

Language Resources to Support Language Diversity – the ELRA Achievements

Valérie Mapelli, Victoria Arranz, Hélène Mazo, Khalid Choukri

ELRA/ELDA

9 rue des Cordelières, F-75013 Paris, France

Tel. +33 1 43 13 33 33 -- Fax. +33 1 43 13 33 30

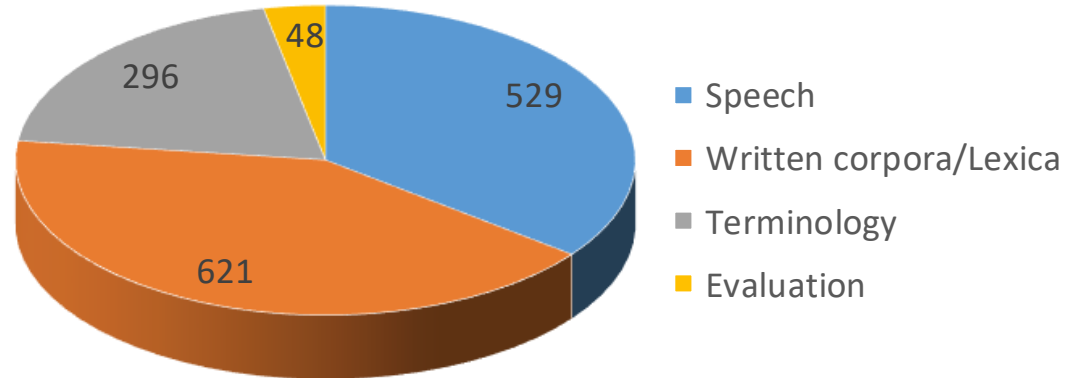
*Email: {**mapelli**; **arranz**; **mazo**; **choukri**}@**elda.org***

<http://www.elra.info>

Overview

- LR Identification & Distribution
- Production projects
- Infrastructure projects
- Events and dissemination
- Future activities

>1400 LRs

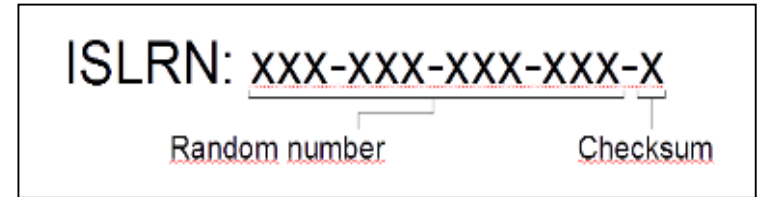


- Large series of datasets thanks to international partnerships, eg. between 2019-2022
 - 50 monolingual / multilingual lexicons from the CJK Dictionary Institute, Inc, specialized in Arabic, Chinese, Japanese and Korean lexicography
 - 115 speech resources for various European and Asian languages from SpeechOcean
 - 32 text corpora and bilingual/trilingual dictionaries for Vietnamese provided by the Kimtudien Multilingual Data Center

- Examples of LRs for low-resourced languages:
 - Ema-lon **Manipuri** Corpus
 - Tham **Khasi** annotated corpus
 - Arbobanko (**Esperanto** Treebank)
 - **Amharic-English** bilingual corpus
 - FAME! Speech Corpus (**Frisian**)
 - Helsinki Corpus of **Swahili**
 - **Mongolian** written corpus
 - PAROLE **Irish** Distributable Corpus
 - **Pashto** monolingual and parallel corpora
 - **Persian** Speech Corpus
 - Speaking atlas of the **regional languages of France**
 - The EMILLE/CIIL Corpus (14 **South Asian languages**)
 - **Welsh** SpeechDat(II) FDB-2000

<http://www.islrn.org>

- ISLRN persistent identifier
- Latest figures:
 - 3223 LRs assigned since January 2014 (31/03/2022)
 - ~250 distinct languages
 - Ongoing contributions, in particular from LREC2022
- New version of ISLRN web portal (February 2022):
 - ported into latest versions of Django and Python frameworks
 - refreshed design and functionalities, including search tool & enhanced submission pages



ISLRN identifier structure

- Community initiative "Share your LRs" since LREC 2014
 - Various types of LRs: corpora, grammar/language models, ontologies, terminology LRs, treebanks, evaluation data/packages
 - Clean version of lists of LRs:
 - 181 LRs available in LREC 2020 list @ <https://bit.ly/3o836kM>
 - 125 LRs available LREC 2018 list @ <https://bit.ly/3dpWdq3>
 - 82 LRs available LREC 2016 list @ <http://ow.ly/KTzf30fXoCe>
 - 116 LRs available LREC 2014 list @ <http://ow.ly/9gDm30fVMiu>
- Share your LRs at LREC 2022!

- OLAC (Open Language Archives Community) archive
<http://www.language-archives.org/archive/catalogue.elra.info>
 - Catalogue metadata export in XML, updated in 2020
- Google Dataset Search
<https://datasetsearch.research.google.com/>
 - ELRA Catalogue update with JSON code (schema.org)
- ELG Language Technology platform
<https://www.european-language-grid.eu/>
 - Catalogue metadata export in Metashare/ELG format
- OpenSLR Mirror hosted by ELRA/ELDA
<https://openslr.elda.org>

Corpora for Speech technologies

- **MGB-5 Moroccan and MGB-3 Tunisian Dialect databases**
Produced for MGB-3 and MGB-5 Challenges (Multi-Genre Broadcast) -
<http://www.mgb-challenge.org/>
14 hours collected and transcribed for Moroccan and Tunisian dialects
- **French transcription corpus**
Development of an automatic subtitling system primarily for deaf people -
Rosetta Project <https://rosettaccess.fr/>
Transcription of 20 hours of TV shows, broadcasts, documentaries, and series
- **Tamasheq Corpus for Automatic Translation**
Building an automatic translation system for African vernacular languages
Translation of a Niger's broadcast news from Tamasheq into French

Corpora for Speech technologies

- **Audio Data Annotation for Speaker Identification (French)**
European research program Chist-ERA - <https://www.chistera.eu/>
Annotating audio data based on video files to identify 5901 unique speakers.
328 hours from French channel LCP
- **Other projects (private demand):**
 - Multilingual conversational telephone speech corpus
Arabic, English, French, German, Italian, Korean, Mandarin and Cantonese
Chinese, Portuguese, Russian
 - Speaker Identification Corpus for French

Corpora for Written technologies

- **Rephrasing a Q&A corpus to improve a conversational system**
Natural questions in English to improve a conversation system
Rephrasing questions from American corpus CoQA (Conversational Question Answering)
- **Annotated tweet corpus in Arabizi, French and English**
INSA Rouen Normandie / SAPHIRS project (System for the Analysis of Information Propagation in Social Networks)
17103 tweet sequences annotated on 3 themes: Hooliganism, Racism, Terrorism
- **Other projects (private demand):**
 - Short messages (SMS) corpus (Arabizi, Moroccan & Tunisian)
 - Corpus of crawled documents
 - English-Arabic Parallel Corpora

Sharing project results

- Objective: capitalization on produced LRs
- Means: making LRs available through the ELRA Catalogue

- Example:

Annotated tweet corpus in Arabizi, French and English:

<http://catalog.elra.info/en-us/repository/browse/ELRA-W0323/>

Other LRs from production project to be released soon...

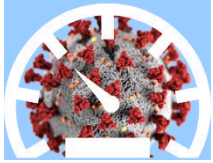


<http://lr-coordination.eu/>

Series of EU SMART contracts since 2015

- Target: support eTranslation with data from public sector across 30 European countries
- Objectives:
 - Improve availability and simplify access to LRs for MT
 - Establish LR observatory across EU Member States and associated countries, including Technology and Public Services National Anchor Points (>50 NAPs)
 - Raise awareness among stakeholders about the value and use of data for MT
 - Clarify legal and commercial issues
- Actions:
 - ELRC and eTranslation Technical and Legal Helpdesk
 - ELRC-SHARE Repository - <https://elrc-share.eu/>: many resources available to the community under open licenses
 - ELRC Conferences and Country Workshops
 - Surveys on Country Profiles and LT Market





<http://eval.covid19-mlia.eu/>
Community Evaluation effort

- **Target:** improve/accelerate access to multilingual information, including but not limited to health-related content, in the context of the pandemic
- **Objectives:**
 - Aggregate and summarize different sources of information
 - Organize three rounds of evaluation
- **Actions:**
 - Data acquisition
 - Evaluation tasks:
 - information extraction
 - multilingual semantic search
 - MT task

=> Resources available on git repositories





EC-funded action under Connecting Europe Facility (CEF)

<https://mapa-project.eu/>

- Target: develop a toolkit for effective and reliable de-identification of texts in the medical and legal domains
- Objectives:
 - Definition of sensitive information to be de-identified
 - Design of annotation guidelines
 - Data collection and annotation in the EU 24 languages
 - Development and evaluation of NER-based de-identification toolkit
- Actions:
 - Data produced available through the ELRA Catalogue (+ ELG & ELRC-Share)
 - Deployable, dock-ready, open-source system
 - ELRA in-house deployment of the MAPA toolkit and anonymisation services offered to the public
 - Use case by the EC: Anonymisation service:

<https://language-tools.ec.europa.eu/NLPServices/NLP>

<https://en.unesco.org/LT4All>



Enabling Linguistic Diversity
& Multilingualism Worldwide

December 5-6, 2019
UNESCO Headquarters
Paris, France



- Organized as part of the International Year of Indigenous Languages 2019 (United Nations)
- December 4-6, 2019, in UNESCO Headquarters in Paris (France)
- Co-organized by several partners including ELRA and SIGUL, under the patronage of UNESCO
- 400 invited participants from 65 countries, all continents represented.
- 100 papers and 105 posters in the Collection of Research Papers of the 1st International Conference on Language Technologies for All
<https://lt4all.elra.info/proceedings/lt4all2019/>
- ELRA involved in the International Decade of Indigenous Languages (2022-2032) with LT4All follow-up (2023)

LREC - Language Resources & Evaluation Conference

- LREC 2020 cancelled due to the pandemic but proceedings online:
<http://www.lrec-conf.org/proceedings/lrec2020/index.html>
- LREC 2022 in Marseille, 20-25 June 2022



www.lrec-conf.org

- **Regular promotion and dissemination activities:**

- **Language Resources and Evaluation Journal (Springer):**

<https://link.springer.com/journal/10579/volumes-and-issues>

Co-edited by Nancy Ide and Nicoletta Calzolari

Published by Springer

Supported by ELRA

Advisory Board: B. Maegaard, K. Choukri, J. Mariani, J. Odijk, C. Cieri, J. Tsujii

ELRA member benefit: Online access for free to ELRA institutional members

- **ELRA Newsletter (New! e-version since February 2022), Members' News, Various websites, Mailing lists...**

- **ELRA Portal** www.elra.info



- Core mission continuation
 - Support **LR sharing** and **re-purposing** of researchers' resources (based on FAIR principles, DMP)
 - Support the **production of new LRs**
 - Support **evaluation campaigns/challenges**
 - Contribute to long-term **European/International initiatives**
- Roadmap
 - **Bridging gaps** by identifying missing LRs ("LR kits")
 - Tackling **multi-harvesting** practice
 - **Less-resourced languages**: SIGUL, next LT4All conference
 - **New AI paradigm**

Thank you for your attention

