectations in terms of content and functionality. This evaluation with the title “Impacts of Open Data in Luxembourg and the Greater Region – 2019” is a satisfaction survey in order to better understand the users of the Open Data portal and stay tuned to their expectations in terms of content and functionality. The visitors of data.public.lu are mainly interested in finding available datasets (see Figure 9). Respondents who had re-used data from data.plublic.lu explained that their main motivation for re-use this data was, e.g. to “find reusable software applied on datasets similar to Kaggle.com”, to “get new uses/derivations to ameliorate my own dataset”, or to support corresponding specific projects for which re-use of the data was necessary.



[22] https://data.public.lu/en/

[23] https://download.data.public.lu/resources/study-impacts-of-open-data-in-luxembourg-and-the-greater-region-2019/20190510-143345/impacts-of-open-data-in-luxembourg-and-the-greater-region-2019-final.pdf

Figure 9. Target of visitors of Open Data portal

The overall user satisfaction when it comes to looking for information are provided in Figure 10 below. We see that users were able to find the information they were looking for even though this was not always easy and the information quality was not always appropriate.

The detailed results are represented in the following graph:

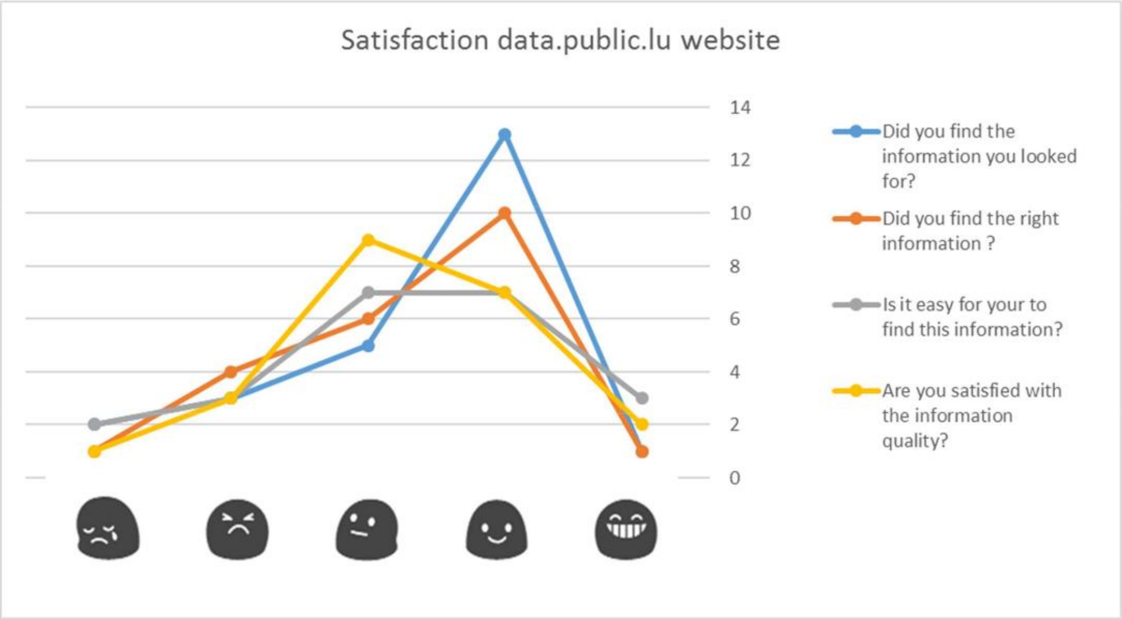


Figure 10. Satisfaction about data

For the first time in its history, a Ministry for Digitalisation headed by the Prime Minister, Minister for Digitalisation, Xavier Bettel, and Minister Delegate for Digitalisation, Marc Hansen[24] was created on 11th December 2018, when the government programme was presented by Prime Minister Xavier Bettel to the Chamber of Deputies. Within this context, there is a recent initiative of the Ministry for Digitalisation and the Government IT Centre (CTIE), called GovTech Lab[25], which is an innovation laboratory that uses open innovation to work with internal (ministries, administrations, public actors) and external actors in the development of innovative solutions (technological or conceptual).

AI is indeed a strategic vision for Luxembourg and Luxembourg intends to remain at the forefront of AI by collaborating across borders, boosting investments, enabling skills training and optimising its data market. The document "Artificial Intelligence: a strategic vision for Luxembourg." [26] has been published by the Government of the Grand Duchy of Luxembourg and digital Luxembourg. It includes a foreword by the Prime Minister, Xavier Bettel, a description of the vision for AI in Luxembourg, the human-centric focus, the regional cluster of AI research in Luxembourg and the focus areas (e.g. data, ethics, skills & lifelong learning, etc.) The strategic vision is not intended as a one-off strategy, but rather the first edition of a policy vision, to be updated on a regular basis and further defined where needed. This policy vision is built on Luxembourg's ambitions as a digital front-runner:

[24] https://digital.gouvernement.lu/en.html

[25] https://govtechlab.public.lu/en.html

[26] https://digital-luxembourg.public.lu/initiatives/artificial-intelligence-strategic-vision-luxembourg

Ambition #1: - To be among the most advanced digital societies in the world, especially in the EU

Ambition #2: - To become a data-driven and sustainable economy

Ambition #3: - To support human-centric AI development[27]

[27] https://digital-luxembourg.public.lu/sites/default/files/2020-09/AI_EN_0.pdf