New Directions in ELRA Activities

Valérie Mapelli, Victoria Arranz, Pawel Kamocki, Hélène Mazo, Vladimir Popescu
ELDA/ELRA
9 rue des Cordelières, 75013 Paris, France
Email: {mapelli, arranz, kamocki, mazo, popescu}@elda.org

Abstract

Beyond the generic activities (cataloguing, producing, distribution of Language Resources, dissemination of information, etc.) that make the overall ELRA mission an indispensable middle-man in the field of Language Resources (LRs), new directions of work are now being undertaken so as to answer the needs of this ever-moving community. This impacts the structure and the operating model of the association *per se* with the creation of a new technical committee dealing with Less-Resourced Languages and the modification of the ELRA membership policy. It also intrinsically impacts the axes of work at all steps of activities: offering new tools for sharing LRs and related information, adapting to new legal requirements, producing and offering field-specific data. This paper addresses these new directions and describes ELRA (and its operational body ELDA) regular activities updates. Future activities are also reported in the last part of the article. They consist in ongoing projects like the ELRC initiative, the start of another CEF-funded project, the European Language Resource Infrastructure (ELRI), the updating of the Review of existing Language Resources for languages of France, and the continuation of the ELRA Catalogue development.

Keywords: Less-Resourced languages, ELRA Membership, Cataloguing, Open Data, ISLRN, LR Production, Events

1. Introduction

Beyond the generic activities (cataloguing, producing, dissemination of information, etc.) that make the overall ELRA[1] mission an indispensable middle-man in the field of Language Resources (LRs), new directions of work are now being undertaken so as to answer the needs of this ever-moving community.

This impacts the structure and the operating model of the association *per se* with the creation of a new technical committee dealing with Less-Resourced Languages and the modification of the ELRA membership policy.

It also intrinsically affects the axes of work at all steps of activities: offering new tools for sharing LRs and related information, adapting to new legal requirements, producing and offering field-specific data.

This paper addresses these new directions and describes ELRA (along with its operational body ELDA) regular activities updates.

2. New operating directions at ELRA

2.1. LRL Committee and SIGUL

In October 2016, ELRA launched the ELRA-LRL TC (Less-Resourced Languages Technical Committee). It focuses on different actions to be undertaken to support the maintenance of linguistic diversity through technology and ICT for minority/less-resourced languages. Defining an agenda for LRTs development, allowing a better sharing of corresponding tools and LRs, sharing and spreading state-of-the art knowledge, exploring techniques for evaluation of LRTs, promoting related management plans, are part of the objectives.

In April 2017, with the support of the LRL Committee, ELRA partnered with the ISCA to create a joint Special Interest Group on Under-Resourced Languages (SIGUL)[2] aiming to bring together a number of professionals involved in the development of Language Resources and technologies for under-resourced languages. Under the auspices of Interspeech 2017 in Stockholm, SIGUL organised its first event at the Special session on Digital Revolution for Under-resourced Languages (DigRev-URL)[3].

On 12th May 2018, in Miyazaki (Japan), in conjunction with LREC 2018, SIGUL will organise the 3rd CCURL Workshop, entitled “Sustaining knowledge diversity in the digital age”[4].

2.2. Enhancing ELRA Membership policy

Driven by the willingness to expand its membership base, the ELRA Board has appointed a Membership Task Force consisting of current and former Board members together with ELDA staff to rethink the current membership model. The Task Force work focused on revising the services proposed to the current institutional members, considering new membership categories and related services, and reassessing the membership fidelity programme.

The final proposal was submitted to the validation of ELRA General Assembly on 20 October 2017 and is being implemented in 2018.

Acknowledging that its membership base needs to be expanded so as to better represent the community, the ELRA Board has decided to bring drastic changes to its membership drive by opening the ELRA membership to individual members and by substituting discounts on membership to miles to reward the loyal institutional members.

The new membership at ELRA now encompasses two (2) types of members:

1. the Institutional Members (formerly called ELRA Members), including the ELRA Subscribers,
2. the Individual Members.

The Individual Members include:

a. the individual researchers and users of LRs who will join the association by paying individual membership fees
b. the employees of institutional ELRA members who are ELRA members by default, without having to pay for the individual membership fees since the institutional membership covers for that.

[1] http://www.elra.info
[2] http://www.elra.info/en/sig/sigul/
[3] http://ahclab.naist.jp/DigRevURL/index.html
[4] http://www.ilc.cnr.it/ccurl2018/

Individual members are represented by a Board member designated among the members and elected by them.

For the Institutional Members, joining ELRA remains the same. The main change lies in the discontinuation of the membership fidelity programme, replaced by a new mechanism rewarding the loyal members. The new membership schema now includes a discount on membership fees on a 3-year basis: the institutional members can cumulate up to 30% discount on the membership fees by renewing their membership for 3 consecutive years or more. Moreover, the fees have been harmonised to attract more subscribers: the EU organisations (above 50 employees) and non-EU organisations (subscribers) now pay the same membership fee/subscription.[5]

The services offered to the ELRA members are also currently being tailored by the Membership task force so as to extend ELRA's offer to the Individual Member statute. Current directions include in particular: LREC reduced fees, access to some Language Resources from the ELRA Catalogue, access to LRE Map and Universal Catalogue restricted services which are currently being defined, monthly members' news.

3. Cataloguing and licensing

3.1. New ELRA Catalogue

A new step forward has been taken in ELRA's ambition to offer an improved catalogue. In order to initiate the path to e-commerce services, the ELRA Catalogue of Language Resources[6] was completely redesigned, with a new interface and an improved navigation so as to allow visitors an easier access to the over 1070 Language Resources (LRs) and their corresponding descriptions. Among the new features, the Catalogue now offers extended metadata to describe the LRs, automatic submission to ISLRN, a refined search on the catalogue data for finding more specific information using criteria such as language, resource or media type, licence, etc. Currently, LRs can be selected and placed in a cart from where the user can send a request for quotation to initiate the order. When logging in, the user selects LRs and obtains distribution details (licensing information, prices) depending on his/her user status: ELRA member/non-member, Research vs Commercial organisation.

To allow this, e-commerce capabilities have been added[7] and aims to provide an end-to-end e-commerce management (order management, delivery, invoicing and payment facilities). Other functionalities such as an e-licensing module (automatic filling in and electronic signature) will also be developed and integrated at a later stage.

3.2. Clean version of Shared LRs available

Striving for constantly making LRs available to the widest community, ELRA had launched an "Open Challenge", known as "Share your LRs" at LREC 2014. All participants to the conference were proposed to share their LRs by giving access to the resource and corresponding description either through an external URL or by depositing them in a specific LREC repository before the main conference.

The Shared LRs set was manually checked and a cleaned version of the lists of LRs gathered at LREC 2014 and 2016 are available online[8]. These lists include LRs complying with the following criteria:

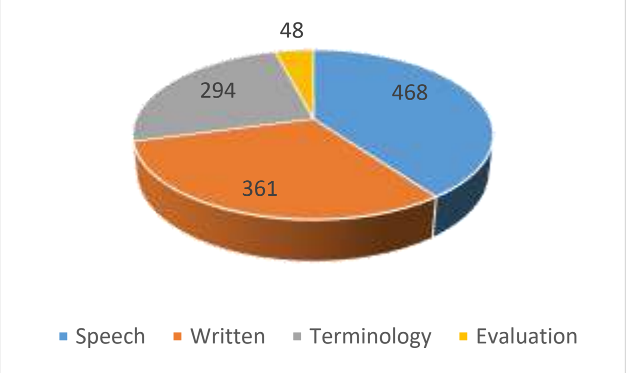
- LRs are accessible either through direct download or provided via an external URL;
- LRs belong to the following categories: corpus, grammar/language model, ontology, terminology, treebank, evaluation data/package.

These efforts also continue for the LREC 2018 edition.

3.3. Latest LRs in the ELRA Catalogue

Every year, new LRs are becoming available through the ELRA Catalogue. From 1 January 2016 to 31 December 2017, 62 LRs were released.

The distribution per type of resources in the catalogue as of the fourth quarter 2017 is given below:



Graph 1: LRs in the ELRA Catalogue (as of 31/12/2017)

As part of the 62 LRs recently announced, a large set of resources produced within the PEA TRAD project, supported by the French Ministry of Defence and packaged by ELDA was released. They consist of Pashto monolingual and parallel corpora (aligned with French and English), Arabic-French and Arabic-English parallel corpora, Chinese-French and Chinese-English parallel corpora.

Beyond the Pashto language represented through PEA TRAD resources, less-resourced languages are also worth quoting in the following LRs: the Helsinki Corpus of Swahili, GlobalPhone Swahili, a Mongolian written corpus, a Persian Speech Corpus, the FAME! Speech Corpus (dealing with Frisian language).

A history of LRs released in the ELRA Catalogue since 2007 (listed in chronological order) can be viewed online[9].

3.4. Open Data Licensing

With ELDA's involvement in the European Language Resource Coordination (ELRC, see specific section further below), significant advances have been achieved towards a better knowledge of the license issues management workflow for language resources labelled as "open data"

[5] http://www.elra.info/en/join-elra/rewarding-our-faithful-inst-members/

[6] http://catalogue.elra.info

[7] http://www.meta-share.eu

[8] http://lrec2014.lrec-conf.org/en/shared-lrs/current-list-shared-lrs/ and http://lrec2016.lrec-conf.org/en/shared-lrs/

[9] http://www.elra.info/en/catalogues/language-resources-announcements/

and aimed to fuel MT@EC, the Machine Translation platform dedicated to public administrations.

Within the ELRC consortium, all the partners identified, collected and made available language resources indicated as “open” coming from public service administrations. To do so, ELDA drafted a workflow supporting anyone wishing to obtain open data (i.e. data that fall under the scope of the PSI Directive) while following the right steps to assess the rights attached to the data (see figure below). Such a workflow aims to be integrated in the IPR clearance section of ELRA’s Data Management Plan (Choukri et al., 2016).

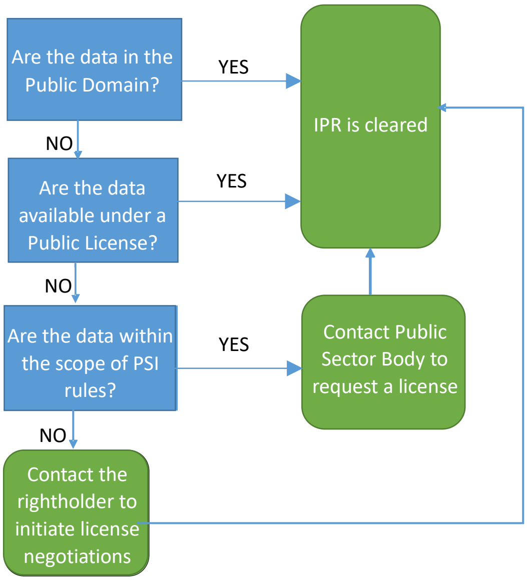

Are the data in the Public Domain?
YES
IPR is cleared
NO
Are the data available under a Public License?
YES
NO
Are the data within the scope of PSI rules?
YES
Contact Public Sector Body to request a license
NO
Contact the rightholder to initiate license negotiations

Figure 1: Workflow for data that fall under the scope of PSI Directive

3.5. ISLRN latest news

As scientific work requires accurate citations of referenced work so as to allow the community to understand the whole context and also replicate the experiments conducted by other researchers, since 2016, LREC endorses the need to uniquely identify LRs through the use of the International Standard Language Resource Number (ISLRN, www.islrn.org), a Persistent Unique Identifier to be assigned to each Language Resource. The assignment of ISLRNs to LRs cited in LREC papers are offered at submission time.

As a follow-up of the LRE Map 2016 initiative, ELRA has processed the information on existing and newly-created Language Resources provided by the authors submitting to LREC 2016 Conference. In order to increase the visibility of these resources, ELRA has allocated an ISLRN to each of the 106 languages resources which distribute as follows: 73 corpora (ca. written, spoken, multimodal), 30 lexicons (including ontologies), 3 evaluation data.

Moreover, in January 2017, the Institute for Applied Linguistics (IULA) at the Universitat Pompeu Fabra (UPF), Spain, became a certified provider to the ISLRN system. This means that IULA can apply for ISLRNs on behalf of the developers of data managed and distributed via the IULA network. IULA has already submitted 107 LRs to the ISLRN. These include monolingual and multilingual written corpora and lexica for the languages spoken in Spain (Aragonese, Asturian, Basque, Castilian Spanish, Catalan, Galician), as well as other European languages (English, French, German, Greek, Italian, Occitan, Portuguese, Romanian), and Esperanto.

The meta-information for these language resources is also available on the ISLRN website[10] with a broad international audience.

As of 31 December 2017, the total number of Language Resources having been allocated an ISLRN persistent identifier amounts to 2583. These LRs include raw and annotated corpora, lexicons and dictionaries, speech resources (conversational, synthesis, etc.), evaluation sets and multimodal resources, and cover 218 distinct languages (including sign languages).

Like in 2016, the ISLRN identifier is part of LRE Map, the LREC submission feature, where the authors can describe the resources(s) cited in their papers or used/developed during their research.

4. Production and infrastructure projects

4.1. Sentiment Annotation in German and French Tweets

Within the framework of two public research projects, uComp, Embedded Human Computation for Knowledge Extraction and Evaluation and Request (funded by the European Union under a CHIST-ERA cooperation action), REcursive QUEry and Scalable Technologies (funded by the French Government under the “Projets Investissements d’Avenir” programme), ELDA has been commissioned by the LIMSI laboratory to work on a deep sentiment and opinion annotation of German and French tweets. The German tweets addressed climate change and hence were rather regular in their structure, and grammatical. On the other hand, the French tweets were expressing virulent stances on transportation issues. Hence, they were much less grammatical and regular than the German tweets. This posed significant challenges to annotators, who nonetheless managed to achieve relatively high average inter-annotator agreements for this kind of projects.

In both annotation tasks the same annotation scheme was used, with really minor modifications to account for the “virulence” aspects of French tweets, hence ELDA has been able to develop an in-house software toolkit based around a LIMSI-provided annotation tool. The toolkit allows users to sample annotations, automatically pre-annotate data, perform various data bookkeeping tasks (tweet origin reconstruction, source file retrieval in samples etc.), and compute different agreement statistics (Cohen’s Kappa measures, F-measures, tag histograms, etc.). An important methodological improvement in the French tweet annotation campaign has been obtained by pipelining the annotation process into two stages:

(i) non-transport-related tweet identification (as these tweets are not further annotated);

(ii) annotation of the transport-related tweets. This process allowed for significant gains in terms of annotation time.

Overall, these two annotation campaigns took ELDA more than two years of almost continuous efforts, as we participated in several aspects of the annotation campaigns: annotator recruitment and management, annotation guide revision / update discussions with the LIMSI, data handling

[10] http://www.islrn.org/

tools implementation, annotation validation and quality control.

4.2. Infrastructure projects

4.2.1. MLi

MLi, Towards a Multilingual Data Services infrastructure, has been working to deliver the strategic vision and operational specifications needed for building a sustainable and comprehensive European MultiLingual data and services Infrastructure. The action has ended mid-2016 and ELDA has been strongly involved in investigating and assessing the Language Technologies, including Machine Translation, in the eCommerce sector in the European Union (Fernández-Barrera et al., 2016).

The group of experts behind this action (INMARK, ATOS, DCU, ELDA, ESTEAM, TILDE and USFD) produced a set of recommendations which was acknowledged by the European Commission.

4.2.2. CRACKER

Within the Cracker (Cracking the Language Barrier) H2020 project funded by the European Commission, ELDA participated in updating the technological base of the pan-European META-SHARE resource cataloguing platform. This represented a significant effort, as several years of maintenance and development lag had to be compensated for. Nonetheless, this effort has been successful, as, together with our partners from the ILSP Athena Research Centre in Greece, we have managed to make META-SHARE evolve, so that it can form the basis of other projects, such as the ELRC-SHARE repository, deployed within the ELRC project, described in Section 4.2.3.

4.2.3. European Language Resource Coordination

The 2-year project SMART 2014/1074 European Language Resource Coordination (ELRC)[11], funded by the European Commission within the (Connecting Europe Facility) programme, ended in April 2017 (Lösch et al., 2018).

Targeted data are those produced by the public sector administrations across all 28 EU Member States countries plus Norway and Iceland, which can be made available for re-use within the CEF Automated Translation platform (CEF.AT) and through the EU Open Data portal, with suitable copyright protection.

Together with the other 3 partners of the ELRC consortium (DFKI, TILDE and ILSP), ELDA contributed to the following achievements of this project:

- Collection, validation and delivery of 225 Language Resources that include 138 bi-/multi-lingual corpora, 50 terminologies and 37 mono-lingual corpora,
- Organisation of 29 workshops and 2 ELRC conferences to raise awareness about the importance of data sets and encourage participants to contribute to the collection of LRs needed,
- setting up and running of the Language Resource Board (LRB), as the governance body of ELRC which totals 60 members (Technology and Public Service National Anchor Points),
- setting up of ELRC infrastructure (website, legal and technical helpdesk, ELRC-Share repository).

The project final report is available through the EC website[12].

Two new projects inheriting from ELRC initiative were launched in December 2016 for a 3-year duration: SMART 2015/1091 ELRC+ L2 and ELRC+ L3. ELRC+ L2 focuses on the collection of LRs through awareness efforts via the organisation of a new batch of workshops and conferences. It also provides the technical infrastructure (ELRC-Share repository) to host collected LRs as well as legal and technical support through an online Helpdesk and dedicated services for IPR clearance. ELRC+ L3 aims to implement the acquisition of additional LRs and related refinement/processing services (e.g. anonymization), also necessitating the use of the ELRC-Share repository and IPR clearance services.

In parallel to the collection task of the first project, ELRC partners initiated a production activity by using the ILSP-Focus Crawler tool in order to crawl a number of public websites and obtain relevant aligned documents in different EU languages. To support the crawling and validation process, ELDA produced a crawling management toolkit, the ELDA_CMTK, available on Github[13]. This toolchain is currently being used in the novel resource production tasks within ELRC+ L3.

4.2.4. European Language Resource Infrastructure

ELRI (European Language Resource Infrastructure) is a 24-month action financed under the Connecting Europe Facility (CEF), Telecommunications Sector (CEF-TC-2016-3), and which started in October 2017. The action aims to deploy a network which allows for the accessible, secure and reliable sharing of LRs. ELRI targets LRs which are produced by translation centres and public institutions in Europe, with the partners' Member States (France, Ireland, Portugal and Spain) as starting point of the network and other countries joining in as future extensions. The Consortium is composed of the following institutions: VicomTech (coordinator, Spain), Linkare IT (Portugal), AMA (Portugal), ELDA (France), FCUL (Portugal), SESIAD (Spain) and DCU (Ireland).

The LRs contributed by the aforementioned stakeholders will be rendered available through different levels of accessibility, depending on the providers' restrictions to share their data. As a first step, the ELRI network will deposit the LRs in the national relay stations of the participating Member States. This first level will only grant access to national contributors and it is the option chosen by those adopters who do not wish to share their data further. It allows for a secure way to store and manage their data and get initiated into sharing. The second level of data sharing offered by ELRI will be the ELRC-Share repository. Those data holders willing to share their data at an EC (European Commission) level will have their data included in such repository, which shares its resources with the DGT (Directorate-General for Translation), so as to improve and customise their eTranslation platform. This platform is at the service of the different Digital Service Infrastructures (DSIs), thus benefiting European citizens

[11] http://lr-coordination.eu/

[12] https://ec.europa.eu/digital-single-market/en/news/european-language-resource-coordination-elrc-final-report-2015-2017

[13] https://github.com/ELDAELRA/elda_cmtk

from their usage. Finally, there is a third level of data sharing which implies opening data use to the community at large through repositories such as ELRA, META-SHARE or the Open Data Portal, all this subject to data providers' wishes.

ELRI will address Intellectual Property Rights (IPRs) for all data and metadata exchanged, making sure these are compliant with both the Public Sector Information (PSI) Directive and DSIs regulations.

Last but not least, ELRI will not only set-up the local relay stations and exchange network for data sharing but will also integrate data processing tools to ensure the creation of quality resources from the data contributed by the different stakeholders and will organise dissemination events to address potential data contributors. Moreover, the usage of such tools will be also IPR cleared by ELRI, who will support and advice stakeholders on the data processing procedures, and on sensitive data handling protocols, such as anonymisation.

All this work is done in close collaboration with other currently-ongoing data-collection and stakeholder awareness-raising actions under the CEF programme, such as ELRC+ L2 and ELRC+ L3. All these actions complement each other, avoiding the doubling of efforts.

5. Events and dissemination activities

5.1. LRE Map

During the last two years, several significant developments have been undertaken by ELRA for LRE Map software. First, a new resource submission form page has been added, which allows conference paper authors to submit resource metadata information along with minimal paper references (authors, paper identifier, conference name and year).

Moreover, the application runtime performance has been vastly improved, its underlying data model simplified and streamlined, and the software of the application has been upgraded to state-of-the-art technologies.

Results from previous conferences can be viewed from the LRE Map dedicated website[14].

5.2. LREC

The last edition of LREC (10th) took place at the Grand Hotel Bernardin in Portorož (Slovenia) on the week of 23 to 28 May, 2016.

The conference was held under the Honorary Patronage of His Excellency Mr. Borut Pahor, President of the Republic of Slovenia. The opening message from Mr Andrus Ansip, Vice-President of the European Commission, conveyed by Mr Zoran Stančič, the Head of the EU Representation in Slovenia, gave a positive and strong sign of the EC commitment on the multilingualism issues.

All in all, 1220 participants from 59 countries registered to the conference and the workshops. The HLT Village gathered 10 booths where projects and initiatives could disseminate their activities and results.

The Antonio Zampolli Prize was awarded to Professor Roger K. Moore, Chair of Spoken Language Processing Dept. Computer Science, University of Sheffield, UK.

The 11th edition of LREC will take place on May 7-12, 2018 at the Phoenix Seagaia Resort in Miyazaki (Japan). The conference website[15] provides general information on the conference. Information will also be circulated on the LREC 2018 Twitter account (@LREC2018) using #LREC2018. The contact email for any question on the conference is lrec@lrec-conf.org.

6. Future work

ELRC+ L2, ELRC+ L3 and ELRI projects will be one of ELDA's major focus for the following 2 years, the aim being to provide a sustainable infrastructure to support the acquisition of appropriate and qualitative LRs for key areas of CEF DSIs (Digital Service Infrastructures) so as to contribute to the customization and improvement of the CEF Automatic Translation system. Thus, these projects will work on Language Resource identification, collection, processing and IPR clearing for the shared data, as well as on the dissemination of the needs and activities through different workshops and conferences.

A follow-up of the Review of the existing Language Resources for languages of France is now being undertaken by ELRA under a new funding by the Délégation Générale à la Langue Française et aux Langues de France (Grouas et al., 2016). It will provide an up-to-date list of identified LRs and will initiate negotiations to make those LRs available.

Finally, the step by step implementation of e-commerce and e-licensing modules for the ELRA Catalogue will continue.

7. Bibliographical References

Choukri K., Mapelli V., Mazo H., Popescu V. (2016). ELRA Activities and Services. In Proceedings of LREC'16. Portorož, Slovenia, 2016.

Grouas T., Mapelli V., Samier Q. (2016). Review on the Existing Language Resources for Languages of France. In Proceedings of LREC'16. Portorož, Slovenia, 2016.

Fernández-Barrera M., Popescu V., Toral A., Gaspari F ., Choukri K. (2016). Enhancing Cross-border EU E-commerce through Machine Translation: Needed Language Resources, Challenges and Opportunities. In Proceedings of LREC'16. Portorož, Slovenia, 2016.

Lösch A., Mapelli V., Piperidis S., Vasiļjevs 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.

8. Language Resource References

GlobalPhone Swahili, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 200-331-212-512-8, ELRA ID: ELRA-S0375.

Helsinki Corpus of Swahili, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 941-187-059-145-7, ELRA ID: ELRA-W0119.

[14] http://lremap.elra.info/

[15] http://lrec2018.lrec-conf.org/en

NUM 5M Mongolian written corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 492-817-146-504-9, ELRA ID: ELRA-W0120.

Persian Speech Corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 068-845-898-304-0, ELRA ID: ELRA-S0393.

The FAME! Speech Corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 340-994-352-616-4, ELRA ID: ELRA-S0391.

TRAD Pashto Broadcast News Speech Corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 918-508-885-913-7, ELRA ID: ELRA-S0381.

TRAD Pashto Monolingual text Corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 394-903-293-388-0, ELRA ID: ELRA-W0092.

TRAD Pashto-French Parallel corpus of transcribed Broadcast News Speech - Training data, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 802-643-297-429-4, ELRA ID: ELRA-W0093.

TRAD Pashto-French Parallel corpus of transcribed Broadcast News Speech - Test data, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 547-897-479-723-3, ELRA ID: ELRA-W0094.

TRAD Pashto-English Parallel corpus of transcribed Broadcast News Speech - Test data, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 006-102-605-738-4, ELRA ID: ELRA-W0095.

TRAD Pashto-French News Articles Parallel corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 649-628-149-051-7, ELRA ID: ELRA-W0096.

TRAD Pashto-English News Articles Parallel corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 612-936-517-010-2, ELRA ID: ELRA-W0097.

TRAD Arabic-French Newspaper Parallel corpus - Test set 1, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 922-732-502-473-8, ELRA ID: ELRA-W0098.

TRAD Arabic-English Newspaper Parallel corpus - Test set 1, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 764-187-795-074-0, ELRA ID: ELRA-W0099.

TRAD Arabic-French Newspaper Parallel corpus - Test set 2, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 722-323-886-920-3, ELRA ID: ELRA-W0100.

TRAD Arabic-French Parallel corpus of transcribed Broadcast News Speech, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 862-201-329-808-4, ELRA ID: ELRA-W0101.

TRAD Arabic-English Parallel corpus of transcribed Broadcast News Speech, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 812-050-111-234-9, ELRA ID: ELRA-W0102.

TRAD Arabic-French Web domain (blogs) Parallel corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 138-395-895-757-7, ELRA ID: ELRA-W0103.

TRAD Arabic-English Web domain (blogs) Parallel corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 762-161-069-435-5, ELRA ID: ELRA-W0104.

TRAD Arabic-French Mailing lists Parallel corpus - Test set, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 895-850-015-188-4, ELRA ID: ELRA-W0105.

TRAD Arabic-English Mailing lists Parallel corpus - Test set, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 858-529-510-480-2, ELRA ID: ELRA-W0106.

TRAD Arabic-French Mailing lists Parallel corpus - Development set, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 333-026-450-858-0, ELRA ID: ELRA-W0107.

TRAD Arabic-English Mailing lists Parallel corpus - Development set, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 213-044-240-074-6, ELRA ID: ELRA-W0108.

TRAD Chinese-French Web domain (blogs) Parallel corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 464-017-697-777-3, ELRA ID: ELRA-W0109.

TRAD Chinese-English Web domain (blogs) Parallel corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 982-341-079-331-4, ELRA ID: ELRA-W0110.

TRAD Chinese-French News Articles Parallel corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 153-566-144-442-2, ELRA ID: ELRA-W0111.

TRAD Chinese-English News Articles Parallel corpus, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 626-096-751-907-7, ELRA ID: ELRA-W0112.

TRAD Chinese-English Email Parallel corpus – Development Set, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 447-281-370-489-0, ELRA ID: ELRA-W0113.

TRAD Chinese-French Email Parallel corpus – Development Set, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 255-358-917-604-3, ELRA ID: ELRA-W0114.

TRAD Chinese-English Email Parallel corpus – Test Set, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 985-956-234-357-3, ELRA ID: ELRA-W0115.

TRAD Chinese-French Email Parallel corpus – Test Set, in ELRA catalogue (http://catalogue.elra.info), ISLRN: 239-027-077-538-0, ELRA ID: ELRA-W0116.