European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.

European Language Resource Coordination
Connecting Europe Facility

Deliverable Task 6

ELRC Workshop Report for Luxembourg

Author(s): **Khalid Choukri**, ELDA
Eric Ras and Sviatlana Höhn, Luxembourg Institute of Science and Technology
Christoph Schommer, University of Luxembourg

Dissemination Level: Public

Version No.: V1.2

Date: 2016-07-19

Contents

1 Executive Summary

This document reports on the *ELRC Workshop Luxembourg*, which took place at the "Salle Tavenas", 102, Avenue Pasteur; L-2310 Luxembourg, at the Campus Limpertsberg, University of Luxembourg, on Tuesday, the 14th of June 2016. The event was held in French and English language and featured 7 presentations and 2 panel sessions.

Overall, 25 registered persons (from the Luxembourg public sector, ministries, institutions, and the university) as well as two translators attended the event. Information about the workshop was disseminated through different information channels (invitation letters, email postings, telephone calls, personal discussions). Invitations to the workshop were sent out to key persons in the area of public information, communication and translation at public agencies and other organizations.

Both the audience and the speakers had a very positive attitude towards the CEF.AT platform and have shown a deep interest in how Luxembourg public administration can provide language resources to the European Commission in order to further develop CEF.AT. The event was organised by Khalid Choukri (ELDA), Sviatlana Höhn and Eric Ras (Luxembourg Institute of Science and Technology), and Christoph Schommer (University of Luxembourg). On-site support, particularly technical aspects, catering, and registration matters, was performed by Siwen Guo, Isabelle Schroeder, and Dimitrios Kampas (all University of Luxembourg). Thanks to the audiovision.lu and aiic.net for the translation service (technical equipment; translation service).

The following chapters include the agenda of the event (Section 2), a list of registered participants (Section 3), and briefly inform about the content of each presentation and panel discussion (Sections 4 and 5).

Further information can be found at http://lr-coordination.eu/luxembourg.

2 Workshop Agenda

08:30 – 09:00 Registration

09:00 – 09:10 **Welcome**

Khalid Choukri (ELDA), Eric Ras (Luxembourg Institute of Science and Technology), Christoph Schommer (University of Luxembourg)

09:10 – 09:30 **Aims and Objectives**

Khalid Choukri (ELDA)

09:30 – 10:00 **Europe and Multilingualism**

Kimmo Rossi (European Commission)

10:00 – 10:30 **Multilingualism in the New Media**

Julia de Brees, Lucas Duane (University of Luxembourg)

10:30 – 11:00 Coffee Break

11:00 – 11:30 **Cultural Aspects of the Multilingual Legislation in the EU**

Rodolfos Maslias (European Parliament)

11:30 – 12:00 **How can public services benefit from CEF.AT?**

Kimmo Rossi (European Commission)

12:00 – 12:45 **Panel Session: Open Data – Best Practices on Open Data**

Moderator: Thibaud Latour (LIST); panellists: Slim Türki (LIST), M. T. Carrasco Benitez (European Commission), John Dann (State Ministry, Central Legislation Service), Yves Maurer (Luxembourg National Library)

12:45 – 14:00 Lunch Break

14:00 – 14:30 **Legal aspects of the preparation of public data**

Meritxell Fernandez Barrera (ELDA)

14:30 – 15:00 **Statistical and Hybrid Machine Translation**

Andreas Eisele (European Commission)

15:00 – 15:15 Coffee Break

15:15 – 15h45 **Data and Language Resources**

Khalid Choukri (ELDA)

15:45 – 16:30 **Panel Session: Multilingualism and Public Services in Luxembourg**

Moderator: Tetyana Karpenko (The Loupe); panellists: Gudrun Ziegler (MultiLEARN Institute & National Council for Foreigners), Andreas Eisele (European Commission)

16:30 – 16:45 Conclusions

Khalid Choukri (ELDA), Eric Ras (Luxembourg Institute of Science and Technology), Christoph Schommer (University of Luxembourg)

3 Workshop Registration List

1. *M. T. Carrasco Benitez*	European Commission
2. *Julia de Bres*	University of Luxembourg
3. *Khalid Choukri*	ELDA
4. *John Dann*	State Ministry, Central Legislation Service
5. *Sneja Dobrosavljevic*	Ministry of Health
6. *Lucas Duane*	University of Luxembourg
7. *Andreas Eisele*	European Commission
8. *Meritxell Fernández Barrera*	ELDA
9. *Thierry Gille*	Art Meta s.a.
10. *Peter Gilles*	University of Luxembourg
11. *Siwen Guo*	University of Luxembourg
12. *Sviatlana Höhn*	Luxembourg Institute for Science and Technology
13. *Dimitrios Kampas*	University of Luxembourg
14. *Tetyana Karpenko*	The Loupe
15. *Madeleine Kayser*	Luxembourg City Administration
16. *Thibaud Latour*	Luxembourg Institute of Science and Technology
17. *Rodolfo Maslias*	European Parliament
18. *Yves Maurer*	Luxembourg National Library
19. *Claude Oé*	City of Berdorf
20. *Eric Ras*	Luxembourg Institute of Science and Technology
21. *Saila Rinne*	European Commission
22. *Kimmo Rossi*	European Commission
23. *Christoph Schommer*	University of Luxembourg
24. *Slim Türki*	Luxembourg Institute of Science and Technology
25. *Gudrun Ziegler*	MultiLEARN institute, National Council for Foreigners

4 Summary of Content of Sessions

Session 1: Aims and Objectives

Khalid Choukri (ELDA)

Khalid Choukri started his talk by highlighting the multilingual nature of Europe and mentioned that, at present, the European Commission supports 24 official languages (plus more than 60 regional languages). “Languages are at the heart of our wealth, of our European cultures and identities” Mr. Choukri said. Furthermore, he explicitly mentioned that the European Commission actively supports multilingualism and that there exists no discrimination through the language only equal chances. “But which actions can be taken to manage the multilingual space?” he asked the audience, and suggested that a continuing translation should be the answer. However, translations by human beings afford highly skilled professionalism, the presence of domain knowledge, and an intensive educational process, for example through many courses of qualification. Moreover, the amount of texts increases heavily on a daily basis, which makes a human performance more than challenging. An automated translation is, therefore, the answer. In this regard, the new **CEF.AT** platform (CEF.AT standing for **C**onnecting **E**urope **F**acility – **A**utomated **T**ranslation) supports to take better account the needs of citizens, public services, and administrations. Mr Choukri demanded good data (« your data will help improve MT output quality, for yourselves and for CEF.AT ») and argued that “without good data, MT cannot work!”

Raised questions:

- Is Luxembourgish used in regulations in law? Is it considered as an official language?
- What about the Irish language?
- Do you consider including languages of non-Europeans in your program?

Session 2: Europe and Multilingualism

Kimmo Rossi (European Commission)

In his talk, Mr. Rossi hypothesized that language barriers affect the use of online services, and particularly, the enormous potential of the Digital Single Market. In this regard, he raised the question of the linguistic challenges of trans-European service networks. He said, the users of the networks have no common language, whether they are from public officials or citizens. 90% of the users dealing with online services want to be served in their native language. However, texts are large, which makes a human translation mostly inappropriate. Moreover, online translation services do not cover all languages and have security and data protection issues. The goal, therefore, should be to offer multilingual services that anyone can use in his/her mother tongue. A platform, served by a secure machine translation system and robust linguistic technologies, is needed. Therefore, the objectives must be to make public services available to all EU citizens, regardless of their native language and language skills. The benefits of the CEF.AT platform are: a) it is free of charge for public services of the member states, b) it is suitable for the translation of documents, messages, online content, c) it facilitates the communication and information exchange, d) it makes the trans-European public services available to citizens, businesses and Luxembourg officials, and opens the French services to other EU countries.

Session 3: Multilingualism in the New Media

Julia de Brees, Lucas Duane (University of Luxembourg)

The talk, which was given by Lucas Duane, concerned multilingualism in the new media. Lucas Duane raised the question of practices, ideologies, and language policies. Specifically, new media such as social networks allow highly flexible language practices. Such practices are characterized by a reduced syntax and morphology, non-standardized orthography, and typographical errors. Inflective forms appear here as well as emoticons, extended letter use, and acronyms. Regarding the ideologies, Mr. Duane mentioned that the new media language is not bad just because mostly non-standard forms and abbreviations are used. “It is more a particular genre, which is characterized by a higher degree of informality“, he said. There exists an emergence of new norms and conventions for the appropriate writing in digital media. Lucas Duane also pointed out that English is reflected as the international language of communication, even when communicating persons are native speakers of another shared language. The talk ended with a statement about the policies as most sociolinguistic research on new media focuses on such practices and ideologies.

Raised Questions:

- How do you want to scale your research?
- Do you think the “bad language” (with symbols or such) increases the semantic information which can be extracted from the new media language?
- Given the variation you have shown, did you see good examples of translations in the new media?

Session 4: Cultural Aspects of the Multilingual Legislation in the EU

Rodolfos Maslias (European Parliament)

Rodolfos Maslias stated at the beginning of the talk that the European Union is the largest union of states with a common body of legislation and that the EU legislation is transposed into national law and applied in the official language of each country. 28 countries are currently member states of the European Union, where 80% of the laws of the member states are based on European legal acts, he said. The legislation of the European Union is based on a highly complex cooperation between representatives of the member states, in particular, commissioners, ministers and parliamentarians. At the institutions, each of them represents the interests of their own country, and also the culture of their own people. Common cultural projects, such as the European Capitals of Culture, subsidy schemes for research projects, e.g. Horizon 2020, and the alteration of educational procedures, such as ‘Bologna’, are examples of efforts to create a common culture. At present, 24 languages are officially supported by the European Union. Translation accompanies every step in the legislative process: from the first draft, which originates with the Commission’s experts, through all the negotiations with national departments and with the Council and Parliament, to the ultimate legal act, on which a vote is taken in each national Parliament and which then becomes an ‘original law’. In implementing the same legal acts in all member states and in all the languages of the European Union, the most important point is that fundamental linguistic concepts should be understood in the same way everywhere. However, this can only be achieved by a shared and consistent terminology, for example with IATE, which is a concept-oriented database covering more than 100 fields.

Raised Questions:

- Do you think in general that machine translation by European parliament is going work for the public rural government?
- Do you consider an evaluation of other languages, for example Russian and Chinese?
- For the language development point of view, do you encounter problems like changed meanings of the terms?
- Is IATE accessible through an API to use it for other language related services?

Session 5: How can public services benefit from CEF.AT?

Kimmo Rossi (European Commission)

Mr. Rossi pointed out that the CEF.AT offers several advanced functionalities for multilingual translations and that it provides a secure and flexible platform for the EU's online services. At present, several online services are served by CEF.AT, he explained, for example SaferInternet (an online service to make the Web safer for children), ODR (an online service for resolving disputes between consumers and suppliers), and ESSI (a system of exchange of information on social security from 32 countries). Advantages of the CEF.AT platform are, among others, a fast and secure automatic translation engine and a 'taking into account' of the specificities of natural language.

Raised Questions:

- Do you have a large amount of data (more than one billion words or phrases), If so, are they accessible publicly?
- Where do you see the future with spoken text? Is translation relevant?

Session 6: Panel: Open Data

Moderator: Thibaud Latour (LIST)

Panel Members:

Slim Türki (LIST)

- M. T. Carrasco Benitez (European Commission)
- John Dann (State Ministry, Central Legislation Service)
- Yves Maurer (Luxembourg National Library)

The following questions have been prepared for the panelists:

Current situation

Which open data do you current provide to the public? In which language?

Do you use any dictionaries, terminologies, ontologies which can be useful for language translation?

What are best practices in your context to work with open data?

What legal framework do you follow with regard to open data? What licenses do you use?

What are current issues you face with regard to using/processing open data?

Sharing open content

Does your organisation produce data that you think as appropriate for CEF.AT?

What are the pre-requisites for data sharing with the EU Commission?

What risks or barriers do you foresee to share open data with the CEF.AT initiative?

Do you think that your organisation can share this data for the CEF.AT purposes?

Are there practical difficulties that could affect your contribution?

Benefits for you and next steps

What do you think are the benefits from your participation in CEF.AT?

Do you have any suggestions on how to move on to support the CEF-AT?

How can we best mobilise the public sector to provide open data?

The panelists contributed the following.

Yves Maurer pointed out that only metadata in French are available for now, also the data From Europeana project. Other data that may be made available are protected by copyright, such as newspaper data and web page data. Only copyrighted material has restrictions in publishing and use. Nevertheless the open data initiative has many objectives. Making the data quality higher and offering good applications based on the open data are examples of specific objectives in this area. Further Yves Maurer explained that the Luxembourgish media data are plurilingual meaning that each article in one newspaper issue might be written in one of the languages commonly used in Luxembourg (French, German or Luxembourgish, sometimes also English). Currently, Bing translation is used to facilitate translation tasks, however employees can envisage to use different MT engines. In addition, Yves Maurer emphasized the importance of languages and translation for the European community as opposed to the US where only English is used for communication, and maybe sometimes Spanish.

M.T. Carrasco Benitez expressed his worries that a huge “gray area” exists in the open data movement. For the translation it is important to have clean, aligned data, which is a hard problem, he said. In addition to data cleaning he also sees issues in data processing illustrating this by an example of accessing a single word by a translator vs. reading multiple terabytes by an MT system.

Slim Türki informed about the history of the Open Data initiative in Luxembourg, which is quite young. The first data types envisaged by the Luxembourgish Open Data initiative were transportation and traffic data, and not textual data.

John Dann sees legislation as one of the most useful data for machine translation. He works on transforming the Luxembourgish Journal Officiel into a structured database. The challenges that he sees in the open data include data descriptions (meta-data) and semantic aspects of the data. He named Euralex as one of the examples of resources that are even used for commercial applications. Being part of the European Legal Identifyer committee, he sees Open Linked Data as one of the priorities. Open Linked Data approaches would help to solve such problems as identification of translations for a given legal record. He expressed his concerns regarding publishing legal data that they are open to everybody, including other countries oversea, and no restrictions can be imposed.

Regarding the suitability of the data owned by the public organization, the following has been expressed in the panel. M. T. Carrasco suggested using rule-based translation for closely related language pairs instead of statistical machine translation. This may be useful in the

case of Luxembourgish-German translation. He characterizes the effort needed for statistical machine translation for the Luxembourgish-German language pair as too huge. John Dann mentioned that all EU directives are already translated in all EU languages. Further translations arise for non-official European languages, such as Catalan. He emphasized again the importance of linking textual information contained in different sites using semantic technology. Yves Maurer who also works for the National Computing Service mentioned that there are a lot of resources in Luxembourg that can be used for machine learning, for instance web sites that are translated or Guichet.lu.

With regard to staring a movement of data sharing in Luxembourg the panelists contributed the following. John Dann sees providing data as a budget issue emphasizing the quality requirements for official publications. Official legislation cannot allow publishing imperfect translations on their web pages because it is also of question of the reputation. Slim Türki referred to a new initiative which aims at boosting resource sharing in Luxembourg. He mentioned the experiences gathered within the Share-PSI project (Best Practices for Sharing Public Sector Information: https://www.w3.org/2013/share-psi/). In his opinion, the users are not interested in raw data, but in meaningful applications. Therefore metadata are also very important. M. T. Carrasco expressed his regrets that there is no interest in machine learning problems in Luxembourg supporting this by his experience in building a team for rule-based machine translation which did not succeed. Finally, Yves Maurer argued that admins who work with data directly do not see the real value of the data; they do not see the benefit of the multilingual information. The data needs to be used for applications beneficial for Luxembourgish citizens in order to be useful.

Session 7 : Legal aspects of the preparation of public data

Meritxell Fernandez Barrera (ELDA)

The talk was motived by the fact that public authorities produce large amounts of data, which become the raw material for new, innovative cross-border applications and services. As a consequence, and in view of the development of a single European market, some clear rules are needed, as Meritxell Fernandez Barrera pointed out; for example, a legal and technical interoperability, a clear policy concerning the re-use of data across the EU, and simple redress mechanisms. Regarding the usage of data by ELRC, Mrs. Fernandez Barrera mentioned that data is not to be re-used as such, but that it is used in order to produce new models to help automatically generate translations of new texts. The stages for releasing data are as follows: a) exclude confidential information, b) obtain prior informed consent, find a legal basis, anonymize or exclude personal data, c) ensure that there is no third party copyrights, that the material is in the public domain or that the necessary licenses have been obtained, d) follow the rules of the national public sector initiative transportation rules, for example the use of the national Open Government Licenses or the standard procedure for releasing the public sector initiative, and e) use a standard open government license or open-public license, or re-use a license. Follow the national or organizational public sector initiative re-use policy. The talk ended with the legal framework in Luxembourg.

Raised Questions:

- Concerning the article 4 of Luxembourg law. What does it mean to identify a copyright?
- If there is no license, data is considered as public. How is this handled in Luxembourg?
- What kind of license is most prominent?

Session 8: Statistical and Hybrid Machine Translation

Andreas Eisele (European Commission)

In this talk, Andreas Eisele presented the CEF.AT initiative, which is to build on the existing Machine Translation services (MT@EC) at the European Commission, and added that CET.AT puts emphasis on a secure, qualitative, and customizable Machine Translation platform for pan-European online services. “This platform is a multilingualism enabler,” he said as well as that “this platform is to serve, among others, public online services, public bodies in the EU member states, and European institutions”. He then presented the technical aspects: the statistical Machine Translation architecture, which included several hubs (language dispatcher, Machine Translation Engines), customized interfaces for users and services, and the usage of user feedback for data modeling. In particular, the question on how to build Machine Translation Engines (for CEF) was communicated as a playing together of technology providers, ‘engine factories’, and data providers. A third point concerned errors in the automatic translation process. Andreas Eisele reported that typical errors depend mainly on the target language: with morphologically simple target languages, the statistical models work reasonably well, he explained, whereas for strongly inflected target languages the word endings (suffices) are often wrong. Some frequent errors can be fixed with simple means, certain types of expressions can be treated with rules, and the normalization of the punctuation helps a lot. Finally, he explained that errors caused by different word order can be reduced. He ended his talk mentioning some ways to improve the translation quality and to better serve the needs of end-users (improve scalability, build models optimized for different use cases, enhanced coverage and robustness) and translators (domain adaptation, learning from streams of corrections, implementation of improvements through language weeks). Recent improvements where shown as well as next steps.

Raised Questions:

- What is the motivation for users to use translation systems, if there are that many errors?
- Concerning the ambiguity of terms, how do you see the quality of a translation?

Session 9: Data and Language Resources

Khalid Choukri (ELDA)

Khalid Choukri stated that the “learning from data” paradigm is a predominant approach and that automatic translation systems should learn from existing data, including documents and other linguistic data and/or various sources like the World Wide Web. He highlighted that an active participation of the public sector is essential. How valuable language resources can be produced from data? Khalid Choukri answered this question using some examples and specified that raw data should be first prepared, then cleaned. Legal issues such as intellectual property licensing should be clarified before that data can finally be shared. He concluded his talk by showing an example of bilingual data management and highlighted that only a minority of data (public web) is visible. “Only 4% of the web content (roughly 8 billion pages) is available, whereas “the Deep Web” (which includes password-protected medical documents, scientific reports, organizational repositories, intranets, etc.) represents approximately 96% of the digital universe.

Session 10: Panel: Multilingualism and Public Services in Luxembourg

Moderator: Tetyana Karpenko (The Loupe),

Panel Members:

- *Gudrun Ziegler (Ministry of Education & National Council for Foreigners)*
- *Andreas Eisele (European Commission)*

The goal of the panel was to bring together potential users of CEF.AT platform (representatives of public services in Luxembourg) and MT service providers with the purpose to discuss users' needs and expectations, and how these could be covered by the MT technology. Gudrun Ziegler (National Council for Foreigners, (CNE)) represented public services. In addition, the voice of the public administration was expressed through questionnaires that were created for the purpose of the panel and discussed during the panel in form of a survey. Andreas Eisele (Project Manager in Machine Translation at the European Commission) represented the technological side.

Andreas Eisele summarized the advantages of the CEF.AT platform for the participants who just arrived. He explained the difference between translation memory and machine translation. Gudrun Ziegler explained the difficulties that employees of public administration face while working with translations and translators. The quality of the translation is subjective and to her it is what people understand in the end, how the translated language is perceived by the target audience (migrant, refugees, etc.). The languages needed when working with migrants include Dari, Farsi and Arabic. Because of the direct relation to the refugee topics, the translation has a high political sensitivity. However, language experts not trained as translators, for instance teachers, perform the translation tasks. In addition, Gudrun Ziegler spoke about language policies at CNE which prescribe using one of the official languages during the meetings of CNE (French, German and Luxembourg). However, English and Portuguese are the two most heavily used languages in practice.

The survey based on a questionnaire was a small-scale study with 6 different participating organizations: Legilux, the National Library of Luxembourg, the City of Luxembourg, CNE, Ministry of Health and the City of Berdorf. The survey showed that there is a difference in translation demand and quality requirements within different public authorities in Luxembourg. Important criteria that we identified were hierarchy level, domain and communication type.

By hierarchy, more central authorities require higher quality criteria and need political decisions for translation. Normally, their documents are only in French because this is the reference version. In contrast, more peripheral institutions emphasize accessibility; translations are seen as help in solving communication tasks in their daily work, although the reference version is still in French. Such organizations frequently use free MT tools to facilitate the translation jobs. In-house translators are not employed, but translation work is done by people who have other duties but speak the target language.

By domain, we found a difference between the legal domain and the other domains. While only the French version of all documents usually exists in the legal domain, all other domains represented in the survey emphasize the importance of accessibility. Even low-quality

machine translation is then seen as helpful. Documents are translated by employees who normally have other duties and are not professional translators.

By communication type, documents directed to citizens and migrants (websites, flyers, information boards) are frequently translated. In addition, documents whose translation facilitates the daily work of the employees are translated (for internal purposes, e.g. drafts of some new regulations that need to be discussed first). As opposed to this, official publications of the law exist only in French.

Finally, organization representatives who deal with translations daily are willing to use the new CEF-AT platform and are willing to share the monolingual and multilingual documents that they have (websites, flyers, translation of notifications to public, letter templates etc.)

Session 11: Conclusions

Khalid Choukri (ELDA), Eric Ras (Luxembourg Institute of Science and Technology), Christoph Schommer (University of Luxembourg)

As a final conclusion, Khalid Choukri expressed his gratitude to the participants, to the keynote speakers, and the organization committee. He then summarized the day and presented the next direct steps to go for:

- Clear all aspects around availability of language data for CEF.AT
- Go on with the requirements of the public service administration in view of CEF.AT
- Continue working together on maintaining multilingual data, particularly, since language evolves over time.
- Support further multilingualism in Europe

Khalid Choukri also asked the participants to fill out the feedback form. The comments are thoroughly very satisfying; the given comments and remarks will be taken into account for the next workshop, which will be repeated in a second round by the end of 2017.

5 Workshop Presentation Material

All workshop materials are available online on the ELRC website: www.lr-coordination.eu/luxembourg