



**European Language
Resource Coordination**
Connecting Europe Facility

Deliverable Task 2

6th Meeting of the Language Resource Board (LRB)



Author(s): Andrea Lösch (DFKI)
Lilli Smal (DFKI)

Dissemination Level: Public

Version No.: <V1.1>

Date: 2018-07-20



Contents

1	<u>Executive Summary</u>	3
2	<u>Meeting Agenda</u>	4
3	<u>Summary of Discussion Points</u>	5
3.1	Update of ELRC results	5
3.2	Workshop updates and lessons learnt	5
3.3	Generic Services: Who they are and what they do	6
3.4	Country profiles: Status quo and vision	11
3.5	Discussion and Outcomes of the Northern Regional WG	11
3.6	Discussion and Outcomes of the South-Eastern Regional WG	12
3.7	Discussion and Outcomes of the South-Western Regional WG	12
3.8	Discussion and Outcomes of the Generic Services WG	14
3.9	Summary and Conclusions	15
	<u>Annex</u>	<u>17</u>
	Presentations LRB Meeting	17

1 Executive Summary

The 6th Meeting of the ELRC Language Resource Board (LRB) took place in the Westminster Hotel (www.westminster-nice.com) in Nice (France) on Thursday, 19th of April 2018. It was chaired by the ELRC consortium and Philippe Gelin (Head of Sector Multilingualism – DG CNECT.G3 Learning Multilingualism and Accessibility). The central goal of the meeting was two-fold: On the one hand, the meeting provided an update on the latest achievements and results of the ELRC project as well as related projects (Generic Services). On the other hand, the meeting aimed to further advance the set-up of a sustainable supply of Language Resources (LR) across all EU Member States, Iceland and Norway.

Following the welcome addresses by the European Commission and the ELRC coordinator, a short update on the ELRC progress was given (in particular regarding the improvement of the ELRC services, the extension of the ELRC network and the advances made in the collection of LR). Subsequently, the key issues (and open questions) regarding the conduct of the ELRC country-specific workshops were addressed and clarified. Then, the six Generic Services projects were introduced in short presentations (detailing their aims and objectives as well as major activities as regards the collection of LR). In the final presentation of the morning, the so-called country profiles, through which ELRC collects information about the national infrastructures for sharing LR in each country were introduced. These country profiles had already been prefilled by the ELRC secretariat with information retrieved through desk research – and had been sent to all NAPs prior to the meeting.

The country profiles were then reviewed by the NAPs in the dedicated regional working groups (North, South-East, South-West) in the afternoon and corresponding results were presented at the end of the meeting. In addition, a working group focusing on the coordination with the Generic Services had been held in order to identify joint activities, to foster collaboration where possible, and to create synergies between all projects.

On the following pages, the report provides the details of the workshop agenda, an overview of the contents and central discussion points of each session, as well as the slides of all presentations held at the meeting (see Annex). The complete list of registered participants and attendees is provided in the Annex.

It is important to note that all presentations held at the meeting are also accessible through the ELRC website: <http://www.lr-coordination.eu/6LRBMeeting/agenda>. Moreover, all information relevant to the organisation of the country-specific workshops are available through the dedicated SharePoint for National Anchor Points.

The Travel Reimbursement Form for the LRB members is available online through the ELRC website: <http://www.lr-coordination.eu/6LRBMeeting>.

2 Meeting Agenda

08:30 – 09:00	Registration Coffee
09:00 – 09:05	Welcome address EC (<i>Philippe Gelin, DG CONNECT</i>)
09:05 – 09:15	Welcome and introduction (<i>Josef van Genabith, DFKI</i>)
09:15 – 09:30	Update on ELRC results (<i>Andrea Lösch, DFKI; Valérie Mapelli, ELDA</i>)
09:30 – 10:00	Workshop updates and lessons learnt 09:30 – 09:40 Workshop updates (<i>Stelios Piperidis, ILSP</i>) 09:40 – 10:00 Practical hints and tips (<i>Maria Giagkou, ILSP; Lilli Smal DFKI</i>)
10:00 – 10:30	Generic services: Who they are and what they do 10:00 – 10:15 ELRI (<i>Thierry Etchegoyhen, Vicomtech</i>) 10:15 – 10:30 EU Council Presidency Translator (<i>Andrejs Vasiljevs, Tilde</i>)
10:30 – 11:00	Coffee Break
11:00 – 12:00	Generic services: Who they are and what they do 11:00 – 11:15 iADAATPA (<i>Amando Estela, Pangeanic</i>) 11:15 – 11:30 eTranslation TermBank Project (<i>Tatjana Gornostaja, Tilde</i>) 11:30 – 11:45 Cross-border eProcurement notifications (<i>Marek Vandak, ANASOFT</i>) 11:45 – 12:00 ParaCrawl (<i>Kenneth Heafield, University of Edinburgh</i>)
12:00 – 12:30	Country profiles – Status quo and vision (<i>Andrea Lösch & Lilli Smal, DFKI</i>)
12:30 – 13:30	Lunch Break
13:30 – 15:00	Regional working groups – Language resource supply
15:00 – 15:30	Coffee Break
15:30 – 16:30	Presentation of results 15:30 – 15:45 Presentation South-Western region (<i>Khalid Choukri, ELDA</i>) 15:45 – 16:00 Presentation South-Eastern region (<i>Stelios Piperidis, ILSP</i>) 16:00 – 16:15 Presentation Northern region (<i>Andrejs Vasiljevs, Tilde</i>) 16:15 – 16:30 Comments and discussion (<i>all</i>)
16:30 – 17:15	Summary and Conclusions
From 17:15	Networking

3 Summary of Discussion Points

3.1 Update of ELRC results

The presentation was jointly held by Andrea Lösch (DFKI) and Valérie Mapelli (ELDA). It first summarized the improvement of services currently provided by ELRC (as opposed to the services that had been available one year before). Furthermore, it illustrated the extension of the ELRC Helpdesk (dedicated technical and legal part, direct support for LR processing, including anonymization), the improvements of the ELRC-SHARE (new metadata schema, integration of eDelivery, functionalities for exporting metadata to the EU ODP etc.), as well as the new ELRC On-site Assistance. The general process for anonymization of LR was introduced, before the extension of the ELRC network (starting with the National Anchor Points (NAPs) up to a network of more than 600 potential data contributors today) was illustrated. Last but not least, the development of LR collection was presented for the previous year, showing a clear increase in unique LR (225 in April 2017 and 392 in April 2018). Details of the developments in terms of coverage per language and legal status (for the reuse of LR) were given (both for April 2017 and April 2018).

3.2 Workshop updates and lessons learnt

The presentation was jointly held by Maria Giagkou (ILSP) and Lilli Smal (DFKI). Following the introduction of the latest schedule for the country-specific workshops, the different supporting materials to guide the NAPs in the preparation of their workshops were highlighted again. The suggested timelines and to do's for the workshop preparation were discussed, in particular highlighting mission critical activities such as the setting of the agenda, the sending of invitations and the publication of the press releases. Based on the feedback from past workshops, additional hints and tips were given for the organization and management of the panel sessions. In particular, it was pointed out that the main goal of the panel was to investigate the needs of digital public services for multilingual communication, and that hence, the digital public services represented in the panel should have a cross-border character. It was also pointed out that the national open data portals were, though being national, already linked to the EU Open Data Portal which reuses their information and which utilizes eTranslation.

Moreover, the most frequently asked questions were presented (typically concerning the access to eTranslation, IPR of outsourced translations, types of language data relevant to ELRC, data protection and personal data, as well as cases for data processing and on-site assistance). Furthermore, the content of the workshop reports and the corresponding deadline for submission were discussed. Most importantly, the role of the DGT Local Field Officers was explained. It was stressed again that the organisers of the workshop should NOT ask the Local FO for any support to the organization of the event, to the sending of invitations or to the mediation of potential

ELRC Report on 6th LRB Meeting

data owners etc. Instead, the support available and provided by ELRC (and corresponding contact points and contact details for each topic and region) was presented.

Last but not least, the key issues relating to the workshop subcontracts were discussed. This concerned in particular the delivery of LR. It was stressed again that prior to any collection efforts, it must be checked with the ELRC regional representatives whether or not the targeted LR have been retrieved yet. In addition to that, the corresponding requirements regarding the invoice and payment were provided and an exemplary invoice was shown for clarification.

3.3 Generic Services: Who they are and what they do

The first presentation of this session was about **ELRI (the European Language Resource Infrastructure)**, which is coordinated by Vicomtech and involves six other partners from four different EU Member States (Portugal, Spain, Ireland, France). The presentation was delivered by Khalid Choukri (ELDA), who is part of the ELRI consortium, since ELRI's coordinator Thierry Etchegoyhen (Vicomtech) could unfortunately not be present himself due to the cancellation of his flight. ELRI's goal is the provision of an infrastructure to help collect, prepare and share language resources, which can improve translation services in Europe. This should be done through automating the creation of translation memories from available data. The main focus of the activities is based on the national level (set-up of a national relay station in each country through which LR can be exchanged within one country), but at the same time, ELRI is also expected to provide flexible means to share data at the European level and at the community level (i.e. among LR providers and users) as well.

The second presentation was about the **EU Council Presidency Translator**, which was presented by Andrejs Vasiljevs (Tilde). The goal of the EU Council Presidency Translator is to enable multilingualism within the EU Council Presidencies 2017 and 2018. The tool is based on eTranslation, which is adapted according to the needs of each EU Council Presidency. It is available to staff members, journalists, delegates and translators through the corresponding website, via desktop, mobile or plugin. It enables quick translation of documents, provides access to information in the local language (e.g. press releases, local news sites etc.) and leads to an increase in translation productivity for translators, as well as reduced costs for the translation of documents through post-edited machine translation. Examples of the different use-cases in Estonia, Bulgaria and Vienna were given and it was explained that the time available for adapting eTranslation to each new EU Council Presidency is about four months.

The third presentation of this session was given by Amando Estela (Pangeanic), who introduced the **iADAATPA Platform** (A Machine Translation Marketplace for Public Administration Bodies). The project is headed by Pangeanic and involves several MT

ELRC Report on 6th LRB Meeting

service providers (KantanMT, Tilde, Prompsit, Pangeanic), Everis, as well as SESIAD and multiple partners at national, local and regional level. It is the goal of iADAATPA to provide (i) MT services between any pair of EU official languages (synchronous and asynchronous), (ii) a series of plug-ins that connect to a wide range of SMS, TMS and Computer Aided Translation (CAT) tools (e.g. TRADOS, MemoQ, Internet Browser etc.), (iii) a programming web service (SDK) interface for domain detection, language detection, quality estimation and translation routing services, plus (iv) 40+ generic and domain adapted machine translation engines. The in-domain trained engines that are currently available include financial, legal, touristic, political, educational and general areas. Envisaged users and use-cases are the Irish Parliament, SEGITTUR and the Lithuanian Parliament.

The fourth presentation of the session was presented by Tatjana Gornostaja (Tilde) and focused on the **eTranslation Term Bank**. First, it was emphasized that terminology is a key enabler for domain adaptation both within neural machine translation (NMT) and statistical machine translation (SMT). Different examples were given to illustrate costly translation mistakes resulting from the use of wrong terminology. The goal of the eTranslation Term Bank project is to provide new terminological resources for three different areas (i.e. health, business legislation, customer protection) in order to support the eHealth, eJustice and Online Dispute Resolution DSIs. Most importantly, the project aims to stimulate the collection and provision of such LR with clear licensing status in general and in the longer term.

The fifth presentation focused on the **Cross-border eProcurement Notifications** project and was presented by Marek Vandak (ANASOFT). The goal of this project is to enable easy cross-border public procurement in selected EU countries (Slovenia, Slovakia, Croatia) by using eTranslation, which will facilitate the distribution of new public procurement announcements to potential bidders across the borders with the provision of basic tender information in the user's preferred language. For the translations, English is used as middle agent and a corresponding quality feedback loop is included.

The sixth and last presentation of this session was held by Kenneth Heafield (University of Edinburgh) and focused on the **ParaCrawl** project. Paracrawl aims to retrieve large, broad corpora from the web. The size of the corpora is between 4 million and 1 billion translated words and they cover more than 500.000 web domains. The estimated quality impacts (BLEU scores) on different language pairs are expected to be considerable, e.g. for Czech-English +0.6, for Finnish-English +2.5, for German-English 1.7, for Romanian-English +3.1 and for Latvian-English +0.8. The resources are publicly available online through tinyurl.com/ybtda3dk. Following the presentation of ParaCrawl, Kenneth Heafield presented several questions on legal issues (in particular personal data and copyright issues), which are discussed in greater detail in the following.

ELRC Report on 6th LRB Meeting

- First of all, it was pointed out that the European Language Resource Association (ELRA) sells webcrawl parallel corpora which, can be considered a normal procedure given the fact that ELRA has obtained the rights for distributing the LR. After that, he referred to a text snippet of a resource that said “Or check the video done by Steve Huff” and stated that some of these resources are not anonymized. (*NB: The lack of anonymization does not pose a problem in this case, because “Steve Huff” is the name of the author of a video, and mentioning the name of the author whenever a work is used in fact is a legal obligation.*) Moreover, Kenneth Heafield showed the extract of a copyright in this resource stating that “All information published on this website is copyright protected and may not be used without written permission from Schaper & Brümmer.” (*NB: This is correct and the resource is still available, because the provider warrants that he obtained all necessary permissions. As such, it is correct to provide the resource this way.*)
- Following this, Kenneth Heafield provided the example of the National Library of the Netherlands, saying that The Netherlands has no legal deposit law and that the National Library archives the web in any case. He further suggested to adopt a pragmatic way to handle copyright issues: The opt-out approach, which assumes implicit permission for web archiving. (*NB: It is very important to note that there is a fundamental difference between web archiving by a national library on the one hand (which is based on the Dutch transposition of the library exception, art. 5.2(c) and 5.3(n) of the InfoSoc Directive) and a body like ELDA which distributes LR commercially or the actions performed within a funded project like ELRC or ParaCrawl (which includes the provision of services) on the other. ParaCrawl resources (like the resources collected by ELDA and ELRC) were not compiled by a publicly accessible library, so no parallel can be drawn at this point.*)
- Kenneth Heafield also claimed that “[e]very EU case which was cited in [ELDA’s legal evaluation] report was won by a crawler”. (*NB: This unfortunately does not correspond to the reality, rather the opposite is the case: e.g. Infopaq and Directmedia showed a clear tendency against crawling. It shall also be remembered that the Court of Justice of the European Union only interprets law and does not apply it to the facts of specific cases (this is done later by national courts), which is why it is not possible to say that a CJEU case had been ‘lost’ or ‘won’ by anyone.*)
- Kenneth Heafield moved on further and illustrated different quotes from ELDA’s legal evaluation report, namely: “not impossible to organize the crawling process in such a way as to comply” vs. “the creation of statistical language models would also qualify as lawful use”. (*NB: Here, it is important to point out that these quotes were taken out of their context. Because it is indeed NOT impossible to structure the crawling process in a way that complies with*

ELRC Report on 6th LRB Meeting

copyright law, providing that no permanent copies of the crawled content are made, and that there is no communication to the public. This process is adapted by ELRC. ParaCrawl, on the other hand, obviously made permanent copies and is now communicating them to the public (see tinyurl.com/ybtda3dk), which may cause problems with regard to copyright law. In addition to that, it can be considered very likely that the creation of language models will qualify as lawful use – it is one of the five elements of the exception for temporary acts of reproduction that need to be met cumulatively. It is evident that ParaCrawl does not meet the first one (see above), which is why arguing about the remaining four is an exercise in futility.)

- Kenneth Heafield also displayed further quotes from ELDA’s legal evaluation report, to demonstrate that, in his opinion, ELRC seems to ignore options and jumps from “it seems that the most viable way” to “only the sources that pass this validation procedure”. (NB: Also in this case, the quotes are taken out of their context, because the report originally says: “**The most viable way** of making sure that the crawling operations are lawful is to perform an a priori clearance of the sources that are to be crawled. (...) [In this approach], **only the sources that pass this validation procedure can be lawfully crawled**”.)
- In the following, Kenneth Heafield suggests as an option for the use of crawled resources the so-called ELRC endorsed Temporary Exemption. (NB: The information about the possibility to apply the exception for temporary acts of reproduction is summarized as follows: “Unfortunately, these exceptions allow for web crawling only in very limited circumstances. This is the case when: the reproductions made in the process are temporary (which is of very limited relevance for crawling activities) (...)” Here, it is very important to note that the temporary character of the reproductions is a *sine qua non* for the application of the abovementioned exception (called rather descriptively: the exception for temporary acts of reproduction). It is essential to highlight that reproductions made by ParaCrawl are not temporary (they are not a part of the process in which they are automatically deleted) and that, as a consequence of that, the exception cannot apply, regardless of whether other conditions are met or not.)
- As a more reasonable approach, Kenneth Heafield suggests to use ROAM (Randomise – shuffle the sentences, Omit – remove data, Anonymise – replace all personal data, Mix – jumble sentences from different sources). (NB: From a legal point, the ROAM approach can by no means be considered more logical due to a number of reasons:
 - Randomise: Shuffle the sentences -- To be able to shuffle the sentences, they need to be downloaded first, which means that reproductions are being made and as explained in the report, in most cases, this requires permission from the right holder (which may be granted upfront, e.g. in a public license, hence the importance of a priori IPR clearance). Moreover, to anticipate the counterargument, the exception for

ELRC Report on 6th LRB Meeting

temporary acts of reproduction does not allow any modifications of the reproduced works.

- *Omit: Remove data -- It is a matter of fact that in Germany, up to 15% of a work can only be communicated to the public for research and teaching purposes, only in a password-protected environment, and subject to the payment of compensation to a German collecting society. This raises the question if ParaCrawl actually meets the remaining requirements and whether they are in fact omitting 85% of the data.*
- *Anonymise: The act of simply removing phone numbers and e-mail addresses etc. as indicated in the slides of Kenneth Heafield is not sufficient to achieve robust anonymisation, which is why the use of appropriate techniques should be the way to go.*
- *Mix: "jumbling sentences" means making derivative works, which requires permission from the right holder. Moreover, in many jurisdictions (e.g. in France) it is likely to be regarded as violation of a moral right of integrity (protecting works against distortion).*
- ⇒ *As such, legally speaking, the ROAM process does not make it legal to obtain the LR, but only conceals it.)*
- *Furthermore, Kenneth Heafield moves towards case law referenced in the ELDA's legal evaluation report. He claims that Google image search provides thumbnails of images from the web and that uploading on the open internet (without any measures to prevent indexing) would yield an implied consent (Vorschaubilder I), even if uploaded without the right holder's permission (Vorschaubilder II). He explains that there is a claim which asks you not to link anything if you "knew or ought to have known" that it infringes and that commercial sites actually ought to know. In Vorschaubilder III, however, Google was not expected to know, despite being commercial. (NB: The presentation slides again misquoted ELDA's legal evaluation report, since the report does not say that the German court "**abandoned** implied consent", but: "[the court] **seems to have abandoned** (...) implied consent". There is a fundamental difference between "abandoned" and "seems to have abandoned". Also, it is important to note that Vorschaubilder III is indeed based on Svensson **and** GS Media (which seems to overrule Vorschaubilder I and II), but it distinguishes between the facts in these cases. In short, the German court ruled that since Google is a search engine provider, the presumption established by the GS Media case (which states that providers of commercial websites are aware of the illegal nature of the content they link to) does not apply. Therefore, the solution of Svensson (stating that linking cannot be seen as communication to the public) prevails. The facts of the Vorschaubilder III case are of very limited relevance for web crawling (unless performed by search engines providers), which is why the report only mentions the case very briefly. Vorschaubilder III is mainly a case about linking, not about crawling data. As a*

ELRC Report on 6th LRB Meeting

consequence, even if the implied license doctrine survived in Germany, it can be expected that it will only apply to openly available search engines – and not to crawling. Until we know of any recent case law that applies this doctrine to other scenarios, we cannot reasonably and responsibly advise anyone to base their crawling activities on implied license/consent. In most countries, implied license/consent is not conceivable, as the law requires the licenses to be in written form.)

3.4 Country profiles: Status quo and vision

The introduction to the country profiles including the vision and their status quo was presented by Andrea Lösch (DFKI) and Lilli Smal (DFKI). In a brief summary of the last year, the key questions to be addressed were explained again. On the European level (ELRC), the main goal is to coordinate and ensure the sustainability of structures for sharing LR on all levels. On the national level, it is now time to further investigate the LR exchange, in particular the relevant stakeholders and the operational frame(s) (process, conditions, constraints etc.) for sharing LR. Then, on the level of the individual institutions in each country, it is important to understand who would be the key people involved in the LR exchange and what the particular institutional frame that governs the LR exchange is. As such, the goal of the country profiles is to gain an understanding of the frame for sustainable sharing of LR in each country. This includes on the one hand a review of all existing stakeholders (LR holders, LR creators, decision makers, existing networks, CEF representatives...) and the investigation of the infrastructure for LR exchange (including also legal frame). The example of Germany was given, followed by an illustration of the collection of information about available NLP tools and services. A preliminary list had been sent out to all participants prior to the meeting together with the country profiles for their region. With the help of this information, all participants were released into their Working Groups (WG) that were scheduled for after lunch.

3.5 Discussion and Outcomes of the Northern Regional WG

This WG was chaired and moderated by Aivars Berzins (Tilde) (Co-chair: Lucia Ingala-Hornung (DFKI)). The discussion showed that translation data is managed differently in each country. While in Finland, the Office of the Prime Minister is responsible, in Latvia it is the Centre of the State Language, Culture Information Systems Centre and in Norway, it is the National Library of Norway. In Norway, there is even a specific License for Open Government Data (NLOD license). Moreover, in all countries of the Northern Region, it could be noticed that there is a centralized public procurement data base, as well as Open Data Portals. There were several notable data opening activities, e.g. in Denmark focusing on dictionaries, in Norway focusing on data from private companies (including LSPs) and in Latvia, there was an order by the Prime Minister to gather information about tmx files. However, there were also several issues identified related to the sharing of data, in particular the lack of anonymization and

ELRC Report on 6th LRB Meeting

existence of sensitive data in the LR, the lack of awareness of the conditions for sharing LR, the identification of the data holder in each organization, and the perception of what open data is (so far, textual data often is not considered as data).

3.6 Discussion and Outcomes of the South-Eastern Regional WG

This WG was chaired and moderated by Stelios Piperidis (ILSP) (Co-chair: Maria Giagkou (ILSP)). In the South-Eastern Regional Working Group, several important issues were noted regarding the organization of the workshops. Based on the experience in Hungary, it turned out that local officials sometimes are not aware of the existence of CEF coordinators in all countries. As a consequence of that, ELRC makes the CEF coordinators' names and contact details available to each country. It was also pointed out that retrieving information on the reuse of eTranslation by DSIs and other services might not be easy, but essential for the preparation of the workshop. ELRC will try to provide as much input on this issue as possible. Currently, the DSIs mainly use eTranslation for translating metadata. However, for the NAPs, it is important to have additional information on the reuse of eTranslation by the DSIs (e.g. if the DSIs will use eTranslation to translate free text). The eProcurement Generic Services project can serve as a good use-case for this scenario.

As regards the data collection, it was noted that there are many different channels and that it is important to identify the stakeholders for LR exchange in each country. In fact, there are many channels through which the EC tries to collect data, e.g. within ELRC, there are donations, workshops and subcontracts, plus data aggregation through crawling. For all ELRC resources, the legal documentation is/will be attached to the data contributed.

As regards the country profiles, it was suggested to make some additions, namely:

- Indication if the LSPs share their TMs (in the case of outsourced translations)
- Indication of people responsible for the Digital Single Market, Information Society

These suggestions will be implemented in the next version of the country profiles. Furthermore, it was pointed out that the stakeholders' position grouping (categorization) is not easy to understand.

Last but not least, the list of tools and services was introduced (mainly to the Technology NAPs), who were asked to review and extend the list based on their local expertise.

3.7 Discussion and Outcomes of the South-Western Regional WG

This WG was chaired by Khalid Choukri (ELDA) (Co-chair: Valérie Mapelli (ELDA)). The discussion in the South-Western Regional WG was mainly focused on the country profiles and emerging differences between all countries. As regards Germany, it was noted that the names within the administrations may change a lot and that it hence is

ELRC Report on 6th LRB Meeting

important to also add the particular position/function of each stakeholder (as it is the case in the current version). Also, Germany is a fairly decentralized country regarding the organization of IT services (and translations). The IT Planning Council is only just starting the centralization work – and MT and the collection of LR is not a high priority. However, having a structure on the national level, which is comparable with the one of ELRI might be a benefit. In this context, the question is how to interact best with the various representatives to raise awareness.

In Spain, Plata is in evolution with the aim to integrate MT@EC, but it is still in the early stages. Most activity in MT can currently be observed on the regional level (translation into regional languages rather than in other EU languages). The situation is similar in Portugal, where translation is mainly relevant on the regional level (Portuguese-Mirandese). Only few institutions have specific translation departments, but there is no real strategy for sharing LR. Over the next months, the Portuguese Public Service NAP (Paulo Vale) and his team will visit the representatives of the translation offices to promote and encourage the sharing of LR; appropriate tools will be developed and shared with the public administrations in return. A corresponding national event is envisaged with the Parliament in this respect.

In Italy, the situation is similar to Portugal and Germany, with a highly fragmented landscape for translations within public services and public administration. Above all, public administrations are not very keen on sharing LR outside their sphere because of legal issues (copyright, personal data). The Italian NAPs are currently contacting decision-makers to promote LT and the sharing of LR.

In contrast to that, France and The Netherlands are largely centralized countries – but this does not support the sharing of LR, as the ownership of the translation memories (TM) remains with the owner of the source. There is also hesitance from the side of LSPs to share their LR (since they do not see the need or benefit in doing so, because they are owners of their translations), which is why there must be a real incentive for both public administrations and LSPs to share their data. For LSPs, it is even more important to have a return on their “investment” as they have no access to eTranslation.

In France, which is also very centralized, only three Ministries have their own translation services: The Ministry of Foreign Affairs, the Ministry of Finances, and the Ministry of Social Security. All remaining translation requests are distributed to LSPs. However, there is the inter-ministerial working group on translation (GIT), which oversees the translation work in the public administration. It has been discovered that the Ministry of Culture holds very important LR (mainly terminological resources, but also 3.000 hours of regional spoken LR), but it needs to be investigated whether and how they could be retrieved.

ELRC Report on 6th LRB Meeting

In Ireland, most translations are outsourced. They follow the aim to create, within the next 6 months, a central translation platform that will enable data collection (Irish-English).

It is important to note that the South-Western Region also includes bi-/tri-lingual countries: Malta, Luxembourg, and Belgium (which was not represented this time). In Malta, translations of all legal texts are available online, but in contrast to that, it is a lot more difficult to get hold of internal translations (internal communications). In Luxembourg, all official documents received are in French, German, English and Portuguese. The follow-up on the recovery of LR will be made in a meeting in Luxemburg on 24 April.

As a major conclusion from this WG, it appears that mainly legal issues (in particular ownership of a resource/copyright and inclusion of personal information) prevent the sharing of LR already on the national level, and that the legal barriers represent the first and most important hurdle to be taken.

3.8 Discussion and Outcomes of the Generic Services WG

This WG was chaired by Philippe Gelin (European Commission) and moderated by Andrea Lösch (DFKI) (Co-chair: Lilli Smal (DFKI)). The goal of this particular WG was to identify overlapping activities and/or events in order to create synergies between the projects (and to minimize the efforts for the consortia as well as the burden for the target group(s)). Moreover, it was important to find out which LR might be available through these projects that could then be shared with the ELRC.

As regards the overlap of activities and events, it was found that the project's activities are quite diverse. Within the eTranslation Term Bank project, the only official event had already taken place. Within ELRI, there will be a conference towards the end of the first year (target group: LR holders, ideally for Spanish, Portuguese and Irish LR, i.e. translations from or into these languages). There might be a chance for collocation not with ELRC events, but with SMART 2017/0103 conference that is scheduled for the end of November 2019. A corresponding concept will be prepared and introduced to the EC. Within the EU Council Presidency Translator, events are bound to the EU Presidency and leave very little room for collaboration because of the dedicated focus of the project.

Moreover, apart from ELRC, several Generic Services projects focus on the collection of LR, namely:

- ELRI (collection of LR on the national level in Ireland, Spain, Portugal),

ELRC Report on 6th LRB Meeting

- eTranslation Term Bank (collection of terminologies in the area of health, business legislation, customer protection (in order to support the eHealth, eJustice and Online Dispute Resolution DSIs),
- ParaCrawl (large parallel corpora from the web).

The sharing of LR from ELRI and eTranslation Term Bank projects can and will be done immediately – all LR that can be shared with the EC will be uploaded in the ELRC-SHARE Repository. As regards the ParaCrawl resources, immediate sharing and use by the EC is not advisable because of the legal uncertainty about the use of webcrawled copyrighted materials.

Most interestingly, LR are also relevant to other projects, in particular:

- EU Council Presidency Translator: While the LR used to train eTranslation in the project cannot be shared, it may be possible to share the language models derived from them with the ELRC.
- Cross-border eProcurement notifications: Here, it is possible to share the improved translations (derived from the quality feedback loop) with the ELRC.
- iADAATPA: Here, it is possible to share the LR from Segitur (Tourism) with ELRC.

This information will be shared with the ELRC Data team, who will follow up on these hints.

3.9 Summary and Conclusions

This session was presented by Andrea Lösch (DFKI). It was pointed out that especially regarding the workshop preparation and conduct (subcontracts, organization, data collection) it is important to stay in close touch with the ELRC team to avoid duplication of efforts and misalignment of activities. Similarly, as regards the Generic Services projects, several activities were identified where it was possible to create synergies between the projects. This did not only concern the organization of events, but largely the recovery and sharing of LR.

Most importantly, it was stressed again that the goal of the creation of country-profiles was a fundamental pre-requisite for the set-up of processes for sustainable LR supply, because these processes need to be based on (and make use of) the infrastructure(s) available in each country. Furthermore, in those cases where there are no infrastructures for the sharing of LR yet, corresponding efforts need to be made to overcome existing barriers and to set-up such structures. For now, the immediate focus is on the national level and will be on the collection of information from the institutional level at a later time. As such, following the initial set-up of the country profiles for and within the LRB Meeting, the country profiles will be updated by the

ELRC Report on 6th LRB Meeting

ELRC accordingly and sent to all NAPs for final feedback and individual follow-up (deadline: 14.05.2018).

Last but not least, several organizational matters directly associated with the LRB Meeting were clarified (i.e. procedure for travel reimbursements, collection of feedback about ELRC website, date and place of next LRB Meeting: 20.09.2018 in Paris).

The LRB Meeting closed with a networking event in the lobby.

ELRC Report on 6th LRB Meeting

Annex

Presentations LRB Meeting

All presentations held at the LRB Meeting are available online through the ELRC website: <http://www.lr-coordination.eu/6LRBMeeting>

The slides presented at the LRB Meeting are also provided in this Annex in the order they were presented in:

- **Update on ELRC results** (*Andrea Lösch, DFKI; Valérie Mapelli, ELDA*)
- **Workshop updates and lessons learnt** (*Maria Giagkou, ILSP; Lilli Smal DFKI*)
- **Generic services: Who they are and what they do**
 - ELRI (*Thierry Etchegoyhen, Vicomtech*)
 - EU Council Presidency Translator (*Andrejs Vasiljevs, Tilde*)
 - iADAATPA (*Amando Estela, Pangeanic*)
 - eTranslation TermBank Project (*Tatjana Gornostaja, Tilde*)
 - Cross-border eProcurement notifications (*Marek Vandak, ANASOFT*)
 - ParaCrawl (*Kenneth Heafield, University of Edinburgh*)
- **Country profiles – Status quo and vision** (*Andrea Lösch & Lilli Smal, DFKI*)
- **Regional working groups – Language resource supply**
- **Presentation of results**
 - Presentation South-Western region (*Khalid Choukri, ELDA*)
 - Presentation South-Eastern region (*Stelios Piperidis, ILSP*)
 - Presentation Northern region (*Andrejs Vasiljevs, Tilde*)

Summary and Conclusions (*Josef van Genabith, DFKI; Andrea Lösch, DFKI*)