Language Resources to Support Language Diversity – the ELRA Achievements

Valérie Mapelli, Victoria Arranz, Hélène Mazo, Khalid Choukri
ELDA/ELRA
9, rue des Cordelières, 75013 Paris, France
Email: {mapelli, arranz, mazo, choukri }@elda.org

Abstract

This article highlights ELRA's latest achievements in the field of Language Resources (LRs) identification, sharing and production. It also reports on ELRA's involvement in several national and international projects, as well as in the organization of events for the support of LRs and related Language Technologies, including for under-resourced languages. Over the past few years, ELRA, together with its operational agency ELDA, has continued to increase its catalogue offer of LRs, establishing worldwide partnerships for the production of various types of LRs (SMS, tweets, crawled data, MT aligned data, speech LRs, sentiment-based data, etc.). Through their consistent involvement in EU-funded projects, ELRA and ELDA have contributed to improve the access to multilingual information in the context of the pandemic, develop tools for the de-identification of texts in the legal and medical domains, support the EU eTranslation Machine Translation system, and set up a European platform providing access to both resources and services. In December 2019, ELRA co-organized the LT4All conference, whose main topics were Language Technologies for enabling linguistic diversity and multilingualism worldwide. Moreover, although LREC was cancelled in 2020, ELRA published the LREC 2020 proceedings for the Main conference and Workshops papers, and carried on its dissemination activities while targeting the new LREC edition for 2022.

Keywords: Language Resources, language diversity, under-resourced languages, Language Technology/Language Resources infrastructures, anonymization

1. Introduction

Despite the large number of Language Resources (LRs) produced and shared for several decades by international data centers like ELRA (European Language Resources Association)[1], the need for them does not diminish. Also, the growth of communication means, with the extended use of social media changes the language schemas and creates new demands. With over 7500 existing languages in the world, and an improved access to technologies for a growing number of communities, expectations are far from being fulfilled.

In that context, ELRA is striving to increase its offer of LRs through its distribution and production activities, always focusing on resources of interest for the ever emerging fields of Language Technologies (LT). ELRA also gets involved in the development of relevant infrastructures, including for under-resourced languages, such as infrastructures for sharing LRs and related technologies like the European Language Grid (ELG) or the European Language Resources Consortium (ELRC), which focuses on open access data for supporting the production of Machine Translation (MT) systems. Moreover, ELRA promotes the use of LRs at an international level, in particular with the regular organization of the now well-known Language Resources and Evaluation Conference (LREC) or within its contribution to the organization of other events such as ELRC Country Workshops.

This article highlights ELRA's major achievements over the past few years, starting with its recent actions in identifying and distributing LRs in an ever-evolving field, then elaborating on the newest expertise acquired in terms of data production, as well as presenting its current infrastructures undertakings, and concluding with an overview of the latest events and other information dissemination activities.

2. LR Identification & Distribution

One of the main missions of ELRA is to make available in a sustainable way LRs that are being produced worldwide. This goes through the maintenance of a repository of LRs, as well as the exploitation of various channels to make those data visible and thus accessible to the widest audience.

2.1. The ELRA Catalogue of Language Resources

Back in 1999, ELRA's Catalogue of Language Resources offered only 155 LRs. Now, this number reaches 1,500 LRs (see Figure 1) covering many modalities (speech/audio, video, sign languages, OCR-images, texts, lexica, and terminology, including data for evaluation purposes).

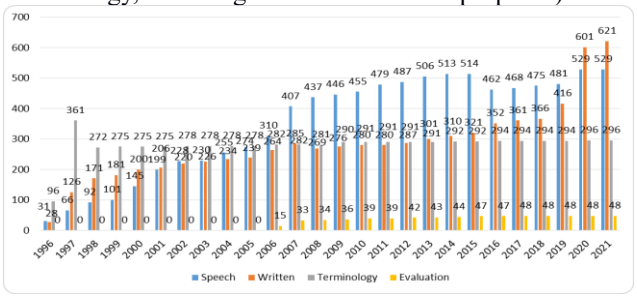

Figure 1: LRs in the ELRA Catalogue (as of 31/12/2021)

The labour-intensive identification work achieved by ELRA, which consists in discovering new datasets all over the world, as well as in participating constantly in various projects that follow the market evolution, has resulted in the diversification of the offer in terms of languages. The ELRA catalogue now covers nearly 80 languages (including language variants and sign languages). Among the new languages covered in the catalogue we can mention less-resourced languages like Manipuri, Khasi or Esperanto. A boost in the offer for written LRs can be explained by the ever-growing interest of multilingual

[1] http://www.elra.info/en/

corpora and lexica for the improvement of MT Systems which proved a highly re-emerging field over the past few years. From time to time, a number of LRs have been removed from the catalogue for various reasons including contract ending with the provider, new packaging or bad quality, which explains some fluctuations over the years.
ELRA can rely on a solid network of international providers to increase the number and the quality of LRs available constantly. For instance, over the past three years, ELRA has concluded an agreement for the release of 50 monolingual and multilingual lexicons from the CJK Dictionary Institute, Inc[2] which specializes in Arabic, Chinese, Japanese and Korean lexicography. In collaboration with SpeechOcean[3], ELRA has incorporated 46 new speech LRs for various European and Asian languages to offer a single-stop shop for its members. ELRA has also added 18 text corpora and bilingual/trilingual dictionaries for Vietnamese provided by the Kimtudien Multilingual Data Center[4].

2.2. Making LRs visible

2.2.1. New approaches to search and find ELRA LRs on the web

To raise awareness about existing LRs, from its earliest years, ELRA adopted a catalogue way system for presenting LR descriptions, further developed as extended metadata. As part of ELRA's collaboration framework, the catalogue metadata is exported as an OLAC (Open Language Archives Community) archive[5], which is a worldwide virtual library dedicated to LRs, as well as into the CLARIN Virtual Language Observatory (VLO)[6], which serves a broader circle of SSH (Social Sciences and Humanities) scholars and beyond - for instance VLO is part of the EOSC (European Open Science Cloud) market places[7]. In 2020, the XML version of the ELRA Catalogue[8] used by such channels was updated to export up-to-date metadata information. This new XML version can be freely used by any additional channels.
Moreover, since the beginning of 2021, the LRs from the ELRA Catalogue can now be searched and found on Google Dataset Search[9] and on the ELG Language Technology platform[10] (developed within the European Language Grid project). To allow the indexing by Google Dataset Search, ELRA has updated the code generating the catalogue pages. The code developed follows the schema.org standard and is publicly available in JSON format so that it can be used for other harvesting purposes.

2.2.2. Fostering the dissemination of LRs worldwide

ELRA's mission is not limited to making LRs available through its own catalogue. Its aim is also to foster the widest dissemination possible of LRs worldwide. For instance, in November 2021, ELDA (Evaluations and Language resources Distribution Agency), working as ELRA's operational body, opened a European mirror of the Open Speech and Language Resources (OpenSLR) website[11] in cooperation with Johns Hopkins University, which hosts speech and other LRs, such as training corpora for speech recognition, and software related to speech recognition. As of today, this corpus comprises 113 LRs representing almost 2 terabytes of data.

2.2.3. ISLRN latest updates

Since 2016, ELRA has encouraged providers/producers of LRs to use a Persistent Unique Identifier specifically created for referencing LRs: the International Standard Language Resource Number (ISLRN)[12]. The ISLRN number is also part of information collected during paper submissions to LREC, ELRA's conference.
At the end of December 2021, the total number of LRs which were allocated an ISLRN number amounts to 3177 which cover nearly 250 languages. Strong collaborations are going on with the Linguistic Data Consortium[13] which manages the ISLRN submissions in partnership with ELRA and has now 888 ISLRN numbers associated with LRs distributed in their Catalogue.

3. Production projects

ELRA, together with its distribution agency ELDA, has a long experience in supporting international enquiries to produce customized LRs for various languages. Relying on a large network of international partners, ELRA has been involved in the production of LRs both in the framework of European and international projects or, in support of companies or institutions. All types of technologies may be addressed: automatic speech recognition, broadcast news transcription, spoken language translation, spoken language understanding, speaker recognition, text-to-speech synthesis, acoustic person tracking, speech activity detection, OCR, annotated text for sentiment analysis, as well as image- and video-oriented technologies. Over 50 languages were compiled from all continents[14]. Furthermore, areas such as data alignment (for MT training, for instance) and data de-identification (for data requiring anonymization in order to be used under GDPR-compliant terms) are also part of ELRA's expertise in re-purposing existing LRs. Both services are offered by ELRA and have also been performed under the umbrella of funded initiatives such as ELRC and MAPA, respectively (both described further down).
In the following sections, we present the extended list of production activities carried out in the past few years. Although we cannot divulge the name of some partners, as some production activities were done under confidential

[2] http://www.cjk.org/cjk/index.htm
[3] https://en.speechocean.com/
[4] https://www.kimtudien.com.vn/
[5] http://www.language-archives.org/archive/catalogue.elra.info
[6] https://vlo.clarin.eu/
[7] https://marketplace.eosc-portal.eu/services/virtual-language-observatory/information
[8] http://catalog.elra.info/elrac/elra_catalogue.xml
[9] https://datasetsearch.research.google.com/
[10] https://www.european-language-grid.eu/
[11] https://openslr.elda.org
[12] http://www.islrn.org
[13] https://www.ldc.upenn.edu/
[14] This figure of 50 languages corresponds to LRs directly produced by ELRA/ELDA and shall not be compared with the 80 languages mentioned in section 2.1, which is the number of languages covered in the ELRA catalogue and gathered from various other producers.

contracts, still they highlight the expertise acquired within such projects.

3.1. Corpora for Speech Technologies

3.1.1. MGB-5 Moroccan and MGB-3 Tunisian Dialect databases

The MGB-5 Moroccan Dialect database was built as part of a project carried out in partnership with QCRI (Qatar Computing Research Institute[15]) to support the MGB-3 and MGB-5 Challenges (Multi-Genre Broadcast) focusing on evaluation of speech recognition systems for dialectal Arabic languages. The first phase targeted Moroccan Arabic dialect for which about 14 hours of data were collected and transcribed. The next phase was dedicated to the collection of a similar amount of data for Tunisian Arabic dialect. The MGB-5 Moroccan Dialect database is under cataloguing and will be available soon via ELRA.

3.1.2. Corpus for the Rosetta Project

The production of this corpus consisted in manually transcribing 20 hours of TV shows, broadcasts, documentaries, and series. Special attention was paid to the exceptional accuracy of compliance with the norms of modern French spelling. This corpus was produced for the benefit of LIMSI-CNRS, now LISN-CNRS (France), within the Rosetta Project[16] for the development of an automatic subtitling system primarily for deaf people, but also for hearing people with a limited command of French (for example students of French language or foreigners immersed in French culture). The system is based on artificial intelligence methods (recurrent neural networks and Deep Learning). Rosetta intends to enable and facilitate the production of captioning and multilingual subtitles on a large scale for all types of video content (television programs, online video, MOOC, YouTube, etc.) and to offer a Sign Language representation of this video content through the animation of virtual or digital signers.

3.1.3. Tamasheq Corpus for Automatic Translation

This corpus was produced as part of a project aiming at building an automatic translation system for some African vernacular languages and conducted in partnership with the LIA (Laboratoire Informatique d'Avignon, France)[17]. Tamasheq, a Touareg language spoken in Niger and Mali, is the first language addressed by the project. As a first step, native speakers have translated the Niger's broadcast news from Tamasheq into French. Tamasheq being a spoken language in which a number of domain-specific terms (economy, science, politics) do not exist, made the process difficult. Native speakers had to substitute those terms with their equivalent in French, Arabic, or other African languages. However, the rigorous objective remained: respecting all the nuances of interpretation on the one hand and trying to enrich the vocabulary of the automatic translation system on the other hand.

3.1.4. Audio Data Annotation for Speaker Identification

The corpus was produced within the framework of the European research program Chist-ERA[18], in collaboration with the LNE (Laboratoire national de métrologie et d'essais, France)[19], the University of Le Mans (France)[20], the Polytechnic University of Catalonia (Spain)[21] and IDIAP (Switzerland)[22]. The production consisted in annotating audio data based on video files to identify 5901 unique speakers. The annotated corpus of 328 hours of audio files from the French channel LCP programs is to be used by the LNE's Artificial Intelligence Systems Evaluation team and its partners for research and evaluation of automatic Lifelong Learning systems for speaker recognition.

3.1.5. Other Speech corpora for private partners

- **Multilingual conversational telephone speech corpus:** Work on this multilingual corpus based on phone call conversations corpus started in 2020 and is still on-going. Arabic, English, French, German, Italian, Korean, Mandarin and Cantonese Chinese, Portuguese, Russian are among the languages addressed in this project. Orthographic transcription for spoken language is always a tricky operation. For Cantonese, it turned out to be even more complicated as the transcription was made in simplified Chinese instead of the traditional version, normally used for this language. As a result, the transcription process also encompassed translation from one language to another since some Cantonese syllables could not be written with simplified characters and had to be translated into standard Chinese.
- **Speaker Identification Corpus for French:** Producing this annotated corpus consisted first in identifying and collecting audio flows selected to cover a wide variety of topics (at least 50 different contexts) for 400 different speakers, then annotating the data to identify the speakers (name, gender) while segmenting the part of speech. Speaking in French was mandatory, and accents were allowed. The recordings were taken from television or radio programs. The annotated corpus is meant to be used for research and evaluation of automatic voice comparison systems.

3.2. Corpora for Written Technologies

3.2.1. Rephrasing a Q&A corpus to improve a conversational system

Commissioned by Orange Innovation[23], this research and development project's objective was to get natural questions in English to improve the partner's conversation system. This was achieved by rephrasing the questions of the American corpus CoQA (Conversational Question Answering), produced by the Stanford NLP Group[24]. This corpus contains 127k+ questions with answers collected from 8k+ conversations. The project started in November 2020 and latest revisions were produced in January 2022. Further details are provided in a distinct paper submitted at

[15] https://www.hbku.edu.qa/
[16] https://rosettaccess.fr/
[17] https://lia.univ-avignon.fr/
[18] https://www.chistera.eu/
[19] https://www.lne.fr/
[20] http://www.univ-lemans.fr/
[21] https://www.upc.edu/
[22] https://www.idiap.ch/
[23] https://orangefabfrance.fr/fr/data-ia/
[24] https://nlp.stanford.edu/

LREC 2022 (Brabant et al. 2022). The corpus will be made available by ELRA.

3.2.2. Annotated tweet corpus in Arabizi, French and English

This corpus was built by ELDA on behalf of INSA Rouen Normandie (Normandie Université, LITIS team)[25], in the framework of the SAPhIRS project (System for the Analysis of Information Propagation in Social Networks), funded by the DGE (Direction Générale des Entreprises, France). The purpose of the corpus constitution, completed in 2020, was to collect and annotate Tweets in 3 languages (Arabizi, French and English) for 3 predefined themes (Hooliganism, Racism, Terrorism).
In total, 17103 sequences were annotated (~ 585 K tweets). Among these sequences, 4578 sequences (~ 127 K tweets) having at least 20 tweets annotated with the 3 predefined themes (Hooliganism, Racism, Terrorism) were obtained, 1866 sequences with an opinion change and 8733 hateful tweets.
More details and the actual corpus are available through the ELRA Catalogue[26].

3.2.3. Written corpora for private partners

- **Short messages (SMS) corpus:** Covering Arabizi (Arabic written with Latin characters), Moroccan and Tunisian Arabic, this corpus consists of several tens of thousands of SMSs collected through an in-house Web application that also allowed data management and quality control.
- **Corpus of crawled documents:** The data collection of a few thousands of documents was ensured via focused crawling, data annotation with respect to the topic(s) and subtopics of each document. Stringent requirements, at both the collection and annotation phases, had to be respected. The ELDA web crawling infrastructure was customized (using ELDA's open-source crawled data management toolkit[27] developed within the ELRC project), and a Web platform tailored to the project's document annotation specifics was developed. The expertise acquired in this project was re-used for other activities such as the COVID19 MLIA initiative (described further down).
- **English-Arabic Parallel Corpora:** The objective of this project, completed in 2020, was to carry out the construction of parallel data (English-Arabic) in 6 different topics in order to train, improve and evaluate a MT system. Using existing tools, the ELDA team crawled parallel web content then aligned English-Arabic parallel sentences in the given topics. Two types of data normalization were carried out: one following the rule-based normalization method and one using an AI model trained on data. The data normalization enabled the deletion of the following information: duplicated sentences, sentences already present in the test and training corpora, sentences with low alignment scores, sentences with alignment score below 1.1 threshold (threshold as defined in Facebook Research evaluation campaigns), sentences in languages different from English and Arabic.

For the named entities, English and Arabic models were used to identify sentences containing at least one named entity. The proposed pre-annotation was then validated or corrected by a language expert team.
In the end, the following corpora were delivered:
Training corpus containing ~ 360K parallel sentences (~ 9 million tokens in English and ~ 8 million tokens in Arabic),
Test Corpus of 6.2k parallel sentences,
Named entities Corpus with ~ 1500 named entities.

4. Infrastructure / European Projects

Below are presented ELRA's various undertakings into cooperative infrastructures for the sake of LRs and LT preservation and sharing.

4.1. European Language Resource Coordination+

The European Language Resource Coordination (ELRC)[28] set up within the CEF.AT Programme of the European Commission (Lösch et al. 2018) continues its actions for the support of the EU eTranslation Machine Translation System through several contracts with the EC since 2015. The latest one, SMART 2019/1083, is ongoing until the end of 2022. This series of contracts led to the implementation of a large repository[29] dedicated to MT which grows continuously and now amounts to above 2,530 LRs.
Through its experience in market analysis as well as in managing and communicating on various aspects of LT issues at the European level, ELDA has become a valuable partner to take part in several EU service contracts on those aspects. Thus, in parallel to ELRC actions, ELDA has also contributed to other EU service contracts on behalf of the EU DG Communications Networks (DG CNECT), including a study on service portfolio development and business case for CEF Automated Translation.

4.1.1. Implementation and maintenance of eTranslation and ELRC Helpdesks

Since December 2017, as part of the Action on CEF Automated Translation Core Service Platform (SMART 2019/1083, handed over from the previous action SMART 2016/0103 LOT 2), ELDA is the lead partner of the eTranslation "Customer Service Desk". As part of this task, ELDA capitalized on its expertise in supporting the LT community on all issues related to data management, legal and technical aspects, and has set up and runs this helpdesk that responds on a day-to-day basis to needs expressed by the users of the CEF Automated Translation services (known as "eTranslation")[30].
eTranslation is available to public administrations in EU countries, Iceland, and Norway, as well as EU institutions and agencies. It can be used either through an online user interface or integrated via an API into information systems to make digital public services and content multilingual.
The main objective of the task is to gain extensive experience in the use and integration of the eTranslation and other MT engines so as to help and assist the public sector technical teams in adopting the eTranslation engine in their operation.

[25] https://www.insa-rouen.fr/recherche/laboratoires/litis
[26] http://catalog.elra.info/en-us/repository/browse/ELRA-W0323/
[27] https://github.com/ELDAELRA/elda_cmtk
[28] http://lr-coordination.eu
[29] https://elrc-share.eu/
[30] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

Since 1st November 2018 and until 31st December 2021, a total of 4,182 requests were managed by the ELDA team. Requests range from registration to eTranslation online user interface, documentation needs, technical issues related to the integrated web service, up to troubleshooting or other support issues. From time to time, ELDA also provided online demonstrations of eTranslation upon demand of some users.

Moreover, ELDA manages the ELRC helpdesk related to the use, production, collection, processing, and sharing of LRs raised by users of ELRC services. Last November, the legal helpdesk[31] provided with a use case document (Rigault et al. 2021) that analyzes under which legal conditions audio, video and dialogue subtitles coming from emergency calls embedded in a German TV show can be re-used for developing AI models[32].

4.1.2. Organization of Country Workshops

Within the framework of the latest contract, 29 workshops are being organized in all the countries taking part in ELRC[33]. Due to the sanitary context, the events that took place in 2020 and 2021 were all held online. Both raising awareness about the importance of domain-specific language data to improve eTranslation performances as well as encouraging participants from public sector to contribute to the collection of LRs have remained strong objectives of the country workshops. The eTranslation latest developments have also been thoroughly presented to the workshops' audience. These include free access extended to SMEs, addition of new languages (Arabic, Chinese, Japanese, Russian, among others), integration into websites or CAT tools through APIs, etc. Also, the focus has been broadened to allow exchanges on both the importance of LT in the daily digital interactions of the Europeans (digital assistants, chatbots, etc.) and the growing place and role of Artificial Intelligence in technology developments. Several Member States, including Spain or France, have already put together national strategies for AI.

With the support of the public and scientific Anchor Points in each country, ELDA is responsible for the country workshops in Belgium, France, Ireland, Italy, the Netherlands, Malta, Portugal, Romania, and Spain.

4.1.3. Study on service portfolio development and business case for CEF Automated Translation

This one-year project (Dec 2017 to Dec 2018), funded by the EC under SMART 2016/0103 – LOT 1 call for tender, aimed at analyzing the LT Market at EU and Member State level, including a competitiveness analysis in at least 3 areas of LT, including Machine Translation, as well as LT Services and solutions currently in use by Public Administrations in the EU. It also had the objective to identify the value proposition of CEF Automated Translation and description of its position in the European LT market/ecosystem. ELDA's task focused on the organization of a survey on LT services and solutions currently in use by Public Administrations in the EU. An online questionnaire was developed and sent to targeted respondents in each of the EU countries, Iceland, and Norway. The target to obtain 50-80 completed questionnaires spread over different countries was finally attained with 79 respondents completing the questionnaire. Final version of the report was submitted to the Contracting Authority on 21 December 2018.

The survey results were gathered in a report published by the Publications Office of the European Union (European Commission et al. 2019)[34]. A related analysis was also produced within this project about the European Language Technology Market (Choukri 2018).

4.2. COVID MLIA, the European initiative to improve the Multilingual Information Access

COVID-19 MLIA is an initiative endorsed by the European Commission's DG CNECT[35]. Coordinated by the University of Padua[36] and ELRA/ELDA, it has been launched in June 2020 to improve the access to multilingual information, including but not limited to health-related content, in the context of the pandemic, when information amount and reliability, as well as their availability in many languages, were challenging.

Overall, the initiative will consider three rounds of evaluation, looking into adding new topics incrementally, improving the language coverage by extending the number of less-resourced EU languages and fostering cross-fertilization between the tasks.

In 2021, two rounds of evaluation were completed[37].

Structured around two main poles: LRs and Evaluation of the MLIA system, the tasks covered are:

- Data acquisition, led by ILSP[38] and JRC[39], whose main objective is to boost the development of datasets and corpora specifically related to Covid-19 issues and make them available to the community from the Data section[40] on the COVID MLIA website.
- Information Extraction, led by DFKI[41] and LISN[42] (former LIMSI), whose main objective is to identify relevant medical information in texts related to the COVID-19 issue with a language coverage consisting of English, German, Modern Greek, Italian and Spanish.
- Multilingual Semantic Search, led by University of Padua[43] and CLARIN ERIC[44], whose main objective task is to collect relevant information for the community, the general public including other stakeholders, when searching for health content in different languages and with different levels of knowledge about the specific topic. The languages addressed within this task are English, French,

[31] https://lr-coordination.eu/helpdesk

[32] http://www.elra.info/media/filer_public/2021/11/19/use_case_-_tv_shows_official-final.pdf

[33] https://lr-coordination.eu/events

[34] https://op.europa.eu/en/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en/format-PDF/source-106906783

[35] http://eval.covid19-mlia.eu/

[36] http://ims.dei.unipd.it/

[37] http://eval.covid19-mlia.eu/meetings/

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

[39] https://ec.europa.eu/jrc/en

[40] http://data.covid19-mlia.eu/

[41] https://www.dfki.de/en/web/

[42] https://www.lisn.upsaclay.fr/

[43] https://www.unipd.it/en/

[44] https://www.clarin.eu/

German, Italian, Modern Greek, Spanish and Swedish. Ukranian was also planned but no Ukrainian submissions took place and there is no ground truth for this language.

- Machine Translation Task, led by Universitat Politècnica de València[45] and Pangeanic[46], whose main objective is to assess the capabilities of the MT systems to translate texts related to Covid-19, comprising new terms and expressions. The first rounds of evaluation covered language pairs from English to each of the following languages: German, French, Spanish, Italian, Modern Greek and Swedish.

Furthermore, the second round of evaluation covered not only the language pairs mentioned above but also the English-Arabic pair.

All the resources produced during the evaluation rounds are available on the git repositories[47] of the initiative, under the CC-BY-SA 4.0 license. The LRs, processed and duly cleared, are made available progressively as evaluation packages from the ELRC-Share repository and the ELRA catalogue. More information can be found on the COVID-MLIA[48] website and COVID-19 MLIA Youtube Channel[49].

4.3. European Language Grid (ELG)

The European Language Grid is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825627. ELG is a 3.5-year project, which started in January 2019 and will run until June 2022. It counts on partners from different expert groups in a variety of languages: DFKI (coordinator - Germany), Charles University (Czech Republic), ELDA (France), Expert System (Spain), HENSOLDT (Austria), ILSP (Greece), Tilde (Latvia), University of Edinburgh and University of Sheffield (UK).

ELG aims at providing a European platform where to access both LRs and services in a joint and functional manner. This platform allows to execute hosted services on the available datasets, simplifying users' tasks when accessing both.

ELDA is responsible for LRs, their conversion, ingestion, and exploitation within ELG. Work has been done for the integration of a large number of repositories, working on conversion mechanisms and with large revision efforts (such as, for instance, for Zenodo[50] and Quantum Stat[51]). This activity is carried out in close collaboration with the work on the ELRA Catalogue, which will be further supporting the HLT community by increasing access to all relevant datasets.

ELDA is also leading the legal activity behind the platform, together with the definition of its Data Management Plan while ensuring GDPR compliance. In this regard, ELDA's in-house legal counsel has carried out a detailed analysis of over 300 licenses (Arranz et al. 2022), more than 100 already in use within the ELG platform, and the remaining ones to complete the analysis for all licenses listed in the SPDX list[52].

This revision has intended to make the conditions of use as clear as possible for the user, focusing on aspects such as:

1. The rights granted by the licensor,
2. The requirements imposed on the user for the redistribution and publication of the resources, and
3. The conditions imposed on users regarding the reuse of the data.

All these legal issues and others relevant to the field of HLT will be discussed at the Legal and Ethical Issues Workshop (Legal2022)[53] that will be held at LREC 2022.

4.4. Multilingual Anonymisation for Public Administrations (MAPA)

The MAPA[54] project is an EC-funded Action under the Connecting Europe Facility (CEF) – Telecommunications Sector with Grant Agreement No INEA/CEF/ICT/A2019/1927065. MAPA started in January 2020, and it has finished in December 2021. The Consortium consists of the following partners: Pangeanic (project coordinator), SEDIA and Vicomtech (Spain), ELDA and LISN-CNRS (France), Tilde (Latvia), University of Malta (Malta). ELDA has led the data collection, sample selection and annotation activities.

MAPA has developed a multilingual toolkit for the de-identification of texts in the legal and medical domains using NER techniques to detect the sensitive entities that require anonymization. The project has addressed the 24 EU languages and has implemented both monolingual and multilingual detection models with success. Models have been trained based on the data that the project has collected and annotated. The MAPA toolkit is a deployable, dock-ready, open-source system that can be easily trained on other languages and domains.

As responsible for the data activities, ELDA has coordinated all work around the project's data:

- Definition of sensitive information to be anonymized and design of annotation guidelines. These guidelines have defined the named entity hierarchy to be considered for both legal and medical domains (e.g., person names, addresses, phone numbers, email addresses, etc.).
- Sample data collection for annotation: legal and medical documents contain very sensitive information and are not generally available for public use. MAPA has made use of alternative data (EUR-LEX)[55] that contains relevant entities and can be used for model training. A further added value is that EUR-LEX is available in 23 languages, with missing Gaelic data being generated through translation with the eTranslation platform and then annotated. The output data are parallel annotated corpora in the 24 languages.
- Data collection and annotation: further annotated datasets have been built from a variety of sources and following different techniques in a subset of MAPA's languages. These have tried to make use of existing monolingual data that were relevant for further training but had to be somehow processed to be used (e.g., French and

[45] https://www.prhlt.upv.es/
[46] https://www.pangeanic.com/
[47] https://bitbucket.org/covid19-mlia/
[48] http://eval.covid19-mlia.eu/
[49] https://bit.ly/3zS1l2j
[50] https://zenodo.org/
[51] https://quantumstat.com/
[52] https://spdx.org/licenses/
[53] https://lrec2022.lrec-conf.org/media/filer_public/ab/77/ab774168-b769-4ca3-8d13-5982839feb93/cfp-lrec-legal2022-workshop.pdf
[54] https://mapa-project.eu/
[55] https://eur-lex.europa.eu/

Greek legal data which has had to be de-anonymized, French clinical cases which have been enriched with personal data).

- 1M sample-sentence raw corpora have also been produced for all 24 languages for machine learning. These have been collected from several sources.
- Production of synthetic data to explore further model training: the automatic creation of parallel annotated data has also been explored in further training.

Both data and software will be made available through the ELRA catalogue, as well as ELG and ELRC infrastructures The MAPA toolkit has also been tested in several use cases, one of them being the EC's Directorate-General for Translation, and it has been integrated as part of their NLP services offer[56]. In that regard, the ELRC initiative has a task devoted to anonymization which is performing the analysis and test of the MAPA toolkit to carry out the potential anonymization of ELRC datasets requiring this service.

Finally, ELRA will be supporting the community in their anonymization needs through the offering of a series of related services deriving from the deployment of the MAPA toolkit in-house.

5. Information dissemination

Facilitating the communication and interaction between LT various stakeholders is also one of ELRA's missions. The sections below present the latest actions in this respect.

5.1. Previous editions of LREC

As it is well known, ELRA organizes the Language Resources and Evaluation Conference (LREC) biennially since 1998[57]. LREC 2020 should have been held in Marseille in May 2020. However, due to the breakout of the COVID-19 pandemic, the conference had to be cancelled. The Programme Committee decided to convene the 13th edition in Marseille in 2022. Many workshops managed to schedule digital events at a later stage. All proceedings, for the main conference and workshops, have been produced and are available online[58]. The 2020 Shared LRs set was also published[59].

5.2. LT4All (International Conference Language Technologies for All)

The International Conference Language Technologies for All (LT4ALL): Enabling Diversity and Multilingualism Worldwide conference[60] took place on December 4-6, 2019 at the UNESCO Headquarters in Paris (France), as part of the UNESCO International Year of Indigenous languages 2019. The International Conference was organized by UNESCO[61], the Government of the Khanty-Mansi Autonomous Area[62] and ELRA, with its Special Interest Group: Under-resourced Languages (SIGUL)[63], in partnership with the UNESCO IFAP and the Interregional Library Cooperation Centre.

Two days of the conference were dedicated to Language Technologies with sessions that addressed the following aspects:

- Innovative applications made possible by LTs,
- Scientific aspects related to the state of the art in LTs, for spoken, written and sign languages,
- Infrastructures and resources,
- Strategies in using LTs for Minority and Indigenous Languages,
- Activities for language documentation, preservation, reclamation, and enhancement,
- Scientific aspects related to handling linguistic diversity and multilingualism, particular of under-resourced and indigenous languages.

In 2020, the LT4All Programme and Editorial Committees put together the set of Research Papers and Posters collected at the occasion of LT4All. This set is available online[64]. New editions are currently under discussion.

5.3. Language Resources and Evaluation Journal

The Language Resources and Evaluation[65] Journal is published by Springer and endorsed by ELRA. The journal is dedicated to the LRs, their acquisition, creation, annotation and use, as well as to the evaluation of LRs, technologies and applications. Generally, four issues are published every year as Regular issues. In 2021, the first issue of the Volume 55 was a Special issue dedicated to the LREC 2018 Selected Papers. Every year, the institutional members of ELRA are granted online access to JLRE for free as an ELRA membership benefit.

6. Perspectives

In this paper, the main highlights can be put on the expertise ELRA has developed in many areas to help serve the community efficiently and cost effectively, including development expertise in all issues related to LRs sharing that has remained the association's core mission for over 25 years: ELRA can help re-purpose existing researchers' resources, package them with the necessary documentation including a data management plan, in compliance with the "FAIR" principles, design the appropriate licenses with the providers and handle all the logistics for distribution and delivery.

ELRA supports the production of new LRs for training, developing, and testing LT systems, and for all possible modalities: speech, audiovisual, image, OCR data, all varieties of textual data (monolingual, bilingual, multilingual, parallel, aligned, comparable) with the necessary annotations and tagging, for instance audio transcriptions, PoS tagging, syntactic and sentiment analysis, etc.

ELRA will also continue to extend its support to the valuable evaluation campaigns and challenges by providing the LRs from its catalogue or data to be newly developed.

---

[56] https://language-tools.ec.europa.eu/NLPServices/NLP
[57] http://lrec-conf.org/
[58] http://www.lrec-conf.org/proceedings/lrec2020/index.html
[59] https://lrec2020.lrec-conf.org/en/shared-lrs/
[60] https://lt4all.elra.info
[61] https://en.unesco.org/
[62] https://admhmao.ru/en/
[63] http://www.elra.info/en/sig/sigul/
[64] https://lt4all.elra.info/proceedings/lt4all2019/
[65] https://www.springer.com/journal/10579

As part of its strategic plans for the years to come, ELRA will continue to encourage its partners to provide resources free of charge to the community.

With this in its roadmap, ELRA will pay a specific attention to the various gaps in the field by:

- identifying existing LRs or "LR Kits" needed for all languages, and clearing IPR/Ethics/GDPR related issues;
- tackling multi-harvesting practices: many inventories harvest each other which leads to more confusion;
- supporting projects but also dissemination events for less-resourced languages: in this regard, the plan is to continue to play an important role in the Indigenous Language Decade set up by the United Nations and UNESCO through the series of LT4All conferences;
- addressing the new AI paradigm (deep learning/machine learning): this requires tremendous amounts of data produced by the community and we need to care for the lack of precautions which led to many biases in the different types of processing, e.g. gender, ethnics, language varieties, etc.

To handle all these issues, ELRA has extended its membership to individuals to ensure that its strategic roadmap complies with the expectations of the community at large.

All these activities are conducted thanks to the large set of partners built over all these years around the LREC community. ELRA is very grateful to all of them for their support to its mission and commits to strengthen its mission and collaborative actions.

7. Bibliographical References

Arranz V., Choukri K., Mapelli V., Rigault M., Labropoulou P., Deligiannis M., Voukoutis L., Piperidis S., Germann U. (2022). Data sets, models, identified gaps, produced resources and their exploitation within ELG (version 3). ELG Project Deliverable D5.3. January 2022.

Brabant Q., Lecorvé G., Rojas-Barahona L. M. (2022). CoQAR: Question Rewriting on CoQA. In Proceedings of LREC'22, Marseille, France, 2022.

Choukri C. (2018). Report on Analysis of European Language Technologies (LT). Internal report. In EU service contract SMART 2016/0103. 22 November 2018.

ELRC Consortium (2019). Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe – Why Language Data Matters. White Paper. 2019.

European Commission, Directorate-General for Communications Networks, Content and Technology, Vasiljevs, A., Choukri, K., Meertens, L., et al., Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem, Publications Office, 2019, https://data.europa.eu/doi/10.2759/142151

Lösch A., Mapelli V., Piperidis S., Vasiljevs A., Smal L., Declerck T., Schnur E., Choukri K. and Van Genabith J. (2018). European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. In Proceedings of LREC'18. Miyazaki, Japan, 2018.

Rigault M., Mapelli V., Mazo, H. (2021). Use Case – Reuse of Emergency Calls embedded in TV Shows. Internal report. In EU service contract SMART 2019/1083. 19 November 2021.

8. Language Resource References

Annotated tweet corpus in Arabizi, French and English, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 482-848-308-105-6, ELRA ID: ELRA-W0323.