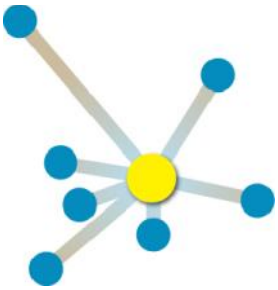
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1091 programme.



European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.11
Task 8

ELRC Workshop Report for Germany

Author(s): Andrea Lösch (DFKI)
Eileen Schnur (DFKI)

Dissemination Level: Public

Version No.: <V1.0>

Date: 2018-12-20

Contents

1 Executive Summary

The second ELRC workshop in Germany took place in the Representation of the European Commission in Berlin on 18th of October 2018. As illustrated through the participant feedback forms, the workshop was very well perceived by the participants with very good ratings (all between 4 and 5).

The workshop started with an official welcome by Prof. Dr. Stephan Busemann, Associate Head of the Language Technology Department at the German Research Centre for Artificial Intelligence (DFKI) and Ms. Gudrun Stock, Deputy Head of the Unit Accessibility, Multilingualism and Safer Internet at the DG CONNECT of the European Commission, who also provided the first presentation of the day about how to connect digital public services in Europe. Insights into open data initiatives in Germany were presented by Christian Horn, the Head of Business and Coordination Center GovData (the German Open Data portal). His presentation included both the legal frame for data sharing as well as corresponding practical activities in Germany. Due to the sudden and unforeseen unavailability of Dr. Georg Rehm, the panel "Multilingualism and open data in public services in Germany" was moderated ad hoc by Dr. Andrea Lösch, the ELRC project manager. It focused on two key questions: (i) the multilingual support required by public services in Germany and Europe and (ii) the challenges faced by public services when enabling multilinguality.

After the coffee break, Ms. Lisa Ribier, DGT Field Officer at the Representation of the European Commission in Berlin, provided a very practical and illustrative presentation about the CEF eTranslation platform. Following a short introduction into the European Language Resource coordination (ELRC) by Dr. Andrea Lösch, the two German National Anchor Points (NAPs) - Alexandra Soska (Public Services NAP and CEF eTranslation Representative) from the Federal Ministry of the Interior and Andreas Witt (Technology NAP), Professor at the University of Cologne – presented their work (both from an organizational and technological perspective) as well as corresponding achievements. This included also practical insights into the sharing of LR in the German public sector.

After the lunch break, Prof. Raue from the University of Trier started the afternoon session with an interesting presentation on the proprietary frame of sharing texts in Germany (focus: copyright law). Following his presentation, Thierry Declerck, Senior Consultant at the Department of Multilingual Technologies of DFKI, gave a practical illustration of how to share data using the ELRC-SHARE repository. In addition, Dr. Khalid Choukri, CEO of ELDA, provided the audience with hands-on hints and tips on managing language resources using a corresponding data management plan (DMP). The subsequent panel (with open discussion round) was moderated by the German Public Services NAP Alexandra Soska. The panelists included legal experts as well as public service representatives. It focused on the legal and organizational questions of sharing language resources in German public services. The final summary of the day and outlook to the future was presented jointly by Prof. Busemann and Dr. Andrea Lösch.

Overall, the workshop gave not only an important impulse for improving the processes and frame for sharing language resources in German public services, it was also well received by the audience with several new offers for future collaboration with the ELRC.

2 Workshop Agenda

Agenda of the 2nd German ELRC Workshop - 18. October 2018

Representation of the European Commission in Berlin

Unter den Linden 78, 10117 Berlin
Großer Konferenzsaal, 1st Floor

09:00 – 09:30 **Registration and Welcome Coffee**

09:30 – 09:50 **Welcome and Introduction**

Prof. Dr. Stephan Busemann, Associate Head of Language Technology Department, German Research Center for Artificial Intelligence (DFKI);

Gudrun Stock, Deputy Head of Unit Accessibility, Multilingualism and Safer Internet, DG CONNECT, European Commission

SESSION 1. OVERCOMING BARRIERS IN DIGITAL EUROPE: EUROPEAN CONTEXT AND LOCAL NEEDS

09:50 – 10:10 **Connecting public services across Europe: ambitions and results so far**

Gudrun Stock, Deputy Head of Unit Accessibility, Multilingualism and Safer Internet, DG CONNECT, European Commission

10:15 – 10:30 **Open Data Initiatives in Germany**

Christian Horn, Head of Business and Coordination Center GovData, Department for IT and Digitalisation, Senate Chancellery Hamburg

10:30 – 11:10 **Panel Session: Multilingualism and Open Data in Public Services in Germany – An Outlook into Current and Future Challenges**

Moderator: Dr. Georg Rehm, Senior Researcher Language Technology, German Research Center for Artificial Intelligence (DFKI)

Panelists:

- *Christian Horn, Head of Business and Coordination Center GovData, Department for IT and Digitalisation, Senate Chancellery Hamburg*
- *Gudrun Stock, Deputy Head of Unit Accessibility, Multilingualism and Safer Internet, DG CONNECT, European Commission*
- *Hartmut Wernich, Member of Business and Coordination Center Federal Information Management, Ministry of Finance, Saxony-Anhalt*

11.10 – 11:30	***Coffee Break***
11:30 – 11:50	**The CEF eTranslation Platform @work** *Lisa Ribier, DGT Field Officer, European Commission*
11:50 – 12:10	**The European Language Resource Coordination (ELRC) Action** *Dr. Andrea Lösch, ELRC Project Manager, German Research Center for Artificial Intelligence (DFKI)*
12:10 – 12:30	**ELRC in Germany: Previous Advances and Achievements** *Alexandra Soska, Translator and CEF Coordinator, Federal Ministry of the Interior, Building and Community (BMI)* *Prof. Dr. Andreas Witt, Professor for Digital Humanities, University of Cologne/ Institute for the German Language (IDS)*
12.30 – 13:30	***Lunch Break***

SESSION 2. ENGAGE: HANDS-ON DATA

13:30 – 14:00	**The Legal Framework of Text Provision – Problems and Proposed Solutions** *Prof. Dr. Benjamin Raue, Professor for Civil Law, Information Society and Intellectual Property, Trier University*
14:00 – 14:15	**Preparing and Sharing Data with the ELRC Repository** *Thierry Declerck, Senior Consultant, Multilinguale Technologien, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)*
14:15 – 14:30	**Data Management: Practical Hints and Tips** *Dr. Khalid Choukri, CEO, Evaluations and Language Resources Distribution Agency (ELDA)*
14:30 – 15:30	**Panel Session with subsequent open discussion: Intelligent Management of Language Resources in Public Administrations** *Moderator: Alexandra Soska, Translator and CEF Coordinator, Federal Ministry of the Interior, Building and Community (BMI)* *Panelists:* • *Anja Hein, Terminologist, Federal Foreign Office (AA)* • *Prof. Dr. Benjamin Raue, Professor for Civil Law, Information Society and Intellectual Property, Trier University* • *Damir Čuljat, Translation Technologist, The Parliament of the Federal Republic of Germany*

- *Patricia Gerecht-Thomsen, Translator, Federal Ministry of Transport and Digital Infrastructure (BMVI)*
- *Dr. Pawel Kamocki, Legal Expert, Evaluations and Language Resources Distribution Agency (ELDA)*

15:30 – 16:00 **Summary and Outlook**

Dr. Andrea Lösch, ELRC Project Manager, German Research Center for Artificial Intelligence (DFKI);

Prof. Dr. Stephan Busemann, Associate Head of Language Technology Department, German Research Center for Artificial Intelligence (DFKI)

16:00 – 16:30 ***Coffee Break and Networking***

3 Summary of Content of Sessions

3.1 Welcome and introduction

The workshop was opened by Prof. Dr. Stephan Busemann, Associate Head of the Language Technology Department at the German Research Centre for Artificial Intelligence (DFKI) showing that the demand for translation and translation support has significantly increased in the past 3 years. The question was raised how we – and translators – can deal with this ever increasing demand. He pointed out that the solution can only lie in combining human expertise with technological support (underlining again that even with AI, it is not possible to replace human translation skills). Before introducing the workshop agenda, he explained again the objectives of the ELRC workshop which are both illustrative and educational, i.e.

- To illustrate how language technology can help with translation in a digitally connected multilingual Europe
- To share experiences and needs of a modern multilingual public administration with regard to multilingual communication
- To jointly identify relevant sources of multi-lingual language resources that can help adapting CEF eTranslation to the needs of different public services
- To address any legal or technical questions with regard to the sharing of language resources
- To define what is needed in the future to achieve a significant increase in translation efficiency and quality

3.2 Welcome by the EC

The workshop participants were also greeted by Ms. Gudrun Stock, Deputy Head of the Unit Accessibility, Multilingualism and Safer Internet at the DG CONNECT of the European Commission. She pointed out that it is the goal of the European Commission that European public services and administrations are able to offer their services in all EU official languages and that this is the reason for founding ELRC. This goal was officially underlined in the declaration of Tallinn from October 2017 and thanks to the progress made in machine translation (MT), it looks as if this goal can be realized. As such, it is more important than ever to gather relevant language resources for public services and administrations in Europe, in order to make eTranslation work well in their particular context. Ms. Stock thanked all participants for their efforts relating to the collection of language resources and wished everyone a successful workshop.

Directly following the welcome by Prof. Busemann and Ms. Stock, a question was raised about the difference between MT@EC and eTranslation (in particular: the difference between statistical machine translation (SMT) and neural machine translation (NMT)). Prof. Busemann explained that even today, it is not entirely clear what exactly happens when a NMT system translates a text into another language. The difference between SMT and NMT lays in the approach. While SMT focusses on the statistical occurrence of e.g. particular words, word forms etc., NMT is able to set parts of the text in connection with the other parts of the texts (they are all part of the neural network). And especially for German, the output of this approach proves to be better than anything that has been offered so far.

3.3 Connecting public services across Europe: ambition and results so far

In her subsequent presentation on the ambitions and results of connecting public services across Europe, Ms. Stock introduced the Connecting Europe Facility (CEF) and explained in detail CEF Automated Translation (CEF AT) as a building block (including CEF AT uptake). Within CEF, the term Building Block refers to a package of technical specifications, services or sample software that can be reused in different policy domains/by different Digital Service Infrastructures (DSIs). Examples of DSIs are Europeana, ODR, Open Data, eHealth, eProcurement, the eJustice portal etc. Each of the DSIs is defined through cross-border use, contribution to EU policies, delivery of services by digital means and additional principles. The Building Blocks (of which CEF Automated Translation is one) help to deliver these services to Europe's citizens in the best possible way.

In her presentation, she explained that the CEF AT building block today includes the eTranslation service – a central machine translation service run by DGT – and the ELRC-SHARE repository which stores all language resources needed to develop multilingual systems and services. Most interestingly, the request for translation with eTranslation has significantly increased: While in 2017, the total number of translated pages was 19 million, only in Q1-Q3 2018, the number had increased to 29 million pages. In fact, several DSIs today are already re-using eTranslation, e.g. ODR, the eJustice portal or the EU Open Data portal, allowing them to provide information to their customers in the various EU official languages. Ms. Stock explained the different ways of using eTranslation and how to obtain financial support for participating in eTranslation integration projects or language resources projects.

She closed her presentation with an outlook to multilingualism in the next Multi-annual Financial Framework. It was shown that Digital will be present in four different parts of the MFF: In Digital Europe – Capacities and roll out (planned with 9.2 billion EUR), in CEF – Digital Connectivity (planned with 3 billion EUR), in Creative Europe – Media (planned with 1.1 billion EUR) and last but not least in Horizon Europe R&D&I. Multilingualism will play an important role both in Horizon Europe as well as in the Digital Europe Programme. While the former will cover research and innovation activities (i) for achieving digital language transparency and (ii) for preserving languages, the Digital Europe Programme focusses on capacity building and deployment of language technologies (LT), including support to the deployment of LT as well as corresponding support to the modernization of public administrations and areas of public interest.

3.4 Open data initiatives in Germany

Christian Horn, the Head of Business and Coordination Center GovData at the Department for IT and Digitalisation at the Senate Chancellery Hamburg gave an introduction to the open data initiative GovData in Germany. GovData is the open data portal for Germany. It serves the goal of open government – making data from administrations and public services transparent, open and re-usable. GovData is an application of the IT Planning Council of Germany and was developed in cooperation with both the Federal Government (Bund) and the federal states (Länder) of Germany. Mr. Horn stressed that there are several important reasons for open data, including in particular:

- The transparency of activities of public services and administrations
- The publication for the purpose of self-interest

- The great economic value of such data (Note: An investigation by the Konrad-Adenauer-Foundation in 2015 showed that in Germany, the economic value of such data would amount to 43 million EUR per annum if calculated in a conservative way – up to a maximum of 136 million EUR!)
- The importance of open data for new developments in AI (like e.g. Smart City)

There are several laws and regulations relevant for open data in Germany, including above all the PSI Directive, but also the Open Data law of the Federal Government and different laws on the level of the federal states (respective open data laws, transparency laws etc.). Mr. Horn gave a comprehensive illustration of different data sets that can and should be shared as open data. He pointed out that currently, 11 federal states participate in the sharing of data within GovData – and that GovData also forwards all relevant data sets to the EU Open Data portal. He stressed again the importance of the legal framework, explaining that the Open Data law of the Federal Government resulted and will continue to result in an increase of data sets from players within the Federal Administration. Last but not least, he emphasized the importance particularly of language resources as part of open data and for training eTranslation: He illustrated how these have been used to train eTranslation and to automatically translate e.g. meta-information within the EU Open Data portal.

Following Mr. Horn's presentation, several questions were raised by the audience.

- The first question was whether it was necessary to demonstrate the different uses if open data are used. Mr. Horn pointed out that this is clearly not the case. On the contrary, it may even be prohibited to trace and track who used open data in which way.
- Another question concerned the actual openness of data. For instance, whether or not it would be possible to share data that is provided through the Bavarian Open Data portal with GovData. The answer provided by Mr. Horn was actually a financial one. Since Barvaria had chosen to go ahead with their own solution and not to get involved in GovData, they may currently not be able to share all data (as they were financed by the federal state as opposed to the Federal Government). However, as part of the new order for the financial relations between Federal Government and federal states, there is an agreement which states that from 2020, all data must be delivered to Gov Data.
- A third question was about the data format that should be used to share data with GovData. Mr. Horn pointed out that above all, any machine-readable format is acceptable. Even formats such as xls or pdf can be accepted, even though they are of limited value as the reusability of such data sets is limited.
- Following a fourth question from the audience, Mr. Horn explained that there is an agreement between GovData and the EU Open Data portal to provide any data sets from Germany that can be shared. This agreement also states that in Germany, only GovData is able to provide data to the EU ODP (and not e.g. federal state based open data initiatives).
- Another question was who from the different public services was actually responsible for providing GovData with data and whether this would be done by the IT Department. Mr. Horn explained that in every federal ministry, there is one person responsible for open data. In addition, there is a new consultation agency within the Federal Authority for Administration

(Bundesverwaltungsamt)[1] that provides corresponding advice. Last but not least, there is a corresponding guide available online[2]. As regards support for the evaluation and processing of potentially open data, there is no support on federal level. With regard to language resources, there is, however, the ELRC Helpdesk and on-site assistance.

3.5 Panel: Multilingualism and open data in public services in Germany – an outlook into current and future challenges

Due to the sudden and unforeseen unavailability of Dr. Georg Rehm, the panel was moderated ad hoc by Dr. Andrea Lösch. The panelists were:

- Mr. Christian Horn, Head of Business and Coordination Center GovData at the Department for IT and Digitalisation at the Senate Chancellery Hamburg
- Gudrun Stock, Deputy Head of Unit Accessibility, Multilingualism and Safer Internet at DG CONNECT at the European Commission
- Hartmut Wernich, Member of Business and Coordination Center Federal Information Management at the Ministry of Finance in Saxony-Anhalt

As a first question, participants were asked which public services, in their opinion, required multilingual support. From a European Commission's perspective, Ms. Stock pointed out that the main goal is to allow DSIs to operate in a multilingual manner. However, she stressed that eTranslation is available for all public services in Europe, no matter on which levels they are situated. If they need multi-lingual support, they can gain access to eTranslation. On the level of Germany, Mr. Wernich explained that he is Member of the Business and Coordination Center for federal Information Management. As such, he is working on the so-called "Leistungskatalog" (LeiKa) – a catalogue that lists all requirements and services of public services in Germany (including their costs, processes, texts produced etc.). He explained that according to the Online Access Law ("Onlinezugangsgesetz" – OZG), all services that can be provided online must be enabled to be online by 2022. This also includes the requirement to be multi-lingual in many cases. For instance, data that is available in data portals of the different federal states are already multi-lingual. About a year ago, Mr. Wernich hence tested eTranslation in order to find out whether it is possible to usefully support multilinguality in Germany with this tool. He pointed out that some translations had to be post-edited – and that new tests will need to be made soon with the improved system.

The second question addressed the major challenges with regard to multilinguality of public services. Ms. Stock explained that right now, for eTranslation to really work in all different domains, the corresponding language resources are required. Furthermore, corresponding language data for non-EU languages (e.g. Turkish) is missing and it is very difficult to generate such data. Nonetheless, there is an increasing demand also for such translations. Mr. Wernich pointed out that from a German point of view, there is a great need for translations from/to English and from/to French – depending on the region also for Russian and Polish. Right now, a few data portals are actually

[1] Please visit: https://www.bva.bund.de/DE/Services/Behoerden/Beratung/Beratungszentrum/_documents/artikel_zentrale_stelle_open_data.html

[2] https://www.verwaltung-innovativ.de/SharedDocs/Publikationen/eGovernment/open_data_handbuch.pdf?__blob=publicationFile&v=2

already offering multilingual information, but in most cases, the resources (both language resources and human resources) are missing to enable multilinguality. With regard to open data, Mr. Horn confirmed that multi-lingual contents are indeed frequently offered, but support to multilinguality is still not widely and systematically used as illustrated in the case of the Brandenburgish Open Data portal. Following a corresponding report of the IT Planning Council of Germany, it becomes clear that there are several public services that would greatly benefit from eTranslation. These are in particular services that are largely used by non-native speakers, i.e. Foreigners' Authority (Ausländerbehörde), citizens' office (Bürgeramt), registry (Standesamt), citizens' phone (Bürgertelefon 115), or central information.

3.6 The CEF eTranslation platform @work

Lisa Ribier, DGT Field Officer at the Representation of the European Commission in Berlin, started her presentation about the CEF eTranslation platform with explaining the distinction between MT@EC (the statistical MT translation service of DGT), eTranslation (the cloud-based, neural translation service provided by DGT) and CEF AT (the translation platform that in addition to eTranslation will provide several other multilingual tools and services, e.g. transliteration, named entities recognition etc.). She highlighted that eTranslation can be used by public services and administrations in Europe and explained how they could access the service. As a translator, she pointed out that eTranslation brings several advantages, in particular time savings and increased productivity, reduction of costs, preservation of privacy and ease of exchanging information. She admitted that in order to translate her slides from English into German, she had actually used eTranslation with very good results (which is also due to the fact that eTranslation is trained with and on EU languages). She also illustrated how eTranslation was used successfully within the N-Lex-portal (a legal application).

Ms. Ribier also explained the differences between neural and statistical machine translation and illustrated these differences with the help of typical mistakes made in translation by the two systems. It was shown that neural MT (NMT) is able to consider the context better than machine translation (which is simply based on the statistical occurrence of words), and that as a result, the translations are more fluent and grammatically less prone to errors than statistical MTs. This is particularly true for morphologically rich languages like German, Hungarian etc.

Ms. Ribier also illustrated the importance of language resources for the quality of translations, showing that the translation of an EU legal act currently works better than the translation of cultural articles simply because the system is trained with EU data. She concluded her presentation with a quote by Arle Lommel "Machine translation will displace only those humans who translate like machines" – again underlining the importance of the human factor. As Prof. Busemann had pointed out in the introduction, the best results are not achieved by machines or by humans – the best results can only be achieved by a team of humans assisted by a machine. As such, human translators are and will not be replaceable by MT. However, MT can help humans to transform better and to overcome existing barriers in terms of time and resources.

3.7 The European Language Resource Coordination (ELRC) action

This session was presented by Dr. Andrea Lösch, ELRC project manager at the German Research Centre for Artificial Intelligence. Following a brief introduction of the heads and organisations behind ELRC and of the ELRC objectives, Ms. Lösch provided an overview of public services in Germany in

need of MT support. Many of the organisations were represented by the attending workshop participants. Again, the importance of language resources (LR) for achieving high quality MT output was illustrated, showing that only with the right training data, eTranslation will be able to deliver high-quality translations. The presentation closed with the achievements of ELRC, showing that by September 2018, the project had managed to collect 450 LR (i.e. an increase of 250 LR within just one year). It was stressed that most of the data collected within ELRC was actually re-usable, and as such, available to the community at large.

3.8 ELRC in Germany: Previous advances and achievements

This presentation was held jointly by the German National Anchor Point Duo Alexandra Soska (Public Services NAP and CEF eTranslation Representative) from the Federal Ministry of the Interior, Building and Community and Andreas Witt (Technology NAP), Professor at the University of Cologne. Alexandra Soska started the presentation by illustrating the work of a National Anchor Point in Germany (including the different activities and meetings to contribute). She also illustrated why it is important to share language resources – and raised all the questions concerning the sharing of LR that typically concern participants from public services and administrations. Key questions include above all, why is it important to exchange LR, am I actually allowed to share my LR, who can actually decide whether I can share my LR or not, who is responsible for implementing the sharing of my LR and last but not least, what do I technically need to do to share my LR?

She pointed out that actually in Germany, the situation is favourable for sharing LR because of the well-organised network of language services in the public sector (which means that the main creators of LR are already in close collaboration. In addition, the use of technology (in particular CAT tools) is advantageous, as it makes it relatively easy to locate the LR. The difficulty, however, lays in the aforementioned legal and technical questions of sharing LR. Until now, the main contributors to the ELRC-SHARE in Germany are the Federal Foreign Office, the Federal Ministry of the Interior, Building and Community and the Federal Ministry of Transport and Digital Infrastructure.

Prof. Witt then followed with a detailed explanation of the approaches to MT, including rule-based approaches, statistical approaches and neural approaches. He explained that both SMT and NMT are approaches which require corresponding text corpora to learn how to translate and that the creation and alignment of such corpora is a resource-intensive job. As such, tmx files are the most valuable resources for training current MT systems. Similar to Ms. Ribier, he explained that the strength and value of the neural approach to MT lies in the language model and in its ability to know/learn the relations between words, phrases and even letters.

3.9 The legal framework of text provision – problems and proposed solutions

The presentation of Prof. Raue from University of Trier focused on the proprietary frame of sharing texts in Germany. Prof. Raue started his presentation with a digression to the structure of the German copyright law. He briefly explained the difference between copyright law and other related protective laws (e.g. for photographs, for audio tapes, for movies, and also for databases). Rights of texts mainly derive from copyright (only in rare cases from data base right). In order to be protected by Germany copyright law, a text must be a personal intellectual creation of the author of particular length. Individual combinations of words are protected only in special cases. Texts can also be protected through the sui generis data base right, but only if there was a significant investment in the collection and creation of the particular texts or a significant change in the data base. It is

important to note that the individual contents of the data base are not protected against a transfer of significant number of contents, but only the data base as a whole.

To be able to use or re-use texts (such as translations), one either needs to have the permission of the copyright holder or there must be a corresponding permission in the law ("Erlaubnistatbestand"). He pointed out that governmental works are typically considered as public domain and as such subject to permission of publication (provided that there are no other legal limitations counteracting a publication). Prof. Raue finished his presentation with an overview of cc-licenses as they allow the re-use of texts within pre-specified conditions. He pointed out that some cc licenses include many rules and that if only one rule is not respected, the right for re-using the particular text is no longer granted. As such, the use of a cc-0-license with the demand of respecting individual special rules or alternatively the negotiation of individual licenses seems advisable.

3.10 Preparing and sharing data with the ELRC repository

This session was presented by Thierry Declerck from DFKI. Mr. Declerck started with an explanation of the notion of data, pointing out that the use of data mainly refers to numerical data, e.g. national statistics, geospatial data, maps and postal codes, public spending, crime statistics, educational data and similar. Translations, he explained, can also be considered as data (hence: language data or language resources). Most importantly, data is described by using metadata (i.e. data about the data such as title, publisher, description of content, URL etc.)

In the context of eTranslation, data typically refers to translated texts and translation memories. Typically, public organisations hold a great variety of such translated texts (language resources), e.g. reports, communication, news, web content, policies, terminologies, archives, FAQs etc. Several examples of data and data formats useful for eTranslation were provided. The preferred domains and topics relevant within the ELRC initiative were listed (namely: ODR - consumer's rights, EESSI - social security, eProcurement - public procurement, eJustice – justice/law, eHealth - health/medicine, BRIS - business, Europeana – culture, Safer Internet and Public Open Data).

Participants were also informed about how to contribute data to CEF eTranslation (including how to access the ELRC-SHARE, how to upload the data etc.). Mr. Declerck explained that if the data does not exist in a format that is useful for MT training (such as tmx or tbx), the ELRC team will process the data, clean it, align it, convert it into the right format and complete the metadata. Moreover, the ELRC team is ready to visit the data donor's institution for a consultation meeting and technical or legal on-site support whenever this is needed, e.g. to assess the usability of data, to support the identification of relevant data sets or the cleaning of data (anonymisation). Corresponding contact details were provided.

3.11 Data management: Practical hints and tips

In this session, Dr. Khalid Choukri, CEO of the European Language Resources Distribution Agency (ELDA) provided practical hints and tips on managing language resources using a corresponding data management plan (DMP). In order to draw up a good DMP, it is important to anticipate and consider all potential legal issues (such as IPR, the ownership of a resource, the removal of any personal information, the consideration of confidentiality). A fundamental part of developing a DMP is also to think about the repurposing of data and to ensure that data is in the right format and is described

with the right metadata. Fundamentally, all data created by public institutions should be considered as data that can be published and shared within the Public Sector Information Directive (PSI).

Dr. Choukri provided typical questions from public sector institutions about managing and sharing their LR – and the corresponding answers to them:

- Question: If a public agency outsources a translation, who owns the copyright of the translated version? Can the translation be shared? – Answer: This depends on what the outsourcing contract establishes with regard to IPR. Public agencies should make sure they keep the right to freely reuse and share translation memories. It also depends on national legislation (in some Member States, this may be regulated by law).
- Question: We have data but we do not have the resources to identify relevant data and to process them. – Answer: ELRC offers language processing services to public administrations (data conversion, tag removal, re-formatting, cleaning, alignment, metadata validation, etc.)
- Question: What do we do with a dataset which contains personal data? – Answer: ELRC can assist in identifying compliance with GDPR. Moreover, ELRC also offers anonymization services.
- Question: What happens if you outsource (some) of your data production (e.g. you outsource a translation to an external translator)? – Answer: Make sure that you keep the right to freely use and re-use the translation and to share it with third parties (Note : Translations are protected by copyright as already presented by Prof. Raue, so if there is no copyright transfer from the translator, you may not be able to reuse the translations). The outcome of the contract with the external translator should be that the translated documents are in your preferred format and that you also obtain translation memories in a reusable format (e.g. tmx).

Dr. Choukri also explained the scope of application of the sui generis database right before focusing on the copyright. He illustrated the differences in German copyright and the copyright law in other countries (e.g. Ireland, France). In contrast to the German copyright, the French copyright for works of public servants belongs ab initio to the State. He also pointed out that computer-generated works (such as machine translations) are not protected by copyright (as they are not their authors' own intellectual creations). Only if there is an original human contribution to the translation, computer-assisted creations are protected by copyright.

3.12 Panel session: Intelligent management of LR in public administrations

This panel session was moderated by Alexandra Soska, ELRC's German Public Services NAP from the Federal Ministry of the Interior, Building and Community (BMI). Her panelists included three representatives of federal public authorities (namely: Ms. Anja Hein – Terminologist at the Federal Foreign Office (AA), Mr. Damir Čuljat – Translation Technologist at The Parliament of the Federal Republic of Germany and Ms. Patricia Gerecht-Thomsen – Translator at the Federal Ministry of Transport and Digital Infrastructure (BMVI). Also, two legal experts (namely: Prof. Dr. Benjamin Raue from Trier University and Dr. Pawel Kamocki – Legal Expert at the Evaluations and Language Resources Distribution Agency (ELDA)) participated as panelists.

The first question addressed one key issue of the copyright in Germany, namely the fact that in contrast to countries like e.g. France, the copyright by default belongs to the author of a particular text. As such, the question was who could actually decide whether a particular translation could be

published – the author or the employer. Prof. Raue pointed out that typically, as part of the civil service law and the contract covering the civil service, the texts are created to fulfill the tasks of the particular public service or authority, and as such, the right for publication lies with the employer. This applies at least to all texts published since 2013. However, he also explained that someone who interprets the copyright law in a more conservative way might argue otherwise. As such, it is always advisable to ask the employer for permission when publishing or sharing language resources.

Another discussion point was how to deal with privacy and how to protect personal data when publishing or sharing a translation. It was pointed out by Dr. Kamocki that any publication would need to respect the GDPR. As such, if a data set has been anonymized, it is not considered to contain personal data anymore and hence does not fall into the scope of the GDPR anymore. However, Dr. Kamocki also pointed out that unfortunately, the legal definition of personal data is extremely broad. In fact, it is any information relating to an identified or identifiable natural person. As such, the key question here to identify personal data in a text is: What are the means reasonably likely to be used to identify a person. 15 years ago, this question was easier to answer than today, since there are now many different social media channels like Facebook and Twitter and a much higher volume of readily available personal data. As such, anonymization is a complex task which is never perfect. However, the law does not require perfection either, but just an indication that all reasonable care was taken. Alternatively, if it seems impossible to go ahead with anonymization, Dr. Kamocki pointed out that it is also always possible to publish or use the text that contains personal information if (i) there is, according to the GDPR, a legitimate interest to publish this data set and (ii) the data subject mentioned agrees with the publication. Prof. Raue pointed out that indeed, laws like copyright or the GDPR are not meant to counteract information transfer rights such as the Freedom of Information Act. As such, there must always be a detailed consideration of the entire context of rights. This statement was also supported by Dr. Kamocki who explained that one of his favourite decisions in German copyright law was the Germania 3 case by the constitutional tribunal in which it was said that copyright exceptions should be interpreted in a way that it does not interfere with constitutional freedom. Prof. Raue further explained that the European Court of Justice increasingly draws on European fundamental rights, too in order to fairly judge and balance conflicts of interest between individual authors and the general public.

From a practical point of view, the representatives of the public authorities agreed that it was very important for public authorities to have a legal contact point who could advise them whether or not a particular translation could be published. At least the Federal Ministries do have a legal counsellor.

At this point, Ms. Soska raised a very interesting question: Suppose that we all are willing and ready to share our data, what could be done in the future to make data sharing easier? What could be changed in the workflows of contributing organisations and public authorities to facilitate the sharing of language resources? It turned out that the ways to facilitate data sharing actually vary depending on the organization and the particular tools used to support the translation process. Mr. Čuljat pointed out that with the CAT-software used by the Parliament of the Federal Republic of Germany, it is not possible to access particular data sets e.g. on sentence level like in the AA of BMVI. The translation is only accessible on the level of the whole document – which makes it more difficult to filter out particular pieces of information. He pointed out that in order to regularly transfer larger amounts of data to ELRC, significant changes in both workflow and system would need to be made and that it is currently questionable whether these investments are justified in view of the perceived benefit of sharing data. Ms. Hein pointed out that within the AA (note: the AA

has the largest translation service of all German federal authorities – apart from the Bundessprachenamt), the maintenance of TM data is already implemented. This means that after the completion of a translation process, all metadata are again checked for their correctness. When doing this, the AA also removes all text segments that should not be kept as part of the TM (e.g. names, phone numbers etc.). In the BMVI, such maintenance is unfortunately not possible due to resource-related constraints. In this case, it would be feasible to simply include an additional meta-data field "can be shared" or similar that is to be completed by the particular translator in order to identify later whether or not a translation can be shared. As it became apparent from earlier discussions, it might also be feasible to already include such information as part of the order management system.

Prof. Raue pointed out that probably, in additional to a proper data management as was just discussed, it would also be necessary to introduce a corresponding rights management in the different organisations. This would include in particular that the employer ensures that all necessary rights for publication have been granted and/or that the translator receives all tools and relevant information to go ahead with the publication. This could be regulated at the internal level of each organization, at least for texts that have been created and translated internally. For texts that come from external organisations (e.g. the Federal Government), which are then translated by the translation service (e.g. of the AA), it would need to be clarified with the right holder (i.e. external organsation) whether or not this text can be published. It was pointed out that such rights management would also require time and personnel / human resources and that it would be best to create a dedicated position for these tasks. In any case, it is of utmost importance to define in each organization who is authorized to decide (and hence indicate in the metadata) that a text can be published or shared.

Following questions from the audience it also became apparent that if a translation service outsources translations to an external language service provider, at least in the case of the AA all usage rights are currently obtained from the external LSP. However, this unfortunately only covers the actual translation – not the TMs. These are not obtained. Following this discussion, the question was raised what "usage rights" actually means and whether a translation can be automatically used for training a MT system if someone has already obtained the rights for this translation.

It was explained by Prof. Raue that copyright does not cover the use of data for MT training. While it is relevant that a text e.g. is saved as part of a translation memory (as opposed to it being saved as original translation on the buyer's organization), it is not relevant whether this text will then be analyzed by a computer programme to train a MT system. As such, it would be of utmost importance to obtain the rights to have the translation as part of a TM / to directly obtain them in form of a TM, it the data should be usable for MT training.

3.13 Summary and outlook

This session was jointly presented by Prof. Busemann and Dr. Andrea Lösch (both from the German Research Centre for Artificial Intelligence). Prof. Busemann briefly summarized the major topics of the day (CEF, eTranslation, ELRC, strategies for sharing data / open data, the PSI and a sketch of the national legal frame for public sector data sharing, and practical issues of managing and sharing LR within the public sector). It was concluded that CEF eTranslation can act as multilingualism enabler for public online services and that in fact, most of the texts generated by the public sector are public open data. For the assessment of data as well as any other legal or technical issues relating to the LR

produced by the German administrations, the ELRC Helpdesk will provide direct support free of charge (including even the possibility of ELRC experts directly visiting the requesting organization on-site to provide necessary assistance and consultation). It was explained again that sharing LR is of benefit – not only because the data can be used to train MT (or any other applications), but also because providers of data have access to new LR, too.

4 Synthesis of Workshop Discussions

This section provides a synthesis of the different panel sessions as well as of the numerous discussions outside these sessions (e.g. as part of corresponding presentations). As such, the main issues, questions and concerns as well as corresponding answers and possible solutions (where possible) are illustrated.

4.1 ELRC and Open language Data in Germany

The open data portal for Germany is called GovData (www.govdata.de). GovData offers uniform, central access to all administrative data from the Federal Government (Bund), the federal states (Länder) and the municipalities. It is the declared goal of GovData to support the re-use of data from the public sector by supporting the use of open licenses and increasing the availability of machine-readable raw data. There are several laws and regulations relevant for supporting open government and open data in Germany, above all the PSI Directive. On the national level, the following laws are of key interest:

- The eGovernment Act ("eGovernment Gesetz": http://www.gesetze-im-internet.de/englisch_egovg/index.html): The goal is to facilitate the electronic communication within and with the administration by overcoming existing barriers on federal level. The law should enable the Federal Government, the federal states and the municipalities to provide electronic services that are easy-to-use, efficient, and easy-to-access.
- The Freedom of Information Act ("Informationsfreiheitsgesetz": http://www.gesetze-im-internet.de/englisch_ifg/index.html): The law grants each person an unconditional right to access official federal information. No legal, commercial, or any other kind of justification is necessary.
- The Law for Improving the Online Access to Public Services ("Onlinezugangsgesetz" – OZG: http://www.gesetze-im-internet.de/ozg/): According to this law, all public services on the level of the Federal Government and on the level of the federal states (including municipalities) are obliged to offer their services in an electronic way through the corresponding public service portal. The public service portals must be connected within an overarching portal group („Portalverbund"). As illustrated by Mr. Horn and Mr. Wernich, the kind of services that need to be provided online are defined in the so-called Services Catalogue („Leistungskatalog" – LeiKa) which has been established by the IT Planning Council of Germany (i.e. in collaboration with the Federal Government and the federal states). The transition to electronic services provision needs to be finalized by the end of 2022 at the latest.

As such, in Germany, the way is paved for the provision of public online services. As regards the data to be made available by public services and authorities, language resources have not played a major role in GovData yet (only exception: terminological resources from terminology data bases). Data sets available through GovData mainly include structural data / geo data, financial/budgetary data, form data, performance data, survey data and statistical data.

Concerning the question of how data from public services can be shared, it was illustrated by Mr. Horn and also Prof. Raue that the wealth of possible terms of use is often overwhelming for data donors and data users. In addition, there is the fundamental question whether this large number of diverse terms of use is actually even applicable to and right for public service data. Hence, the project "Open Government" pursued the goal to foster the use of few simple and uniform conditions of use: In collaboration with the Federal Government, the federal states and municipalities, the so-called Data License Germany 2.0 ("Datenlizenz Deutschland 2.0":) was developed specifically to cover the use and re-use of data from public services in Germany. The license exists in two versions:

- Version "Namensnennung" ("mentioning of name": https://www.govdata.de/dl-de/by-2-0) makes it necessary for data users to name the provider of the data set.
- Version "Zero" (https://www.govdata.de/dl-de/zero-2-0) allows for unrestricted re-use of the data.

This provides a sound and easy to understand legal basis for the sharing of language resources in the public sector. As pointed out by Mr. Horn, it is important to first check whether or not public services data may be restricted by other legal constraints, in particular:

- Privacy constraints (personal information included in the data set)
- Copyright constraints (author's rights)
- Confidentiality constraints (confidential information is present in the data set)

From a technical perspective, in June 2018, the IT Planning Council of Germany announced the DCAT-AP standard[3] as the basis of the uniform and free exchange of public data in Germany. This standard is compatible with the EU standard as it also allows exchange e.g. with the EU Open Data portal.

4.2 Success stories and lessons learnt

As became evident especially following Mr. Horn's presentation on open data and Dr. Choukri's presentation on data management, most public services and administrations have not yet been even aware of (i) the fact that public data can be shared and (ii) their data actually has a value that goes beyond the individual translation.

Regarding the fact that public data can be shared, it was important for the workshop participants to see that there are actually only a few restrictions that would prevent them from sharing their translations (namely: privacy constraints, copyright constraints and confidentiality constraints). It was concluded that actually most of the data that is produced every day could be shared, provided that it was anonymized. This was a realization that was very positive. The key question and remaining problem to solve is how to enable the anonymization of language resources. As pointed out by the Federal Foreign Office, at least the meta-data would need to be anonymized in all cases. Thierry Declerck from DFKI offered to provide a corresponding tool/script through which meta-data could be anonymized automatically and through which even obvious mistakes, which frequently occur (in particular wrong provision of target and source language) could automatically be corrected.

[3] https://www.it-planungsrat.de/SharedDocs/Sitzungen/DE/2018/Sitzung_26.html?pos=9

As regards the value of language resources, the workshop showed that even though language resources had been stored by most public services in some way through the use of corresponding CAT environments and even though each data set is typically delivered with a minimal set of metadata (title, author, date of creation etc.), the participating organisations actually had no idea how to systematically categorize or describe data to facilitate their re-use (e.g. by providing information about whether the translation contains any personal data, any confidential data or any copyright limitations). Especially through the recent advances in machine translation, the translation services had started to re-assess the importance and management of their LR with the goal of potential re-use for the training of MT systems. As part of the discussion round, a major step ahead was planned: To start adjusting translation workflows (and translation environments) to provide metadata indicating whether or not particular data sets can be re-used or shared. This, however, means not only adaptation on the technical side (i.e. modification of the meta-data fields), but of course also additional workload for the translators (additional analysis and provision of additional information about the translation). On the technical side, it turned out that for most participating translation services, it is technically possible to add corresponding metadata fields for describing the translation. Only in the case of the Deutsche Bundestag, it proves technically difficult to implement such a change because of the particular tool used there. As regards the additional workload for translators, it was discussed whether it would actually be feasible to ask the provider/author of the original text to add relevant metadata (indicating the legal status and ability to share the text) as part of the order information within the order database. This would clearly minimize the translators' efforts to check the text from a legal point of view.

As regards the future work of ELRC, emphasis hence must be placed on:

- Providing practical and wherever necessary on-site support to freeing existing language resources that have not been contributed (in particular through anonymization)
- Providing further assistance in the promotion and implementation of good data management practices
- Illustrating the benefits of the sharing of language resources (for MT training purposes as well as from a wider, pan-European re-use perspective)
- Providing further assistance and consultation for making data ready for MT training (including corresponding MT workshops with relevant organisations).