European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management

Andrea Lösch[1], Valérie Mapelli[2], Stelios Piperidis[3], Andrejs Vasiļjevs[4], Lilli Smal[1], Thierry Declerck[1], Eileen Schnur[1], Khalid Choukri[2] and Josef van Genabith[1]

[1]DFKI GmbH, [2]ELDA, [3]ILSP/Athena RC, [4]Tilde

[1]Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany

[2]9 rue des Cordelières, 75013 Paris, France

[3]Epidavrou & Artemidos, Maroussi, Athens, Greece

[4]Vienibas gatve 75a, LV1004, Riga, Latvia

[1]{andrea.loesch|lilli.smal|declerck|eileen.schnur|Josef.Van_Genabith}@dfki.de

[2]{mapelli|choukri}@elda.org, [3]spip@ilsp.gr, [4]andrejs@tilde.com

Abstract

In order to help improve the quality, coverage and performance of automated translation solutions for current and future Connecting Europe Facility (CEF) digital services, the European Language Resource Coordination (ELRC) consortium was set up through a service contract operating under the European Commission's CEF SMART 2014/1074 programme to initiate a number of actions to support the collection of Language Resources (LRs) within the public sector in EU member and CEF-affiliated countries. The first action focused on raising awareness in the public sector through the organisation of dedicated events: 2 international conferences and 29 country-specific workshops to engage national as well as regional/municipal governmental organisations, language competence centres, relevant European institutions and other potential holders of LRs from public service administrations and NGOs. In order to gather resources shared by the contributors, the ELRC-SHARE Repository was set up together with services supporting the sharing of LRs, such as the ELRC Helpdesk and Intellectual Property Rights (IPR) clearance support. All collected LRs pass a validation process developed by ELRC. The collected LRs cover all official EU languages, plus Icelandic and Norwegian.

Keywords: ELRC, Public Sector Information directive, LR evaluation, LR validation

1. Language Resources in and for the Public Sector

1.1 The European Public Sector Information Context

With the European Language Resource Coordination (ELRC), the European Commission (EC) has taken a decisive step towards supporting a truly multilingual Digital Single Market by enabling public services for Europe's citizens and businesses to operate freely across language barriers. The Connecting Europe Facility (CEF[1]) has various building blocks, one of them being "eTranslation", which is offering an automated translation platform[2] to facilitate multilingual communication and the exchange of documents and other linguistic content in Europe between national public administrations on the one hand and between these administrations and EU and CEF-affiliated country citizens and businesses (European Commission, 2017) on the other hand.

The CEF eTranslation platform provides machine translation (MT) services in all official languages of the EU as well as CEF-affiliated countries to address public administration scenarios in the areas of consumer rights, health, public procurement, social security, culture and others. eTranslation is technically managed by the EC Directorate-General for Translation (DGT) and will empower Europe's public online services such as the Online Dispute Resolution platform (ODR), eJustice, the Electronic Exchange of Social Security Information (EESSI), Business Registry Information System (BRIS), eProcurement and many others. Table 1 provides an overview of existing sector-specific Digital Service Infrastructures (DSIs) and their domains.

1.2 The PSI Directive

In accordance with the Public Sector Information (PSI) Directive 2003/98/EC (modified in 2013 by the Directive 2013/37/UE), Member States should ensure that documents, which are held by public sector bodies and accessible according to national access regimes, are re-usable for commercial or non-commercial purposes.

This obligation ensures that data produced by public administrations will be easier to share and made available for different types of usage.

CEF DSI	Domain
Online Dispute Resolution	Consumers' rights
Electronic Exchange of Social Security Information	Social security, insurance
eProcurement	Public procurement, contractual agreements
European e-Justice Portal	Justice, Law
eHealth	Health, Medicine
Business Registers Interconnection System	Business, market
Safer Internet	ICT
Cybersecurity	ICT
Public Open Data	Multiple domains
Europeana	Culture

Table 1: CEF Digital Service Infrastructures (DSIs) and their domains

[1] For more details on CEF, see: https://ec.europa.eu/digital-single-market/en/connecting-europe-facility.

[2] For more details on CEF eTranslation, see: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation.

Even though the PSI Directive is an important instrument to open up public sector data, there are many challenges in collecting LRs from public services, such as lack of awareness, lack of technical or legal competence, poor data management, etc. (Vasiļjevs et al., 2018).

1.3 Setting up a European Language Resource Coordination (ELRC)

European, national and regional public administrations deal with a huge amount of multilingual textual information in original and translated form. By sharing this linguistic data and turning it into language resources (LRs), they can improve the quality, coverage and performance of CEF eTranslation that needs multilingual LRs to train MT systems.

In April 2015, the ELRC Consortium was set up through EC's Connecting Europe Facility SMART 2014/1074 programme to initiate a number of actions with the aim to support the collection of such LRs. ELRC is coordinated by DFKI[3] (*Deutsches Forschungszentrum für Künstliche Intelligenz*, Germany), in partnership with ELDA[4] (Evaluations and Language Resources Distribution Agency, France), ILSP/Athena RC[5] (Institute for Language and Speech Processing/Athena Research Centre, Greece) and TILDE[6] (Latvia).

A Language Resource Board (LRB) was set up as governance and oversight body for the ELRC effort, consisting of National Anchor Points (NAPs), i.e. leading technological and public service representatives for each CEF country[7].

Through CEF eTranslation and the data collected by ELRC, European public services and public administrations across Europe will be one step closer towards operating without language barriers. By helping to improve eTranslation public service providers will gain access to better quality MT systems. These increase the efficiency of human translators and, in addition, the MT systems can be integrated in various online services.

This provides a major motivation for public services across Europe to donate and share data through ELRC: increased data provided for the content and languages relevant to particular administrations and/or services produces increased translation speed based on better accuracy of the machine translation output.

The more organisations contribute, the better their translated content will be. In addition, organisations benefit from the data donated by other public administrations, because they will also contribute content that may be relevant.

ELRC is responsible for the online coordination tool – ELRC website http://lr-coordination.eu – that integrates data sharing facilities, access to the LR catalogue, and information on ELRC events, Helpdesk and up-to-date information on the services provided by ELRC.

[3] http://dfki.de/en

[4] http://elda.org/en

[5] http:// www.ilsp.gr/en

[6] http://tilde.com

[7] http://lr-coordination.eu/anchor-points

2. Sharing Language Resources within ELRC

2.1 Raising Awareness in the Public Sector

In order to address the Public Sector and raise awareness of the need for data and sharing LRs to improve CEF eTranslation a large information dissemination initiative has been put into action by ELRC.

Two international ELRC conferences were organised: 1) as part of the Riga Summit 2015 on the Multilingual Digital Single Market[8] on 29 April 2015 and 2) as a satellite event of the Translating Europe Forum (TEF) in Brussels, on 26 October 2016. At each conference, more than 120 key stakeholders of the ELRC network (in particular potential data donors) participated.

In addition, in order to interact closely with national administrations, ELRC organised 29 country-specific workshops to meet with national and regional/municipal governmental organisations, language competence centres, relevant European institutions, other potential holders of LRs from the respective national public service administrations as well as important NGOs. Bringing ELRC to each country and getting engaged on the national level was essential to foster local ownership and local responsibility on which ELRC is built. Through the workshops, ELRC managed to identify more than 1.000 potential data sources. ELRC workshops established contacts to potential data holders, who are central to the subsequent data collection process.

The ELRC website provides detailed reports on these events, including presentations and many video recordings[9]. Vasiļjevs et al. (2018) details key discussions and findings at the workshops organised in Nordic and Baltic countries.

2.2 Collection Process

The process of sharing LRs with ELRC is straightforward and simple. Depending on the size of the data set, participating organisations may:

- Send the data to ELRC via email to data@lr-coordination.eu, as a zip file.
- Upload the data directly onto the ELRC-SHARE Repository, as a zip file, at https://www.elrc-share.eu (see Figure 1 below).
- Contact ELRC for other support such as setting up FTP services for specific deliveries.

ELRC-SHARE (Piperidis et al, 2018) covers the whole life cycle of LR sharing: uploading, documentation, uploading of accompanying documents, monitoring and reporting, updating, browsing, delivery and downloading.

The process is built on and inspired by META-SHARE (Piperidis, 2012) and is essentially an adaptation of its latest version v3.1.1, mainly in terms of the employed metadata schema, the identified user roles and the largely simplified operational workflow.

[8] http://www.rigasummit2015.eu/

[9] http://lr-coordination.eu/events

Figure 1: The ELRC-SHARE Repository

For all donations, ELRC needs to ensure that the material is in the public domain and follows the Public Sector Information (PSI) Directive transposition rules or that the necessary licenses have been obtained. Confidential and personal information is excluded from data to be shared.

2.3 Services to Facilitate the Sharing of LRs

In order to minimize efforts on the side of the donating institutions due to technical and legal obstacles, ELRC offers a range of support services for donating institutions. The ELRC Helpdesk[10] provides assistance and support to both technical and legal questions involved in the use, production, collection, processing and sharing of LRs.
The Helpdesk can be reached freely through several channels (including a telephone number associated to a web-conferencing desktop (Skype) and an email address). Moreover, a web forum platform[11] was set up so as to compile relevant questions and answers.
ELRC also supports donating organisations with different data processing services, including:

- Data conversion (e.g. to plain text or XML)
- Tag removal
- Re-formatting
- Data extraction
- Cleaning and alignment
- Meta-data validation
- Anonymization
- Etc.

If a public administration has potentially usable language data that needs assessment and processing, but that cannot be accessed outside the institution due to technical or legal reasons (e.g. data privacy, confidentiality), ELRC can provide on-site assistance. A member of the ELRC consortium with special knowledge of the particular language processing issue will travel to the organization on-site to provide technical/legal assistance required.
For public administrations, these services are provided for free in order to facilitate their data sharing efforts.

2.4 Data Processing

Each LR is analysed and processed by ELRC experts to ensure compliance with the *Language Resources Data Formats Specification* agreed with EC. According to this specification, resulting parallel data should be provided in the TMX format in UTF-8 encoding, without optional data fields (e.g. translator id, adjacent segments) and without non-printable control characters.
Monolingual corpora are to be delivered in plain text format without any additional annotation, in UTF-8 encoding, single file by language and resource, segmented into paragraphs. Terminology resources should be provided in the TBX format.
Several workflows including automated and manual tasks have been developed to ensure efficient processing of data and compliance to the required specifications. In a typical example, a public institution donates documents and translations in DOCX or PDF format. ELRC experts perform document level alignment of source and translated files, convert documents to plain text format, remove tags, perform sentence level alignment and also data cleaning by removing sentence pairs that include non-printable characters or text in language other than respective source or target language, and convert this data to TMX format. TMX files are validated using TMXValidator[12]. Resulting LRs are stored in the ELRC-SHARE repository and the corresponding metadata fields are filled.
Among the tools used in the data processing are DictMetric (Su & Babych, 2012) for document alignment, Microsoft Bilingual Sentence Aligner[13], language detection tool PYCLD2[14], and many others.

2.5 IPR Clearance

For a number of LRs, ELRC interacted with providers to clarify legal issues linked to IPR and licensing constraints. Licenses were drafted to address e.g. country-specific rules in relation to the PSI Directive, limitations in use within the CEF eTranslation platform, etc.
The ELRC-Repository legal part of the metadata was adapted in order to include open data licenses (such as Open Licence/Licence Ouverte, for France). LRs provided can be classified and viewed depending on the licence available: public domain, open under PSI, open licenses, standard licenses, and non-standard licenses.

3. Validation of Language Resources within ELRC

3.1 Validation Guidelines Implementation

Validation can be understood as the quality control of a LR against a list of relevant criteria (Schneller et al., 2017). Due to the high number of LRs required within the project, the ELRC consortium decided to complement the donated LRs with additional LRs produced from scratch through a website crawling process (Papavassiliou et al, 2018). Web crawling was conducted using ILSP-FC, a comprehensive end-to-end solution for the acquisition of domain-specific monolingual and bilingual corpora from the web.

[10] http://lr-coordination.eu/helpdesk
[11] http://helpdesk.lr-coordination.eu/overview
[12] https://www.maxprograms.com/products/tmxvalidator.html
[13] https://www.microsoft.com/en-us/download/details.aspx?id=52608
[14] https://github.com/aboSamoor/pycld2

The ELRC partners (the National Anchor Points) initially identified and documented public administration websites (e.g. websites of ministries, local authorities, museums, etc.) as candidate sources for the extraction of content relevant to the CEF Digital Service Infrastructures (DSIs) and subsequently deployed ILSP-FC to acquire language resources for specific (EN-X) language pairs, where X stands for official EU languages in CEF-affiliated countries.

All gathered LRs needed to be submitted to the ELRC validation process to check their conformity with the original requirements of the project. The goal of the ELRC validation procedures is to provide a methodology, known as "Validation Guidelines" to validate donated and crawled data, which was collected within the project. The validation of donated and crawled data had to be conducted in different ways.

- It was assumed that the donated data consist of high quality data in terms of content (in particular translations for multilingual data, data produced by human experts), but require a technical- and legal-oriented evaluation. Here validation consists of:
    - checking compliance of data with the ELRC objectives and scope,
    - checking the format of provided data, and
    - checking whether the legal information provided is compliant with the ELRC scope.
- As crawled data come from automatic processing, their validation necessitates deeper content validation, while crawling already delivers the format that corresponds to the Language Resource production requirements. Legal validation of crawled data is also required and had to be carried out according to the following steps:
    - checking whether crawled websites are under the scope of the Public Sector Information (PSI) Directive to make sure that the content can be re-used, and
    - estimating the quality of translation with a team of language expert validators.

All types of data were uploaded to the ELRC-Share Repository and corresponding metadata was validated and completed. Finally, a Validation Report was provided for each data set, and all available legal related documentation were asserting the quality of all data.

Full Validation Guidelines, including the Validation Report template are made available online: http://www.lr-coordination.eu/sites/default/files/common/ELRC Data Validation Guidelines.pdf.

3.2 Issues Regarding the Validation of Language Resources

As indicated above, the validation of donated data was limited to a "Quick Quality Check" that includes:

- compliance with the ELRC scope (relevant language, not data from or already available to EC), and
- checking the metadata elements against a number of minimal requirements in terms of technical information (i.e. format, content, alignment), as well as legal information (whether required information on licensing has been well filled in, e.g. PSI-compliant data, available licenses, attribution, etc.

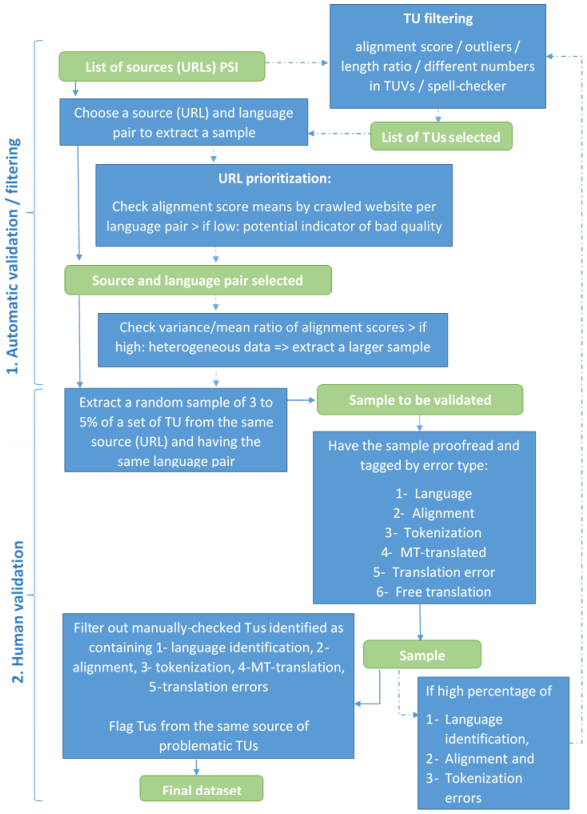

Figure 2: Workflow for Content Validation of Crawled Language Resources

For crawled data, the validation process was much more comprehensive, e.g. the list of crawled URLs was manually checked to assess if the websites are under the PSI scope. Content from websites that did not fall under the PSI Directive was excluded. Errors in Translations Units (TU) were reported and TUs marked up as containing errors were automatically removed; the remaining TUs were annotated with an indication on the probability of finding the same errors. The validation of the content consisted of both an automatic and a manual procedure. Figure 2 illustrates the workflow employed within ELRC for the validation of crawled data.

4. Impact and Results

Having received the first resources starting in Spring 2016, ELRC has managed to collect 225 LRs within one year, covering all official EU languages, plus Icelandic and both varieties of Norwegian, Bokmål and Nynorsk. This includes bi- or multi-lingual contents in digital editable formats ranging from reports, publications and other materials for internal and external use, web contents and brochures, but also terminologies and glossaries.
In 2017, ELRC processed 100 of these 225 LRs to make them compliant with the ELRC technical specification and be fully ready to be used in MT training. 82 new additional LRs were collected in 2017.

More than 58 public sector organisations across Europe have shared their language data with ELRC, including in particular national ministries, governmental bodies and public services.

5. Conclusion and future Work

This paper provides an overview of the major challenges faced by ELRC and how they have been addressed. We are in the process of identifying clear recommendations for future actions regarding Language Resources sharing and collection, in particular with regard to:

- The continuation of personal support structures provided by ELRC, complementary to the ELRC Helpdesk
- The organization of future conferences
- The organization of workshops (including in particular the further development of stakeholder involvement and data pipeline sustainability)
- The support of future work of the Language Resource Board, with for example additional country-specific reach-out events

6. Acknowledgements

The European Language Resource Coordination (ELRC) is a service contract operating under the EC's Connecting Europe Facility SMART 2014/1074 programme from April 2015 to April 2017 and will continue until December 2019 under SMART 2015/1091 LOT 2 "Language Resource coordination and collection with related legal and technical work" and SMART 2015/1091 LOT 3 "Acquisition of additional Language Resources and related refinement/processing services and their provision of the Language Resource Repository of CEF Automated Translation Platform".

7. Bibliographical References

European Commission (2017). eTranslation – Making European Digital Public Services Multilingual. Available at: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation [last accessed: 22.02.2018]

Schneller, Priscille ; Fernandez-Barrera, Meritxell ; Mapelli, Valérie ; Popescu, Vladimir, Choukri, Khalid ; Prokopidis, Prokopis ; Papavassiliou, Vassilis (2017). European Language Resource Coordination – Validation Guidelines. Available at http://www.lr-coordination.eu/sites/default/files/common/ELRC Data Validation Guidelines.pdf [last accessed: 22.02.2018]

CEF: https://ec.europa.eu/digital-single-market/en/connecting-europe-facility [last accessed: 22.02.2018]

CEF eTranslation: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

Directive 2013/37/UE: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013L0037 [last accessed: 22.09.2017]

ELRC-SHARE: http://www.lr-coordination.eu/resources [last accessed: 22.02.2018]

Piperidis, S.; Labropoulou, P.; Deligiannis, M.; Giagkou, M. (2018). Managing Public Sector Data for Multilingual Applications Development, In Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Papavassiliou, V.; Prokopidis, P.; Piperidis, S. (2018) Discovering parallel language resources for training MT engines. In Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Proceedings of the Eighth International Language Resources and Evaluation (LREC 2012), Istanbul, Turkey. European Language Resources Association (ELRA).

Su, F.; Babych, B. (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-) parallel translation equivalents. In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra) (pp. 10-19). Association for Computational Linguistics.

Vasiļjevs, A.; Rozis, R.; Kalniņš, R.; Bērziņš, A. (2018). Collecting Language Resources from Public Administrations in the Nordic and Baltic Countries. In Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).