

# Collection and Curation of Language Data within the European Language Resource Coordination (ELRC)

Andrea Lösch<sup>1</sup>, Valérie Mapelli<sup>2</sup>, Khalid Choukri<sup>2</sup>, Maria Giagkou<sup>3</sup>, Stelios Piperidis<sup>3</sup>, Prokopis Prokopidis<sup>3</sup>, Vassilis Papavassiliou<sup>3</sup>, Miltos Deligiannis<sup>3</sup>, Aivars Berzins<sup>4</sup>, Andrejs Vasiljevs<sup>4</sup>, Eileen Schnur<sup>1</sup>, Thierry Declerck<sup>1</sup>, and Josef van Genabith<sup>1</sup>

<sup>1</sup> DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

<sup>2</sup> ELDA, 9 rue des Cordelières, 75013 Paris, France

<sup>3</sup> ILSP/Athena RC, Epidavrou & Artemidos, Maroussi, Athens, Greece

<sup>4</sup> Tilde, Vienibas gatve 75a, LV1004, Riga, Latvia

**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) was set up in 2015 through a service contract operating under the European Commission's CEF SMART 2014/1074 programme. Since then, ELRC initiated a number of actions to support the collection of Language Resources (LRs) within the public sector in EU member and CEF-affiliated countries. All resources shared by the contributors were gathered and curated in the ELRC-SHARE Repository, after having passed the validation process developed by ELRC. This paper provides insights into the overall data collection and curation process (including both technical and legal validation of resources) employed within ELRC. The ELRC Helpdesk provides both technical and legal guidance (e.g. Intellectual Property Rights (IPR) clearance support) to potential data contributors, thus enabling the sustainable sharing of language data.

**Keywords:** ELRC, language data, LR evaluation, LR validation, LR curation, PSI Directive, Open Data

## 1 The Value of Language Data

Language-centric Artificial Intelligence (AI) and the fast pace of development in language technologies in recent years (especially as a result of the deployment of the neural network paradigm) offer a great variety of new opportunities for digital and technology-enabled communication - and hence excellent market opportunities: According to the Slator Language Industry Market Report 2019<sup>1</sup>, the global language services and technology industry was a 23.2 billion USD market in 2018 which is expected to grow to 28.2 billion USD by 2022 with Europe being among the key players. According to the latest CEF Market Study<sup>2</sup>, the expected growth rate for the European Language Technology (LT) market is even higher (i.e., 10% annual growth until the end of 2021),

<sup>1</sup> <https://slator.com/data-research/slator-2019-language-industry-market-report/>

<sup>2</sup> <https://op.europa.eu/en/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en/format-PDF/source-106906783>

including also public service applications. Especially Natural Language Understanding (NLU) (in particular chatbot applications) as well as domain- and application-specific machine translation (MT) are expected to be of increasing importance in the European LT-Market.

In a report issued in September 2018 on a resolution passed by the European Parliament (“Language equality in the digital age”<sup>3</sup>), an explanatory statement stresses that especially smaller or minority languages are the ones to gain most from language technologies, tools and resources. A clause in the adopted motion calls on the EC to “make as a priority of language technology those Member States which are small in size and have their own language” (European Parliament (2018), p.13) which coincides with the recommendation of the CEF Market Study, according to which EU developers have a clear competitive advantage with regard to building strong experience for such languages thanks to the multilingual market. As such, investing in AI-driven LT research for smaller European languages does not only yield significant impact and potential for the European LT industry, but also for European public services and citizens who are enabled to participate in and contribute to the European Digital Single Market.

## **2 The importance of a European Language Resource Coordination (ELRC)**

The development of language technologies in general – and machine translation systems in particular – require substantial amounts of language data which for most domains and smaller languages simply do not exist in sufficient volumes. On the other hand, every day, European, national and regional public administrations in all EU Member States deal with a huge amount of multilingual textual information in original and translated form. With the European Language Resource Coordination (ELRC)<sup>4</sup>, the European Commission (EC) has taken a decisive step towards minimizing language barriers across Europe and enabling the development of European language technologies by supporting the collection of this language data for all EU official languages, Norwegian, Icelandic, and other languages of interest to the EU Member States (e.g., Chinese, Russian, Turkish etc.).

ELRC was set up through the Connecting Europe Facility’s SMART 2014/1074 programme in April 2015 and since then is coordinated by DFKI<sup>5</sup> (Deutsches Forschungszentrum für Künstliche Intelligenz, Germany), in partnership with ELDA<sup>6</sup> (Evaluations and Language Resources Distribution Agency, France), ILSP/Athena RC<sup>7</sup> (Institute for Language and Speech Processing/Athena Research Centre, Greece), TILDE<sup>8</sup> (Latvia) and recently CrossLang<sup>9</sup> (Belgium). It is governed by the Language

<sup>3</sup> [http://www.europarl.europa.eu/doceo/document/A-8-2018-0228\\_EN.html](http://www.europarl.europa.eu/doceo/document/A-8-2018-0228_EN.html)

<sup>4</sup> <http://www.lr-coordination.eu>

<sup>5</sup> <http://dfki.de/en>

<sup>6</sup> <http://elda.org/en>

<sup>7</sup> <http://www.ilsp.gr/en>

<sup>8</sup> <http://tilde.com>

<sup>9</sup> <http://www.crosslang.com>

Resource Board (LRB) – an oversight body consisting of National Anchor Points (NAPs), i.e., leading technological and public service representatives for each CEF affiliated country<sup>10</sup>. The main activities of ELRC include the collection of language resources, the provision of corresponding language data sharing facilities, in principle through the ELRC-SHARE repository for language data<sup>11</sup>, the support of data sharing through awareness-raising events (country-specific workshops, European conferences<sup>12</sup>) and the ELRC Technical and Legal Helpdesk<sup>13</sup>. By supporting the sharing of this language data and by turning it into standardized, machine-readable formats and actionable language resources (LRs), ELRC directly contributes to improving the quality, coverage and performance of CEF eTranslation and other MT systems that need multilingual LRs as training data.

### 3 Data Collection and Curation Process within ELRC

#### 3.1 Data Collection Process

Language data refers to any textual, audio or audio-visual data produced using human language or data about human language (such as lexica, raw or annotated corpora, language models etc.). The collection of one or more language data sets grouped together according to certain criteria constitutes a language resource. ELRC collects in particular the following types of language resources:

- Corpora, a set of documents or a text in one or more languages such as: official documents in the official administration (decisions, legal acts etc.); articles, reports, magazines, newsletters, etc.; sets of documents and their translations (parallel/comparable); Translation Memories (i.e. aligned text segments in the source and target language)
- Language or translation models
- Lexical and Conceptual resources, such as terminologies, glossaries, thesauri, wordnets, ontologies.

As mentioned above, ELRC mainly seeks to unlock language data that reside in public organisations across Europe, but also to collect sharable data from other potential data owners, such as research institutions, NGOs, etc. For instance, the Press and Information Office of Cyprus uploaded, among others, two datasets<sup>14</sup>: a) “Bilingual publications of the Press and Information Office of Cyprus” consisting of 9 pairs of EN-EL PDF documents, and b) “Press Releases (01.2018-01.2019) of the PIO” consisting of 13 XML files, which contain (in its internal specific structure) the Press Releases in Greek and the translation of some of them in English.

Considering them as raw datasets, ELRC exploited its internal pipeline to process the first dataset (i.e. text extraction from PDF files, identification of sentence pairs and

<sup>10</sup> <http://lr-coordination.eu/anchor-points>

<sup>11</sup> <https://elrc-share.eu>

<sup>12</sup> <http://www.lr-coordination.eu/events>

<sup>13</sup> <http://www.lr-coordination.eu/helpdesk>

<sup>14</sup> PIO publications and press releases

parallel corpus filtering) with the purpose of constructing a precision-high parallel corpus<sup>15</sup> in TMX format<sup>16</sup>. Similarly, ELRC developed custom scripts to parse the XML files of the second raw dataset, and then generated a parallel corpus of 5162 translation units (TUs).

ELRC additionally acts as the focal one-stop point where the language resources created by relevant CEF-funded projects are gathered, documented and made available to the EC or to the wider public, depending on their conditions of use.

Given the diversity of the potential contributors and stakeholders, ELRC put in place a multifaceted, yet straightforward and simple process for collecting and sharing LRs (Lösch et al., 2018). Depending on the size of the data set, the technological readiness, and the needs of potential contributors, participating organisations and individuals may:

- Upload the data directly to the ELRC-SHARE Repository<sup>17</sup>, as a zip file, through the corresponding online contribution form (see <https://www.elrc-share.eu/repository/contribute>).
- Send data and metadata files through the ELRC-SHARE-client API.
- Send data and metadata files through the ELRC-SHARE CEF eDelivery access point.

ELRC-SHARE (Piperidis et al., 2018) is a web-based platform designed to cover 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 extension and adaptation of its latest version, mainly in terms of the employed metadata schema, the user management module and the largely simplified operational workflow.

In addition to collecting contributions from external stakeholders, ELRC collects and processes language data from scratch with web crawling techniques (Papavassiliou et al., 2013; Papavassiliou et al., 2018). Web crawling is conducted using ILSP Focused Crawler (ILSP-FC), a comprehensive end-to-end solution for the acquisition of domain-specific monolingual and bilingual corpora from the web.

The ELRC partners 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 any official EU languages in CEF-affiliated countries.

Starting from a list of seed URLs (i.e., the homepages of the identified websites), the crawler fetches the web pages, extracts links from fetched web pages, adds the links to the list of pages to be visited and so on. During this process, modules for page fetching,

---

<sup>15</sup> Processed datasets: PIO publications and press releases

<sup>16</sup> TMX stands for “Translation Memory eXchange” - see <https://www.maxprograms.com/articles/tmx.html> for further details

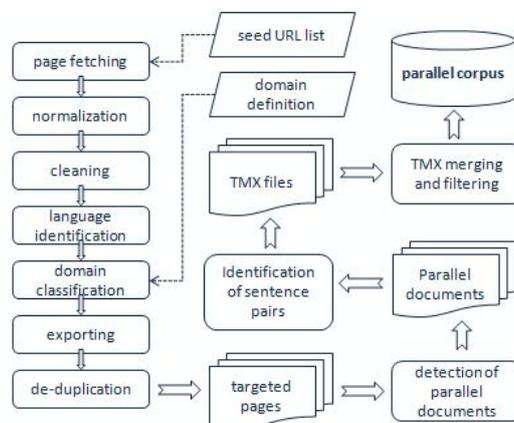
<sup>17</sup> <https://elrc-share.eu>

content normalization (i.e., conversion to UTF-8), boilerplate removal (i.e. "noisy" elements like navigation headers, advertisements, disclaimers, etc.) and language identification are used.

The content of each page is then compared to a user-provided domain definition, which consists of term triplets (<relevance weight, (multi-word) term, subdomain>) that describe the targeted domain. If the page is classified as relevant to the targeted domain, an XML file is generated containing basic metadata (e.g. title, URL, language, domain, etc.) and the content split into paragraphs. The set of XML files corresponding to domain-relevant webpages is then forwarded to further processing.

As the web contains many near-duplicate documents, a module for (near) de-duplication is exploited to eliminate the negative effect of duplicates in creating a representative corpus. Having collected the in-domain, de-duplicated sets of pages in the targeted languages, the next steps concern the detection of bitexts (i.e. pairs of documents that could be considered parallel) and the identification of sentence pairs in each document pair.

Finally, a battery of criteria is applied with the purpose of filtering out sentence pairs with potential alignment or translation issues, or of limited use for training MT systems. As an example, after crawling the "Science in Poland" website<sup>18</sup> of the Ministry of Science and Higher Education of Poland, an EN-PL parallel corpus in the "education" domain of about 28K TUs (Translation Units)<sup>19</sup> was automatically constructed. **Fig. 1** below summarizes the focused crawling process using the ILSP-FC within ELRC.



**Fig. 1.** ELRC focused crawling process using the ILSP-FC

Besides crawling websites of National Agencies, we also target websites of International organisations, and broadcast websites, which make their multilingual content available for use and process. To this end, the modules of the ILSP-FC toolkit (data

<sup>18</sup> <https://scienceinpoland.pap.pl/en>

<sup>19</sup> EN-PL "Science in Poland" corpus

acquisition, web page cleaning and normalization, detection of pairs of parallel documents, identification of sentence pairs) were applied on VoxEurop and constructed a multilingual (EN, DE, FR, ES, IT, PT, NL, CS, PL, RO) dataset of 927707 TUs in total<sup>20</sup>.

In case of any technical or legal questions around the preparation and/or submission of language resources, potential contributors can contact the ELRC Helpdesk (email: help@lr-coordination.eu, Skype: ELRC Helpdesk, phone: +33 970 440522).

### 3.2 Data Processing and Validation

To engage as many data holders as possible and to facilitate the sharing of language data, ELRC encourages contributions of language data in a digital format. The **processing of the contributed data** is then taken up by the ELRC language technologists in order to convert them to machine readable data, ready to be used for training LT systems. Each LR is analysed and processed by ELRC experts to ensure compliance with the *Language Resources Data Formats Specification* agreed with the EC. According to this specification, resulting parallel data should be provided in the TMX<sup>21</sup> 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.<sup>22</sup>

As indicated earlier, the ELRC Helpdesk team can support providers with a wide range of processing services, going from some basic data cleaning to more sophisticated text extraction from problematic formats such as PDFs or OCR-requiring scanned documents, etc. Processing has been performed for “The Udáras na Gaeltachta Corpus of bilingual PDFs and Word documents (Processed)”<sup>23</sup>, which was provided by the ADAPT Centre, DCU, Dublin, Ireland, and has undergone the following processing steps:

1. Text extraction from PDF and DOC documents.
2. Automatic document pair detection.
3. Automatic sentence alignment.

This has resulted in a clean and aligned Irish-English corpus which can be used by the community at large. All processing information about a particular resource is detailed in the corresponding Processing Report that is provided with the processed version of the corpus (together with the validation report, see below), so that ELRC can keep track

---

<sup>20</sup> VoxEurop corpus

<sup>21</sup> TMX stands for „Translation Memory eXchange”. TMX is an XML specification for the exchange of translation memories.

<sup>22</sup> TBX stands for “TermBase eXchange” TBX is an XML-based format for the representation and exchange of terminology data

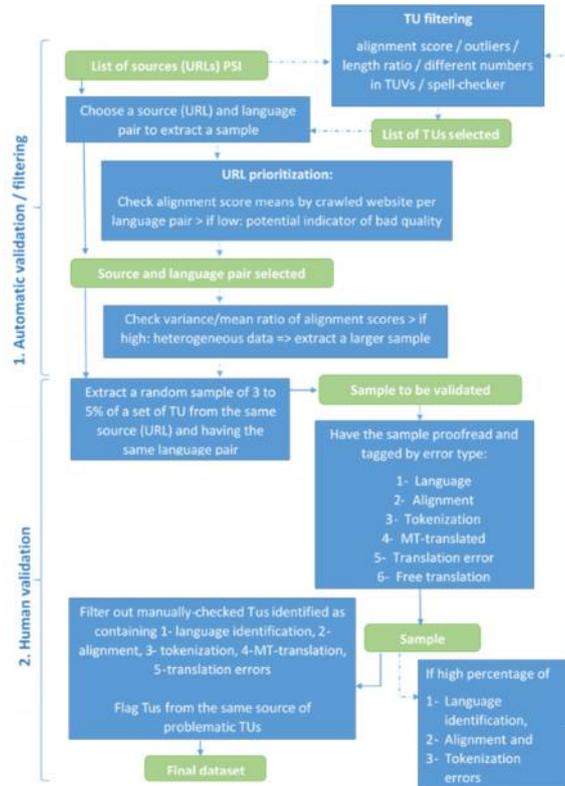
<sup>23</sup> <https://elrc-share.eu/repository/browse/the-udaras-na-gaeltachta-corpus-of-bilingual-pdfs-and-word-documents-processed/ed8a4632c35711e8b7d400155d026706a233557ae9d246eb8a7a0dec13f35e9a/>

of all data management steps performed and data users can check the processing steps followed.

ELRC also take care of the validation of their language resources. In the context of ELRC, **validation** is understood as the quality control of a LR against a list of relevant criteria (Schneller et al., 2018). It is important to note that due to the different processes of gathering the data and their quality level, the validation may be conducted in two different ways:

- **Quick Content Check (QCC):** it can be assumed that some 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 includes:
  - checking compliance of data with the ELRC objectives and scope,
  - checking the format of provided data,
  - checking that the metadata fields have been correctly filled in and are compliant with the data content, and
  - checking whether the legal information provided is compliant with the ELRC scope.
- **Extended content validation:** a deeper content validation may be considered necessary, for instance for data that derive from automatic processing (like crawling). On a legal point of view, even though crawling already delivers the format that corresponds to the LR production requirements, the list of crawled URLs was manually checked to assess if the websites are under the scope of the PSI (Public Sector Information Directive 2003/98/EC (modified in 2013 by the Directive 2013/37/UE)), for details see **Fig. 3** below). Content from websites that do not fall under the PSI Directive, or content that is not explicitly marked as open with a permissive license, must be excluded. On a technical point of view, the content validation may be followed along two steps:
  - automatic procedure: a series of processing steps may be considered necessary, such as cleaning the data, removing TUs whose quality can be deemed as poor by automated means.
  - manual procedure: this step may be undertaken only when the Editor(s) deem it necessary, e.g. for high-priority under-resourced languages where data quality should compensate for the lack of quantity. Errors in Translation Units (TU) need to be reported by human annotators. Depending on the acceptance percentage (eg. 10%) decided by the Editor(s), the LR that is declared as non-acceptable is discarded. In other particular cases, e.g. for high-volume high-priority LRs, the corresponding LR may be kept with indicators showing the probability of finding the same characteristics to help maximize the TU recall (e.g. by taking TUs marked as “Machine-translated text” or “Free translation” into account).

**Fig. 2** below summarizes the workflow employed within ELRC for a fine-grained content validation.



**Fig. 2.** ELRC Workflow for fine-grained content validation

In order to illustrate this, the aforementioned “The Udáras na Gaeltachta Corpus of bilingual PDFs and Word documents (Processed)” has also undergone validation. The procedure chosen for this specific case has consisted in performing both a QCC and an automatic validation performing spell-checking based TU filtering and TU length ratio-based filtering. As expected, the quality of the translated data constituting the corpus has been considered as high and the corpus has been validated for publication. The corresponding Validation Report has been produced. This Validation Report also includes the legal information checking that is based on the IPR clearance work (see following section for details).

Among the tools used in the data processing are DictMetric (Su & Babych, 2012) for document alignment, Microsoft Bilingual Sentence Aligner<sup>24</sup>, language detection tool PYCLD2<sup>25</sup>, and many others. TMX files are validated using TMXValidator<sup>26</sup>.

All types of language data are uploaded to the ELRC-Share Repository and corresponding metadata must be completed and validated. Finally, a Validation Report needs

<sup>24</sup> <https://www.microsoft.com/en-us/download/details.aspx?id=52608>

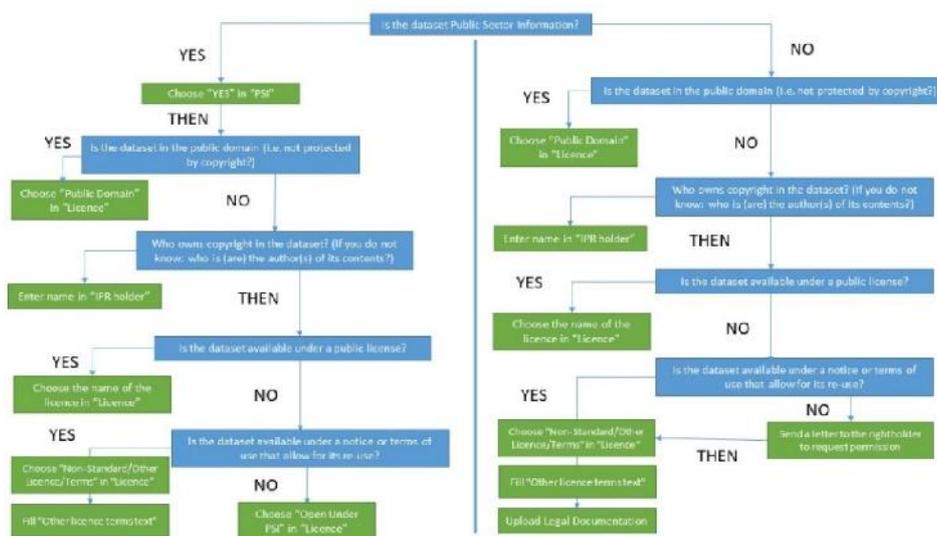
<sup>25</sup> <https://github.com/aboSamoor/pyclد2>

<sup>26</sup> <https://www.maxprograms.com/products/tmxvalidator.html>

to be provided for each data set, and all available legal related documentation asserts the quality of the data. ELRC's Validation Guidelines are available online through the ELRC website<sup>27</sup>.

### 3.3 IPR Clearance

**Fig. 3** below illustrates the overall workflow for legal clearance developed and employed by ELRC.



**Fig. 3.** ELRC Workflow for Legal Clearance

As shown in **Fig. 3**, in order to determine the appropriate license for a particular LR, several questions need to be assessed:

- Does the data fall within the scope of the PSI (Public Sector Information Directive 2003/98/EC (modified in 2013 by the Directive 2013/37/UE)?
- Is the data protected by copyright? (National laws may contain rules excluding certain works from copyright protection)
- If the data is protected by copyright, can I identify the owner of the copyright or the author of the work? (see IPR Holder field)
- Is the data available under a public license? For example, certain datasets are made available by the owner of copyright under a license that allows reuse or redistribution free of charge (e.g., cc licenses, NCGL 1.0, OGL 3.0 etc.)

<sup>27</sup> [http://www.lr-coordination.eu/sites/default/files/common/ELRC-SHARE%20repository\\_Guidelines%20for%20generic%20services%20projects\\_merged%20-%20FINAL%2020200227.pdf](http://www.lr-coordination.eu/sites/default/files/common/ELRC-SHARE%20repository_Guidelines%20for%20generic%20services%20projects_merged%20-%20FINAL%2020200227.pdf).

- If no public license is clearly marked on the document, you should check the terms of use or if any documentation may help you determine the conditions of reuse of the material.

It is important to point out that most of the aforementioned legal issues are debated on fora organized by the ELDA team in charge of the legal helpdesk. For instance the legal workshops organized as satellite events of LREC are major sources of knowledge and input that help share the information about these issues within the community and allow to get a clear picture of the various IPR contexts in different countries (Choukri et. al, 2020).

If we go back to the example of the “The Udáras na Gaeltachta Corpus of bilingual PDFs and Word documents (Processed)” presented in the Data processing and Validation section, this resource was released under the legal status “Open under PSI” and was then considered as a valid resource that could be published and shared widely. The validation report also provides the *Attribution text* and the pre-existing rights to be considered.

## 4 Summary and Conclusions

Having received the first resources starting in Spring 2016, ELRC has managed to collect almost 2,500 LRs and corresponding tools, covering all official EU languages, plus Icelandic and both varieties of Norwegian, Bokmål and Nynorsk. This amounts to more than 200 billion words in all EU languages, including 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. More than 60 public sector organisations across Europe have shared their language data with ELRC, including in particular national ministries, governmental bodies and public services. However, in order to make all this data available and re-usable for the development of MT systems, a dedicated validation and clearing process is necessary involving both technical (manual and automatic) and legal evaluation. The ELRC workflows and infrastructure that are in place facilitate a sustainable language data sharing, storing, documenting and rendering cycle, thus unlocking data for training Language Technology systems, for the benefit of the LT community across Europe.

## 5 Acknowledgements

The European Language Resource Coordination (ELRC) is a service contract operating under the EC’s Connecting Europe Facility SMART programme (starting from SMART 2014/1074 in April 2015, continued 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” until end of 2021 within SMART 2019/1083 “Action on Automated Translation Core Service Platform (CSP)”).

## Bibliographical References

1. 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]
2. Schneller, P.; Fernandez-Barrera, M.; Mapelli, V; Popescu, V.; Choukri, K.; Arranz, V.; Giagkou, M. Prokopidis, P; Papavassiliou, V., Rozis, R. (2018). European Language Resource Coordination – Validation Guidelines. Available at [https://lr-coordination.eu/sites/default/files/common/Validation\\_guidelines\\_CEF-AT\\_v6.2\\_20180720.pdf](https://lr-coordination.eu/sites/default/files/common/Validation_guidelines_CEF-AT_v6.2_20180720.pdf) [last accessed: 11.01.2020]
3. CEF: <https://ec.europa.eu/digital-single-market/en/connecting-europe-facility> [last accessed: 22.02.2018]
4. CEF eTranslation: <https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation>
5. Directive 2013/37/UE: <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013L0037> [last accessed: 22.09.2017]
6. Choukri, K.; Lindén, K., Rigault, M.; Siegert, I. (2020): Proceedings of the LREC2020 Workshop on Legal and Ethical Issues, Marseille, France, European Language Resources Association (ELRC). Available at: <https://www.aclweb.org/anthology/2020.legal2020-1>.
7. European Parliament: Report on Language Equality in the Digital Age (2018/2018(INI)), 2018.
8. ELRC-SHARE: <http://www.lr-coordination.eu/resources> [last accessed: 22.02.2018]
9. Lösch, A.; Mapelli, V.; Piperidis, S.; Vasi jevs, A.; Smal, L.; Declerck, T.; Schnur, E.; Choukri, K.; van Genabith, J. (2018): “European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management”, In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, European Language Resources Association (ELRA).
10. 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).
11. Papavassiliou, V., Prokopidis, P. & G. Thurmair. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, pages 43-51. Sofia, Bulgaria: Association for Computational Linguistics
12. 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).
13. Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey. European Language Resources Association (ELRA).
14. Smal, L.; Lösch, A.; van Genabith, J.; Giagkou, M.; Declerck, T.; Busemann, S. (2020): “Language Data Sharing in European Public services – Overcoming Obstacles and Creating Sustainable Data Sharing Infrastructures” in: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, , Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.): Proceedings of the Twelfth International Conference on Language Resources

- and Evaluation (LREC 2020), Pages 3443-3448, Marseille, France, European Language Resources Association (ELRA).
15. 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.
  16. Vasiļevs, A.; Rozis, R.; Kalniņa, R.; Bērziņa, 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).