EUROPEAN LANGUAGE GRID

D5.2
Data sets, identified gaps, produced resources and models (version 2)

Authors:	Victoria Arranz, Khalid Choukri, Valérie Mapelli, Mickaël Rigault (ELDA); Jan Hajic, Ondrej Kosarko (CUNI); Cristian Berrio, Andrés Garcia-Silva (EXPSYS); Rémi Calizzano, Nils Feldhus (DFKI); Miltos Deligiannis, Penny Labropoulou, Stelios Piperidis (ILSP); Ulrich Germann (UEDIN)
Dissemination Level:	Public
Date:	31-01-2021

About this document

Project	ELG – European Language Grid
Grant agreement no.	825627 – Horizon 2020, ICT 2018-2020 – Innovation Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2019, 42 months (GA amendment version: AMD-825627-7)
Deliverable number	D5.2
Deliverable title	Data sets, identified gaps, produced resources and models (version 2)
Type	Other (Report + Data Sets)
Number of pages	98
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 31-01-2021 – Actual: 31-01-2021
WP number and title	WP5: Grid Content – Language Resources, Datasets, and Models
Task number and title	Task 5.1: Identification and collection of existing data sets and resources to make them available through the ELG; Task 5.2: Identification of severe LR gaps and creation of LRs; Task 5.3: Dynamic training of new models in ELG
Authors	Victoria Arranz, Khalid Choukri, Valérie Mapelli, Mickaël Rigault (ELDA); Jan Hajic, Ondrej Kosarko (CUNI); Cristian Berrio, Andrés Garcia-Silva (EXPSYS); Rémi Calizzano, Nils Feldhus (DFKI); Miltos Deligiannis, Penny Labropoulou, Stelios Piperidis (ILSP); Ulrich Germann (UEDIN)
Reviewers	Katrin Marheinecke (DFKI); José Manuel Gomez-Pérez (EXPSYS)
Consortium	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Institute for Language and Speech Processing (ILSP), Greece
	University of Sheffield (USFD), United Kingdom
	Charles University (CUNI), Czech Republic
	Evaluations and Language Resources Distribution Agency (ELDA), France
	Tilde SIA (TILDE), Latvia
	Hensoldt Analytics GmbH (formerly SAIL LABS Technology GmbH), Austria
	Expert System Iberia SL (EXPSYS), Spain
	University of Edinburgh (UEDIN), United Kingdom
EC project officers	Philippe Gelin, Alexandru Ceausu
For copies of reports and other ELG-related information, please contact:	DFKI GmbH European Language Grid (ELG) Alt-Moabit 91c D-10559 Berlin Germany Dr. Georg Rehm, DFKI GmbH georg.rehm@dfki.de Phone: +49 (0)30 23895-1833 Fax: +49 (0)30 23895-1810 http://european-language-grid.eu © 2021 ELG Consortium

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
CC	Creative Commons
CSV	Comma-Separated Values
ELE	European Language Equality
ELG	European Language Grid
FBK	Fondazione Bruno Kessler
HTTP	Hypertext Transfer Protocol
ILC	Institute for Computational Linguistics “A. Zampolli”
ILSP	Institute for Language and Speech Processing
IPIPAN	Institute of Computer Science, Polish Academy of Sciences
LR/LRs	Language Resources or Language Resource
LT/LTs	Language Technology or Technologies
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OCR	Optical Character Recognition
R2	Release 2 of ELG
REST	Representational State Transfer
RPC	Remote Procedure Call
SDK	Software Development Kit
SPDX	Software Package Data Exchange
TBX	TermBased eXchange
TMX	Translation Memory Exchange
WMT	Workshop/Conference on Machine Translation
XML	Extensible Markup Language
XSD	XML Schema Definition
XSLT	Extensible Stylesheet Language Transformations
VLO	Virtual Language Observatory

Abstract

Deliverable D5.2 is the report on "Data sets, identified gaps, produced resources and models (version 2)", describing the work carried out to populate the European Language Grid since Release 1 of the platform (April 2020, M16) to present, and with a delivery date for Release 2 in M26. The document goes through a) the repositories whose ingestion had already been initiated for R1, b) the new repositories addressed, c) the conversion, harvesting, description and ingestion work carried out so far. It also presents the status of the technical, legal, financial and distributional issues reported in D5.1, which have been undertaken appropriately. The report introduces the work done in T5.3 towards the leveraging of LRs in ELG, with the specifications and setting-up of the model training and use cases that are going to take place. Finally, we describe the different steps taken for the gap analysis and subsequent detection of severe gaps. This gap analysis builds upon the results of the study of the ELG content, the output from the pilot projects and demands and needs from technology experts.

1 Introduction

D5.2 "Data sets, identified gaps, produced resources and models (version 2)" reports on the work carried out towards the continued population of the ELG. Following up on the repositories and datasets described in D5.1[1], D5.2 describes the LRs ingested into ELG after Release 1 (Section 2). These derive from:

- Repositories already under ingestion for R1: resources remaining from the ELRA catalogue[2], ELRC-SHARE[3] and selected META-SHARE nodes (those managed by DFKI[4], ELDA[5] and ILSP[6]).
- Repository priorities for ELG Release 2: LINDAT/CLARIAH-CZ Repository[7], LREC SHARED LRs (2014, 2016, 2018 and 2020)[8], META-SHARE Network, Zenodo[9] and Quantum Stat[10].

The report documents the next steps of conversion, harvesting and description (Section 3) required by the new repositories, as well as improvements implemented for this second phase of repositories. Section 4 focuses on the work done to solve the issues identified in Release 1. These issues are of different types: technical, legal, financial and distribution issues. Section 5 provides the initial steps towards leveraging ELG resources for offline model training to tackle NLP tasks, with the setting up of specifications for upcoming experiments and use cases. Section 5 elaborates on the gap analysis, providing details on the contents of the ELG catalogue in terms of datasets as well as identified needs and gaps. These needs may lead to a later production of LRs if some gaps are considered critical and no corresponding datasets are available. For that purpose, Deliverable D5.5 (M24) provides guidelines for the production of LRs within a sustainable lifecycle. Section 7 draws some conclusions on the work performed in WP5 so far and provides some insights on the upcoming work. The deliverable concludes with several annexes listing the datasets ingested into the ELG so far.

[1] ELG Deliverable: "D5.1: Identification and collection of existing datasets, models, identified gaps and plans (version 1) (April 2020).
[2] http://catalogue.elra.info
[3] https://elrc-share.eu
[4] http://metashare.dfki.de
[5] http://metashare.elda.org
[6] http://metashare.ilsp.gr:8080
[7] https://lindat.mff.cuni.cz/repository/xmlui/
[8] These resources have been reported so far (D5.1) as ELRA-SHARE LRs, but they have been recently renamed as LREC Shared LRs.
[9] https://zenodo.org
[10] https://quantumstat.com

2 Provision of Language Resources for ELG Release 2

As reported in D5.1, a first prioritisation of repositories to be ingested into ELG was done towards R1 following a strategic decision: inserting repositories managed by the ELG consortium partners (ELRA catalogue, ELRC-SHARE and three META-SHARE nodes). This enabled us to test procedures and methodologies as well as to identify issues addressed for R2 in a more manageable manner.

The target LR provision for R2 has been five-fold:

1. Continue and conclude with the repositories initiated for R1.
2. Focus on further repositories managed by ELG consortium partners such as LREC shared LRs and LINDAT/CLARIAH-CZ.
3. Move ahead with remaining nodes from the META-SHARE network.
4. Work on Quantum Stat and Zenodo.
5. Define upcoming priorities for R3.

2.1 Remaining Resources from Release 1 Repositories

R1 helped set up the basis for the conversion, description, and ingestion of LRs into the ELG. 280 LRs were ingested from the three targeted repositories (ELRA catalogue, ELRC-SHARE, META-SHARE), which enabled us to work on several technical, legal and financial-distributional issues that have been solved to continue and improve the work towards R2. Such continuous work, together with the incremental expertise acquired have allowed us to perform much larger ingestion actions, which have resulted in the provision of 2284 LRs from these initial repositories (Table 1).

	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	Total
ELRA	635	545	–	1180
ELRC-SHARE	989	41	–	1030
META-SHARE	53	14	7	74

Table 1: Remaining LRs ingested from initiated R1 repositories

The conversion, harvesting and description work done to ingest these repositories is detailed in Section 3. Even if these procedures were implemented for R1, updates, corrections and improvements have taken place for R2.

2.2 Repository Priorities for ELG Release 2

As mentioned in Section 2, the next steps in LR provision for R2 were as follows:

1. Focus on further repositories under Consortium partners' management: LREC Shared LRs and LINDAT/CLARIAH-CZ.
2. Move ahead with remaining nodes from the META-SHARE network.
3. Work on Quantum Stat and Zenodo.

The reasons behind these choices combined strategy and diversity, looking at repositories which:

- Were easier to deal with as they were directly managed by ELG consortium partners.
- Were large and offered many LRs.
- Offered many LRs under open-sharing licensing schemas.
- Presented a wide range of language resources types, populating the ELG catalogue in a richer way.

Even under such favourable conditions, LR provision is still complex and sometimes tedious work, since large repositories such as Quantum Stat and Zenodo require considerable work in terms of content checking and metadata description (see below). Ingestion has been concluded for LINDAT/CLARIAH-CZ, resulting in 309 included LRs. Protocols have been put into place to automatically harvest LR updates once a week. This is also the case for the ELRC-SHARE repository. With regard to the LREC Shared LRs, all those that have been filtered as shareable from LREC 2014, 2016 and 2018 have already been converted, described and ingested. This represents 71 LRs available through the ELG catalogue. The LREC Shared LRs 2020 are currently being prepared (see Section 3.5). Work on Zenodo and Quantum Stat is currently ongoing (see Sections 3.6 and 3.7). However, a first set of 73 language resources from Zenodo has been ingested. The ingested resources from LINDAT/CLARIAH-CZ, LREC Shared LRs and Zenodo are summarised in Table 2.

	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	Total
LINDAT/CLARIAH-CZ	243	66	–	309
LREC Shared LRs	46	25	–	71
Zenodo	36	37	–	73

Table 2: LRs ingested from R2 repository priorities

At the time of writing, the ELG catalogue hosts 2737 LRs. Approx. 1,000 additional datasets are currently being worked on (filtered and described) towards R2.

3 Language Resource Metadata Conversion, Harvesting and Description

This section describes the ingestion procedure carried out for R2 regarding the repositories listed in Section 2. Some of these repositories use harvesting protocols which combined with metadata conversion have allowed to insert large numbers of LRs into the ELG. Other repositories follow a combination of metadata conversion and manual uploads. Others use automatic content extraction for manual analysis and ingestion.

3.1 ELRA Catalogue

Only a subset of LRs from the ELRA catalogue were integrated into the ELG catalogue for R1. This was caused by technical issues regarding licensing and pricing conditions (see D5.1) that have now been addressed. For R2, the aim is to add the full set of LRs currently available in the ELRA catalogue. It was decided that metadata for LRs that combine free and for-a-fee licenses (e.g., free for research purposes and for-a-fee for commercial purposes) would also be added into the ELG by linking them directly to the ELRA catalogue entries. This would be the alternative to having the LRs downloaded directly from ELG while waiting for a solution for managing their distribution (in particular while setting up the billing functionality) from the ELG platform.

The first procedures for R1 included a conversion protocol that required several manual actions to support the insertion of the original metadata into ELG. Based on that expertise, and in order to avoid such manual operations, the conversion tool required some improvements before submitting the metadata of the remaining resources into the ELG catalogue. A Python script has been written to facilitate this task[11]. The list of improvements includes the following changes (corrections, additions, deletions):

[11] This new version of the converter will be made available through the ELG Gitlab for R2.

- The “ms:” prefix was added for all metadata elements (e.g., "resourceName" -> "ms:resourceName").
- "xml:lang="en-US"" was changed into "xml:lang="en"" and "xml:lang="und" has been changed into "xml:lang="en"" for all multilingual elements.
- The “annotation” component was removed for raw corpora.
- The “distributionAudioFeature” component must be filled in for audio corpora.
- The “distributionTextFeature” component must be filled in for text and lexical conceptual resources.
- “dataFormat” must be added outside <audioFormat> and <videoFormat>.
- The “languageVariety” element was removed for Castilian, Flemish and Valencian.
- “languageId” was added for language and metalanguage elements.
- “multilingualityType” was added after ms:lingualityType if lingualityType is bi- or multilingual.
- ELRA’s website was added whenever ELRA is mentioned as a distributionRightsHolder.
- When the licencesTerms correspond to one of the CC licences, only one datasetDistribution is kept so as to avoid redundant information.
- “distributionLocation” was added whenever LRs are not downloadable.
- The URL linking to ELRA licences was corrected.
- The value for DatasetDistributionForm elements was corrected (it should be http://w3id.org/meta-share/meta-share/other when no value is specified in the original metadata).
- The <ms:LicenceIdentifier ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-share/SPDX"> element was deleted whenever the original metadata is empty.
- For any information that cannot be inferred from the description, the “unspecified” value was added.
- The "MembershipInstitution" element was removed when distribution information is meant for non-members of ELRA (e.g., ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0). For these, the "MembershipInstitution" element makes no sense, so it can be removed.
- The metalanguage component was removed as this is not a mandatory component.

Once the conversion work was finalised, the new metadata XML files were produced and ingested into the ELG. The import was facilitated by the new editor dashboard, which allowed the ingestion of several XML files at the same time. Further corrections were also done manually directly through the editor dashboard.

3.2 ELRC-SHARE

The ELRC-SHARE repository is used for documenting, storing, browsing and accessing LRs collected through the European Language Resource Coordination and considered useful for feeding into CEF e-translation. The infra-structure implements a metadata export functionality in ELRC-SHARE schema compliant XML files and JSON. Recently, an OAI-PMH v2[12] server was developed to facilitate metadata harvesting by other infrastructures. The OAI-PMH endpoint[13] provides a subset of the metadata records published on ELRC-SHARE to the ELG platform. The selection of resources, to be available for harvesting by ELG, is based on the following criteria:

1. Monolingual, bilingual data resources (corpora, lexical/conceptual resources, language descriptions)
2. TMX, CSV, TBX format
3. Permissive license for download
4. Cleared intellectual property rights (IPR)
5. Use by third parties besides DGT (Directorate-General for Translation) is allowed

[12] The Open Archives Initiative Protocol for Metadata Harvesting, 2015.
[13] https://elrc-share.eu/repository/oai-pmh/?verb=Identify

The conversion of the metadata from ELRC-SHARE to ELG uses an XSLT stylesheet[14] that maps information from ELRC-SHARE to the appropriate ELG fields. Both schemas belong to the META-SHARE family of metadata profiles (see D2.3[15]). The transformation was direct for elements used in both schemas, with minor structural changes. Further mappings were required for the following elements into their respective IRI values from the MS-OWL ontology[16] used in ELG (for elements whose data type differs, e.g., free text values vs. controlled vocabularies and for transforming values of controlled vocabularies, which in ELRC are specified as free text).

3.3 META-SHARE

As described in D5.1, the selection of LRs for ingestion done for the three META-SHARE nodes needed to be revised due to licensing restrictions. These involved proprietary licenses such as MS-C-NoReD, MS-NC-NoReD and MS-Commons-BY-SA, as well as other licenses that required negotiation with data providers.

To address this issue, a study of the licenses was performed by the ELDA legal team. We took the opportunity to draft a proposal for license mapping (Table 3) where non-restrictive licenses are invited to move to CC licenses, and restrictive licenses are encouraged to move into more open licenses, too. Some restricted-use licenses require further discussions with the data providers and are thus displayed as N/A for the time being. In addition, some resources are labelled as UnderNegotiation as the procedure of sharing is still in progress. For the two latter, proposals need to be done in a personalised manner and have not been ingested into the ELG catalogue so far. The impact of this mapping in ELG's context is twofold:

1. It allows to proceed with the unblocking of those LRs which remain to be ingested from the three nodes targeted for R1 (this is planned for R2 in M26).
2. It allows us to have a comparative legal basis to progress with the remaining META-SHARE network.

Table 3 illustrates the mapping. It should be added, though, that this mapping does not replace discussions and consultations with the data providers, unless it just normalises the license name (e.g., CC-BY changed into CC-BY-4.0). The provider's consent needs to be obtained in that matter.

Used license	Proposed License
CC-BY	CC-BY-4.0
CC-BY-NC	CC-BY-NC-4.0
CC-BY-NC-ND	CC-BY-NC-ND-4.0
CC-BY-SA	CC-BY-SA-4.0
CC-BY-NC-SA	CC-BY-NC-SA-4.0
CC-ZERO	CC0
MS-C-NoReD	CC-BY-SA-4.0
MSCommons-BY-SA	CC-BY-SA-4.0
MS-NC-NoReD	CC-BY-NC-4.0
Other	CC-BY-NC-4.0
Other (Restricted Use)	N/A
Proprietary	CC-BY-NC-4.0
Proprietary (Restricted Use)	N/A
UnderNegotiation	(No license available, proposals done on a case-by-case basis: N/A, CC-BY-NC-SA-4.0, CC-BY-NC-4.0)

Table 3: License mapping for the META-SHARE nodes

[14] https://gitlab.com/ilsp-nlpli-elrc-share/elrc-share-repository/-/raw/master/misc/tools/ELGConverters/ELRC2ELG.xsl, https://gitlab.com/ilsp-nlpli-elrc-share/elrc-share-repository/-/raw/master/misc/tools/ELGConverters/elrc2ELGMap.xml
[15] ELG deliverable D2.3: Metadata schema (August 2019).
[16] https://link.springer.com/chapter/10.1007/978-3-319-25639-9_42#enumeration

Additionally, the whole META-SHARE network is being analysed for its inclusion in ELG, which has proven more complex than expected. Our analysis prepares the strategy to optimise the procedure by looking at licensing issues (resources with one or more than one license), downloading conditions, statistics on resource types, linking with CLARIN nodes, harvesting directions (which node is harvested by which) and managing node status. Representatives of the nodes are being approached and tests are being planned for either full migration of datasets (e.g., this is the case for the FBK node[17]) or integration of metadata records linked to the datasets in their original repository (for IPIPAN[18] and ILC[19], among others). Different procedures and approaches need to be foreseen for each node. The integration of the whole META-SHARE network is planned for R3.

3.4 LINDAT-CLARIAH-CZ

The LINDAT/CLARIAH-CZ repository makes its metadata available for harvesting through its OAI-PMH v2[20] endpoint[21]. Means for ingesting metadata in the META-SHARE schema[22] were already in place on the ELG side and the repository did provide a mapping from its internal metadata storage to META-SHARE. An attempt was made at reusing this conversion, but the result was deemed unacceptable. The existing mapping has provided only the necessary elements to pass schema validation. It was clear from looking at the LINDAT/CLARIAH-CZ UI (i.e., the LINDAT web pages), that there was much more metadata available that was not being mapped.

After a few iterations we arrived at a mapping between concepts that are important and required in the ELG schema and the metadata stored in the LINDAT/CLARIAH-CZ repository. The mapping described in Annex A.G served as a basis for the implementation, which was first tested offline. After several rounds of feedback and improvements the exchange went online to also test the OAI-PMH endpoints.

Based on the feedback, the scope of metadata (or items) exported by LINDAT changed from “everything” to only those having files which were downloadable and had a media type specified. The reason behind this decision: for historical reasons, LINDAT curates a collection of “metadata only” resources with a rather minimal set of descriptive metadata. In addition, in its about nine years of service the repository has evolved its standards for metadata quality provided during submission (the metadata descriptions are provided by the submitter). There are some older records where the metadata lack in quality and it is difficult to update them.

LINDAT updated the metadata for several of its resources following the feedback received from ELG. Also, based on the feedback from LINDAT some changes were made on the ELG schema. For example, a new “identifier type” was defined for (funding) projects (the info:eu-repo/grantAgreement namespace used by OpenAIRE[23]), where some fields were made optional, etc. The implementation of the LINDAT mapping represents around 1200 changed lines of code[24], including some tooling to reflect some of the metadata issues discovered.

3.5 LREC Shared LRs

LREC, the Language Resources and Evaluation Conference[25] organised by ELRA, launched an initiative in 2014 called “Share your LRs”. Conference participants can share their LRs either by uploading them in a special LREC

[17] The Fondazione Bruno Kessler (FBK), Trento, Italy runs the http://metashare.fbk.eu node.
[18] The Institute of Computer Science Polish Academy of Sciences runs the http://metashare.nlp.ipipan.waw.pl node.
[19] The Institute for Computational Linguistics «A. Zampolli» of the Italian CNR runs the http://metashare.ilc.cnr.it node.
[20] The Open Archives Initiative Protocol for Metadata Harvesting, 2015
[21] http://lindat.mff.cuni.cz/repository/oai/request?verb=Identify
[22] http://www.meta-share.org/p/93/Documentation
[23] https://www.openaire.eu
[24] https://github.com/ufal/clarin-dspace/pull/930
[25] http://www.lrec-conf.org

repository or linking them to their original download location by filling in an online form. LRs are then made available the corresponding LREC website:

- **Shared LRs 2014**: http://lrec2014.lrec-conf.org/en/shared-lrs/
- **Shared LRs 2016**: http://lrec2016.lrec-conf.org/en/shared-lrs/
- **Shared LRs 2018**: http://lrec2018.lrec-conf.org/en/shared-lrs/

All of these have been made available already through the ELG catalogue. The **Shared LRs 2020** will be published online in late January 2020. They will be ingested into ELG Release 2.

A selection of LRs was done by checking the compliance of licenses with the ones accepted in ELG. Licenses that remained too vague (e.g., "Open Source", "Creative Commons" without further specification) were left aside.

Given that the original metadata was available in the form of an Excel file, we adapted the Excel template and conversion tool produced to gather Zenodo metadata (see below). As the Shared LRs metadata contained only a minimal set of information, missing information was added manually into this Excel file to comply with the mandatory ELG metadata (e.g., type of LR, linguality, annotation, data format, license, etc.). Then, the Excel file was converted into XML and ingested into the ELG platform through the editor dashboard.

3.6 Zenodo

Zenodo[26] is a digital library launched in May 2013 within the OpenAire[27] project, to enable the compilation of research artefacts, such as publications, images, datasets, software, etc. A good number of those artefacts consist of LRs that may be of interest to the LT community. At the time of writing, Zenodo has compiled 1,686,291 artefacts. This high number and the incompatibility between Zenodo metadata and ELG metadata makes the identification of relevant LRs a big challenge. This is why we opted for a semi-automatic approach to collect what ELG considers as LRs. We first used the Zenodo search[28] and filtering tool.

1. We filtered with the following selection of criteria:
   - Access right=Open,
   - File Type=Xml, Docx and Txt
   - Type=Dataset
2. We used two keywords to make two distinct extractions: "corpus" and "lexicon". The keyword "corpus" produced a list of 69 datasets, "lexicon" produced a list of 45 datasets.

To compile metadata and facilitate the information collection task, we created a template which enabled us to register the minimal set of metadata elements mandatory for ELG for each type of resource to be ingested in ELG in Excel format. A Python script was developed to export the metadata from Excel to XML, which could then be ingested into ELG. This process had its limitations due to the high number of entries in the Zenodo Library, thus we moved towards a more automatised process. Zenodo offers a Rest API that allows harvesting metadata, available under Creative Commons CC-0 license[29]. To obtain more LRs for ELG, we decided to query the Zenodo database directly using a script. At the time of writing, work is at an early stage, but we expect to collect over 600 LRs using this method and to make them available through ELG in the coming weeks (by M27).

[26] http://zenodo.org
[27] https://www.openaire.eu
[28] https://zenodo.org/search?page=1&size=20&q=
[29] See here for details: https://developers.zenodo.org/#records

3.7 Quantum Stat

Quantum Stat[30] enables LR producers to identify their datasets through "The Big Bad NLP Database". The procedure for identifying, describing and ingesting these datasets is as follows: we started by exporting a table that lists 481 datasets and made an analysis of those LRs that could be worth ingesting into ELG. The analysis consists in checking licensing information (whether licenses are well identified), the type of dataset (if it concerns an LR or another type of dataset), and whether the LR is downloadable. Then, as for LREC Shared LRs and Zenodo, we compiled the minimal set of metadata information, while also adding missing information not present in the Quantum table (e.g., type of resource, annotation, etc.) into our Excel template which was then converted into XML and ingested into ELG. The ingestion of the datasets is expected to be completed for R2 (M26).

4 Technical, Legal and Financial Issues Addressed for Release 2

4.1 Technical Issues

The technical issues reported in D5.1 have been resolved during this period, with the improvement of the ELG platform, and the adoption of protocols and best practices for the population procedures, both manual and automatic. ELG Release 2 (M26) includes a number of functionalities addressing these requirements:

- The introduction of the OAI-PMH harvesting server allows for mass population of the ELG and regular updates of metadata records from external repositories. Problems identified during this process have already been resolved at the source repository or at the ELG side (see Sections 3.2 and 3.4).
- Automatic validation mechanisms in the batch upload procedure of XML metadata records have been improved to tackle issues reported in D5.1 (e.g., duplicates).
- The addition of the interactive editor supports the manual creation of metadata records, guiding users to fill in the various elements as required and, thus, reducing the creation of invalid metadata records.
- The introduction of the fallback value "unspecified" for mandatory metadata elements relaxes the requirements on source metadata records imported with automatic procedures, thus allowing mass population, albeit with some loss on the informativeness of the metadata.

The implementation of the upload mechanism for content files together with the creation of metadata records empowers their interlinking without resorting to naming conventions. Mappers and converters have been gathered and are available for re-use through the ELG GitLab repository[31] facilitating conversions from new sources.

One issue which cannot be handled in a general way, is the normalisation of values and deduplication of entries. For instance, the name element of all entities is a free text field and can, thus, be filled in with similar strings (e.g., the same licence can be represented with the names "CC-BY 4.0", "CC_BY_4", "Creative Commons – Attribution 4", etc.), which results in multiple entries for the same object. Where possible, we use additional fields, such as identifiers from authority lists (e.g., for licences, the SPDX list[32]), or fields that can help us identify entries represented with similar names. For these, we created a lookup mechanism parametrized with the unique values of each entry type.

[30] https://datasets.quantumstat.com
[31] https://gitlab.com/european-language-grid/platform/ELG-SHARE-schema/-/tree/master/Support%20tools
[32] https://spdx.org/licenses/

4.2 Legal Issues

Thanks to the ELG's rich metadata schema, licensing information can be documented in a detailed way. For LRs that are downloadable directly from ELG, this licensing information is linked to the actual text of the license. Thus, when one downloads directly from ELG and the licensing terms require the explicit consent on the user side, the user must accept the license (Figure 1). With regard to licenses like CC or public domain datasets, the user will no longer be required to accept them after R2, simplifying the process to download a LR.

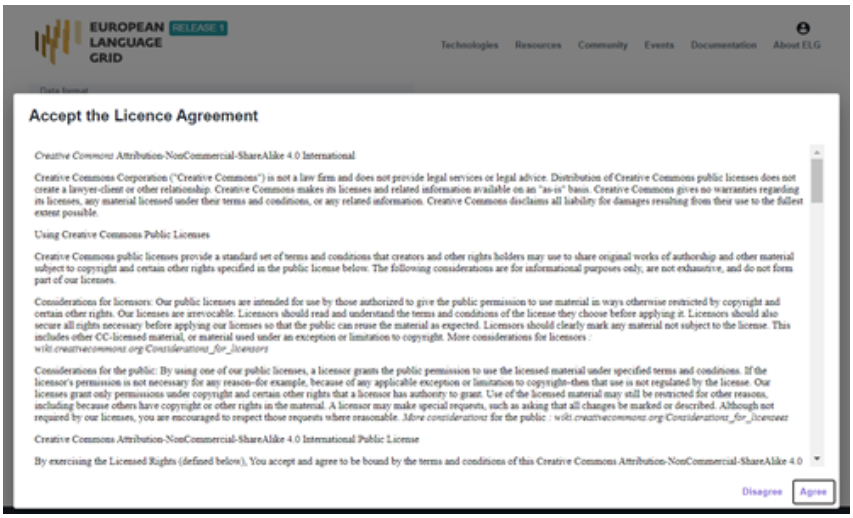

Figure 1: Example of license acceptance

Beyond the legal issues mentioned in D5.1, the identification of various repositories proved the importance of legal checking all throughout the information compilation process. In some cases (e.g., Zenodo), licenses are well identified and can usually be integrated in the ELG metadata without further analysis. However, for other cases (e.g., LREC Shared LRs, Quantum), legal information does not always comply with ELG requirements or is simply missing. Consequently, legal expertise is needed to either check and confirm the accuracy of present legal information, or to search for and gather the appropriate legal information.

The purpose of use shall also be taken into consideration distinctively for each repository. We have chosen several ways for ingesting LRs into the ELG platform depending on the authorised use of LRs:

- A harvesting process was chosen for LINDAT and ELRC-SHARE repositories in agreement with providers.
- Metadata and data were ingested directly into ELG for a subset of the ELRA catalogue, i.e., only for freely available resources and upon the agreement of ELRA, whereas only metadata was ingested for the rest of the ELRA catalogue due to pricing and billing requirements (see below).
- Only metadata with links to the download information are ingested into ELG for Shared LRs, Zenodo and Quantum Stat, while having no agreement with providers to include the downloadable data in ELG.
- A combination of the above-mentioned processes is being explored for each META-SHARE node depending on the agreement discussed with node managers (Section 3.3).

4.3 Financial and Distributional Issues

Thanks to the rich ELG metadata, financial and distribution information are reported for each LR. This includes information about costs, distribution or access location, membership institution and dataset distribution form. This information allows any ELG visitor to either obtain the resources through direct download or from the original provider. Moreover, direct download is available which allows visitors to get the resources hosted by

ELG simply by clicking on the Download button and accepting the corresponding license terms. See below some examples showing the variety of financial and distribution information managed in ELG.

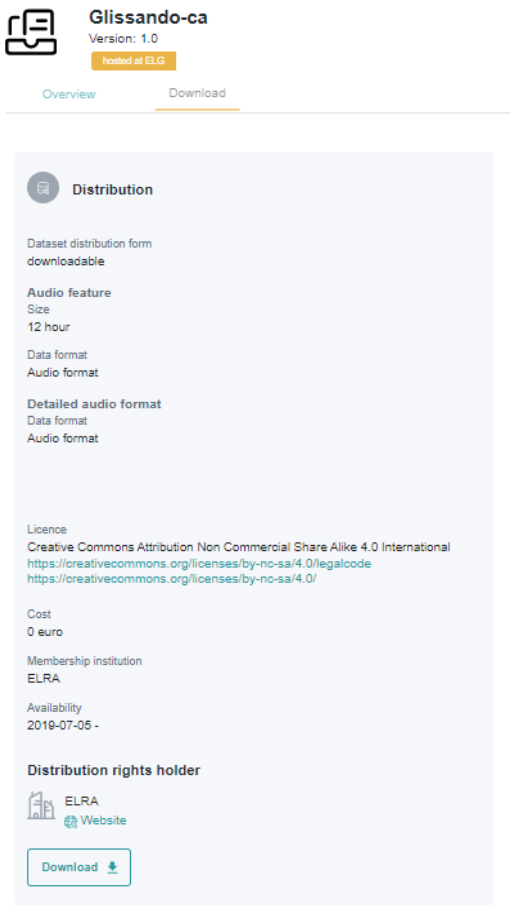

Figure 2: Distribution metadata (1)

As described in the previous section, the possibility to distribute directly from the ELG platform depends on the authorisation obtained from the original provider. Those authorisations are also linked with the financial and distribution process currently offered by ELG. For the time being, only free LRs can be downloaded directly from ELG whereas others are linked to the original distribution page at the provider's site. In the future, a solution for managing the LR distribution (in particular billing) from the ELG platform will be integrated and should solve those issues. This is planned for ELG Release 3.

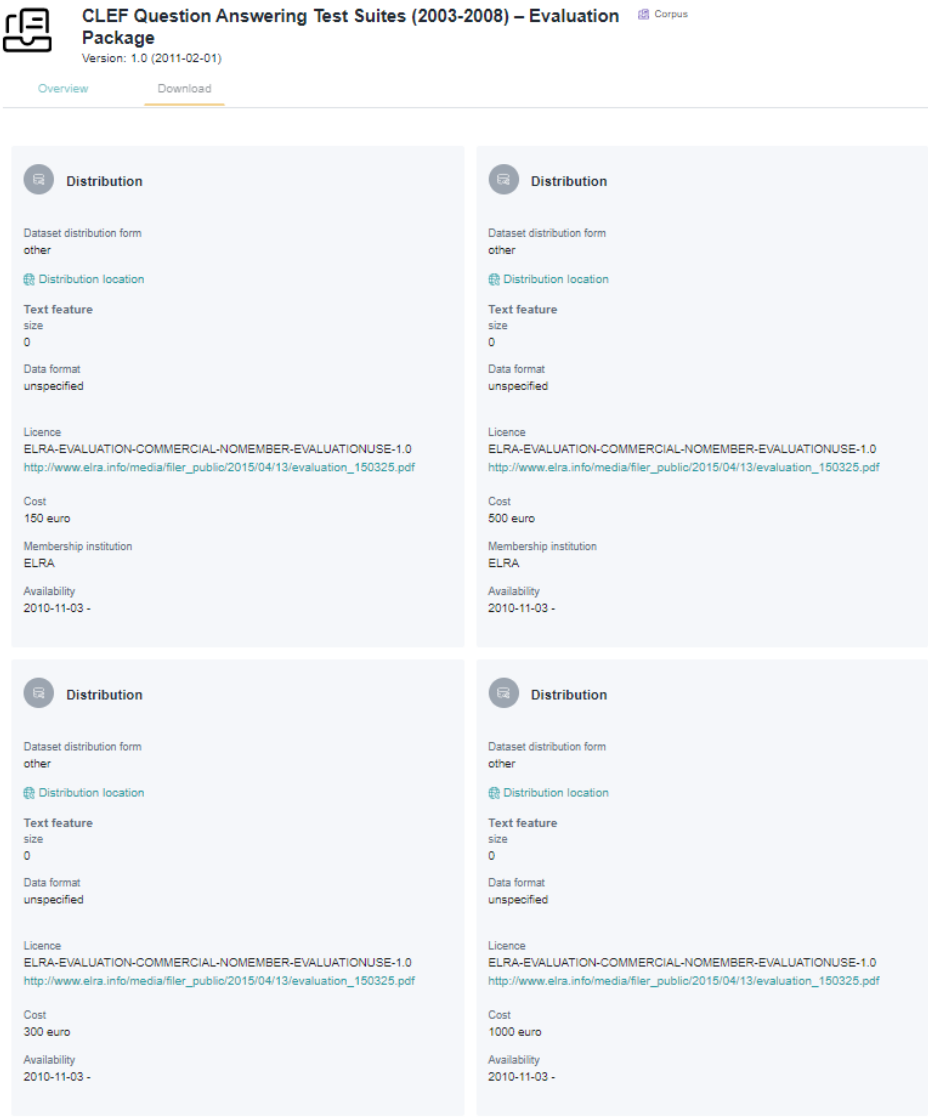

Figure 3: Distribution metadata (2)

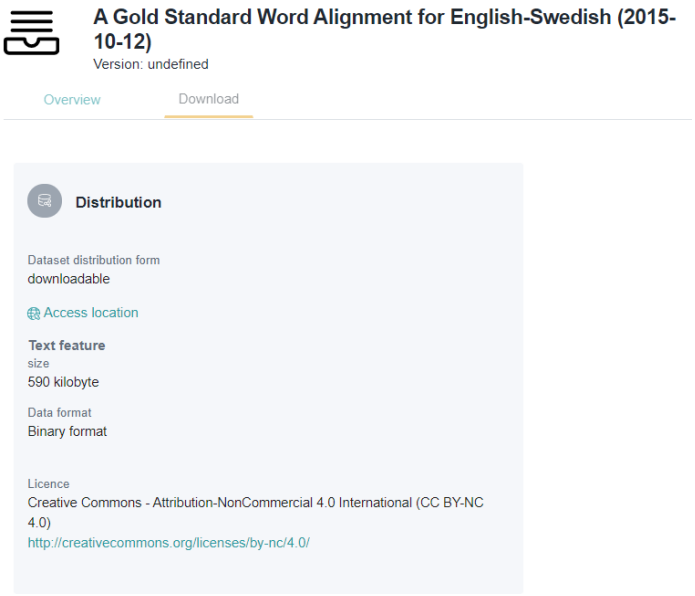

Figure 4: Distribution metadata (3)

5 Gap Analysis for Language Resources

With its 2737 datasets and following the ingestion of several repositories, the ELG catalogue is at a compelling stage to establish the next steps in its dataset provision strategy. Analysing the ingestion statistics (see below) and the content that goes with them will be one of the decisive parameters for the work to be done in the months to come. This will be only one of the parameters to be taken into account for the gap analysis that will contribute to our priority setting. Among these other parameters, we have the collaborations with the pilot projects, which will contribute datasets as well as share their needs; we will also consider feedback from technology developers and data users who share their needs with ELG; we will analyse the results from our LR identification work, with regard to datasets as well as technology and services. We will also review the substantial identification work that the ELE (European Language Equality) project will carry out, and in general, ELG will benefit from, and make use of, the ELE results which will be in line with ELG's interests and work.

5.1 Analysis of the ELG Content and Detected Gaps

This section offers an overview of what is available in the ELG catalogue with regard to LRs. The first step in this content/gap analysis has been looking at the number of LRs per resource type (see Table 4). The large majority of resources fall within the corpus or lexical/conceptual resource type categories, with 2002 and 728 entries, respectively. There are only seven language descriptions (resource type that also contains the language models), which provides some kind of indication that stablishing the means to create its own models will be of great use to the LT community. This is the objective of Task 5.3 (see Section 5).

Resource Types and their Linguality	Number of Resource Types
Corpus	**2002**
bilingual	1095
monolingual	798
multilingual	105
N/A	4
Language description	**7**
monolingual	7
Lexical/Conceptual resource	**728**
bilingual	324
monolingual	291
multilingual	113
Total	**2737**

Table 4: ELG Content – Types of resources split according to linguality

There is a certain unbalance between some linguality types, i.e., there are more bilingual and monolingual datasets (both for corpus and lexical/conceptual resources) than multilingual ones, but this can be expected as fewer initiatives focus on (and have the financial means for) developing LRs in a wide range of languages.

With regard to the languages covered by the ELG, Table 5 lists the 99 languages for which datasets are currently available. As anticipated, English (1677 LRs) outnumbers by far even other very frequent languages such as Spanish (433 LRs), German (304 LRs) or French (294 LRs). The reason for this large difference is that most of the bilingual or multilingual LRs have English as one of their languages.

We have datasets in some minority languages and we also have languages from several continents, but statistics for many of these are very low (one to seven LRs), which matches the trend of data production in general.

This is expected to change with the integration of the META-SHARE network. Furthermore, the CLARIN VLO will be also looked at if relevant for the gaps detected and if a potential collaboration can be established.

Besides under-resourced languages, these statistics also show the coverage of the ELG catalogue for the 24 official EU languages and we can see that all of them are well represented, even if some of them may require some further effort (languages like Slovak, Slovenian, etc. are represented in fewer than 40 LRs).

For the time being, ELG has targeted rather large repositories for the reasons explained in D5.1, one of them being the fact of optimizing the conversion and ingestion efforts by targeting large data blocks. Some labor will be set aside to consider these less represented languages.

Languages	Number of Language Resources
Afrikaans	2
Akkadian	1
Albanian	1
Amharic	3
Arabic	72
Azerbaijani	1
Bantu languages	1
Basque	15
Belarusian	1
Bengali	3
Bosnian	1
Brokpake	1
Bulgarian	54
Burmese	1
Catalan	21
Chinese	107
Chug	2
Croatian	52
Czech	202
Czech Sign Language	3
Danish	65
Dutch	61
Egyptian Arabic	1
English	1677
Esperanto	2
Estonian	41
Finnish	46
French	294
Galician	5
German	304
German Sign Language	1
Hausa	3
Hebrew	5
Hindi	11
Hungarian	26
Icelandic	23
Indonesian	5
Irish	54
Italian	130
Japanese	52
Judeo-Persian	1
Kam	1

Kanuri	1
Khmer	1
Korean	34
Latin	7
Latvian	33
Lithuanian	29
Luxembourgish	1
Macedonian	7
Malagasy	1
Malay (macrolanguage)	1
Malayalam	3
Maltese	23
Mbre	2
Modern Greek (1453-)	83
Mongolian	1
Multiple languages	77
N/A	3
Nepali (macrolanguage)	3
No linguistic content	3
Northern Kurdish	1
Norwegian	4
Norwegian Bokmål	50
Norwegian Nynorsk	6
Nyanja	1
Occitan (post 1500)	1
Oriya (macrolanguage)	2
Oromo	1
Persian	18
Polish	91
Portuguese	96
Pushto	9
Romance languages	1
Romanian	41
Russian	21
Sanskrit	1
Serbian	11
Serbo-Croatian	1
Slovak	37
Slovenian	38
Somali	1
Songhai languages	1
Spanish	433
Standard Moroccan Tamazight	2
Swahili (macrolanguage)	4
Swedish	57
Swiss German	1
Tagalog	1
Tamashek	1
Tamil	4
Telugu	1
Thai	4
Tibetan	1
Tigrinya	1
Turkish	7
Ukrainian	3

Uncoded languages	1
Urdu	4
Vietnamese	7
Welsh	2
Western Frisian	1
Total	4633

Table 5: ELG catalogue – languages covered by ingested datasets

Table 5 also contains two lines with resources not represented by any language:

- **N/A** refers to three human evaluations sets with MT evaluation rankings and assessments (WMT 2015 Human Evaluations, WMT 2016 Human Evaluations and WMT 2017 Human Evaluations).
- **No linguistic content**: this is used for three resources with no language data. One of them (MUSCIMA++) is a dataset of handwritten music notation for musical symbol detection, while the other two (UPC-TALP database of isolated meeting-room acoustic events and FBK-Irst database of isolated meeting-room acoustic events) are databases of acoustic events (noises) used in areas of speech processing.

The table also contains an uncoded-LR which refers to a Komnzo-English lexicon, as well as an entry for "Multiple languages" (77 LRs) which contain data in different languages that have not been specified at the source repository. This will be studied so as to see how this can be better specified for users to reach this information.

The reason behind the total number of 4633 LRs is that many resources contain several languages, and the table represents the number of language occurrences, where one LR in three languages counts as three entries.

5.2 Contributions and Gaps from Pilot Projects and Users

Different means are used to detect gaps and set up the priorities for future ingestion work. This section looks into the contributions from the pilot projects and other platform users. The pilot projects are intended to demonstrate the usefulness of the ELG by contributing datasets or services to the platform. These contributions benefit both the community that will have access to the assets provided as well as the pilot projects that will gain visibility with their work and by displaying it in the ELG. The projects that provide datasets often target severe gaps. Among these, we report on the following:

- **E3C European Clinical Case Corpus**[33]: This project aims to create of a corpus of clinical cases in five European Languages (English, French, Italian, Spanish, Basque). This is a relevant project that will produce NE annotated data in several layers and harmonise annotation. Data in the medical domain is highly searched for in technologies like Information Extraction. This corpus collects open data which will allow its developers to share it widely (under CC-BY-NC). Besides, it targets four official EU languages and a co-official under-resourced language (Basque).
- **EVALITA4ELG – Italia EVALITA Bechmark Linguistic Resources, NLP Services and Tools for the ELG Platfom**[34]: This project aims to broaden the ELG portfolio with more than 50 linguistic resources for Italian, capitalizing on the work done in the last more than ten years within EVALITA[35], a reference on evaluation for Italian. This represents further support for underrepresented languages as well as knowledge and benchmarking sharing. EVALITA4ELG supports several language technologies and tasks.

[33] https://e3c.fbk.eu
[34] http://evalita4elg.di.unito.it
[35] http://www.evalita.it

- **Open Translation Models, Tools and Services**: This project aims at developing an extension of the OPUS-MT project, creating NMT models and tools that will be shared through ELG. Focus is on European minority languages, which are highly underrepresented, and which are of great interest to ELG (as we have seen in Section 5.1). Furthermore, we lack models in the ELG catalogue, and this project brings a very good opportunity to enrich this resource type, too.
- **Extracting Terminological Concept Systems from Natural Language Text**: This is a technology-oriented project that allows to build terminologies and ontological learning. Even if not directly presented as a dataset contribution initiative, the technology deployed could be used in ELG for LR creation.
- **Textual Paraphrase dataset for deep language modelling**: The outcome of this project will be a large Finnish paraphrase dataset, it also creates a dataset for model training, finetuning and testing.

These projects are an excellent proof of concept for the ELG platform and we look forward to their output to further enrich ELG.

Platform users may also provide feedback about their interaction with the platform or about unmet expectations of datasets or LT services. With regard to the latter, and in close connection with the use cases (Section 5), users have raised the need for data in relation to specific technologies. These are listed below and will be targeted shortly so that the experts working on model training and preparing the use cases can use them.

- Text Classification:
    - The Multilingual Amazon Reviews Corpus:
      Supported Languages: English, Japanese, German, French, Spanish and Chinese
      URL: https://registry.opendata.aws/amazon-reviews-ml/
- Named Entity Recognition:
    - CoNLL-2002 Language-Independent Named Entity Recognition (II):
      Supported Languages: Spanish and Dutch
      URL: https://www.clips.uantwerpen.be/conll2002/ner/
    - CoNLL-2003 Language-Independent Named Entity Recognition (II):
      Supported Languages: English and German
      URL: https://www.clips.uantwerpen.be/conll2003/ner/
    - Named Entity Recognition data for Europeana Newspapers:
      Supported Languages: Dutch, French, German
      URL: https://github.com/EuropeanaNewspapers/ner-corpora
    - EVALITA 2009 Named Entity Recognition (NER):
      Supported Languages: Italian
      URL: http://www.evalita.it/2009/tasks/entity

Finally, a survey will be carried out with the collaboration of the recent META-FORUM 2020 participants to query about gaps and needs. As members of the LT community, they are a relevant audience to talk to.

6 Leveraging ELG Resources for Model Training

Task 5.3 explores the ELG as a resource for building and improving NLP models for tasks such as Machine Translation and Text Classification. Due to the exploratory nature of this task, we decided to perform the actual

training offline, i.e., not on the ELG's rented computing infrastructure but on local infrastructure. This puts the focus on the ELG's utility as a catalogue of resources and potentially as a service back-end for the pipeline.

6.1 Programmatic Access to the ELG Catalogue

Model training typically takes place on a server infrastructure through terminal connections with a command line interface. As part of Task 5.3, we developed scripts and packages to access the ELG remotely. We currently have two Python packages related to programmatic access to the ELG, one (developed by UEDIN) focussing on finding and downloading specific datasets, the other (developed by DFKI) focussing on accessing the ELG's services remotely. We are in the process of merging these two efforts into one package.

6.1.1 Offline Access to Datasets

One obvious way to access the catalogue is through the search API. However, we found that the information returned by the REST API for search does not quite meet the needs of the researcher looking for data for model training. In particular, we found that there is no straightforward way to look for specific types of corpora through the search API, such as parallel corpora or annotated corpora for training NER or text classifiers. This information is available in the complete metadata record for each resource but not in the custom view of information returned by the search interface (tailored to the browser-based catalogue GUI). In order to avoid having to request metadata records through HTTP calls for every resource for every search, the data access package caches metadata records locally (similar to known software package managers such as Debian's apt-get).

Currently, the data access managers offer the following actions:

- **list**: list all available resources; these can be filtered by language and corpus type. For example, *./elg list --parallel -L de,en* will list all parallel German/English corpora available through the ELG.
- **info**: show more information about a specific resource (e.g., license and cost)
- **download**: download the resource if a direct download link is available; otherwise, show the link to the access location.
- **metadata**: show the complete metadata record of a given resource (for development and debugging).
- **update**: update the cached metadata records

6.1.2 Programmatic Access to Services Running in ELG

ELG services are executable via REST APIs. This allows client programs to access services offered by ELG as remote procedure calls (RPC), independent of the programming language used. To support potential users in integrating ELG services into their workflow, we developed a software development kit (SDK) for Python. We chose Python as it is the most popular programming language in the field of LT, especially for exploratory work. The ELG Python SDK is available via the Python Package Index (PyPI), i.e., it is installable via pip, and makes use of the ELG REST API for calling LT services.

The ELG Python SDK allows users to execute LT services from Python scripts. The process of using a LT service is as follows. First, users with an ELG account load a LT service through its ID. During this step, users need to authenticate in a browser window and copy and paste the obtained token to their command line interface at execution time. At that moment, the obtained identification tokens are saved in a JSON file, and automatically reused the next time users load an LT service to not have to authenticate again. Once the LT service is loaded, users can execute it by passing either a file or a string. The obtained result is in a JSON format and is the same result as one gets from calling the REST API directly.

As mentioned above, the ELG data manager and the ELG SDK will eventually be merged into one project.

6.2 Use Cases

Below we describe the two use cases for which the usefulness of the ELG will be tested.

6.2.1 Machine Translation

MT has made considerable progress in recent years and continues to be a very active area of research. Many of the research prototypes performing best in terms of translation quality (see, e.g., the results of the most recent WMT Shared Task on News Translation[36]) are expensive to train (often, multiple, very large models are trained and ensembled, with adaptation steps to the specific task). Joining this race and building top-performing systems is not the focus of our work in this task; we aim to develop recipes to build good models using resources available through ELG. Our training framework is the Marian toolkit. The workflow is as follows:

1. Identify and download suitable training data through the ELG.
2. Extract the data.
3. Filter the data to remove poor training samples. Unlike statistical phrase-based machine translation, neural machine translation is fairly sensitive to poor data, and data available through the ELG comes without data quality guarantees. In the first step, we train models on data filtered only by a language identification check. We use a pre-trained language identification model for the FastText [37]text classification system and use a threshold on the certainty of the classifier about each sentence in question to be in the language that the data set claims them to be. By cursory inspection, we found that some datasets (especially crawled data) sometimes contain misclassified sentence pairs. We then train basic models on this filtered data and use those to score all available training data, using Dual Conditional Cross-entropy Filtering[38] to produce a better training set.
4. Train the final model on filtered data.

In the course of this work, we aim to develop reliable recipes with hyperparameters that perform reasonably well for the general case; we are not aiming for a fully automated pipeline as different types of computing resources are required for different steps in the process.

6.2.2 Text Classification

Expert System is interested in releasing a polarity classifier for product reviews in multiple languages. Product reviews are an important feature for branding and marketing of online and offline stores. Positive product reviews can increase sales, and negative reviews might have the opposite effect. In the end these reviews are a credibility indicator of either a product or brand. Most reviews are accompanied with a user rating about the product or brand which summarizes, often in a scale from 1 to 5, the sentiment of the review. Thus, we can leverage these ratings to supervise and evaluate the classifier.

The interest in this classifier is not on the commercial side, since as we mentioned before often the reviews are associated with ratings, but in the internal use within Expert System as part of an evaluation initiative. This evaluation initiative aims at using fractions of the dataset to generate and evaluate the classifiers. The dataset fractions are gradually incremented, and the corresponding classifiers evaluated. The goal is to understand the relation of the effort required to develop the classifiers and the gains in the classifier performance by increasing the dataset size. Expert System NLP tools rely on an extensive knowledge graph of linguistic information

[36] http://www.statmt.org/wmt20/translation-task.html
[37] https://fasttext.cc
[38] https://www.aclweb.org/anthology/W18-6478/

that enables generating high-performant classifiers with small data, as opposed to other approaches that require large datasets to achieve good results.

We plan to use resources already in ELG such as Webis-CLS-10[39]. This dataset comprises about 800,000 Amazon product reviews for three product categories – books, DVDs, music – written in two different languages: English and French. In addition, we are interested in other languages such as German, Spanish and Italian which are the core languages supported by Expert System technology. We have already identified datasets outside of ELG for German and Spanish that we will suggest being uploaded or referenced from ELG if their license allows it. To implement the classifier, we use Jupyter notebooks[40] so that we can document the code, work collaboratively on it, and enable reproducibility. The main steps to be included in the Jupyter notebook are:

- Search and download the datasets of product reviews from ELG using the ELG Python client.
- Inspect the dataset structure to identify where the data is and how it is split.
- Generate the subsets of the dataset (dataset fractions incremented gradually).
- For each subset:
    - Generate training, validation, and test data if required.
    - Train the classifiers.
    - Evaluate the classifiers.
- Report evaluation results.

7 Summary and Conclusions

D5.2 reports on the work done since D5.1 (ELG R1) with regard to the provision of LRs to the ELG. The procedure for identification, conversion, harvesting and ingestion has matured considerably, which has allowed us to enrich the catalogue considerably. The ELG catalogue hosts 2737 LRs at present and is ready to host many more. Ingestion work has concluded for several repositories (several LREC Shared LRs); protocols have been set up for the automatic ingestion of some others (ELRC-SHARE, LINDAT-CLARIAH-CZ); and conversion and analysis are still on the way for the other repositories described in this deliverable. Moreover, the different phases of gap analysis carried out pave the way for some future work, in terms of language and resource type coverage. The contributions from the pilot projects and platform users are also very useful and good indicators of gaps and priorities. Finally, Task 5.3 has started its work and has designed some experiments for model training and use case development, leveraging ELG language resources. Their contribution will be of great significance for the many technology and service developers that will be interested and willing to use the platform.

The annexes below contain all the datasets that have been ingested in ELG so far, classified by source repository. They are listed with their content languages.

A. Annexes

A.A. ELRA Catalogue (1180 LRs)

This annex provides an overview of the LRs ingested into the ELG from the ELRA catalogue.

[39] https://live.european-language-grid.eu/catalogue/#/resource/service/corpus/1392
[40] https://jupyter.org

Resource name	Language(s)
"Le Monde Diplomatique" Arabic tagged corpus	Arabic
"Le Monde Diplomatique" Text corpus in Arabic	Arabic
"Le Monde Diplomatique" Text corpus in English	English
"Le Monde Diplomatique" Text corpus in French - archives 1980-1998	French
"Le Monde Diplomatique" Text corpus in French - archives from 1999	French
2006 CoNLL Shared Task - Ten Languages	Multiple languages
2006 CoNLL Shared Task - Arabic & Czech	Arabic | Czech
2007 CoNLL Shared Task - Arabic & English	English | Arabic
2007 CoNLL Shared Task - Basque, Catalan, Czech & Turkish	Czech | Basque | Catalan | Turkish
2007 CoNLL Shared Task - Greek, Hungarian & Italian	Italian | Modern Greek | Hungarian
88milSMS. A corpus of authentic text messages in French	French
A "scientific" corpus of modern French ("La Recherche" magazine) - Complete version	French
A "scientific" corpus of modern French ("La Recherche" magazine) - Raw data	French
Accented English GlobalPhone	English
ACCOR - English	English
ACL RD-TEC: A Reference Dataset for Terminology Extraction and Classification Research in Computational Linguistics	English
Acoustic database for Polish concatenative speech synthesis	Polish
Acoustic database for Polish unit selection speech synthesis	Polish
aGender	German
Albayzin corpus	Spanish
Alcohol Language Corpus (BAS ALC)	German
Al-Hayat Arabic Corpus	Arabic
Amaryllis Corpus - Evaluation Package	French
American English Conversational Speech Recognition Corpus (Multi-Channel)	English
American English Speech Recognition Corpus (Desktop)	English
American English Speech Recognition Corpus (Mobile)	English
American Spanish Recognition Corpus (Desktop+Mobile)	English
American/Canadian English Speech Recognition Corpus (headset+mobile)	English
Amharic-English bilingual corpus	English | Amharic
ANITA (Audio eNhancement In Telecom Applications)	German | English | French | Spanish
An-Nahar Newspaper Text Corpus	Arabic
APASCI	Italian
Arabic dictionary of inflected words	Arabic
Arabic dictionary of inflected words with recognition of agglutinated clitics and inflection system	Arabic
Arabic Morphological Dictionary	Arabic
Arabic Speech Corpus	Arabic
Arbobanko (Esperanto Treebank)	Esperanto
Arboretum treebank	Danish
ARCADE II Evaluation Package	Multiple languages
ARCADE/ROMANSEVAL corpus	Italian | English | French
A-SpeechDB	Arabic
AURORA Project Database - Aurora 4a - Evaluation Package	English
AURORA Project Database - Aurora 4b - Evaluation Package	English

AURORA Project database - Subset of SpeechDat-Car - Danish database - Evaluation Package	Danish
AURORA Project database - Subset of SpeechDat-Car - Finnish database - Evaluation Package	Finnish
AURORA Project database - Subset of SpeechDat-Car - German database - Evaluation Package	German
AURORA Project database - Subset of SpeechDat-Car - Italian database - Evaluation Package	Italian
AURORA Project database - Subset of SpeechDat-Car - Spanish database - Evaluation Package	Spanish
AURORA Project Database 2.0 - Evaluation Package	English
AURORA-5	English
Australian English Kids Speech Recognition Corpus (Desktop)	English
Australian English Speech Recognition Corpus (Desktop)	English
Australian English Speech Recognition Corpus (Mobile)	English
Austrian SpeechDat(AT) FDB-1000 database	German
Austrian SpeechDat(AT) MDB-1000 database	German
Automobile Engineering	German | French | Spanish | English
BABEL Bulgarian Database	Bulgarian
BABEL Estonian Database	Estonian
BABEL Hungarian Database	Hungarian
BABEL Polish database	Polish
BABEL Romanian database	Romanian
BAS GEO1	German
BAS PHATT 1.0.X (sub-set)	German
BAS PHATT 1.1.X (complete corpus)	German
Basic multilingual lexicon (MEMODATA)	French | English | Italian | German | Spanish
Basque FDB-1060 database (SpeechDat-like)	Basque
Basque Spoken Corpus, by Jon Aske (Department of Foreign Languages, Salem State College - Salem, Massachusetts, USA)	Basque
Basque WordNet	Basque
BDBRUIT	French
BDLEX	French
BDSONS Base de données des sons du français	French
Bilingual (Spanish-English) Speech synthesis HTS models	English | Spanish
Bilingual Collocational Dictionary (Horst Bogatz)	German | English
Bilingual English-Russian Russian-English Dictionaries	English | Russian
BioLexicon	English
Biology Database	Korean | English
BITS Logatome Synthesis Corpus - BITS-LG	German
BITS Unit Selection Synthesis Corpus	German
Bizkaifon (Bizkaieraren Fonoteka)	Basque
BrasiLEX Brazilian Portuguese lexicon	Portuguese
Brazilian Portuguese Speech Recognition Corpus (Desktop)	Portuguese
BREF-120 - A large corpus of French read speech	French
BREF-80	French
BREF-POLYGLOT	French
British English Kids Speech Recognition Corpus (Desktop)	English

British English Source Lexicon (BESL) version 2.2	English
British English Speech Recognition Corpus (Desktop)	English
British English Speech Recognition Corpus (Mobile)	English
British English SpeechDat(II) FDB-4000	English
British English SpeechDat(II) MDB-1000	English
British English SpeechDat(II) SDB-2400	English
British-English SpeechDat-Car	English
Bulgarian Linguistic Database	Bulgarian
Bulgarian Morphological Dictionary	Bulgarian
Bulgarian WordNet	Bulgarian
Canadian English Speech Recognition Corpus (Desktop)	English
Canadian English Speech Recognition Corpus (Telephone) - person name	English
Canadian English Speech Recognition Corpus (Telephone) - place name	English
Canadian English Speech Recognition Corpus (Telephone) - sentences	English
Canadian English Speech Recognition Corpus (Telephone) - spell words	English
Canadian French Speech Recognition Corpus (Mobile)	French
Cantonese Readings Database	Chinese
Cantonese Speecon database	Chinese
CAREGIVER Corpus	English | Finnish | Dutch
Catalan Corpus of News Articles	Catalan
Catalan SpeechDat-Car database	Catalan
Catalan Speecon database	Catalan
Catalan-Spanish Parallel Corpus	Spanish | Catalan
Catalan-SpeechDat For the Fixed Telephone Network Database	Catalan
Catalan-SpeechDat for the Mobile Telephone Network Database	Catalan
CELEX Dutch lexical database - Complete set	Dutch
CELEX Dutch lexical database - Derivational Morphology Subset	Dutch
CELEX Dutch lexical database - Frequency Subset	Dutch
CELEX Dutch lexical database - Inflectional Morphology Subset	Dutch
CELEX Dutch lexical database - Orthography Subset	Dutch
CELEX Dutch lexical database - Phonology Subset	Dutch
CELEX Dutch lexical database - Syntax Subset	Dutch
CEPLEXicon	Portuguese
CESART Evaluation Package	French
CESTA Evaluation Package	English | French | Arabic
CHIEDE Corpus: a spontaneous child language corpus of Spanish	Spanish
CHIL 2004 Evaluation Package	English
CHIL 2005 Evaluation Package	English
CHIL 2006 Evaluation Package	English
CHIL 2007 Evaluation Package	English
CHIL 2007+ Evaluation Package	English
Chinese English Speech Recognition Corpus (Desktop)	English
Chinese Lexical Database	Chinese
Chinese Mandarin (North) database	Chinese
Chinese Mandarin (South) database	Chinese
Chinese Mandarin Speech Recognition Corpus (Mobile)	Chinese

Chinese Morphological Database	Chinese
Chinese Phonological Database	Chinese
Chinese-English Database of Proper Nouns	English | Chinese
Chinese-English Database of Proverbs and Idioms (Chengyu)	English | Chinese
Chinese-Japanese Database of Proper Nouns	Japanese | Chinese
Chinese-Japanese Technical Terms Dictionary	Japanese | Chinese
CINTIL-DeepBank	Portuguese
CINTIL-DependencyBank	Portuguese
CINTIL-PropBank	Portuguese
CINTIL-TreeBank	Portuguese
CLE Pakistan Urdu Speech Corpus	Urdu
CLEF AdHoc-News Test Suites (2004-2008) - Evaluation Package	Multiple languages
CLEF Domain Specific Test Suites (2004-2008) - Evaluation Package	German | English | Russian
CLEF QAST (2007-2009) - Evaluation Package	English | French | Spanish
CLEF Question Answering Test Suites (2003-2008) - Evaluation Package	Multiple languages
CLEFeHealth 2013 Task 3 Evaluation Package	English
CLEFeHealth 2014 Task 3 Evaluation Package	English
CLIPS_MT_MANUAL	Italian
Collins Multilingual database (MLD) - PhraseBank	Multiple languages
Collins Multilingual database (MLD) - WordBank	Multiple languages
Collins Multilingual database (MLD) - PhraseBank with audio files	Multiple languages
Collins Multilingual database (MLD) - WordBank with audio files	Multiple languages
Comprehensive Word List of Japanese	Japanese
Comprehensive Word List of Korean	Korean
Comprehensive Word List of Traditional Chinese	Chinese
Comprehensive Word Lists for Chinese, Japanese, Korean and Arabic	Japanese | Korean | Chinese | Arabic
Comprehensive Wordlist of Simplified Chinese	Chinese
Computer Science Database	Korean | English
Concise Oxford Dictionary - Audio Files	English
Concise Oxford Spanish Dictionary	English | Spanish
Concise Oxford-Duden German Dictionary	German | English
CORAL Corpus	Portuguese
C-ORAL-ROM - Integrated reference corpora for spoken romance languages. Multi-media edition; tools of analysis; standard linguistic measurements for validation in HLT	Spanish | French | Portuguese | Italian
Corpus of Contemporaneous Spanish Novels	Spanish
COST232	English
CRATER 2 Corpus	English | French | Spanish
CRATER corpus	English | French | Spanish
Czech Audio-Visual Speech Corpus for Recognition with Impaired Conditions	Czech Sign Language
Czech Sign Language Corpus for Recognition - Amateur Signer	Czech Sign Language
Czech Sign Language Corpus for Recognition - Professional Signer	Czech Sign Language
Czech SpeechDat(E) Database	Czech
Czech Speecon database	Czech
Czech WordNet	Czech
Danish EUROM1	Danish
Danish Propbank	Danish

Danish SpeechDat(II) FDB-1000	Danish
Danish SpeechDat(II) FDB-4000	Danish
Danish SpeechDat(M) database - DB1	Danish
Danish SpeechDat(M) database - DB2	Danish
Danish SpeechDat-Car - Full database	Danish
Danish SpeechDat-Car - GSM recordings - GSM recordings only	Danish
Danish SpeechDat-Car - In-car recordings	Danish
Database of Arab Names	Arabic
Database of Arab Names in Arabic	Arabic
Database of Arabic Place Names	Arabic
Database of Arabic Plurals	Arabic
Database of Chinese Full Names	Chinese
Database of Chinese Name Variants	Chinese
Database of Chinese Names	Chinese
Database of Foreign Names in Arabic	Arabic
Database of Japanese Name Variants	Japanese
Database of Persian Names	Persian
DEFT'08 Evaluation Package	French
deL1L2IM corpus	German
DICO-MORPH_Collocation	French
DICO-MORPH_Lemme	French
DICO-SYNT	French
Dictionary of Law	English
Dictionary of Medicine	English
DixAF (Bilingual Dictionary French Arabic, Arabic French)	French | Arabic
Dutch Lexicon	Dutch
Dutch PAROLE Distributable Corpus	Dutch
Dutch PAROLE lexicon	Dutch
Dutch Polyphone Database	Dutch
Dutch-French Lexicon (LanTmark)	French | Dutch
EASy Evaluation Package	French
ECI/MCI (European Corpus Initiative/Multilingual Corpus I)	Multiple languages
ECI-ELSNET Italian & German tagged sub-corpus	Italian | German
ECPC Corpus (European Comparable and Parallel Corpora of Parliamentary Speeches Archive) - set 1	English | Spanish
Egyptian Arabic Speecon database	Arabic
Electrical Engineering	German | French | Spanish | English
Eleftherotypia Journal Speech database	Modern Greek (1453-)
Emotional speech synthesis database	Spanish
Energy Technology	French | English | Italian | German | Spanish
English lexicon with morphological information	English
English SpeechDat Polyphone database DB1	English
English SpeechDat(M) Polyphone database DB2	English
English-Japanese Dictionary	Japanese | English
English-French Lexicon (LanTmark)	English | French
English-Nepali Parallel Corpus	English | Nepali (macrolanguage)

English-Persian database of idioms and expressions	English | Persian
English-Persian parallel Corpus	Persian | English
English-Persian terminology database of computer and IT	English | Persian
English-Persian terminology database of management and economics	English | Persian
English-to-Simplified Chinese Dictionary	English | Chinese
English-Vietnamese Parallel Corpus	Vietnamese | English
EnToFrNE - a Parallel English-French Lexicon of Named Entities	English | French
EnToSSLNE - a Lexicon of Parallel Named Entities from English to South Slavic Languages	Multiple languages
EPAC Corpus: orthographic transcriptions	French
EQueR Evaluation Package	French
Erlanger Bahnansage - ERBA	German
ESTER 2 Corpus	French
ESTER Corpus	French
ESTER Evaluation Package	French
ETAPE Evaluation Package	French
euLEX (Lexical Database for Basque)	Basque
EUROM1e English	English
EUROM1f French	French
EUROM1g German	German
EUROPARL Corpus Parallel Corpora: Portuguese-English	English | Portuguese
European Parliament Interpretation Corpus (EPIC)	Italian | English | Spanish
EuroWordNet Czech	Czech
EuroWordNet Dutch	Dutch
EuroWordNet English Addition to English WordNet	English
EuroWordNet Estonian	Estonian
EuroWordNet French	French
EuroWordNet German	German
EuroWordNet Spanish	Spanish
Euskararen Datu-Base Lexikala (EDBL) - Lexical Database for Basque	Basque
EvaSy Evaluation Package	French
Farsdat (Farsi Speech Database)	Persian
FASiL combined unimodal “fasil-all” corpus	English | Portuguese | Swedish
FASiL English unimodal “fasil-uk” corpus	English
FASiL multimodal “fasil-mm” corpus	English | Portuguese | Swedish
FASiL Portuguese unimodal “fasil-pt” corpus	Portuguese
FASiL Swedish unimodal “fasil-sv” corpus	Swedish
FBK-Irst database of isolated meeting-room acoustic events	No linguistic content
FESTCAT Catalan TTS baseline female speech database	Catalan
FESTCAT Catalan TTS baseline male speech database	Catalan
FESTCAT Catalan TTS baseline speech database - 8 speakers	Catalan
Finnish Speechdat(II) FDB-1000	Finnish
Finnish Speechdat(II) FDB-4000	Finnish
Finnish SpeechDat-Car	Finnish
Finnish Speecon database	Finnish
Finnish-Swedish Speechdat(II) FDB-1000	Swedish

FoxPersonTracks: a Benchmark for Person Re-Identification from TV Broadcast Shows	French
France French Speech Recognition Corpus (Desktop)	French
French dictionary of definitions (SYNAPSE)	French
French Lexicon	French
French lexicon with morphological information	French
French Source Lexicon	French
French Speecon database	French
French-Canadian Speecon database	French
French-Dutch Lexicon (LanTmark)	French | Dutch
French-English Lexicon (LanTmark)	English | French
Fundamental Portuguese Corpus	Portuguese
GeFRePaC - German French Reciprocal Parallel Corpus	German | French
GEOBASE	English | French
German Kids Speech Recognition Corpus (Desktop)	German
German Pronunciation Rules Set - PHONRUL 9.0	German
German Speech Recognition Corpus (Desktop)	German
German SpeechDat(II) MDB-1000	German
German SpeechDat-Car	German
German Speecon database	German
GLiCom Spanish Wordform list - Regular word-forms	Spanish
GLiCom Spanish Wordform list - Regular word-forms + verb-clitic combinations	Spanish
GLiCom Spanish Wordform list - Verb-clitic combinations	Spanish
Glissando-ca	Catalan
Glissando-sp	Spanish
GlobalPhone 2000 Speaker Package	Multiple languages
GlobalPhone Arabic	Arabic
GlobalPhone Arabic Pronunciation Dictionary	Arabic
GlobalPhone Bulgarian	Bulgarian
GlobalPhone Bulgarian Pronunciation Dictionary	Bulgarian
GlobalPhone Bulgarian Pronunciation Dictionary 260k entries (extended version)	Bulgarian
GlobalPhone Chinese-Mandarin	Chinese
GlobalPhone Chinese-Mandarin Pronunciation Dictionary	Chinese
GlobalPhone Chinese-Shanghai	Chinese
GlobalPhone Croatian	Croatian
GlobalPhone Croatian Pronunciation Dictionary	Croatian
GlobalPhone Czech	Czech
GlobalPhone Czech Pronunciation Dictionary	Czech
GlobalPhone French	French
GlobalPhone French Pronunciation Dictionary	French
GlobalPhone German	German
GlobalPhone German Pronunciation Dictionary	German
GlobalPhone Hausa	Hausa
GlobalPhone Hausa Pronunciation Dictionary	Hausa
GlobalPhone Japanese	Japanese
GlobalPhone Japanese Pronunciation Dictionary	Japanese
GlobalPhone Korean	Korean

GlobalPhone Korean Pronunciation Dictionary	Korean
GlobalPhone Multilingual Model Package	Multiple languages
GlobalPhone Polish	Polish
GlobalPhone Polish Pronunciation Dictionary	Polish
GlobalPhone Portuguese (Brazilian)	Portuguese
GlobalPhone Portuguese (Brazilian) Pronunciation Dictionary	Portuguese
GlobalPhone Russian	Russian
GlobalPhone Russian Pronunciation Dictionary	Russian
GlobalPhone Spanish (Latin American)	Spanish
GlobalPhone Spanish (Latin American) Pronunciation Dictionary	Spanish
GlobalPhone Swahili	Swahili (macrolanguage)
GlobalPhone Swahili Pronunciation Dictionary	Swahili (macrolanguage)
GlobalPhone Swedish	Swedish
GlobalPhone Swedish Pronunciation Dictionary	Swedish
GlobalPhone Tamil	Tamil
GlobalPhone Thai	Thai
GlobalPhone Thai Pronunciation Dictionary	Thai
GlobalPhone Turkish	Turkish
GlobalPhone Turkish Pronunciation Dictionary	Turkish
GlobalPhone Ukrainian	Ukrainian
GlobalPhone Ukrainian Pronunciation Dictionary	Ukrainian
GlobalPhone Vietnamese	Vietnamese
GlobalPhone Vietnamese Pronunciation Dictionary	Vietnamese
Gram Vaani data set	Hindi
Greek SpeechDat(II) FDB-5000	Modern Greek (1453-)
Greek SpeechDat-Car	Modern Greek (1453-)
GRONINGEN	Dutch
GVLEX tales corpus	French
Hanzi Pinyin Database for Simplified Chinese	Chinese
Hebrew Speecon database	Hebrew
Helsinki Corpus of Swahili	Swahili (macrolanguage)
Hempel	German
Hong Kong Cantonese Speech Recognition Corpus (Desktop)	Chinese
Hungarian Speecon database	Hungarian
Hydrogeology database	English | French
IBNC - An Italian Broadcast News Corpus	Italian
ICE-GB (British English component of the International Corpus of English)	English
IDIOLOGOS 1 “Bootstrap” (NEOLOGOS Project)	French
IDIOLOGOS 2 “Eingenspeakers” (NEOLOGOS Project)	French
ILC Italian Morphological Lexicon	Italian
ILE: Italian LExicon	Italian
ILPho phonetic lexicon	French
ILSP/ELEFTHEROTYPIA Corpus (Greek corpus)	Modern Gree
Insurance (Termcat)	English | Spanish | Catalan
ISLE Speech Corpus	English
Italian Kids Speech Recognition Corpus (Desktop)	Italian

Italian lexicon with morphological information	Italian
Italian lexicon with morphological information and clitic verbs	Italian
Italian Speech Corpus 1 (Appen)	Italian
Italian Speech Recognition Corpus (Desktop)	Italian
Italian SpeechDat-Car database	Italian
Italian Speecon database	Italian
Italian Syntactic-Semantic Treebank (ISST)	Italian
Italian TTS Speech Corpus (Appen)	Italian
ItalWordNet (Italian WordNet)	Italian
Japanese - English Dictionary of Technical Terms	Japanese | English
Japanese Companies and Organizations	Japanese
Japanese English Speech Recognition Corpus (Desktop)	English
Japanese English Speech Recognition corpus (Mobile)	English
Japanese Kids Speech database (Lower Grade)	Japanese
Japanese Kids Speech database (Upper Grade)	Japanese
Japanese Lexical Database	Japanese
Japanese Morphological Database	Japanese
Japanese Orthographical Database	Japanese
Japanese Phonological Database	Japanese
Japanese Speech Recognition Corpus (Desktop)	Japanese
Japanese Speech Recognition Corpus (desktop) - name, digit string, place, sentences (200 people)	Japanese
Japanese Speech Recognition Corpus (Desktop) - sentences (200 people)	Japanese
Japanese Speech Recognition Corpus (desktop) - digit string (200 people)	Japanese
Japanese Speech Recognition Corpus (desktop) - Japanese person name (200 people)	Japanese
Japanese Speech Recognition Corpus (desktop) - Japanese place name (200 people)	Japanese
Japanese Speech Recognition Corpus (Mobile)	Japanese
Japanese - English Database of Proper Nouns	Japanese | English
Japanese - English Dictionary	Japanese | English
JV_TDM Corpus	French
Karl May Korpus (KMK)	German
Khresmoi manually annotated reference corpus	English
Korean English Speech Recognition Corpus (Mobile)	English
Korean Lexical Database	Korean
Korean Lexicon	Korean
Korean Speech Recognition corpus (Desktop) - name, digit string, place, sentences	Korean
Korean Speech Recognition Corpus (desktop) - digit string (110 people)	Korean
Korean Speech Recognition Corpus (desktop) - person name (150 people)	Korean
Korean Speech Recognition Corpus (desktop) - place name (150 people)	Korean
Korean Speech Recognition Corpus (desktop) - single Korean sentences (40 people)	Korean
Korean Speech Recognition Corpus (Desktop+Mobile)	Korean
Korean Speecon database	Korean
Korean-Chinese Database of Proper Nouns	Korean | Chinese
Korean-English Database of Proper Nouns	Korean | English
Korean-Japanese Database of Proper Nouns	Japanese | Korean
Korean-Japanese Dictionary of Technical Terms	Japanese | Korean
KORLEX - Croatian Lexicon	Croatian

KORLEX - Serbian Lexicon	Serbian
LABEL-LEX (MW)	Portuguese
LABEL-LEX (SW)	Portuguese
Labelling of WordNet 1.6 with semantic fields (WordNet Domains)	Italian | English
Laboratory Conditions Czech Audio-Visual Speech Corpus	Czech
Large Farsdat	Persian
LC-STAR Catalan phonetic lexicon	Catalan
LC-STAR English-Finnish Bilingual Aligned Phrasal lexicon	English | Finnish
LC-STAR English-German Bilingual Aligned Phrasal lexicon	German | English
LC-STAR English-Hebrew (Israel) Bilingual Aligned Phrasal lexicon	English | Hebrew
LC-STAR English-Italian Bilingual Aligned Phrasal lexicon	Italian | English
LC-STAR English-Slovenian Bilingual Aligned Phrasal lexicon	English | Slovenian
LC-STAR Finnish Phonetic lexicon	Finnish
LC-STAR German Phonetic lexicon	German
LC-STAR German Phonetic lexicon in the Touristic Domain	German
LC-STAR Greek Phonetic lexicon	Modern Greek
LC-STAR Hebrew (Israel) phonetic lexicon	Hebrew
LC-STAR Italian Phonetic lexicon	Italian
LC-STAR Mandarin Chinese Phonetic lexicon	Chinese
LC-STAR Slovenian Phonetic lexicon	Slovenian
LC-STAR Spanish phonetic lexicon	Spanish
LC-STAR Standard Arabic Phonetic lexicon	Arabic
LC-STAR US English phonetic lexicon	English
LECTRA (LECture TRAnscriptions in European Portuguese)	Portuguese
LexIn 2:e Swedish Lexicon	Swedish
LEX-MWE-PT - Word Combination in Portuguese	Portuguese
LILA Korean database	Korean
Linguistics (Termcat)	English | Spanish | Catalan
LORETO Thesaurus	Multiple languages
LT Corpus	Portuguese
LusoLEX European Portuguese Lexicon	Portuguese
M2VTS Speaker Verification Database	French
Macedonian lexicon of compound words (MACPLEX_COMP)	Macedonian
Macedonian lexicon of derived adjectives (MACPLEX_ADJDERV)	Macedonian
Macedonian lexicon of participles (MACPLEX_ADJPARTIC)	Macedonian
Macedonian lexicon of proper nouns (MACPLEX_PROPERS)	Macedonian
Macedonian lexicon of toponyms (MACPLEX_TOPO)	Macedonian
Macedonian Morphological Lexicon (MACPLEX)	Macedonian
Mandarin Chinese Desktop Speech Recognition Corpus - Digit String (120 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Digit String (200 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Digit String (849 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Digit String (98 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Monosyllabic (94 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Person name (849 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Person name, Place Name (10 people)	Chinese

Mandarin Chinese Desktop Speech Recognition Corpus - Person Name, Place Name (70 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Simple Chinese sentences (850 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - SMS (120 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - SMS (200 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Spontaneous Speech (50 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Spontaneous Speech (849 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Stock (70 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Stock (849 people)	Chinese
Mandarin Chinese Desktop Speech Recognition Corpus - Stock, Person Name, String, Simple Chinese sentences, Spontaneous Speech (50 people)	Chinese
Mandarin Chinese high clarity Speech Recognition Corpus (in recording studio) - single Chinese sentence (200 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (desktop) - digit string (119 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (desktop) - digit string (200 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (desktop) - person name (120 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (desktop) - place name (120 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (desktop) - place name (200 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (desktop) - short message (120 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (desktop) - single Chinese sentence (200 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (desktop)- person name (200 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (in the car) - person name, place name in Beijing, stocks, digit string (20 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (telephone channel) - Chinese single sentence (100 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (telephone channel) - digit string (100 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (telephone channel) - person name (100 people)	Chinese
Mandarin Chinese Speech Recognition Corpus (telephone channel) - place name (100 people)	Chinese
Mandarin Chinese Speecon database	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - Digit String	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - Digit String (649 people)	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - Person Name (649 people)	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - Person Name, Place Name (Mobile telephone 265)	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - Simple Chinese sentences (649 people)	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - Spontaneous Speech (648 people)	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - Stock	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus -Person Name, Place Name	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - SMS (Fixed phone 86)	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - SMS (Mobile telephone 64)	Chinese
Mandarin Chinese Telephone Speech Recognition Corpus - Stock (649 people)	Chinese
MAURDOR Evaluation Package	English | French | Arabic

Mbochi speech corpus	French | Bantu languages
MCL - Multifunctional Computational Lexicon of Contemporary Portuguese	Portuguese
MDT Mandarin Chinese Conversational Recognition Corpus - 1 channel	Chinese
MDT Mandarin Chinese Conversational Recognition Corpus - 2 channels	Chinese
MDT Mandarin Chinese Conversational Recognition Corpus - 3 channels	Chinese
MDT Mandarin Chinese Conversational Recognition Corpus - Complete set	Chinese
Mechanical Engineering	German | English | French | Spanish
MEDAR Evaluation Package	English | Arabic
MEDIA Evaluation Package	French
MEDIA speech database for French	French
Metalogue Multi-Issue Bargaining Dialogue	English
Mexican Spanish Kids Speech Recognition Corpus (Desktop)	Spanish
MHATLex	French
MICROAES	Spanish
MIST Multi-lingual Interoperability in Speech Technology database	German | French | English | Dutch
MLCC Multilingual and Parallel Corpora	Multiple languages
Modern French Corpus including Anaphors Tagging	French
Monolingual Greek corpus	Modern Greek (1453-)
MoveOn Speech and Noise Corpus	English
MTP Annotated German corpus - tagged version	German
MTP Annotated German corpus - untagged version	German
MULTEXT JOC Corpus	French | English | Italian | German | Spanish
MULTEXT Lexicons	French | English | Italian | German | Spanish
MULTEXT Prosodic database	French | English | Italian | German | Spanish
Multilingual Corpus	Korean | English | Chinese
Multilingual Database of Japanese Points-of-Interest 1	Korean | English | Chinese | Japanese
Multilingual Dictionary of Sports - English-French bilingual database	English | French
Multilingual Dictionary of Sports - English-French-Arabic trilingual database	English | French | Arabic
Multilingual Dictionary of Sports - English-French-German trilingual database	German | English | French
Multilingual Dictionary of Sports - English-French-Greek trilingual database	Modern Greek | English | French
Multilingual Dictionary of Sports - English-French-Greek-Arabic-German-Spanish-Portuguese multilingual database	Multiple languages
Multilingual Dictionary of Sports - English-French-Portuguese trilingual database	English | French | Portuguese
Multilingual Dictionary of Sports - English-French-Spanish trilingual database	English | French | Spanish
Multilingual Phrasebank	French | Portuguese | English | Italian | German | Spanish
Multilingual Proper Noun Database	Japanese | Korean | Chinese | English
Multilingual Wordbank	French | Portuguese | English | Italian | German | Spanish
MultiWordNet database (included semantic fields) (MultiWordNet)	Italian | English
N4 (NATO Native and Non-Native) database	English
Nautilus Speaker Characterization (NSC) Corpus	German
NE3L named entities Arabic corpus	Arabic
NE3L named entities Chinese corpus	Chinese

NE3L named entities Russian corpus	Russian
NEMLAR Broadcast News Speech Corpus	Arabic
NEMLAR Speech Synthesis Corpus	Arabic
NEMLAR Written Corpus	Arabic
Nepali Monolingual written corpus	Nepali (macrolanguage)
Nepali Spoken Corpus	Nepali (macrolanguage)
NetDC Arabic BNSC (Broadcast News Speech Corpus)	Arabic
New Oxford Dictionary of English, 2nd Edition	English
New Oxford Thesaurus of English	English
NEWBASE - Extended version of ELRA-T0090 GEOBASE	English | French
NODE+DIMAP	English
Normalized Arabic Fragments for Inestimable Stemming (NAFIS)	Arabic
Norwegian EUROM1	Norwegian
Norwegian SpeechDat(II) FDB-1000	Norwegian
NPChunks	Portuguese
NUM 5M Mongolian written corpus	Mongolian
Offensive Word Filter 1	English
Offensive Word Filter 2	English
Olympic Sports (Termcat)	English | French | Spanish | Catalan
ONOMASTICA-COPERNICUS DATABASE	Multiple languages
OrienTel Arabic as spoken in Israel database	Arabic
OrienTel Egypt MCA (Modern Colloquial Arabic) database	Arabic
OrienTel Egypt MSA (Modern Standard Arabic) database	Arabic
OrienTel English as spoken in Egypt database	English
OrienTel English as spoken in Jordan database	English
OrienTel French as spoken in Morocco database	French
OrienTel French as spoken in Tunisia database	French
OrienTel Hebrew database	Hebrew
OrienTel Jordan MCA (Modern Colloquial Arabic) database	Arabic
OrienTel Jordan MSA (Modern Standard Arabic) database	Arabic
OrienTel Morocco MCA (Modern Colloquial Arabic) database	Arabic
OrienTel Morocco MSA (Modern Standard Arabic) database	Arabic
OrienTel Tunisia MCA (Modern Colloquial Arabic) database	Arabic
OrienTel Tunisia MSA (Modern Standard Arabic) database	Arabic
Oxford Business French Dictionary	English | French
Oxford Business Spanish Dictionary	French
Oxford English phonetics files	English
Oxford French Minidictionary	English | French
Oxford Paperback Thesaurus, 2nd edition	English
PAIDIALOGOS (NEOLOGOS Project)	French
PANACEA English-French and English-Greek parallel corpus acquired for Environment domain	Modern Greek | English | French
PANACEA English-French and English-Greek parallel corpus acquired for Labour Legislation domain	Modern Greek | English | French
PANACEA Environment English monolingual corpus	English
PANACEA Environment French monolingual corpus	French
PANACEA Environment Greek monolingual corpus	Modern Greek

PANACEA Environment Italian monolingual corpus	Italian
PANACEA Environment Spanish monolingual corpus	Spanish
PANACEA Labour English monolingual corpus	English
PANACEA Labour French monolingual corpus	French
PANACEA Labour Greek monolingual corpus	Modern Greek
PANACEA Labour Italian monolingual corpus	Italian
PANACEA Labour Spanish monolingual corpus	Spanish
Parallel EMG-Acoustic English GlobalPhone	English
PAROLE English lexicon	English
PAROLE French Corpus	French
PAROLE Greek Lexicon	Modern Greek
PAROLE Irish Distributable Corpus	Irish
PAROLE Italian Corpus	Italian
PAROLE Portuguese Corpus - complete version	Portuguese
PAROLE Portuguese Corpus - tagged subset	Portuguese
PAROLE Portuguese Lexicon	Portuguese
PAROLE Spanish Lexicon	Spanish
PAROLE-SIMPLE-CLIPS PISA Italian Lexicon - Full lexicon	Italian
PAROLE-SIMPLE-CLIPS PISA Italian Lexicon - Morphological layer	Italian
PAROLE-SIMPLE-CLIPS PISA Italian Lexicon - Phonetic layer	Italian
PAROLE-SIMPLE-CLIPS PISA Italian Lexicon - Semantic layer	Italian
PAROLE-SIMPLE-CLIPS PISA Italian Lexicon - Syntactic layer	Italian
Pashto phonetic lexicon	Pushto
Pashto Speech Recognition corpus (Desktop)	Pushto
Pedology database	English | French
Persian 1984 corpus (Multext-East framework)	Persian
Persian Audio Dictionary	Persian
Persian Lexicon	Persian
Persian Multext-East framework lexicon	Persian
Persian Speech Corpus	Persian
PHONDAT 1 - PD1 (2nd edition)	German
PHONDAT 2 - PD2 (2nd edition)	German
Phonetically Balanced Sentences	Korean
Phonetically Balanced Words (1)	Korean
Phonetically Balanced Words (2)	Korean
Phonetically Balanced Words (3)	Korean
Phonetically Balanced Words (4)	Korean
Phonetically Balanced Words (5)	Korean
Phonetically Rich Words	Korean
PHONOLEX (BAS/DFKI)	German
Pocket Oxford Italian Dictionary	Italian | English
Polderland Dutch General Lexicon	Dutch
Polderland Dutch Lexicon of Abbreviations and Acronyms	Dutch
Polderland Dutch Lexicon of Business Terminology	Dutch
Polderland Dutch Lexicon of Legal Terminology	Dutch
Polderland Dutch Lexicon of Medical Terminology	Dutch

Polderland Dutch Lexicon of Names	Dutch
Polderland Dutch Lexicon of Social Terminology	Dutch
Polderland Dutch Lexicon of Technical Terminology	Dutch
POLEX Polish Lexicon	Polish
Polish Speecon database	Polish
POLYCOST	English
PolyVar	French
PortMedia French and Italian corpus	Italian | French
Portuguese Speech Recognition Corpus (Desktop)	Portuguese
Portuguese SpeechDat(II) FDB-4000	Portuguese
Portuguese SpeechDat(M) database	Portuguese
Portuguese Speecon database	Portuguese
PRESS 65	Swedish
Pronunciation lexicon of British place names, surnames and first names	English
PTPARL Corpus	Portuguese
Quaero Broadcast News Extended Named Entity corpus	French
Quaero Old Press Extended Named Entity corpus	French
Qualified POS Tagged Corpus	Korean
REPERE Evaluation Package	French
ROCO Romanian journalistic corpus	Romanian
ROMBAC - Romanian balanced corpus	Romanian
Russian Speech Database	Russian
Russian Speech Kids Recognition Corpus (Desktop)	Russian
Russian Speech Recognition Corpus (Desktop)	Russian
RVG1 (Regional Variants of German 1, Part 1)	German
RVG-J (Regional Variants of German J)	German
SALA II Spanish from Costa Rica database	Spanish
SALA II Spanish from Mexico database	Spanish
SALA II Spanish Mobile Network Database collected in Venezuela	Spanish
SALA II US English database (2000 speakers)	English
SALA Spanish Mexican Database	Spanish
SALA Spanish Venezuelan Database	Spanish
SCI-AN-ALL English-German Bilingual Dictionary	German | English
SCI-ANES English-Spanish Bilingual Dictionary	English | Spanish
SCI-FRAL-EURADIC French-German Bilingual Dictionary	German | French
SCI-FRAN-EURADIC French-English Bilingual Dictionary	English | French
SCI-FRES-EURADIC French-Spanish Bilingual Dictionary	French | Spanish
SCI-FRIT-EURADIC French-Italian Bilingual Dictionary	Italian | French
SCIPER-AL-EURADIC German Monolingual Dictionary	German
SCIPER-AN-EURADIC English Monolingual Dictionary	English
SCIPER-ES-EURADIC Spanish Monolingual Dictionary	Spanish
SCIPER-FR-EURADIC French Monolingual Dictionary	French
SCIPER-IT-EURADIC Italian Monolingual Dictionary	Italian
SecuVoice	Spanish
Serbian emotional speech database (GEES)	Serbian

Shorter Oxford English Dictionary - Audio Files	English
SIEMENS 100 - SI100	German
SIEMENS 1000 - SI1000	German
Siemens Synthesis Corpus - SI1000P	German
SIGNUM Database	German Sign Language
Simplified Chinese-to-English Dictionary	English | Chinese
Simplified Chinese<->English Technical Terms	English | Chinese
Simplified to Traditional Chinese Conversion	Chinese
Slovak SpeechDat(E) Database	Slovak
Slovenian BNSI Broadcast News Speech Corpus	Slovenian
SmartKom Audio	German
SmartKom Home	German
SmartKom Mobil	German
SmartKom Public	German
SmartWeb Handheld Corpus (SHC)	German
SmartWeb Motorbike Corpus (SMC)	German
SmartWeb Video Corpus (SVC)	German
Spain Spanish Kids Speech Recognition Corpus (Desktop)	Spanish
Spain Spanish Speech Recognition Corpus (Desktop)	Spanish
Spanish EUROM.1	Spanish
Spanish Festival HTS models - female speech	Spanish
Spanish Festival HTS models - male speech	Spanish
Spanish Festival voice female	Spanish
Spanish Festival voice male	Spanish
Spanish Full-form Lexicon (Bilingual)	English | Spanish
Spanish Full-form Lexicon (Monolingual)	Spanish
Spanish gilcUB-M Dictionary	Spanish
Spanish lexicon with morphological information	Spanish
Spanish Speech Corpus 1 (Appen)	Spanish
Spanish SpeechDat Database for the Mobile Telephone Network	Spanish
Spanish SpeechDat(II) FDB-1000	Spanish
Spanish SpeechDat(II) FDB-4000	Spanish
Spanish SpeechDat(M) - DB1	Spanish
Spanish SpeechDat(M) - DB2	Spanish
Spanish SpeechDat-Car database	Spanish
Spanish TTS Speech Corpus (Appen)	Spanish
Speaking atlas of the regional languages of France	Multiple languages
SpeechDat Catalan FDB database	Catalan
SpeechDat Galician Database for the Fixed Telephone Network	Galician
SpeechDat Speaker Verification database	French
Speechtera Pronunciation Dictionary	Portuguese
SPINA Corpus ("Robots Commands")	German
SPK	Italian
Spoken Portuguese Corpus	Portuguese
Statistics (Termcat)	English | Spanish | Catalan
STO SprogTeknologisk Ordbase (Danish Lexicon for NLP/HLT Applications)	Danish

Strange Corpus 1 - SC1 (ACCENTS)	German
Strange Corpus 10 - SC10 ('Accents II')	German
Strange Corpus 2 - SC2 (Noises)	German
Swedish EUROM1	Swedish
Swedish SpeechDat(II) FDB-1000	Swedish
Swedish SpeechDat(II) FDB-5000	Swedish
Swedish SpeechDat(II) MDB-1000	Swedish
Swedish Speecon database	Swedish
Swiss-French Polyphone Database 1000 speakers	French
Swiss-French Polyphone Database 4000 speakers	French
Swiss-French SpeechDat(II) FDB-3000	French
Swiss-French SpeechDat(M)	French
Swiss-German SpeechDat(II) FDB-2000	German
Tagged text in French (MEMODATA) with rules of morphological disambiguation	French
Tagged text in French (MEMODATA) with typographic tags	French
Taiwan Mandarin Speecon database	Chinese
Taiwanese Speech Recognition Corpus (Desktop)	Chinese
TAXI - Multilingual telephone dialog database	German | English
TC-STAR 2005 Evaluation Package - ASR English	English
TC-STAR 2005 Evaluation Package - ASR Mandarin Chinese	Chinese
TC-STAR 2005 Evaluation Package - ASR Spanish	Spanish
TC-STAR 2005 Evaluation Package - SLT Chinese-to-English	English | Chinese
TC-STAR 2005 Evaluation Package - SLT English-to-Spanish	English | Spanish
TC-STAR 2005 Evaluation Package - SLT Spanish-to-English	English | Spanish
TC-STAR 2006 Evaluation Package - ASR English	English
TC-STAR 2006 Evaluation Package - ASR Mandarin Chinese	Chinese
TC-STAR 2006 Evaluation Package - ASR Spanish - CORTES	Spanish
TC-STAR 2006 Evaluation Package - ASR Spanish - EPPS	Spanish
TC-STAR 2006 Evaluation Package - SLT Chinese-to-English	English | Chinese
TC-STAR 2006 Evaluation Package - SLT English-to-Spanish	English | Spanish
TC-STAR 2006 Evaluation Package - SLT Spanish-to-English - CORTES	English | Spanish
TC-STAR 2006 Evaluation Package - SLT Spanish-to-English - EPPS	English | Spanish
TC-STAR 2006 Evaluation Package β€“ End-to-End	English | Spanish
TC-STAR 2007 Evaluation Package - ASR English	English
TC-STAR 2007 Evaluation Package - ASR Mandarin Chinese	Chinese
TC-STAR 2007 Evaluation Package - ASR Spanish - CORTES	Spanish
TC-STAR 2007 Evaluation Package - ASR Spanish - EPPS	Spanish
TC-STAR 2007 Evaluation Package - SLT Chinese-to-English	English | Chinese
TC-STAR 2007 Evaluation Package - SLT English-to-Spanish	English | Spanish
TC-STAR 2007 Evaluation Package - SLT Spanish-to-English - CORTES	English | Spanish
TC-STAR 2007 Evaluation Package - SLT Spanish-to-English - EPPS	English | Spanish
TC-STAR 2007 Evaluation Package - End-to-End	English | Spanish
TC-STAR Bilingual Expressive Speech Database	English | Spanish
TC-STAR Bilingual Voice-Conversion English Speech Database	English
TC-STAR Bilingual Voice-Conversion Spanish Speech Database	Spanish
TC-STAR English Test Corpora for ASR	English

TC-STAR English Training Corpora for ASR: Recordings of EPPS Speech	English
TC-STAR English Training Corpora for ASR: Transcriptions of EPPS Speech	English
TC-STAR English-Spanish Training Corpora for Machine Translation: Aligned Final Text Editions of EPPS	English | Spanish
TC-STAR Spanish Baseline Female Speech Database	Spanish
TC-STAR Spanish Baseline Male Speech Database	Spanish
TC-STAR Spanish Test Corpora for ASR	Spanish
TC-STAR Spanish Training Corpora for ASR: Recordings of EPPS Speech	Spanish
TC-STAR Transcriptions of Spanish Parliamentary Speech	Spanish
TED Translanguage English Database	English
TEDphone (Polyphone-like Translanguage English Database)	English
Terminology database of expressions	English | French
Terminology database of finance	English | French
Terminology database of natural sciences	English | French | Latin
Terminology database of telecommunication	English | French
Text corpus of "Le Monde"	French
Thai Speecon database	Thai
THAMUS Bilingual dictionaries - Aeronautics (1)	Italian | English
THAMUS Bilingual dictionaries - Aeronautics (2)	Italian | English
THAMUS Bilingual dictionaries - Computer Science (1)	Italian | German
THAMUS Bilingual dictionaries - Computer Science (2)	Italian | German
THAMUS Bilingual dictionaries - Computer Science (3)	Italian | German
THAMUS Bilingual dictionaries - Computer Science (4)	Italian | German
THAMUS Bilingual dictionaries - Computer science (5)	Italian | English
THAMUS Bilingual dictionaries - Computer science (6)	Italian | English
THAMUS Bilingual dictionaries - Computer science (7)	Italian | English
THAMUS Bilingual dictionaries - Computer science (8)	Italian | English
THAMUS Bilingual dictionaries - Economics (1)	Italian | English
THAMUS Bilingual dictionaries - Economics (2)	Italian | English
THAMUS Bilingual dictionaries - Economics (3)	Italian | English
THAMUS Bilingual dictionaries - Economics (4)	Italian | English
THAMUS Bilingual dictionaries - Engineering (1)	Italian | English
THAMUS Bilingual dictionaries - Engineering (2)	Italian | English
THAMUS Bilingual dictionaries - Engineering (3)	Italian | English
THAMUS Bilingual dictionaries - Engineering (4)	Italian | English
THAMUS Bilingual dictionaries - Law (1)	Italian | English
THAMUS Bilingual dictionaries - Law (2)	Italian | English
THAMUS Bilingual dictionaries - Law (3)	Italian | English
THAMUS Bilingual dictionaries - Law (4)	Italian | English
THAMUS Bilingual dictionaries - Medicine (1)	Italian | English
THAMUS Bilingual dictionaries - Medicine (2)	Italian | English
THAMUS Generic Italian Dictionary - canonical forms	Italian
THAMUS. Generic Italian Dictionary - canonical forms - technical domain	Italian
THAMUS. Generic Italian Dictionary - inflected forms	Italian
THAMUS. Generic Italian Dictionary - inflected forms - technical domain	Italian
The "SIVA" Speech Database for Speaker Verification and Identification	Italian
The CINTIL Corpus - International Corpus of Portuguese	Portuguese

The CLEF Test Suite for the CLEF 2000-2003 Campaigns - Evaluation Package	Multiple languages
The EMILLE Lancaster Corpus	Multiple languages
The EMILLE/CIIL Corpus	Multiple languages
The FAME! Speech Corpus	Western Frisian
The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication	English
The Lancaster Corpus of Mandarin Chinese (LCMC)	Chinese
The MWN.PT - MultiWordnet of Portuguese	Portuguese
The Oxford Spanish Dictionary	Spanish
Toponymic Geography	French
TRAD Arabic-English Mailing lists Parallel corpus - Development set	English | Arabic
TRAD Arabic-English Mailing lists Parallel corpus - Test set	English | Arabic
TRAD Arabic-English Newspaper Parallel corpus - Test set 1	English | Arabic
TRAD Arabic-English Parallel corpus of transcribed Broadcast News Speech	English | Arabic
TRAD Arabic-English Web domain (blogs) Parallel corpus	English | Arabic
TRAD Arabic-French Mailing lists Parallel corpus - Development set	French | Arabic
TRAD Arabic-French Mailing lists Parallel corpus - Test set	French | Arabic
TRAD Arabic-French Newspaper Parallel corpus - Test set 1	French | Arabic
TRAD Arabic-French Newspaper Parallel corpus - Test set 2	French | Arabic
TRAD Arabic-French Parallel corpus of transcribed Broadcast News Speech	French | Arabic
TRAD Arabic-French Web domain (blogs) Parallel corpus	French | Arabic
TRAD Chinese-English Email Parallel corpus - Development Set	English | Chinese
TRAD Chinese-English Email Parallel corpus - Test Set	English | Chinese
TRAD Chinese-English News Articles Parallel corpus	English | Chinese
TRAD Chinese-English Web domain (blogs) Parallel corpus	English | Chinese
TRAD Chinese-French Email Parallel corpus - Development Set	French | Chinese
TRAD Chinese-French Email Parallel corpus - Test Set	French | Chinese
TRAD Chinese-French News Articles Parallel corpus	French | Chinese
TRAD Chinese-French Parallel Text - Blog	French | Chinese
TRAD Chinese-French Web domain (blogs) Parallel corpus	French | Chinese
TRAD Pashto Broadcast News Speech Corpus	Pushto
TRAD Pashto Monolingual text Corpus	Pushto
TRAD Pashto-English News Articles Parallel corpus	English | Pushto
TRAD Pashto-English Parallel corpus of transcribed Broadcast News Speech - Test data	English | Pushto
TRAD Pashto-French News Articles Parallel corpus	Pushto | French
TRAD Pashto-French Parallel corpus of transcribed Broadcast News Speech - Test data	Pushto | French
TRAD Pashto-French Parallel corpus of transcribed Broadcast News Speech - Training data	Pushto | French
Training and test data for Arabizi detection and transliteration	English | Arabic
Translanguage English Database (TED) Transcripts database	English
TSNLP (Test Suites for NLP Testing)	German | French | English
TUNA Corpus	English
Turkish Continuous and Isolated Word Speech Database	Turkish
Turkish Speecon database	Turkish
Twin database - TWINDB1	French
UK English Speecon database	English
UPC-TALP database of isolated meeting-room acoustic events	No linguistic content

US Spanish Speecon database	Spanish
Venice Italian Treebank (VIT)	Italian
VERBA Polytechnic and Plurilingual Terminological Database - A-QA Mathematics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - A-QG Metrology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-AA General Chemistry	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-AB Analytical Chemistry	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-AC Inorganic Chemistry	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-AD Organic Chemistry	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-AE Physical Chemistry	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MA Acoustics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MB Electricity	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MC Electromechanics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MD Spectrography	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-ME Solid State Physics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MF General Physics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MG Atomic Physics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MH Particle Physics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MI Plasma Physics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MJ Nuclear Physics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MK General Mechanics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-ML Quantum Mechanics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MM Statistical Mechanics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MN Fluid Mechanics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MO Nucleonics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MP Optics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MQ Relativity	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - B-MR Thermodynamics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - C-AB Geography	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - C-AC Geology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - C-AG Petrology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - C-GB Climate Studies	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - C-GC Weather Studies	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - C-LA Hydrology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - C-LB Oceanography	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - C-RE Energy Resources	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-AE Climate Control	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-GA Control of Industrial Pollution	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-GB Air Pollution	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-GC Chemical Pollution	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-GD Marine Pollution	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-GE Soil Contamination	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-GH Structures	English | Spanish

VERBA Polytechnic and Plurilingual Terminological Database - D-GI Environmental Law	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-GK Noise Pollution	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-KA Water Cycle	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - D-KB Solid Waste Treatment	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-KC Laboratory Techniques	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-KD Sewage Plant equipment	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - D-KE Environmental Technology	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - E-AA Health Materials and Equipment	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - E-AB Hospital Services	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - E-AC Hospital Management	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - E-AD Pharmacology	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - E-AF General Medicine	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - F-AA Agrarian Economics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-AB Farming Activities and Techniques	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-AC Edafology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-AD Drainage and Irrigation	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-AE Fertilizers	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-AF Pest Protection	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-AJ Food Plants	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - F-AL Arboriculture and Viticulture	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-AM Trees and Bushes	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - F-AQ Agriculture-General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-AR Tobacco Industry	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - F-HA Cattle Breeding	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-HB Aviculture, Cuniculture, Apiculture	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-HD Animal Health and Nutrition	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - F-MA Meat Industry	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - G-AA Computing-General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-AB Peripherals	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-AE Applications and Services	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-AH Data Processing	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-AN Data Transmission	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-AS Software Quality and Engineering	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-AU General Terminology	English | Spanish

VERBA Polytechnic and Plurilingual Terminological Database - G-BC Essays	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GF Microelectronics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GH Cybernetics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GJ Cathode Rays	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GM Semi- and Super-Conductors	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GP Electronics-General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GR Ionics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GU Magnetics Recording and Playback	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GY Integrated Circuits	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-GZ Electronic Office	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-HL Components and Material	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NA Radioelectric Broadcasting	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NB Radar	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NC Space Communications	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NF Cables and Conductors	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NH Radiocommunications	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NM T.V.	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NQ Telecomms Lines and Devices	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NR Telephone and Telegraph	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-NZ Telecommunications-General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-OB Control	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-OF Signalling	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-OG Switching Devices	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-SA Electrical Systems	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - G-SB Instrumentation	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - H-AB Reinforced Concrete	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - H-GA Architecture	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - H-GE Construction-General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - H-GG Town Planning	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-AA Metal and Steel Foundries	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - I-AB Oil Industry	English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-AC Automobile Industry	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-AD Textile Industry	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - I-MA Aerospace Engineering	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-MB Engineering Design	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-MC Mechanical Engineering	English | Spanish

VERBA Polytechnic and Plurilingual Terminological Database - I-MG Control Systems	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-MJ Hydraulic Engineering	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-MM Air-Conditioning	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-MN Outfitting	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-MO Tools	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-MY Machine Tools	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-QN Industry, General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-TA Paints	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-TB Products and Ingredients	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - I-TC Manufacturing, General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - L-AA Transport, General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - L-AD Air Transport	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - L-AG Sea Shipping	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - L-AH Infrastructure	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - L-MA Transport Vehicles	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AA Law, General Topics	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - M-AB Criminal Law	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AC Civil Law	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AD Politics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AF Financial Law	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AI International Law	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AL Marine Law	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AM Company Law	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AP Court Procedure	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-AR Roman Law	Spanish | Latin
VERBA Polytechnic and Plurilingual Terminological Database - M-AT Labour Law	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-KD State Administration	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-RA Politics, General Topics	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - M-RB Diplomacy	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - M-RC Politics and International Co-operation	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - M-RD International Conferences	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - M-RE International Treaties	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - M-RF International Institutions	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - M-RG International Courts	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - M-RH Armed Conflicts	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AA Economic Growth	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AB Economic Cycles	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AC Economic Policy	English | Spanish

VERBA Polytechnic and Plurilingual Terminological Database - N-AD Macroeconomics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AE Microeconomics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AF History of Economics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AG Economic Structure	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AH Accounting	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AI State Exchequer	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AJ Natural Resources and Environment	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AK Statistics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AL European Union	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AM Regional and Urban	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AN Labour Economy	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AO Agricultural Economy	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AU Demography	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AW Economic Institutions	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-AX Economics of Real Estate	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-BB Social Welfare	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-BC Economics, General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-LA Trade, General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-LB Foreign Trade	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-LC Measures and Currencies	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-LF Marketing	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-LG Import-Export	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-LH Distribution	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - N-LP Business Correspondence	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - N-PA Banking	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-PB Stock Markets	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-PC International Finance	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-PD Money and Currencies	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-PE Insurance	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-PF Financial Services	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-TA Business Management	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-TB Human Resources	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-TC Quality Control	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - N-TD Manufacturing and Logistics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-TF Business Finance	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - N-TH Business Computing	English | Spanish

VERBA Polytechnic and Plurilingual Terminological Database - S-AA Anatomy	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AC Biology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AE Biochemistry	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AF General Botany	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AG Cytology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AH Ecology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AI Embryology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AK General Physiology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AL Genetics	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - S-AM Histology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AN Mycology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AO Microbiology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AQ Palaeontology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AR Plant Pathology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AS Taxonomy	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AT Virology	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AU Zoology, General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AV Zoology of Invertebrates	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-AW Zoology of Vertebrates	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - S-BJ Flora	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - S-BK Fauna	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - T-AB Press	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - T-AC Radio	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - T-AD TV	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - T-AE Cinema	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - T-AG Communications, General Topics	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - T-AH Photography	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - T-AI Printing Industry	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - T-MB Documentation	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-AA Trampoline Jumping	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AB Target Shooting	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-AC Skating	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AD Skiing	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AE Table Tennis	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-AF Lawn Tennis	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AG Volleyball	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AH Weight-Lifting	German | French | Spanish | English

VERBA Polytechnic and Plurilingual Terminological Database - V-AI Waterpolo	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AJ Wrestling	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-AS American Football	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AT Field and Track Athletics	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AU Tenpin Bowling	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-AV Boxing	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AW Baseball	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AX Basketball	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-AY Handball	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-AZ Cricket	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-BA Cycling	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BC Fencing	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-BD Swimming	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BE Aquatic Choreography	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BF Football	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BG Mountaineering	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BH Golf	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-BI Gymnastics	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BJ Hockey	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BK Ice Hockey	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-BL Judo	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BM Canoeing	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BN Modern Pentathlon	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BO Polo	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BP Rugby	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-BQ Riding	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BR Rowing	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - V-BS Sailing	German | English | French | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-BT Sports, General Topics	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - VERBA Polytechnic and Plurilingual Terminological Database	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - V-TA Leisure	English | Spanish

VERBA Polytechnic and Plurilingual Terminological Database - W-AA Weapons	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - W-WA Specialised terminology without field coding	English | Spanish
VERBA Polytechnic and Plurilingual Terminological Database - W-WB Specialised terminology without field coding	German | French | English
VERBA Polytechnic and Plurilingual Terminological Database - Z-ZA General Vocabulary	German | French | Spanish | English
VERBA Polytechnic and Plurilingual Terminological Database - Z-ZB General Vocabulary	French | English | Italian | German | Spanish | Portuguese
VERBMOBIL - VM CD 1.1 (new edition)	German
VERBMOBIL - VM CD 12.1 (new edition)	German
VERBMOBIL - VM CD 13.1 (new edition)	English
VERBMOBIL - VM CD 14.1 (new edition)	German
VERBMOBIL - VM CD 2.1 (new edition)	German
VERBMOBIL - VM CD 3.1 (new edition)	German
VERBMOBIL - VM CD 4.1 (new edition)	German
VERBMOBIL - VM CD 5.1 (new edition)	German
VERBMOBIL - VM CD 6.1 (new edition)	English
VERBMOBIL - VM CD 7.1 (new edition)	German
VERBMOBIL - VM CD 8.1 (new edition)	English
VERBMOBIL - VM CD S 1.0 (original edition)	German
VERBMOBIL II - VM Bonus CD - VMBONUS (BAS edition)	German
VERBMOBIL II - VM CD 15.1 - VM15.1 (new edition)	German | English
VERBMOBIL II - VM CD 16.1 - VM16.1 (new edition)	Japanese
VERBMOBIL II - VM CD 17.1 - VM17.1 (new edition)	Japanese
VERBMOBIL II - VM CD 18.1 - VM18.1 (new edition)	Japanese
VERBMOBIL II - VM CD 19.1 - VM19.1 (new edition)	Japanese
VERBMOBIL II - VM CD 22.1 - VM22.1 (BAS edition)	German
VERBMOBIL II - VM CD 23.1 - VM23.1 (BAS edition)	English
VERBMOBIL II - VM CD 24.1 - VM24.1 (BAS edition)	German
VERBMOBIL II - VM CD 25.1 - VM25.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 26.1 - VM26.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 27.1 - VM27.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 28.1 - VM28.1 (BAS edition)	English
VERBMOBIL II - VM CD 29.1 - VM29.1 (BAS edition)	German
VERBMOBIL II - VM CD 30.1 - VM30.1 (BAS edition)	English
VERBMOBIL II - VM CD 31.1 - VM31.1 (BAS edition)	English
VERBMOBIL II - VM CD 32.1 - VM32.1 (BAS edition)	English
VERBMOBIL II - VM CD 33.1 - VM33.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 34.1 - VM34.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 35.1 - VM35.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 38.1 - VM38.1 (BAS edition)	German
VERBMOBIL II - VM CD 39.1 - VM39.1 (BAS edition)	German
VERBMOBIL II - VM CD 42.1 - VM42.1 (BAS edition)	English
VERBMOBIL II - VM CD 43.1 - VM43.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 44.1 - VM44.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 45.1 - VM45.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 46.1 - VM46.1 (BAS edition)	Japanese | German

VERBMOBIL II - VM CD 47.1 - VM47.1 (BAS edition)	German | English
VERBMOBIL II - VM CD 48.1 - VM48.1 (BAS edition)	German
VERBMOBIL II - VM CD 49.1 - VM49.1 (BAS edition)	German
VERBMOBIL II - VM CD 50.1 - VM50.1 (BAS edition)	English
VERBMOBIL II - VM CD 51.1 - VM51.1 (BAS edition)	German | English
VERBMOBIL II - VM CD 52.1 - VM52.1 (BAS edition)	German | English
VERBMOBIL II - VM CD 53.1 - VM53.1 (BAS edition)	German
VERBMOBIL II - VM CD 55.1 - VM55.1 (BAS edition)	German | English
VERBMOBIL II - VM CD 56.1 - VM56.1 (BAS edition)	German | English
VERBMOBIL II - VM CD 57.1 - VM57.1 (BAS edition)	Japanese | German
VERBMOBIL II - VM CD 58.1 - VM58.1 (BAS edition)	Japanese | German
VERBMOBIL II - VM CD 59.1 - VM59.1 (BAS edition)	Japanese | German
VERBMOBIL II - VM CD 60.1 - VM60.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 61.1 - VM61.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 62.1 - VM62.1 (BAS edition)	Japanese
VERBMOBIL II - VM CD 63.0 - VM63.0 (original edition)	German
VERBMOBIL II - VM CD 64.0 - VM64.0 (original edition)	German
VERBMOBIL II - VM CD 65.0 - VM65.0 (original edition)	German
VERBMOBIL II - VM CD20.1 - VM20.1 (new edition)	German
VERBMOBIL II - VM CD21.1 - VM21.1 (new edition)	German
VERBMOBIL II - VM Lexicon database - VMLEX (BAS edition)	German
VERIF1DE	German
WEBCOMMAND	English | French
Welsh SpeechDat(II) FDB-2000	Welsh
Wolverhampton Business English Corpus	English
ZipTel	German

Table 6: LRs from ELRA catalogue

A.B. ELRC-SHARE (1030 LRs)

This annex provides an overview of the LRs that have been ingested into ELG from ELRC-SHARE.

Resource name	Language(s)
2017 Activity Report Hohe Tauern National Park (Processed)	German | English
ACM news items 2017 and 2018	English | Dutch
AECOSAN	English | Spanish
AECOSAN (Processed)	English | Spanish
Agencia Tributaria	English | Spanish
Agencia Tributaria (Processed)	English | Spanish
Anonymised ParaCrawl release 7 Basque-Spanish	Spanish | Basque
Anonymised ParaCrawl release 7 Bulgarian-English	English | Bulgarian
Anonymised ParaCrawl release 7 Catalan-Spanish	Spanish | Catalan
Anonymised ParaCrawl release 7 Croatian-English	English | Croatian
Anonymised ParaCrawl release 7 Czech-English	English | Czech
Anonymised ParaCrawl release 7 Danish-English	Danish | English
Anonymised ParaCrawl release 7 Dutch-English	English | Dutch
Anonymised ParaCrawl release 7 Estonian-English	English | Estonian
Anonymised ParaCrawl release 7 Finnish-English	English | Finnish

Anonymised ParaCrawl release 7 French-English	English | French
Anonymised ParaCrawl release 7 Galician-Spanish	Galician | Spanish
Anonymised ParaCrawl release 7 German-English	German | English
Anonymised ParaCrawl release 7 Greek-English	Modern Greek | English
Anonymised ParaCrawl release 7 Hungarian-English	English | Hungarian
Anonymised ParaCrawl release 7 Icelandic-English	English | Icelandic
Anonymised ParaCrawl release 7 Irish-English	English | Irish
Anonymised ParaCrawl release 7 Italian-English	Italian | English
Anonymised ParaCrawl release 7 Latvian-English	English | Latvian
Anonymised ParaCrawl release 7 Lithuanian-English	English | Lithuanian
Anonymised ParaCrawl release 7 Maltese-English	English | Maltese
Anonymised ParaCrawl release 7 Norwegian Bokmål-English	Norwegian Bokmål | English
Anonymised ParaCrawl release 7 Norwegian Nynorsk-English	Norwegian Nynorsk | English
Anonymised ParaCrawl release 7 Polish-English	English | Polish
Anonymised ParaCrawl release 7 Portuguese-English	English | Portuguese
Anonymised ParaCrawl release 7 Romanian-English	English | Romanian
Anonymised ParaCrawl release 7 Slovak-English	English | Slovak
Anonymised ParaCrawl release 7 Slovenian-English	English | Slovenian
Anonymised ParaCrawl release 7 Spanish-English	English | Spanish
Anonymised ParaCrawl release 7 Swedish-English	English | Swedish
ANR translation memory containing major publications, as well as several administrative documents and news (Processed)	English | French
Audioguide for the Military History Museum in Vienna (Processed)	Italian | German
Austrian Research and Technology Report 2015 (Processed)	German | English
Belgian government bilingual parallel corpus (Processed)	French | Dutch
Belgian parallel corpus about Belgium and the justice system (Processed)	French | Dutch
Bilingual Bulgarian-English corpus from the 2018 Proposal for a National Climate Change Adaptation Strategy and Action Plan from the website of the Bulgarian Ministry of Environment and Water (Processed)	English | Bulgarian
Bilingual Bulgarian-English corpus from the National Revenue Agency (BG) (Processed)	English | Bulgarian
Bilingual Bulgarian-English corpus in the field of Information society (Processed)	English | Bulgarian
Bilingual collection of documents about the Cyprus Problem (Processed)	Modern Greek | English
Bilingual collection of reports of the Greek Public Power Corporation (Processed)	Modern Greek | English
Bilingual corpus from the European Vaccination Portal (BG-EN)	English | Bulgarian
Bilingual corpus from the European Vaccination Portal (CS-EN)	English | Czech
Bilingual corpus from the European Vaccination Portal (DA-EN)	Danish | English
Bilingual corpus from the European Vaccination Portal (DE-EN)	German | English
Bilingual corpus from the European Vaccination Portal (EL-EN)	Modern Greek | English
Bilingual corpus from the European Vaccination Portal (ES-EN)	English | Spanish
Bilingual corpus from the European Vaccination Portal (ET-EN)	English | Estonian
Bilingual corpus from the European Vaccination Portal (FI-EN)	English | Finnish
Bilingual corpus from the European Vaccination Portal (FR-EN)	English | French
Bilingual corpus from the European Vaccination Portal (GA-EN)	English | Irish
Bilingual corpus from the European Vaccination Portal (HR-EN)	English | Croatian
Bilingual corpus from the European Vaccination Portal (HU-EN)	English | Hungarian
Bilingual corpus from the European Vaccination Portal (IT-EN)	Italian | English
Bilingual corpus from the European Vaccination Portal (LT-EN)	English | Lithuanian

Bilingual corpus from the European Vaccination Portal (LV-EN)	English | Latvian
Bilingual corpus from the European Vaccination Portal (MT-EN)	English | Maltese
Bilingual corpus from the European Vaccination Portal (NL-EN)	English | Dutch
Bilingual corpus from the European Vaccination Portal (PL-EN)	English | Polish
Bilingual corpus from the European Vaccination Portal (PT-EN)	English | Portuguese
Bilingual corpus from the European Vaccination Portal (RO-EN)	English | Romanian
Bilingual corpus from the European Vaccination Portal (SK-EN)	English | Slovak
Bilingual corpus from the European Vaccination Portal (SL-EN)	English | Slovenian
Bilingual corpus from the European Vaccination Portal (SV-EN)	English | Swedish
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-BG)	English | Bulgarian
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-CS)	English | Czech
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-DA)	Danish | English
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-DE)	German | English
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-EL)	Modern Greek | English
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-ES)	English | Spanish
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-ET)	English | Estonian
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-FI)	English | Finnish
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-FR)	English | French
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-GA)	English | Irish
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-HR)	English | Croatian
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-HU)	English | Hungarian
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-IT)	Italian | English
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-LT)	English | Lithuanian
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-LV)	English | Latvian
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-MT)	English | Maltese
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-NL)	English | Dutch
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-PL)	English | Polish
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-PT)	English | Portuguese
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-RO)	English | Romanian
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-SK)	English | Slovak
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-SL)	Slovenian | English
Bilingual corpus from the Publications Office of the EU on the medical domain (EN-SV)	English | Swedish
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-BG)	English | Bulgarian
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-CS)	English | Czech
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-DA)	Danish | English
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-DE)	German | English
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-EL)	Modern Greek | English
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-ES)	English | Spanish
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-ET)	English | Estonian

Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-FI)	English | Finnish
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-FR)	English | French
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-GA)	English | Irish
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-HR)	English | Croatian
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-HU)	English | Hungarian
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-IT)	Italian | English
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-LT)	English | Lithuanian
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-LV)	English | Latvian
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-MT)	English | Maltese
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-NL)	English | Dutch
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-PL)	English | Polish
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-PT)	English | Portuguese
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-RO)	English | Romanian
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-SK)	English | Slovak
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-SL)	Slovenian | English
Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-SV)	English | Swedish
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-BG).	English | Bulgarian
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-CS).	English | Czech
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-DA).	Danish | English
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-DE).	German | English
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-EL).	Modern Greek | English
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-ES).	English | Spanish
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-ET).	English | Estonian
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-FI).	English | Finnish
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-FR).	English | French
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-HR).	English | Croatian
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-HU).	English | Hungarian
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-IS).	English | Icelandic
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-IT).	Italian | English
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-LT).	English | Lithuanian
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-LV).	English | Latvian

Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-MT).	English | Maltese
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-NL).	English | Dutch
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-NO).	English | Norwegian
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-PL).	English | Polish
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-PT).	English | Portuguese
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-RO).	English | Romanian
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-SK).	English | Slovak
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-SL).	Slovenian | English
Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-SV).	English | Swedish
Bilingual Croatian-English Parallel Corpus	English | Croatian
Bilingual Croatian-English Parallel Corpus (Processed)	English | Croatian
Bilingual Danish-English parallel corpus from the State Audit Office (Rigsrevisionen) website	Danish | English
Bilingual documents Bulgarian-English in the field of ICT and Transport (Processed)	English | Bulgarian
Bilingual documents Bulgarian-English in the field of open data, broadband and information society (Processed)	English | Bulgarian
Bilingual documents Bulgarian-English in the field of transport (Processed)	English | Bulgarian
Bilingual English-Danish parallel corpus from Aarhus 2017 - European Capital of Culture website	Danish | English
Bilingual English-Danish parallel corpus from Danish FSA website	Danish | English
Bilingual English-Danish parallel corpus from Danish Maritime Authority website	Danish | English
Bilingual English-Danish parallel corpus from Danish Ministry of Finance website	Danish | English
Bilingual English-Danish parallel corpus from Danish Ministry of Foreign Affairs website	Danish | English
Bilingual English-Danish parallel corpus from Danish Ministry of Higher Education and Science website	Danish | English
Bilingual English-Danish parallel corpus from Danish Ministry of Transport, Building and Housing website	Danish | English
Bilingual English-Danish parallel corpus from Danish Working Environment Authority website	Danish | English
Bilingual English-Danish parallel corpus from Danmarks Statistik website	Danish | English
Bilingual English-Danish parallel corpus from Denmark National Space Institute website	Danish | English
Bilingual English-Danish parallel corpus from Denmark Prosecution Service website	Danish | English
Bilingual English-Danish parallel corpus from Holstebro Kunstmuseum website	Danish | English
Bilingual English-Danish parallel corpus from National Gallery of Denmark website	Danish | English
Bilingual English-Danish parallel corpus from National Museum of Denmark website	Danish | English
Bilingual English-Danish parallel corpus from Odense Municipality website	Danish | English
Bilingual English-Danish parallel corpus from Royal Danish Library website	Danish | English
Bilingual English-Danish parallel corpus from The Agency for Culture and Palaces website	Danish | English
Bilingual English-Danish parallel corpus from The Danish Environmental Protection Agency website	Danish | English
Bilingual English-Danish parallel corpus from The Danish Gambling Authority website	Danish | English
Bilingual English-Danish parallel corpus from The Danish Medicines Agency website	Danish | English
Bilingual English-Danish parallel corpus from The Danish Nature Agency website	Danish | English

Bilingual English-Danish parallel corpus from The Geological Survey of Denmark and Greenland (GEUS) website	Danish | English
Bilingual English-Danish parallel corpus from the official Nordic cooperation website	Danish | English
Bilingual English-Danish parallel corpus from The Viking Ship Museum website	Danish | English
Bilingual English-Danish parallel corpus from Visit Vejle website	Danish | English
Bilingual English-Danish parallel corpus from VisitDenmark - The official tourism site of Denmark website	Danish | English
Bilingual English-Finnish parallel corpus from the official Nordic cooperation website	English | Finnish
Bilingual English-Icelandic parallel corpus from Harpa Reykjavik Concert Hall and Conference Centre website	English | Icelandic
Bilingual English-Icelandic parallel corpus from Icelandic Financial Supervisory Authority	English | Icelandic
Bilingual English-Icelandic parallel corpus from Icelandic Post and Telecom Administration website	English | Icelandic
Bilingual English-Icelandic parallel corpus from Nordisk eTax website	English | Icelandic
Bilingual English-Icelandic parallel corpus from the official Nordic cooperation website	English | Icelandic
Bilingual English-Lithuanian parallel corpus from Seimas of the Republic of Lithuania website	English | Lithuanian
Bilingual English-Lithuanian parallel corpus from the Bank of Lithuania website	English | Lithuanian
Bilingual English-Lithuanian parallel corpus from the Ministry of National Defence Republic of Lithuania website	English | Lithuanian
Bilingual English-Norwegian (Nynorsk) parallel corpus from the Courts of Norway website	English | Norwegian Nynorsk
Bilingual English-Norwegian (Nynorsk) parallel corpus from The Norwegian Directorate of Immigration's website	English | Norwegian Nynorsk
Bilingual English-Norwegian (Nynorsk) parallel corpus from the Norwegian Industrial Property Office website	English | Norwegian Nynorsk
Bilingual English-Norwegian parallel corpus from Altinn website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Avinor company website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from BarentsWatch website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Geological survey of Norway website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Institute of Marine Research website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from KORO / Public Art Norway website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Nofima institute website	Norwegian Bokmål | English
Bilingual English-Norwegian parallel corpus from Nordisk eTax website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Norges Bank Investment Management website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Norway's central bank website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Norwegian Academy of Music website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Norwegian Institute of Public Health website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Norwegian Maritime Authority website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Norwegian Polar Institute website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Oslo and Akershus University College of Applied Sciences website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Petoro AS website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Petroleum Safety Authority Norway website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from Statistics Norway website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the Accident Investigation Board Norway website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the Courts of Norway website	English | Norwegian Bokmål

Bilingual English-Norwegian parallel corpus from The Financial Supervisory Authority of Norway website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the Immigration Appeals Board website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the National Contact Point For Responsible Business Norway website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from The Norway's Governments website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from The Norwegian Directorate of Immigration's website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the Norwegian Industrial Property Office website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from The Norwegian Petroleum Directorate website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the Office of the Auditor General (Riksrevisjonen) website	Norwegian Bokmål | English
Bilingual English-Norwegian parallel corpus from the official Nordic cooperation website	Norwegian Bokmål | English
Bilingual English-Norwegian parallel corpus from the Statoil energy company website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the University College of South-Eastern Norway website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the University of Agder website	English | Norwegian Bokmål
Bilingual English-Norwegian parallel corpus from the University of Bergen website	English | Norwegian Bokmål
Bilingual English-Swedish parallel corpus from the official Nordic cooperation website	English | Swedish
Bilingual en-sk parallel corpus of annual reports from the Statistical Office of the Slovak Republic (Processed)	English | Slovak
Bilingual hr-en parallel corpus from Croatian Mine Action website	English | Croatian
Bilingual hr-en parallel corpus from Croatian Mine Action website (Processed)	English | Croatian
Bilingual hr-en parallel corpus from Croatian National Bank website	English | Croatian
Bilingual hr-en parallel corpus from Croatian National Bank website (Processed)	English | Croatian
Bilingual hr-en parallel corpus from the Journal of the Croatian Association of Civil Engineers website	English | Croatian
Bilingual hr-en parallel corpus from the Journal of the Croatian Association of Civil Engineers website (Processed)	English | Croatian
Bilingual hr-en parallel corpus from the National and University Library in Zagreb website	English | Croatian
Bilingual hr-en parallel corpus from the National and University Library in Zagreb website (Processed)	English | Croatian
Bilingual Icelandic-English parallel corpus from Statistics Iceland website	English | Icelandic
Bilingual is-en parallel corpus from Icelandic Medicines Agency website	English | Icelandic
Bilingual is-en parallel corpus from National Gallery of Iceland website	English | Icelandic
Bilingual is-en parallel corpus from The Icelandic Directorate of Immigration website	English | Icelandic
Bilingual is-en parallel corpus from THE LITERATURE WEB website	English | Icelandic
Bilingual nb-en parallel corpus from the Norwayβ€™s petroleum website	English | Norwegian Bokmål
Bilingual nb-en parallel corpus from The Norwegian Directorate for Education and Training website	English | Norwegian Bokmål
Bilingual Norwegian-English parallel corpus from Hafslund AS website	English | Norwegian Bokmål
Bilingual resource with Bulgarian strategic documents in the field of innovations and digital growth (Bulgarian - English) (Processed)	English | Bulgarian
Bilingual resource with Bulgarian strategic documents in the field of telecommunications and broadband (Bulgarian - English) (Processed)	English | Bulgarian
BMI Brochure Civil Protection	German | English
BMI Brochure Civil Protection (Processed)	German | English
BMI Brochures 2011-2015	German | English
BMI Brochures 2011-2015 (Processed)	German | English
BMI Brochures and Website 2016	German | English

BMI Brochures and Website 2016 (Processed)	German | English
BMVI Publications	German | English
BMVI Publications (Processed)	German | English
BMVI Website	German | English
BMVI Website (Processed)	German | English
Bulgarian-English corpus of legislation from the Republic of Bulgaria Ministry of Energy website (Processed)	English | Bulgarian
Bulgarian-English parallel corpora from State Administration web sites	English | Bulgarian
Bulgarian-English Parallel corpus from Tatoeba project	English | Bulgarian
Bulgarian-English glossary of diseases (Processed)	English | Bulgarian
Central Statistical Office Dataset (Processed)	English | Polish
Citizens Information Bilingual Web-Corpus (Processed)	English | Irish
Civil Aviation Regulations (Processed)	English | Polish
CNIO	English | Spanish
CNIO (Processed)	English | Spanish
Comhdháil Cumann Idirnáisiúnta na gCoimisineirí Teanga 2019	English | Irish
Compendium The Social Insurance Institution (Processed)	English | Polish
Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment - United Nations (French-English-Greek) (Processed)	Modern Greek | English | French
Convention on the transfer of sentenced persons (English - Greek) (Processed)	Modern Greek | English
Coronavirus and the Law in Poland (Processed)	English | Polish
Corpora of legal text (Processed)	Italian | English
Corpus of Icelandic texts from the Central Bank of Iceland (Processed)	Icelandic
Corpus of State-related content from the Latvian Web (Processed)	English | Latvian
Corpus on Finance and Economics from Bank of Latvia	English | Latvian
Corpus on Finance and Economics from Bank of Latvia (Processed)	English | Latvian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-AF)	English | Afrikaans
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-AR)	English | Arabic
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-AZ)	English | Azerbaijani
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-BE)	English | Belarusian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-BG)	English | Bulgarian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-BN)	English | Bengali
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-BS)	English | Bosnian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-CA)	English | Catalan
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-CS)	English | Czech
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-DA)	Danish | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-DE)	German | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-EL)	Modern Greek | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-EO)	English | Esperanto
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-ES)	English | Spanish
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-ET)	English | Estonian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-EU)	English | Basque
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-FA)	Persian | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-FI)	English | Finnish
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-FR)	English | French
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-GL)	Galician | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-HE)	Hebrew | English

COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-HI)	English | Hindi
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-HR)	English | Croatian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-HU)	English | Hungarian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-ID)	English | Indonesian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-IT)	Italian | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-KO)	Korean | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-LT)	English | Lithuanian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-LV)	English | Latvian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-MK)	English | Macedonian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-ML)	English | Malayalam
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-MS)	English | Malay (macrolanguage)
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-NL)	English | Dutch
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-NO)	English | Norwegian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-PL)	English | Polish
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-PT)	English | Portuguese
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-RO)	English | Romanian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-RU)	English | Russian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-SH)	English | Serbo-Croatian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-SK)	English | Slovak
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-SL)	Slovenian | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-SQ)	English | Albanian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-SR)	Serbian | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-SV)	English | Swedish
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-SW)	English | Swahili (macrolanguage)
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-TA)	English | Tamil
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-TE)	English | Telugu
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-TH)	Thai | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-TL)	English | Tagalog
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-TR)	English | Turkish
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-UK)	English | Ukrainian
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-VI)	Vietnamese | English
COVID-19 - HEALTH Wikipedia dataset. Bilingual (EN-ZH)	English | Chinese
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-BG)	English | Bulgarian
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-CS)	English | Czech
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-DA)	Danish | English
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-DE)	German | English
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-EL)	Modern Greek | English
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-ES)	English | Spanish
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-ET)	English | Estonian
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-FI)	English | Finnish
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-FR)	English | French
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-GA)	English | Irish
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-HR)	English | Croatian
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-HU)	English | Hungarian
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-IS)	English | Icelandic
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-IT)	Italian | English

COVID-19 ANTIBIOTIC dataset. Bilingual (EN-LT)	English | Lithuanian
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-LV)	English | Latvian
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-MT)	English | Maltese
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-NB)	Norwegian Bokmal | English
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-NL)	English | Dutch
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-PL)	English | Polish
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-PT)	English | Portuguese
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-RO)	English | Romanian
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-SK)	English | Slovak
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-SL)	Slovenian | English
COVID-19 ANTIBIOTIC dataset. Bilingual (EN-SV)	English | Swedish
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-BG)	English | Bulgarian
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-CS)	English | Czech
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-DA)	Danish | English
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-DE)	German | English
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-EL)	Modern Greek | English
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-ES)	English | Spanish
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-ET)	English | Estonian
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-FI)	English | Finnish
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-FR)	English | French
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-GA)	English | Irish
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-HR)	English | Croatian
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-HU)	English | Hungarian
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-IT)	Italian | English
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-LT)	English | Lithuanian
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-LV)	English | Latvian
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-MT)	English | Maltese
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-NL)	English | Dutch
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-PL)	English | Polish
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-PT)	English | Portuguese
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-RO)	English | Romanian
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-SK)	English | Slovak
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-SL)	Slovenian | English
COVID-19 EC-EUROPA v1 dataset. Bilingual (EN-SV)	English | Swedish
COVID-19 EN-ET parallel corpus from www.kriis.ee	English | Estonian
COVID-19 EN-LV parallel corpus from covid19.gov.lv	English | Latvian
COVID-19 EU presscorner v1 dataset. Bilingual (EN-BG)	English | Bulgarian
COVID-19 EU presscorner v1 dataset. Bilingual (EN-CS)	English | Czech
COVID-19 EU presscorner v1 dataset. Bilingual (EN-DA)	Danish | English
COVID-19 EU presscorner v1 dataset. Bilingual (EN-DE)	German | English
COVID-19 EU presscorner v1 dataset. Bilingual (EN-EL)	Modern Greek | English
COVID-19 EU presscorner v1 dataset. Bilingual (EN-ES)	English | Spanish
COVID-19 EU presscorner v1 dataset. Bilingual (EN-ET)	English | Estonian
COVID-19 EU presscorner v1 dataset. Bilingual (EN-FI)	English | Finnish
COVID-19 EU presscorner v1 dataset. Bilingual (EN-FR)	English | French
COVID-19 EU presscorner v1 dataset. Bilingual (EN-GA)	English | Irish

COVID-19 EU presscorner v1 dataset. Bilingual (EN-HR)	English | Croatian
COVID-19 EU presscorner v1 dataset. Bilingual (EN-HU)	English | Hungarian
COVID-19 EU presscorner v1 dataset. Bilingual (EN-IT)	Italian | English
COVID-19 EU presscorner v1 dataset. Bilingual (EN-LT)	English | Lithuanian
COVID-19 EU presscorner v1 dataset. Bilingual (EN-LV)	English | Latvian
COVID-19 EU presscorner v1 dataset. Bilingual (EN-MT)	English | Maltese
COVID-19 EU presscorner v1 dataset. Bilingual (EN-NL)	English | Dutch
COVID-19 EU presscorner v1 dataset. Bilingual (EN-PL)	English | Polish
COVID-19 EU presscorner v1 dataset. Bilingual (EN-PT)	English | Portuguese
COVID-19 EU presscorner v1 dataset. Bilingual (EN-RO)	English | Romanian
COVID-19 EU presscorner v1 dataset. Bilingual (EN-SK)	English | Slovak
COVID-19 EU presscorner v1 dataset. Bilingual (EN-SL)	Slovenian | English
COVID-19 EU presscorner v1 dataset. Bilingual (EN-SV)	English | Swedish
COVID-19 EU presscorner v2 dataset. Bilingual (EN-BG)	English | Bulgarian
COVID-19 EU presscorner v2 dataset. Bilingual (EN-CS)	English | Czech
COVID-19 EU presscorner v2 dataset. Bilingual (EN-DA)	Danish | English
COVID-19 EU presscorner v2 dataset. Bilingual (EN-DE)	German | English
COVID-19 EU presscorner v2 dataset. Bilingual (EN-EL)	Modern Greek | English
COVID-19 EU presscorner v2 dataset. Bilingual (EN-ES)	English | Spanish
COVID-19 EU presscorner v2 dataset. Bilingual (EN-ET)	English | Estonian
COVID-19 EU presscorner v2 dataset. Bilingual (EN-FI)	English | Finnish
COVID-19 EU presscorner v2 dataset. Bilingual (EN-FR)	English | French
COVID-19 EU presscorner v2 dataset. Bilingual (EN-GA)	English | Irish
COVID-19 EU presscorner v2 dataset. Bilingual (EN-HR)	English | Croatian
COVID-19 EU presscorner v2 dataset. Bilingual (EN-HU)	English | Hungarian
COVID-19 EU presscorner v2 dataset. Bilingual (EN-IT)	Italian | English
COVID-19 EU presscorner v2 dataset. Bilingual (EN-LT)	English | Lithuanian
COVID-19 EU presscorner v2 dataset. Bilingual (EN-LV)	English | Latvian
COVID-19 EU presscorner v2 dataset. Bilingual (EN-MT)	English | Maltese
COVID-19 EU presscorner v2 dataset. Bilingual (EN-NL)	English | Dutch
COVID-19 EU presscorner v2 dataset. Bilingual (EN-PL)	English | Polish
COVID-19 EU presscorner v2 dataset. Bilingual (EN-PT)	English | Portuguese
COVID-19 EU presscorner v2 dataset. Bilingual (EN-RO)	English | Romanian
COVID-19 EU presscorner v2 dataset. Bilingual (EN-SK)	English | Slovak
COVID-19 EU presscorner v2 dataset. Bilingual (EN-SL)	Slovenian | English
COVID-19 EU presscorner v2 dataset. Bilingual (EN-SV)	English | Swedish
COVID-19 EUR-LEX dataset. Bilingual (EN-BG)	English | Bulgarian
COVID-19 EUR-LEX dataset. Bilingual (EN-CS)	English | Czech
COVID-19 EUR-LEX dataset. Bilingual (EN-DA)	Danish | English
COVID-19 EUR-LEX dataset. Bilingual (EN-DE)	German | English
COVID-19 EUR-LEX dataset. Bilingual (EN-EL)	Modern Greek | English
COVID-19 EUR-LEX dataset. Bilingual (EN-ES)	English | Spanish
COVID-19 EUR-LEX dataset. Bilingual (EN-ET)	English | Estonian
COVID-19 EUR-LEX dataset. Bilingual (EN-FI)	English | Finnish
COVID-19 EUR-LEX dataset. Bilingual (EN-FR)	English | French
COVID-19 EUR-LEX dataset. Bilingual (EN-GA)	English | Irish

COVID-19 EUR-LEX dataset. Bilingual (EN-HR)	English | Croatian
COVID-19 EUR-LEX dataset. Bilingual (EN-HU)	English | Hungarian
COVID-19 EUR-LEX dataset. Bilingual (EN-IT)	Italian | English
COVID-19 EUR-LEX dataset. Bilingual (EN-LT)	English | Lithuanian
COVID-19 EUR-LEX dataset. Bilingual (EN-LV)	English | Latvian
COVID-19 EUR-LEX dataset. Bilingual (EN-MT)	English | Maltese
COVID-19 EUR-LEX dataset. Bilingual (EN-NL)	English | Dutch
COVID-19 EUR-LEX dataset. Bilingual (EN-PL)	English | Polish
COVID-19 EUR-LEX dataset. Bilingual (EN-PT)	English | Portuguese
COVID-19 EUR-LEX dataset. Bilingual (EN-RO)	English | Romanian
COVID-19 EUR-LEX dataset. Bilingual (EN-SK)	English | Slovak
COVID-19 EUR-LEX dataset. Bilingual (EN-SL)	Slovenian | English
COVID-19 EUR-LEX dataset. Bilingual (EN-SV)	English | Swedish
COVID-19 EUROPARL dataset v1. Bilingual (EN-BG)	English | Bulgarian
COVID-19 EUROPARL dataset v1. Bilingual (EN-CS)	English | Czech
COVID-19 EUROPARL dataset v1. Bilingual (EN-DA)	Danish | English
COVID-19 EUROPARL dataset v1. Bilingual (EN-DE)	German | English
COVID-19 EUROPARL dataset v1. Bilingual (EN-EL)	Modern Greek | English
COVID-19 EUROPARL dataset v1. Bilingual (EN-ES)	English | Spanish
COVID-19 EUROPARL dataset v1. Bilingual (EN-ET)	English | Estonian
COVID-19 EUROPARL dataset v1. Bilingual (EN-FI)	English | Finnish
COVID-19 EUROPARL dataset v1. Bilingual (EN-FR)	English | French
COVID-19 EUROPARL dataset v1. Bilingual (EN-GA)	English | Irish
COVID-19 EUROPARL dataset v1. Bilingual (EN-HR)	English | Croatian
COVID-19 EUROPARL dataset v1. Bilingual (EN-HU)	English | Hungarian
COVID-19 EUROPARL dataset v1. Bilingual (EN-IT)	Italian | English
COVID-19 EUROPARL dataset v1. Bilingual (EN-LT)	English | Lithuanian
COVID-19 EUROPARL dataset v1. Bilingual (EN-LV)	English | Latvian
COVID-19 EUROPARL dataset v1. Bilingual (EN-MT)	English | Maltese
COVID-19 EUROPARL dataset v1. Bilingual (EN-NL)	English | Dutch
COVID-19 EUROPARL dataset v1. Bilingual (EN-PL)	English | Polish
COVID-19 EUROPARL dataset v1. Bilingual (EN-PT)	English | Portuguese
COVID-19 EUROPARL dataset v1. Bilingual (EN-RO)	English | Romanian
COVID-19 EUROPARL dataset v1. Bilingual (EN-SK)	English | Slovak
COVID-19 EUROPARL dataset v1. Bilingual (EN-SL)	Slovenian | English
COVID-19 EUROPARL dataset v1. Bilingual (EN-SV)	English | Swedish
COVID-19 EUROPARL v2 dataset. Bilingual (EN-BG)	English | Bulgarian
COVID-19 EUROPARL v2 dataset. Bilingual (EN-CS)	English | Czech
COVID-19 EUROPARL v2 dataset. Bilingual (EN-DA)	Danish | English
COVID-19 EUROPARL v2 dataset. Bilingual (EN-DE)	German | English
COVID-19 EUROPARL v2 dataset. Bilingual (EN-EL)	Modern Greek | English
COVID-19 EUROPARL v2 dataset. Bilingual (EN-ES)	English | Spanish
COVID-19 EUROPARL v2 dataset. Bilingual (EN-ET)	English | Estonian
COVID-19 EUROPARL v2 dataset. Bilingual (EN-FI)	English | Finnish
COVID-19 EUROPARL v2 dataset. Bilingual (EN-FR)	English | French
COVID-19 EUROPARL v2 dataset. Bilingual (EN-GA)	English | Irish

COVID-19 EUROPARL v2 dataset. Bilingual (EN-HR)	English | Croatian
COVID-19 EUROPARL v2 dataset. Bilingual (EN-HU)	English | Hungarian
COVID-19 EUROPARL v2 dataset. Bilingual (EN-IT)	Italian | English
COVID-19 EUROPARL v2 dataset. Bilingual (EN-LT)	English | Lithuanian
COVID-19 EUROPARL v2 dataset. Bilingual (EN-LV)	English | Latvian
COVID-19 EUROPARL v2 dataset. Bilingual (EN-MT)	English | Maltese
COVID-19 EUROPARL v2 dataset. Bilingual (EN-NL)	English | Dutch
COVID-19 EUROPARL v2 dataset. Bilingual (EN-PL)	English | Polish
COVID-19 EUROPARL v2 dataset. Bilingual (EN-PT)	English | Portuguese
COVID-19 EUROPARL v2 dataset. Bilingual (EN-RO)	English | Romanian
COVID-19 EUROPARL v2 dataset. Bilingual (EN-SK)	English | Slovak
COVID-19 EUROPARL v2 dataset. Bilingual (EN-SL)	Slovenian | English
COVID-19 EUROPARL v2 dataset. Bilingual (EN-SV)	English | Swedish
COVID-19 Parallel Global Voices dataset. Bilingual (EN-AR)	English | Arabic
COVID-19 Parallel Global Voices dataset. Bilingual (EN-BN)	English | Bengali
COVID-19 Parallel Global Voices dataset. Bilingual (EN-CS)	English | Czech
COVID-19 Parallel Global Voices dataset. Bilingual (EN-DE)	German | English
COVID-19 Parallel Global Voices dataset. Bilingual (EN-EL)	Modern Greek | English
COVID-19 Parallel Global Voices dataset. Bilingual (EN-ES)	English | Spanish
COVID-19 Parallel Global Voices dataset. Bilingual (EN-FR)	English | French
COVID-19 Parallel Global Voices dataset. Bilingual (EN-IT)	Italian | English
COVID-19 Parallel Global Voices dataset. Bilingual (EN-MG)	English | Malagasy
COVID-19 Parallel Global Voices dataset. Bilingual (EN-NL)	English | Dutch
COVID-19 Parallel Global Voices dataset. Bilingual (EN-PL)	English | Polish
COVID-19 Parallel Global Voices dataset. Bilingual (EN-PT)	English | Portuguese
COVID-19 Parallel Global Voices dataset. Bilingual (EN-RO)	English | Romanian
COVID-19 Parallel Global Voices dataset. Bilingual (EN-RU)	English | Russian
COVID-19 Parallel Global Voices dataset. Bilingual (EN-SR)	Serbian | English
Criminal Intelligence Service Austria (Processed)	German | English
Croatian-English corpus with Acts on Biological and Landscape Diversity and Environmental Protection (Processed)	English | Croatian
Croatian-English corpus with statistical reports and studies from the Croatian Bureau of Statistics website (Processed)	English | Croatian
Croatian-English corpus with studies on the challenges to the Croatian Accession to the European Union from the Croatian Institute of Public Finance website (Processed)	English | Croatian
Croatian-English corpus with the Rural Development Programme for the Period 2014-2020 from the Croatian Rural Development Programme website (Processed)	English | Croatian
Croatian-English glossary of statistical terms (Processed)	English | Croatian
Croatian-English parallel corpus from the website of the Croatian Journal of Fisheries (Processed)	English | Croatian
Croatian-English parallel corpus from the website of the Embassy of Finland, Zagreb (Processed)	English | Croatian
Croatian-English parallel corpus from the website of the Government Office for Cooperation with NGOs (Processed)	English | Croatian
Croatian-English parallel corpus from the website of the Ministry of Foreign and European Affairs, Republic of Croatia (Processed)	English | Croatian
Croatian-English translation memory from the Ministry of Agriculture (Part 1) (Processed)	English | Croatian
Croatian-English translation memory from the Ministry of Agriculture (Part 2) (Processed)	English | Croatian
Croatian-English translation memory from the Ministry of Regional Development and EU Funds (Part 1) (Processed)	English | Croatian

Croatian-English translation memory from the Ministry of Regional Development and EU Funds (Part 2) (Processed)	English | Croatian
Czech Republic Supreme Audit Office: 2003-2017 press releases (Processed)	English | Czech
Czech Republic Supreme Audit Office: 2008-2017 reports (Processed)	English | Czech
Czech Republic Supreme Audit Office: 2018 press releases (Processed)	English | Czech
Czech Republic Supreme Audit Office: 2018 reports (Processed)	English | Czech
Czech-English Parallel corpus from Tatoeba project	English | Czech
DA-EN Danish Ministry of Higher Education and Science (Processed)	Danish | English
DA-EN Danish Ministry of Higher Education and Science 2 (Processed)	Danish | English
DA-EN Danish Ministry of Higher Education and Science 3 (Processed)	Danish | English
DA-EN Danish Ministry of Higher Education and Science 4 (Processed)	Danish | English
Descripciones de vulnerabilidades de la BBDD NVD	English | Spanish
DireГ§Г£o-Geral do Consumidor	English | Portuguese
Documents concerning Federal Constitutional Law in Austria (Processed)	German | English
Documents from the Ministry of Agriculture, Forestry and Food of the Republic of Slovenia (EN-SL) (Processed)	English | Slovenian
Dutch Government Website	English | Dutch
EIR Romanian-English Newsletter (2009-March 2011) (Processed)	English | Romanian
EIR Romanian-English SPOS (2011-2017) (Processed)	English | Romanian
EIR Romanian-English TM (ECHR-33234/12) (Processed)	English | Romanian
EIR terminology (banking) (RO-EN) (Processed)	English | Romanian
EIR terminology (legal) (RO-EN) (Processed)	English | Romanian
EJTN Handbook (Processed)	English | Bulgarian
Electronic Exchange of Social Security Information documents in Czech-English (Processed)	English | Czech
Employment in Poland 2009 report in EN-PL (Processed)	English | Polish
Energy Report of the City of Vienna (Processed)	German | English
English - Croatian parallel corpus from texts of the Swedish Crime Victim Compensation and Support Authority (Brottsoffermyndigheten) web site (Processed)	English | Croatian
English - Swedish parallel corpus from texts of the Swedish Crime Victim Compensation and Support Authority (Brottsoffermyndigheten) web site (Processed)	English | Swedish
English to Norwegian Bokmål translation memories from Amesto Translations	Norwegian Bokmål | English
ENGLISH/POLISH PHRASE BOOK FOR ADMINISTRATIVE STAFF of LOCAL GOVERNMENT UNITS (Processed)	English | Polish
English-Bulgarian Computer Terms (Processed)	English | Bulgarian
English-Bulgarian Economy Terms (Processed)	English | Bulgarian
English-Bulgarian Legal Terms (Processed)	English | Bulgarian
English-Croatian Parallel corpus from Tatoeba project	English | Croatian
English-Croatian translation memory from the Ministry of Agriculture (Processed)	English | Croatian
English-Croatian translation memory from the Ministry of Regional Development and EU Funds (Processed)	English | Croatian
English-Danish EASTIN-CL Multilingual Ontology of Assistive Technology (Processed)	Danish | English
English-Danish Parallel corpus from Tatoeba project	Danish | English
English-Danish Parallel corpus from Tatoeba project (Processed)	Danish | English
English-Dutch Parallel corpus from Tatoeba project	English | Dutch
English-Estonian corpus from Finnish Information Bank	English | Estonian
English-Estonian corpus from Finnish Information Bank (Processed)	English | Estonian
English-Estonian EASTIN-CL Multilingual Ontology of Assistive Technology (Processed)	English | Estonian
English-Estonian Parallel corpus compiled from translated annual reports from Estonian Academy of Sciences	English | Estonian

English-Estonian parallel corpus from press releases of Ministry of Foreign Affairs of Estonia	English | Estonian
English-Estonian Parallel corpus from Tatoeba project	English | Estonian
English-Estonian parallel corpus from the web site of Arnold RĽOĽOtel, President of the Republic of Estonia, 2001-2006	English | Estonian
English-Estonian parallel corpus from the web site of Kersti Kaljulaid, President of the Republic of Estonia, 2016-	English | Estonian
English-Estonian parallel corpus from the web site of Lennart Meri, President of the Republic of Estonia, 1992-2001	English | Estonian
English-Estonian parallel corpus from the web site of Toomas Hendrik Ilves, President of the Republic of Estonia, 2006-2016	English | Estonian
English-Estonian parallel corpus from the www.visitestonia.com web site	English | Estonian
English-Finnish corpus from Finnish Information Bank	English | Finnish
English-Finnish corpus from Finnish Information Bank (Processed)	English | Finnish
English-Finnish parallel corpus from National Audit Office of Finland	English | Finnish
English-Finnish Parallel corpus from Tatoeba project	English | Finnish
English-Finnish parallel corpus from the contents of City of Turku web site	English | Finnish
English-Finnish parallel corpus from the Finnish Government web site	English | Finnish
English-Finnish parallel corpus from the Prime Minister's Office of Finland web site	English | Finnish
English-Finnish parallel corpus from the web site of Finnish Tax Administration	English | Finnish
English-Finnish parallel corpus from the www.visitestonia.com web site	English | Finnish
English-French parallel corpus from CORDIS Project News	English | French
English-French parallel corpus from CORDIS Project Results in Brief	English | French
English-French Parallel corpus from Tatoeba project	English | French
English-Hungarian Parallel corpus from Tatoeba project	English | Hungarian
English-Icelandic parallel corpus from Statistics Iceland	English | Icelandic
English-Icelandic parallel corpus from Statistics Iceland (Processed)	English | Icelandic
English-Icelandic Parallel corpus from Tatoeba project	English | Icelandic
English-Irish Parallel corpus from Tatoeba project	English | Irish
English-Irish website parallel corpus (Processed)	English | Irish
English-Italian parallel corpus from CORDIS Project News	Italian | English
English-Italian parallel corpus from CORDIS Project Results in Brief	Italian | English
English-Italian Parallel corpus from Tatoeba project	Italian | English
English-Latvian EASTIN-CL Multilingual Ontology of Assistive Technology (Processed)	English | Latvian
English-Latvian Parallel corpus from Tatoeba project	English | Latvian
English-Latvian parallel corpus from the www.visitestonia.com web site	English | Latvian
English-Lithuanian EASTIN-CL Multilingual Ontology of Assistive Technology (Processed)	English | Lithuanian
English-Lithuanian Parallel corpus from Tatoeba project	English | Lithuanian
English-Maltese Parallel corpus from Tatoeba project	English | Maltese
English-Norwegian Nynorsk Parallel corpus from Tatoeba project	Norwegian Nynorsk | English
English-Norwegian parallel corpus from Forbruker Europa, 2017 release (Processed)	Norwegian Bokmål | English
English-Norwegian Parallel corpus from Tatoeba project	English | Norwegian Bokmål
English-Norwegian Translation memory from Standard Norge	English | Norwegian Bokmål
English-Polish parallel corpus from CORDIS Project News	English | Polish
English-Polish parallel corpus from CORDIS Project Results in Brief	English | Polish
English-Polish Parallel corpus from Tatoeba project	English | Polish
English-Portuguese Parallel corpus from Tatoeba project	English | Portuguese
English-Portuguese website parallel corpus (Processed)	English | Portuguese

English-Romanian Parallel corpus from Tatoeba project	English | Romanian
English-Slovak corpus of annual reports from the Slovak National Centre for Human Rights website (Processed)	English | Slovak
English-Slovak corpus of annual reports on immigration and asylum policies from the EMN National Contact Point for the Slovak Republic website (Processed)	English | Slovak
English-Slovak Parallel corpus from Tatoeba project	English | Slovak
English-Slovak parallel corpus of texts from The Ministry of Culture of the Slovak Republic (Processed)	English | Slovak
English-Slovak parallel corpus of texts from The Ministry of Justice of the Slovak Republic (Processed)	English | Slovak
English-Slovene glossary of defence terminology (Processed)	Slovenian | English
English-Slovene glossary of engineering (Processed)	English | Slovenian
English-Slovenian Parallel corpus from Tatoeba project	English | Slovenian
English-Spanish parallel corpus from CORDIS Project News	English | Spanish
English-Spanish parallel corpus from CORDIS Project Results in Brief	English | Spanish
English-Spanish Parallel corpus from Tatoeba project	English | Spanish
English-Spanish website parallel corpus (Processed)	English | Spanish
English-Swedish corpus from Finnish Information Bank	English | Swedish
English-Swedish corpus from Finnish Information Bank (Processed)	English | Swedish
English-Swedish parallel corpus from Annual Reports of the Swedish Pension System (Processed)	English | Swedish
English-Swedish parallel corpus from National Audit Office of Finland	English | Swedish
English-Swedish Parallel corpus from Tatoeba project	English | Swedish
English-Swedish parallel corpus from the Annual Overview of Swedenβ€™s Official aid Agency SIDA Activities (Processed)	English | Swedish
English-Swedish parallel corpus from the contents of City of Turku web site	English | Swedish
English-Swedish parallel corpus from the Finnish Government web site	English | Swedish
English-Swedish parallel corpus from the Prime Minister's Office of Finland web site	English | Swedish
English-Swedish parallel corpus from the translation of 'Sweden a Pocket Guide' book (Processed)	English | Swedish
English-Swedish parallel corpus from the web site of Finnish Tax Administration	English | Swedish
English-Swedish parallel corpus from the web site of the Swedish Migration Board - Migrationsverket (Processed)	English | Swedish
English-Swedish parallel corpus from the www.visitestonia.com web site	English | Swedish
English-Swedish parallel texts from The Swedish Agency for Economic and Regional Growth - Tillväxtverket (Processed)	English | Swedish
Estatuto da VΓtima em PT e ENG	English | Portuguese
Estonian-English parallel corpus from the Estonian Classification of Economic Activities (EMTAK).	English | Estonian
EUIPO - IP case law French-English (Processed)	English | French
EUIPO - IP case law German-English (Processed)	German | English
EUIPO - IP case law Italian-English (Processed)	Italian | English
EUIPO - IP case law Spanish-English (Processed)	English | Spanish
EUIPO - list of goods and services French and English (Processed)	English | French
EUIPO - list of goods and services German and English (Processed)	German | English
EUIPO - list of goods and services German and French (Processed)	German | French
EUIPO - list of goods and services German and Italian (Processed)	Italian | German
EUIPO - list of goods and services German and Spanish (Processed)	German | Spanish
EUIPO - list of goods and services Italian and English (Processed)	Italian | English
EUIPO - list of goods and services Italian and French (Processed)	Italian | French
EUIPO - list of goods and services Italian and Spanish (Processed)	Italian | Spanish

EUIPO - list of goods and services Spanish and English (Processed)	English | Spanish
EUIPO - list of goods and services Spanish and French (Processed)	French | Spanish
EUIPO - Trade mark Guidelines (October 2017) (English-Bulgarian) (Processed)	English | Bulgarian
EUIPO - Trade mark Guidelines (October 2017) (English-Croatian) (Processed)	English | Croatian
EUIPO - Trade mark Guidelines (October 2017) (English-Czech) (Processed)	English | Czech
EUIPO - Trade mark Guidelines (October 2017) (English-Danish) (Processed)	Danish | English
EUIPO - Trade mark Guidelines (October 2017) (English-Dutch) (Processed)	English | Dutch
EUIPO - Trade mark Guidelines (October 2017) (English-Estonian) (Processed)	English | Estonian
EUIPO - Trade mark Guidelines (October 2017) (English-Finnish) (Processed)	English | Finnish
EUIPO - Trade mark Guidelines (October 2017) (English-French) (Processed)	English | French
EUIPO - Trade mark Guidelines (October 2017) (English-German) (Processed)	German | English
EUIPO - Trade mark Guidelines (October 2017) (English-Greek) (Processed)	Modern Greek | English
EUIPO - Trade mark Guidelines (October 2017) (English-Hungarian) (Processed)	English | Hungarian
EUIPO - Trade mark Guidelines (October 2017) (English-Italian) (Processed)	Italian | English
EUIPO - Trade mark Guidelines (October 2017) (English-Latvian) (Processed)	English | Latvian
EUIPO - Trade mark Guidelines (October 2017) (English-Lithuanian) (Processed)	English | Lithuanian
EUIPO - Trade mark Guidelines (October 2017) (English-Maltese) (Processed)	English | Maltese
EUIPO - Trade mark Guidelines (October 2017) (English-Polish) (Processed)	English | Polish
EUIPO - Trade mark Guidelines (October 2017) (English-Portuguese) (Processed)	English | Portuguese
EUIPO - Trade mark Guidelines (October 2017) (English-Romanian) (Processed)	English | Romanian
EUIPO - Trade mark Guidelines (October 2017) (English-Slovak) (Processed)	English | Slovak
EUIPO - Trade mark Guidelines (October 2017) (English-Slovenian) (Processed)	Slovenian | English
EUIPO - Trade mark Guidelines (October 2017) (English-Spanish) (Processed)	English | Spanish
EUIPO - Trade mark Guidelines (October 2017) (English-Swedish) (Processed)	English | Swedish
EuroPat release 1 English-French	English | French
EuroPat release 1 English-German	German | English
European single procurement document hu (Processed)	Hungarian
Expression of interest (Processed)	Modern Greek | English
Faisnéis faoi IDS	English | Irish
Financial Stability Reports from the National Bank of Poland (2013-14) (Processed)	English | Polish
Financial Stability Reports from the National Bank of Poland (2015-16) (Processed)	English | Polish
Foirm FSS Iarratais Duine ar a Shonraí	English | Irish
Forbruker Europa 2017 NO-EN glossary (Processed)	English | Norwegian Bokmål
General Romanian-English bilingual corpus (Processed)	English | Romanian
Genetics Termbase (processed)	English | Maltese
German-English parallel corpus from CORDIS Project News	German | English
German-English parallel corpus from CORDIS Project Results in Brief	German | English
German-English Parallel corpus from Tatoeba project	German | English
German-English parallel data by the Presidency of the Council of the EU held by Austria in 2006	German | English
German-English parallel data by the Presidency of the Council of the EU held by Luxembourg in 2015	German | English
German-English Presidency related parallel data	German | English
German-English website parallel corpus from the Federal Foreign Office Berlin	German | English
German-English website parallel corpus from the Federal Foreign Office Berlin (Processed)	German | English
German-French website parallel corpus from the Federal Foreign Office Berlin	German | French

German-French website parallel corpus from the Federal Foreign Office Berlin (Processed)	German | French
German-Portuguese website parallel corpus from the Federal Foreign Office Berlin	German | Portuguese
German-Portuguese website parallel corpus from the Federal Foreign Office Berlin (Processed)	German | Portuguese
Glossary City of Vienna (Processed)	German | English
Glossario bilingue PT EN	English | Portuguese
Greek anti-corruption legislation and National Anti-Corruption Plan (greek-english) (Processed)	Modern Greek | English
Greek-English parallel corpus from EQF Referencing Report (Processed)	Modern Greek | English
Greek-English parallel corpus from governmental documents about Migration Policy (Processed)	Modern Greek | English
Greek-English Parallel corpus from Tatoeba project	Modern Greek | English
Greek-English parallel corpus from the Hellenic Gaming Commission (Processed)	Modern Greek | English
Greek-English parallel corpus from the website of the Prime Minister of the Hellenic Republic (Processed)	Modern Greek | English
Greek-English glossary of diseases (Processed)	Modern Greek | English
Hallituskausi 2007-2011 -- Finnish-English Translation Memory (Processed)	English | Finnish
Hallituskausi 2007-2011 fi-en	English | Finnish
Hallituskausi 2011-2015 -- Finnish-English Translation Memory (Processed)	English | Finnish
Hallituskausi 2011-2015 fi-en	English | Finnish
Hellenic Ministry of Foreign Affairs Greek-English announcements corpus (Processed)	Modern Greek | English
Information Portal of the Czech President and Czech Castle	English | Czech
Information Portal of the German State Chancellery	German | English
Inteliterm EN-ES	English | Spanish
International Agreements (Processed)	English | Latvian
International Statistical Classification of Diseases and Related Health Problems - ICD-10 (EN-PL) (Processed)	English | Polish
Laws of Malta (Processed)	English | Maltese
Leabhrán d'Aonad Altranais Pobail Teach Uí Riada	English | Irish
Legal texts from Estonian Ministry of Justice (Processed)	English | Estonian
Legislation PT (Processed)	Portuguese
Legislação bilingue	English | Portuguese
Lei 20 de 2008 - PT e ENG	English | Portuguese
Lei 25 de 2009 - PT e ENG	English | Portuguese
Lei orgânica 2 de 2008	English | Portuguese
Letter of rights for persons arrested and or detained (Processed)	Multiple languages
Letter of rights for persons arrested on the basis of a European Arrest Warrant (Processed)	Multiple languages
List of names for substances for pharmaceutical use and preparations presented in European Pharmacopoeia (processed)	English | Lithuanian
Litir ó Oifig an Choimisinéara Teanga	English | Irish
Luxembourg Museum Websites (de-en) (Processed)	German | English | French
LáithreánGréasáinOÉG	English | Irish
Macroeconomic Developments (Processed)	Modern Greek | English
Malta Government Gazette (Processed)	English | Maltese
Maltese-English website parallel corpus	English | Maltese
Maltese-English website parallel corpus (Processed)	English | Maltese
Manufactured data based on ParaCrawl release 7 Bulgarian-English	English | Bulgarian
Manufactured data based on ParaCrawl release 7 Czech-English	English | Czech

Manufactured data based on ParaCrawl release 7 Danish-English	Danish | English
Manufactured data based on ParaCrawl release 7 Estonian-English	English | Estonian
Manufactured data based on ParaCrawl release 7 Finnish-English	English | Finnish
Manufactured data based on ParaCrawl release 7 Greek-English	Modern Greek | English
Manufactured data based on ParaCrawl release 7 Icelandic-English	English | Icelandic
Manufactured data based on ParaCrawl release 7 Latvian-English	English | Latvian
Manufactured data based on ParaCrawl release 7 Lithuanian-English	English | Lithuanian
Manufactured data based on ParaCrawl release 7 Portuguese-English	English | Portuguese
Manufactured data based on ParaCrawl release 7 Romanian-English	English | Romanian
Manufactured data based on ParaCrawl release 7 Russian-English	English | Russian
Manufactured data based on ParaCrawl release 7 Slovak-English	English | Slovak
Manufactured data based on ParaCrawl release 7 Slovenian-English	English | Slovenian
Manufactured data based on ParaCrawl release 7 Swedish-English	English | Swedish
Marketing de Influência	English | Portuguese
Memorandum for a ESM programme (Processed)	Modern Greek | English
Methodological Reconciliation (Processed)	Modern Greek | English
Monolingual documents from the Government of Lithuania (Processed)	Lithuanian
Multilingual documents in the field of health care and social policy (Processed)	English | Bulgarian
National Health Fund Dataset (Processed)	English | Polish
National Security and Defence. English-Estonian parallel data	English | Estonian
Natolin European Centre Dataset (Processed)	English | Polish
Norwegian Bokmål to English translation memories from Amesto Translations	English | Norwegian Bokmål
Oifigí Ombudsman in Éirinn	English | Irish
Orossimo Terminological Resource - Photography, film & video (Processed)	Modern Greek | English
ParaCrawl release 4 Bulgarian-English	English | Bulgarian
ParaCrawl release 4 Croatian-English	English | Croatian
ParaCrawl release 4 Czech-English	English | Czech
ParaCrawl release 4 Danish-English	Danish | English
ParaCrawl release 4 Dutch-English	English | Dutch
ParaCrawl release 4 Estonian-English	English | Estonian
ParaCrawl release 4 Finnish-English	English | Finnish
ParaCrawl release 4 French-English	English | French
ParaCrawl release 4 German-English	German | English
ParaCrawl release 4 Greek-English	Modern Greek | English
ParaCrawl release 4 Hungarian-English	English | Hungarian
ParaCrawl release 4 Irish-English	English | Irish
ParaCrawl release 4 Italian-English	Italian | English
ParaCrawl release 4 Latvian-English	English | Latvian
ParaCrawl release 4 Lithuanian-English	English | Lithuanian
ParaCrawl release 4 Maltese-English	English | Maltese
ParaCrawl release 4 Polish-English	English | Polish
ParaCrawl release 4 Portuguese-English	English | Portuguese
ParaCrawl release 4 Romanian-English	English | Romanian
ParaCrawl release 4 Slovak-English	English | Slovak
ParaCrawl release 4 Slovenian-English	Slovenian | English
ParaCrawl release 4 Spanish-English	English | Spanish

ParaCrawl release 4 Swedish-English	English | Swedish
ParaCrawl release 5 Bulgarian-English	English | Bulgarian
ParaCrawl release 5 Croatian-English	English | Croatian
ParaCrawl release 5 Czech-English	English | Czech
ParaCrawl release 5 Danish-English	Danish | English
ParaCrawl release 5 Dutch-English	English | Dutch
ParaCrawl release 5 Estonian-English	English | Estonian
ParaCrawl release 5 Finnish-English	English | Finnish
ParaCrawl release 5 French-English	English | French
ParaCrawl release 5 German-English	German | English
ParaCrawl release 5 Greek-English	Modern Greek | English
ParaCrawl release 5 Hungarian-English	English | Hungarian
ParaCrawl release 5 Irish-English	English | Irish
ParaCrawl release 5 Italian-English	Italian | English
ParaCrawl release 5 Latvian-English	English | Latvian
ParaCrawl release 5 Lithuanian-English	English | Lithuanian
ParaCrawl release 5 Maltese-English	English | Maltese
ParaCrawl release 5 Polish-English	English | Polish
ParaCrawl release 5 Portuguese-English	English | Portuguese
ParaCrawl release 5 Romanian-English	English | Romanian
ParaCrawl release 5 Slovak-English	English | Slovak
ParaCrawl release 5 Slovenian-English	Slovenian | English
ParaCrawl release 5 Spanish-English	English | Spanish
ParaCrawl release 5 Swedish-English	English | Swedish
ParaCrawl release 6 Bulgarian-English	English | Bulgarian
ParaCrawl release 6 Croatian-English	English | Croatian
ParaCrawl release 6 Czech-English	English | Czech
ParaCrawl release 6 Danish-English	Danish | English
ParaCrawl release 6 Dutch; Flemish-English	English | Dutch
ParaCrawl release 6 Estonian-English	English | Estonian
ParaCrawl release 6 Finnish-English	English | Finnish
ParaCrawl release 6 French-English	English | French
ParaCrawl release 6 German-English	German | English
ParaCrawl release 6 Hungarian-English	English | Hungarian
ParaCrawl release 6 Icelandic-English	English | Icelandic
ParaCrawl release 6 Irish-English	English | Irish
ParaCrawl release 6 Italian-English	Italian | English
ParaCrawl release 6 Latvian-English	English | Latvian
ParaCrawl release 6 Lithuanian-English	English | Lithuanian
ParaCrawl release 6 Maltese-English	English | Maltese
ParaCrawl release 6 Modern Greek (1453-)-English	Modern Greek | English
ParaCrawl release 6 Polish-English	English | Polish
ParaCrawl release 6 Portuguese-English	English | Portuguese
ParaCrawl release 6 Romanian; Moldavian; Moldovan-English	English | Romanian
ParaCrawl release 6 Slovak-English	English | Slovak
ParaCrawl release 6 Slovenian-English	English | Slovenian

ParaCrawl release 6 Spanish; Castilian-English	English | Spanish
ParaCrawl release 6 Swedish-English	English | Swedish
ParaCrawl release 7 Basque-Spanish	Spanish | Basque
ParaCrawl release 7 Bulgarian-English	English | Bulgarian
ParaCrawl release 7 Catalan-Spanish	Spanish | Catalan
ParaCrawl release 7 Croatian-English	English | Croatian
ParaCrawl release 7 Czech-English	English | Czech
ParaCrawl release 7 Danish-English	Danish | English
ParaCrawl release 7 Dutch-English	English | Dutch
ParaCrawl release 7 Estonian-English	English | Estonian
ParaCrawl release 7 Finnish-English	English | Finnish
ParaCrawl release 7 French-English	English | French
ParaCrawl release 7 Galician-Spanish	Galician | Spanish
ParaCrawl release 7 German-English	German | English
ParaCrawl release 7 Greek-English	Modern Greek | English
ParaCrawl release 7 Hungarian-English	English | Hungarian
ParaCrawl release 7 Icelandic-English	English | Icelandic
ParaCrawl release 7 Irish-English	English | Irish
ParaCrawl release 7 Italian-English	Italian | English
ParaCrawl release 7 Latvian-English	English | Latvian
ParaCrawl release 7 Lithuanian-English	English | Lithuanian
ParaCrawl release 7 Maltese-English	English | Maltese
ParaCrawl release 7 Norwegian Bokmål-English	English | Norwegian Bokmål
ParaCrawl release 7 Norwegian Nynorsk-English	Norwegian Nynorsk | English
ParaCrawl release 7 Polish-English	English | Polish
ParaCrawl release 7 Portuguese-English	English | Portuguese
ParaCrawl release 7 Romanian-English	English | Romanian
ParaCrawl release 7 Slovak-English	English | Slovak
ParaCrawl release 7 Slovenian-English	Slovenian | English
ParaCrawl release 7 Spanish-English	English | Spanish
ParaCrawl release 7 Swedish-English	English | Swedish
Parallel Bulgarian-English from the Official Tourism Portal of Bulgaria (Processed)	English | Bulgarian
Parallel corpus (Bulgarian - English) in the public administration domain	English | Bulgarian
Parallel corpus (Bulgarian - English) in the public administration domain (Processed)	English | Bulgarian
Parallel corpus (en-pl) from the Export Promotion Portal of Poland (Processed)	English | Polish
Parallel corpus (Greek - English) in the law domain (Processed) (Part1)	Modern Greek | English
Parallel corpus (Greek - English) in the public administration domain	Modern Greek | English
Parallel corpus (Greek - English) in the public administration domain (Processed)	Modern Greek | English
Parallel corpus (Polish - English) from the website of the Polish Investment and Trade Agency (Processed)	English | Polish
Parallel corpus from Bank of Estonia (Processed)	English | Estonian
Parallel corpus from Estonian Cabinet of Ministers (Processed)	English | Estonian
Parallel corpus from Estonian Ministry of Foreign Affairs (Processed)	English | Estonian
Parallel corpus from Parliament of Estonia (Processed)	English | Estonian
Parallel corpus from Social Insurance Agency -- Försäkringskassan (Sweden)	English | Swedish
Parallel corpus from Social Insurance Agency -- Försäkringskassan (Sweden) (Processed)	English | Swedish

Parallel Corpus from the Web Site of the the MFA of Latvia	English | Latvian
Parallel Corpus from the Web Site of the the MFA of Latvia (Processed)	English | Latvian
Parallel corpus from the website of the Chancellery of the Prime Minister of Poland (Processed)	English | Polish
Parallel English-Icelandic corpus from the contents of Icelandic National Debt Management Agency website	English | Icelandic
Parallel English-Icelandic corpus from the Icelandic Directorate for International Development Cooperation website	English | Icelandic
Parallel Global Voices (Bulgarian - English) (Processed)	English | Bulgarian
Parallel Global Voices (English - Czech) (Processed)	English | Czech
Parallel Global Voices (English - Dutch) (Processed)	English | Dutch
Parallel Global Voices (English - French) (Processed)	English | French
Parallel Global Voices (English - German) (Processed)	German | English
Parallel Global Voices (English - Hungarian) (Processed)	English | Hungarian
Parallel Global Voices (English - Italian) (Processed)	Italian | English
Parallel Global Voices (English - Polish) (Processed)	English | Polish
Parallel Global Voices (English - Portuguese) (Processed)	English | Portuguese
Parallel Global Voices (English - Romanian)	English | Romanian
Parallel Global Voices (English - Romanian) (Processed)	English | Romanian
Parallel Global Voices (English - Spanish) (Processed)	English | Spanish
Parallel Global Voices (English - Swedish) (Processed)	English | Swedish
Parallel Global Voices (Greek - English)	Modern Greek | English
Parallel Global Voices (Greek - English) (Processed)	Modern Greek | English
Parallel Global Voices (Greek - French)	Modern Greek | French
Parallel Global Voices (Greek - French) (Processed)	Modern Greek | French
Parallel Global Voices (Greek - Spanish)	Modern Greek | Spanish
Parallel Global Voices (Greek - Spanish) (Processed)	Modern Greek | Spanish
Parallel texts from Swedish Labour market agency (Processed)	Multiple languages
Parallel texts from Swedish Labour market agency. Part 2 (Processed)	Multiple languages
Parallel texts from Swedish National Food Agency (Processed)	French | Polish | Swedish | English | Finnish | Spanish
Parallel texts from Swedish Social Security Authority (Processed)	Multiple languages
Parallel texts from Swedish Work environment Authority (Processed)	Multiple languages
Parallel texts from the Swedish Competition Authority - Konkurrensverket (Processed)	English | Swedish
Parallel texts from the Swedish Migration Board - Migrationsverket (English-Croatian part) (Processed)	English | Croatian
Pathomorphological Diagnose (processed)	English | Estonian
PKN Orlen Dataset (Processed)	English | Polish
Plan Nacional e Integral de Turismo (PNIT)	English | Spanish
Pleananna ITBÁC le comóradh a dhéanamh ar 1916	English | Irish
Polish Food 4 & Food Policy Dataset (Processed)	English | Polish
Polish Food Dataset (Processed)	English | Polish
Polish Food Dataset 2 (Processed)	English | Polish
Polish Food DataSet 3 (Processed)	English | Polish
Polish Ministry of Foreign Affairs Historical Dataset (Processed)	English | Polish
Polish Ministry of Foreign Affairs Regional Dataset (Processed)	English | Polish
Polish Ministry of Foreign Affairs reports in EN and PL (Processed)	English | Polish
Polish Ministry of Foreign Affairs Youth 2011 Report (Processed)	English | Polish
Polish-English Internal Aviation Glossaries (Processed)	English | Polish

Polish-English parallel corpus from the website "Business in Poland" (Processed)	English | Polish
Polish-English parallel corpus from the website "geoportal.gov.pl" (Processed)	English | Polish
Polish-English parallel corpus from the website "Polish Aid" (Processed)	English | Polish
Polish-English parallel corpus from the website "Science in Poland" (Processed)	English | Polish
Polish-English parallel corpus from the website of Public Employment Services in Poland (member of EURES network) (Processed)	English | Polish
Polish-English parallel corpus from the website of the Central Statistical Office (Processed)	English | Polish
Polish-English parallel corpus from the website of the Citizens Information Board (Processed)	English | Polish
Polish-English parallel corpus from the website of the ING Polish Art Foundation (Processed)	English | Polish
Polish-English parallel corpus from the website of the Institute of Mathematics of the Polish Academy of Sciences (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Agriculture and Rural Development (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Culture and National Heritage (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Development (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Digital Affairs (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Digitization (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Foreign Affairs (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Justice (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of National Defence (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Regional Development (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of Science and Higher Education (Processed)	English | Polish
Polish-English parallel corpus from the website of the Ministry of the Interior and Administration (Processed)	English | Polish
Polish-English parallel corpus from the website of the National Audiovisual Institute (Processed)	English | Polish
Polish-English parallel corpus from the website of the National Centre for Nuclear Research (Processed)	English | Polish
Polish-English parallel corpus from the website of the National Centre for Research and Development (Processed)	English | Polish
Polish-English parallel corpus from the website of the National Digital Archives (Processed)	English | Polish
Polish-English parallel corpus from the website of the National Science Centre (Processed)	English | Polish
Polish-English parallel corpus from the website of the National Security Bureau (Processed)	English | Polish
Polish-English parallel corpus from the website of the Office of the Commissioner for Human Rights (Processed)	English | Polish
Polish-English parallel corpus from the website of the Polish Tourism Organisation (Processed)	English | Polish
Polish-English parallel corpus from the website of the State Marine Accident Investigation Commission (Processed)	English | Polish
Polish-English parallel corpus from the website of the U.S. EMBASSY and CONSULATE IN POLAND (Processed)	English | Polish
PolsasГ ar FheiniЃlacht agus LГ©iriЃl Inscne Ollscoil MhГ¯ Nuad 2019	English | Irish
Portuguese legislation in English and French (Processed)	English | French
Portuguese legislation in FR (Processed)	French

Portuguese-English bilingual corpus from Legislation concerning the Portuguese Parliament	English | Portuguese
Portuguese-English bilingual corpus from Legislation concerning the Portuguese Parliament (Processed)	English | Portuguese
Portuguese-English bilingual corpus from the Portuguese Constitution	English | Portuguese
Portuguese-English bilingual corpus from the Portuguese Constitution (Processed)	English | Portuguese
Portuguese-French bilingual corpus from Portuguese law on referendum	French | Portuguese
Portuguese-French bilingual corpus from Portuguese law on referendum (Processed)	French | Portuguese
Preasráiteas faoi foirgneamh nua scoile	English | Irish
Preasráiteas faoi Uachtarán nua	English | Irish
Preasráiteas: Mí Iúil	English | Irish
Preasráitis Gaois, Fiontar & Scoil na Gaeilge, DCU (1)	English | Irish
Preasráitis Gaois, Fiontar & Scoil na Gaeilge, DCU (2)	English | Irish
Preasráitis Oifig an Choimisinéara Teanga	English | Irish
Preasráitis Ollscoil Mhá Nuad Earrach 2019	English | Irish
Preasráitis Ollscoil Mhá Nuad Samhradh 2019	English | Irish
Press Releases (01.2018-01.2019) of the PIO (Processed)	Modern Greek | English
Press Releases from Department of Children January - May 2019	English | Irish
Programme for Government Annual Report 2013	English | Irish
Public Procurement Dataset 1 (Processed)	English | Polish
Public Procurement Dataset 2 (Processed)	English | Polish
Póstaer faoi scoil ag clárú	English | Irish
Quarterly Reports of the Parliamentary Budget Office (Hellenic Parliament) (Processed)	Modern Greek | English
RELATÓRIO GAM	French | Portuguese
Romanian - English literature corpus (Processed)	English | Romanian
Romanian - English news corpus (Processed)	English | Romanian
Romanian Ombudsman archive (Processed)	English | Romanian
Romanian - English New Criminal Procedure Code (Processed)	English | Romanian
Romanian-English parallel wordlists (Processed)	English | Romanian
Romanian-English corpus with studies, reports and statistical data in the field of culture from the National Institute for Cultural Research and Training website (Processed)	English | Romanian
Rural Development Programme of Romania (Processed)	English | Romanian
Ráitis Airgeadais Oifig an Choimisinéara Teanga	English | Irish
Ráitis Airgeadais Ollscoil Mhá Nuad 2017 - 2018	English | Irish
Secretariat-General parallel corpus SL-EN and EN-SL (part 1)	English | Slovenian
Secretariat-General parallel corpus SL-EN and EN-SL (part 1) (Processed)	English | Slovenian
Secretariat-General parallel corpus SL-EN and EN-SL (part 2)	English | Slovenian
Secretariat-General parallel corpus SL-EN and EN-SL (part 2) (Processed)	English | Slovenian
SEKO - Finnish Performance Vocabulary (processed)	Finnish
SIP Internal dictionary (Processed)	German | French | English
SIP Publications (Processed)	German | English | French
Slovak-English glossary of diseases (Processed)	English | Slovak
Slovene-English chemical glossary (Processed)	Slovenian | English
Slovenian-English corpus with chapters from the "Youth 2010, the social profile of young people in Slovenia" publication (Processed)	Slovenian | English
Slovenian-English corpus with statistical reports from the Statistical Office of the Republic of Slovenia website (Processed)	English | Slovenian
Spanish-English website parallel corpus	English | Spanish

Spanish-English website parallel corpus (Processed)	English | Spanish
Spanish-French website parallel corpus	French | Spanish
Spanish-French website parallel corpus (Processed)	French | Spanish
Spanish-German website parallel corpus	German | Spanish
Spanish-German website parallel corpus (Processed)	German | Spanish
Spanish-Italian website parallel corpus	Italian | Spanish
Spanish-Italian website parallel corpus (Processed)	Italian | Spanish
Spanish-Portuguese website parallel corpus	Portuguese | Spanish
Spanish-Portuguese website parallel corpus (Processed)	Spanish | Portuguese
Statens Vegvesen Translation Memories	Norwegian Bokmål | English
Statistical / Economic Administrative Terminology (Processed)	English | Icelandic
Terminology (Italian legal system) German-Italian	Italian | German
Terminology (Italian legal system) Italian-German	Italian | German
Terminology in the domain of Information and Communication Technology (ICT)	English | Latvian
Terminology_of_international_contracts_Dutch_(Processed)	Dutch
Terminology_of_international_contracts_German_(Processed)	German
Terminology_of_international_contracts_Italian_(Processed)	Italian
Terminology_of_international_contracts_Portuguese_(Processed)	Portuguese
Terminology_of_the _German_Foreign_Office_Dutch_(Processed)	Dutch
Terminology_of_the _German_Foreign_Office_German_(Processed)	German
Terminology_of_the _German_Foreign_Office_Italian_(Processed)	Italian
Terminology_of_the _German_Foreign_Office_Portuguese_(Processed)	Portuguese
TERMIS: Slovene-English terminology in the field of public relations (Processed)	English | Slovenian
TERMITUR EN-ES	English | Spanish
The Coimisineir Teanga Bilingual Corpus of Reference Documents (Processed)	English | Irish
The Coimisineir Teanga Bilingual Corpus of Reports and Press Releases (Processed)	English | Irish
The Coimisineir Teanga Bilingual Web Corpus (Processed)	English | Irish
The Constitution of Greece (English-Greek) (Processed)	Modern Greek | English
The Croatian-English corpus with the nature protection strategy of Croatia (Processed)	English | Croatian
The Gaois bilingual corpus of English-Irish legislation (Processed)	English | Irish
The UCD Bord na Gaeilge Corpus of bilingual PDFs and Word documents (Processed)	English | Irish
The Udáras na Gaeltachta Corpus of bilingual PDFs and Word documents (Processed)	English | Irish
TMX ó Chonradh don Aonad Chumarsáide den Roinn Gnóthaí Eachtracha agus Trádála	English | Irish
TMX ó Chonradh Ginearálta den Roinn Gnóthaí Eachtracha agus Trádála	English | Irish
Toiliú don Scagthástáil Scoile um Amhairc & Éisteachta	English | Irish
Translation memories from Forbruker Europa	Norwegian Bokmål | English
Translation memories from Forbruker Europa (Processed)	English | Norwegian Bokmål
Translation memories from The Ministry of Foreign Affairs of Norway	Norwegian Bokmål | English
Translation memories from The Ministry of Foreign Affairs of Norway (Processed)	Norwegian Bokmål | English
Translation memory from Swedish National Audit Office (NAO) - Riksrevisionen (Processed)	English | Swedish
Translations of Lithuanian legislation from Seimas of the Republic of Lithuania	English | Lithuanian
Translations of Lithuanian legislation from Seimas of the Republic of Lithuania (Processed)	English | Lithuanian
Trilingual Documents related to International Judicial Cooperation in Civil Matters (Greek-English-French) (Processed)	Modern Greek | English | French
Tuairisc a thug Máire Nic Shiubhlaigh, Aisteoir Tionscanta de chuid Amharlann na Mainistreach, ar ghéilleadh gharastún mhonarcha Jacob	English | Irish

Tuarascáil Bhliantúil Chomhairle Chontae Longfoirt (2017)	English | Irish
Tuarascálaca Bliantúla na Roinne Leanaí agus Gnóthaí Óige	English | Irish
Téarmaíocht agus aistriúcháin a bhaineann le fógraí poist, folúntais, ábhair chomórtha 1916 agus eolas ginearálta ar Oifig na Gaeilge.	English | Irish
Vienna Environmental Report2004/2005 (Processed)	German | English
Webpage_Foreign_Office_AA_de-fr_2016-2018 (Processed)	German | French
Website of the President of the Republic of Lithuania (Processed)	English | Lithuanian
ZZP. Cisco Academy terminology (Processed)	English | Croatian

Table 7: LRs from ELRC-SHARE

A.C. META-SHARE (74 LRs)

This annex provides an overview of the LRs ingested into ELG from META-SHARE.

Resource name	Language(s)
Czech Association of Medical Physicists - Physics Glossary (Processed)	English | Czech
Czech Banking Association Terminology (Processed)	English | Czech
DeepBankDE	German
English ontology lexicon	English
eSCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing	Italian | English | German
European Clinical Case Corpus	French | English | Italian | Basque | Spanish
Finance domain ontology	English
Finance English corpus	English
Finance English grammar	English
Finance English web corpus, automatically harvested	English
Greek Textual Entailment Corpus	Modern Greek (1453-)
Greek-Bulgarian Bul-TM parallel corpus	Modern Greek | Bulgarian
ILSP PsychoLinguistic Resource	Modern Greek (1453-)
INTERA Corpus - the Bulgarian POS annotated part of the BG-EN pair	Bulgarian
INTERA Corpus - the Bulgarian structurally annotated part of the BG-EN pair	Bulgarian
INTERA Corpus - the Bulgarian-English part	English | Bulgarian
INTERA Corpus - the Bulgarian-English terms from the BG-EN pair	English | Bulgarian
INTERA Corpus - the English POS annotated part of the BG-EN pair	English
INTERA Corpus - the English POS annotated part of the EL-EN pair	English
INTERA Corpus - the English POS annotated part of the English-Slovene SVEZ ACQUIS Corpus	English
INTERA Corpus - the English POS annotated part of the SR-EN pair	English
INTERA Corpus - the English structurally annotated part of the BG-EN pair	English
INTERA Corpus - the English structurally annotated part of the EL-EN pair	English
INTERA Corpus - the English structurally annotated part of the EN-SL SVEZ ACQUIS corpus	English
INTERA Corpus - the English structurally annotated part of the SR-EN pair	English
INTERA Corpus - the English-Slovene terms from the EN-SL SVEZ ACQUIS Corpus	English | Slovenian
INTERA Corpus - the Greek POS annotated part of the EL-EN pair	Modern Greek
INTERA Corpus - the Greek structurally annotated part of the EL-EN pair	Modern Greek
INTERA Corpus - the Greek-English part	Modern Greek | English
INTERA Corpus - the Greek-English terms from the EL-EN pair	Modern Greek | English
INTERA Corpus - the Serbian POS annotated part of the SR-EN pair	Serbian
INTERA Corpus - the Serbian structurally annotated part of the SR-EN pair	Serbian

INTERA Corpus - the Serbian-English part	English | Serbian
INTERA corpus - the Serbian-English terms from the SR-EN pair	English | Serbian
INTERA Corpus - the Slovene structurally annotated part of the EN-SL SVEZ ACQUIS corpus	Slovenian
INTERA Corpus - the Slovene SVEZ ACQUIS POS annotated part of the EN-SL SVEZ ACQUIS Corpus	Slovenian
INTERA English-Slovene SVEZ ACQUIS Corpus	Slovenian | English
IT helpdesk Italian web corpus, manually harvested	Italian
IT helpdesk Spanish web corpus, automatically harvested	Spanish
IWSLT 2015 Human Post-Editing data	German | Vietnamese | English
IWSLT 2016 Human Post-Editing data	German | English | French
IWSLT 2017 Human Post-Editing data	Italian | German | Dutch | Romanian
KELLY word-list Greek.	Modern Greek
PANACEA Environment Corpus n-grams EL (Greek)	Modern Greek
PANACEA Labour Legislation Corpus n-grams EL (Greek)	Modern Greek
Parallel Global Voices	Multiple languages
POETICON Multisensory and Multimedia Recordings of Everyday Interaction	English
SemEval-2016 ABSA Museum Reviews-French: Test Data-GOLD (Subtask 3)	French
SemEval-2016 ABSA Museum Reviews-French: Test Data-Phase A (Subtask 3)	French
SemEval-2016 ABSA Museum Reviews-French: Test Data-Phase B (Subtask 3)	French
SemEval-2016 ABSA Restaurant Reviews-French: Test Data-GOLD (Subtask 1)	French
SemEval-2016 ABSA Restaurant Reviews-French: Test Data-Phase A (Subtask 1)	French
SemEval-2016 ABSA Restaurant Reviews-French: Test Data-Phase B (Subtask 1)	French
SemEval-2016 ABSA Restaurant Reviews-French: Train Data (Subtask 1)	French
TermCymru	English | Welsh
The TaraXÜ Corpus of Human-Annotated Machine Translations - 4th evaluation round	French | English | German | Spanish | Czech
Tourism English grammar	English
Tourism Italian grammar	Italian
Travel domain ontology	English | Modern Greek | German | Italian | Spanish
Travel English crowdsourced corpus	English
Travel English grammar	English
Travel English ontology lexicon	English
Travel English web corpus, automatically harvested	English
Travel English web corpus, manually harvested	English
Travel German ontology lexicon	German
Travel Greek grammar	Modern Greek (1453-)
Travel Greek web corpus, automatically harvested	Modern Greek (1453-)
WMT 2015 Human Evaluations	N/A
WMT 2015 Translation Task Submissions	French | Russian | English | German | Finnish | Czech
WMT 2016 Human Evaluations	N/A
WMT 2016 Translation Task Submissions	Multiple languages
WMT 2017 Human Evaluations	N/A
WMT 2017 Translation Task Submissions	Multiple languages
WMT18 Quality Estimation Task: Product Reviews	English | French

Table 8: LRs from META-SHARE

A.D. LINDAT-CLARIAH-CZ (309 LRs)

This annex provides an overview of the LRs ingested into ELG from the LINDAT-CLARIAH-CZ repository.

Resource name	Language(s)
"Al wassit" Arabic dictionary	Arabic
"Al wassit" LMF Arabic dictionary	Arabic
A Gold Standard Word Alignment for English-Swedish (2015-10-12)	English | Swedish
A morphological layer for the German part of the SMULTRON corpus	German
A Small Dataset for English-to-Czech Speech Translation in the Travel Domain	English | Czech
A Speech Test Set of Practice Business Presentations with Additional Relevant Texts	English
Additional German-Czech reference translations of the WMT'11 test set	German | Czech
AdjDeriNet: Words Derived from Adjectives in Czech	Czech
Air Traffic Control Communication	English
AKCES 1	Czech
AKCES 2	Czech
AKCES 2 ver. 2	Czech
AKCES 3	Czech
AKCES 4	Czech
AKCES 5 (CzeSL-SGT)	Czech
AKCES 5 (CzeSL-SGT) Release 2	Czech
AKCES-GEC Grammatical Error Correction Dataset for Czech	Czech
Alex Context NLG Dataset	English
Amharic Web Corpus	Amharic
Amharic WIC Corpus	Amharic
Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.0)	Multiple languages
Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1)	Multiple languages
Annotated corpora and tools of the PARSEME Shared Task on Semi-Supervised Identification of Verbal Multiword Expressions (edition 1.2)	Multiple languages
Annotated Corpus of Czech Case Law for Reference Recognition Tasks	Czech
Annotated Corpus of Czech Case Law for Reference Recognition Tasks (2019-06-25)	Czech
Annotated Corpus of Czech Case Law for Segmentation Tasks	Czech
APE Shared Task WMT17: Human Post-edits Test Data DE-EN	English
APE Shared Task WMT17: Human Post-edits Test Data EN-DE	German
APE Shared Task WMT18: Human Post-edits and References Test Data EN-DE PBSMT	German
Arabic characters lexicon	Arabic
Arabic Enclitics Lexicon	Arabic
Arabic Morphological evaluation corpus	Arabic
Arabic Particles Lexicon	Arabic
Arabic Proclitics Lexicon	Arabic
Arabic Special verbs Lexicon	Arabic
Artificial Treebank with Ellipsis	Russian | English | Finnish | Slovak | Czech
Aspect-Term Annotated Customer Reviews in Czech	Czech
ATCC: Pronunciation lexicon and n-gram counts for ASR module	English
Automatic Paraphrases of Czech Reference Sentences for WMT11, 13 and 14	Czech
Automatically generated spelling correction corpus for Czech (Czech-SEC-AG)	Czech
Balaxan Corpus of Kurmanji	Northern Kurdish

BushBank	Czech
C4Corpus (CC BY-NC part)	Multiple languages
C4Corpus (CC BY-NC-ND part)	Multiple languages
C4Corpus (CC BY-NC-SA part)	Multiple languages
C4Corpus (CC BY-ND part)	Multiple languages
C4Corpus (CC BY-SA part)	Multiple languages
C4Corpus (CC-BY part)	Multiple languages
C4Corpus (publicdomain part)	Multiple languages
CEHugeWebCorpus	German
CoNLL 2009 Shared Task - Czech Data	Czech
CoNLL 2009 Shared Task Czech Trial Set	Czech
CoNLL 2017 and 2018 Shared Task Blind and Preprocessed Test Data	Multiple languages
CoNLL 2017 Shared Task System Outputs	Multiple languages
CoNLL 2018 Shared Task System Outputs	Multiple languages
Contemporary Arabic dictionary	Arabic
Corpus for training and evaluating diacritics restoration systems	Multiple languages
Corpus of contemporary blogs	Czech
COSTRA 1.0: A Dataset of Complex Sentence Transformations	Czech
COSTRA 1.1: A Dataset of Complex Sentence Transformations and Comparisons	Czech
CsEnVi Pairwise Parallel Corpora	Vietnamese | English | Czech
CWC2011	Czech
Czech and English abstracts of Г?FAL papers	English | Czech
Czech Court Decisions Corpus (CzCDC 1.0)	Czech
Czech Court Decisions Dataset	Czech
Czech Grammar Agreement Dataset for Evaluation of Language Models	Czech
Czech Legal Text Treebank	Czech
Czech Legal Text Treebank 2.0	Czech
Czech Malach Cross-lingual Speech Retrieval Test Collection	French | English | German | Spanish | Czech
Czech Multiword Expressions	Czech
Czech Named Entity Corpus 1.0	Czech
Czech Named Entity Corpus 1.1	Czech
Czech Named Entity Corpus 2.0	Czech
Czech Parliament Meetings	Czech
Czech Relationship Extraction Dataset	Czech
Czech restaurant information dataset for NLG	Czech
Czech Senior COMPANION Expressive Speech Corpus	Czech
Czech Sociological Review 1993-2016	Czech
Czech SubLex 1.0	Czech
Czech Television News Broadcasting Faces	Czech
Czech Text Document Corpus v 2.0	Czech
Czech Translation of SQuAD 2.0 and 1.1	Czech
Czech translation of the EBUContentGenre thesaurus	English | Czech
Czech Verbal MWEs	Czech
Czech WordNet 1.9 PDT	Czech
Czech-English Manual Word Alignment	English | Czech
Czech-English Parallel Corpus 1.0 (CzEng 1.0)	English | Czech

Czech-Slovak Parallel Corpus	Slovak | Czech
CzeDLex 0.5	Czech
CzeDLex 0.6	Czech
CzEng 0.7	English | Czech
CzEngClass 0.1	English | Czech
CzEngClass 0.2	English | Czech
CzEngClass 0.3	English | Czech
CzEngVallex	English
czes	Czech
Czesl - Universal Dependencies Release 0.5	Czech
CzeSL Grammatical Error Correction Dataset (CzeSL-GEC)	Czech
czTenTen12 v9 subcorpus of problematic phenomena	Czech
Database of speech corpora of Czech laryngectomy patients	Czech
Deep Universal Dependencies 2.4	Multiple languages
Deep Universal Dependencies 2.5	Multiple languages
Deep Universal Dependencies 2.6	Multiple languages
Deltacorpus	Multiple languages
Deltacorpus 1.1	Multiple languages
DeriNet 1.0	Czech
DeriNet 1.2	Czech
DeriNet 1.5	Czech
DeriNet 1.6 (2018-09-24)	Czech
DeriNet 2.0	Czech
Digital Humanities Courses at Czech Colleges 2017/2018	Czech
DiscoMT 2015 Shared Task on Pronoun Translation	English | French
DiscoMT 2016 Shared Task on Cross-lingual Pronoun Prediction	German | French | English
DiscoMT 2017 Shared Task on Cross-lingual Pronoun Prediction	German | French | Spanish | English
DOESTE v0.5	Portuguese
Engineering job ads corpus	Spanish
English TTS speech corpus of air traffic (pilot) messages - Czech accent	English
English TTS speech corpus of air traffic (pilot) messages - German accent	English
English TTS speech corpus of air traffic (pilot) messages - Serbian accent	English
English TTS speech corpus of air traffic (pilot) messages - Taiwanese accent	English
English-Czech Corpus from Wikipedia	English | Czech
English-Hindi Parallel Corpus	English | Hindi
English-Slovak Parallel Corpus	English | Slovak
English-Urdu Religious Parallel Corpus	Urdu | English
EngVallex - English Valency Lexicon	English
Enriched Discourse Annotation of PDiT Subset 1.0 (PDiT-EDA 1.0)	Czech
EnTam: An English-Tamil Parallel Corpus (EnTam v2.0)	English | Tamil
enTenTen	English
Europarl QTLeap WSD/NED corpus	English | Spanish | Bulgarian | Basque | Portuguese | Czech
Extended CLEF eHealth 2013-2015 IR Test Collection	Multiple languages
Extended Morphosyntactic Testset for Word2Vec	English
Extended Textual Coreference and Bridging Relations in PDT 2.0	Czech

Eye-Tracking Recordings from a Pilot Study of WMT-style MT Outputs Ranking	English | Czech
Facebook Data for Sentiment Analysis	Czech
FAspell	Persian
FicTree 1.0	Czech
Gold Standard Reference Data for Multiword Expression Extraction: Czech Dependency Bigrams from the Prague Dependency Treebank	Czech
HamleDT 2.0	Multiple languages
HamleDT 3.0	Multiple languages
High-Coverage Multi-Level Text Corpus for Non-Professional Voice Conservation	Czech
HindEnCorp 0.5	English | Hindi
Hindi Visual Genome 1.0	English | Hindi
Hindi Visual Genome 1.1	English | Hindi
Hindi Web Texts	Hindi
HindMonoCorp 0.5	Hindi
IDENTICv1.0	English | Indonesian
IDENTICv1.0-raw	English | Indonesian
Indonesian web corpus	Indonesian
Indonesian web corpus (idWac)	Indonesian
Italian Content Words	Italian
Italian Content Words v2	Italian
Italian Content Words v3	Italian
Italian Function Words	Italian
Italian Function Words v2	Italian
Italian Function Words v3	Italian
IWPT 2020 Shared Task Data and System Outputs	Multiple languages
KAMOKO: KAsseler MOrgenstern KOrpus	French
Khresmoi Query Translation Test Data 1.0	German | French | English | Czech
Khresmoi Query Translation Test Data 2.0	Multiple languages
Khresmoi Summary Translation Test Data 1.1	German | English | French | Czech
Khresmoi Summary Translation Test Data 2.0	Multiple languages
Large Corpus of Czech Parliament Plenary Hearings	Czech
Large-Scale Colloquial Persian 0.5	Hindi | English | Italian | German | Persian | Czech
LatinISE corpus	Latin
LatinISE corpus (version 4)	Latin
Lexicon of Czech and German Anaphoric Connectives	German | Czech
Lexico-Semantic Annotation of PDT using Czech WordNet	Czech
LMF Arabic characters lexicon	Arabic
LMF Contemporary Arabic dictionary	Arabic
Manual Re-evaluation of Translation Quality of WMT 2018 English-Czech systems	English | Czech
Manually Classified Errors in Cs->Sk Translation	Slovak | Czech
Manually Classified Errors in En->Sk Translation	English | Slovak
Manually Ranked Translation Outputs	Slovak | Czech
Many Czech References for 50 Sentences Selected from WMT11 Data	Czech
Medieval Charter Sections Corpus	Latin | Czech
MorfFlex CZ	Czech

MorfFlex CZ 160310	Czech
MorfFlex CZ 161115	Czech
MorfFlex SK 170914	Slovak
Multilingual corpus of literal occurrences of multiword expressions	Polish | Modern Greek | German | Basque | Portuguese
Multiword expressions in the Prague Dependency Treebank 2.0	Czech
MUSCIMA++	No linguistic content
NAFIS Arabic Stemming Gold Standard Corpus	Arabic
NomVallex I.	Czech
OAGK Keyword Generation Dataset	English
OAGKX Keyword Generation Dataset	English
OAGL Paper Length Dataset	English
OAGS Title Generation Dataset	English
OAGSX Title Generation Dataset	English
OdiEnCorp 1.0	English | Oriya (macrolanguage)
OdiEnCorp 2.0	English | Oriya (macrolanguage)
Open SDP	English | Czech
Open SDP 1.2	English | Czech
ORAL2006: Corpus of informal spoken Czech	Czech
ORAL2008: Balanced corpus of informal spoken Czech	Czech
ORAL2013: balanced corpus of informal spoken Czech (transcriptions & audio)	Czech
ORAL2013: balanced corpus of informal spoken Czech (transcriptions)	Czech
Oromo web corpus	Oromo
ORTOFON v1: balanced corpus of informal spoken Czech with multi-tier transcription (transcriptions & audio)	Czech
ORTOFON v1: balanced corpus of informal spoken Czech with multi-tier transcription (transcriptions)	Czech
OVM β€“ OtГ¨zky VГ¨clava Moravce	Czech
ParaCrawl Corpus version 1.0	Multiple languages
ParaDi 2.0	Czech
ParaDi 2.0 (2018-01-24)	Czech
ParaDi: Dictionary of Paraphrases of Czech Complex Predicates with Light Verbs	Czech
ParCorFull: A Parallel Corpus Annotated with Full Coreference	German | English
ParCzech PS7 1.0	Czech
PAWS	Czech | English | Polish | Russian
PDT-Vallex: Czech Valency lexicon linked to treebanks	Czech
Persian Morphologically Segmented Lexicon 0.5	Persian
Plaintext Wikipedia dump 2018	Multiple languages
Posts of German PC Games Online Forum	German
Prague Arabic Dependency Treebank 1.0	Arabic
Prague Czech-English Dependency Treebank 2.0	English | Czech
Prague Czech-English Dependency Treebank 2.0 - Russian translation	Czech | English | Russian
Prague Czech-English Dependency Treebank 2.0 Coref	English | Czech
Prague DaTabase of Spoken Czech 1.0	Czech
Prague Dependency Treebank 2.0 - sample data	Czech
Prague Dependency Treebank 2.0 (PDT 2.0)	Czech
Prague Dependency Treebank 2.5	Czech
Prague Dependency Treebank 3.0	Czech

Prague Dependency Treebank 3.5	Czech
Prague Dependency Treebank of Spoken Language (PDTSL) 0.5	English | Czech
Prague Discourse Treebank 1.0	Czech
Prague Discourse Treebank 2.0	Czech
QT21 Data	German | English | Latvian | Czech
QTLeap WSD/NED corpus	English | Spanish | Bulgarian | Basque | Portuguese | Czech
Question Dialogs Dataset	English
Restaurant Reviews CZ ABSA corpus v2	Czech
Retrograde Morphemic Dictionary of Czech	Czech
Retrograde Morphemic Dictionary of Czech - verbs	Czech
sholva-0.6	Czech
skTenTen	Slovak
Slovak Dependency Treebank	Slovak
SLΓNDa	Swedish
Somali Web Corpus	Somali
Special Nouns Lexicon	Arabic
Speech databases of typical children and children with SLI	Czech
Spoken corpus of Karel Makoň	Czech
Spoken corpus of Karel Makoň (2020-11-16)	Czech
SQAD	Czech
sqad 2.1	Czech
sqad 3.0	Czech
SQAD v2	Czech
STAZKA - Speech recordings from vehicles	Czech
STYX 1.0	Czech
STYX 1.0 (2017-10-03)	Czech
SumeCzech	Czech
SYN v4: large corpus of written Czech	Czech
SYN2005: balanced corpus of written Czech	Czech
SYN2006PUB: corpus of Czech newspapers	Czech
SYN2009PUB: corpus of Czech newspapers	Czech
SYN2010: balanced corpus of written Czech	Czech
SYN2013PUB: corpus of written Czech newspapers	Czech
SYN2015: representative corpus of written Czech	Czech
SynSemClass 1.0	English | Czech
SynSemClass2.0	English | Czech
SynTagRus gapping test set	Russian
Tamil Dependency Treebank v0.1	Tamil
Test Data DE-EN APE Shared Task WMT17	German | English
Test Data EN-DE APE Shared Task WMT17	German | English
Test Data EN-DE MT_NMT APE Shared Task WMT18	German | English
Test Data EN-DE MT_PBSMT APE Shared Task WMT18	German | English
The ACL RD-TEC 2.0	English
Tigrinya Web Corpus	Tigrinya
UMC 0.1: Czech-Russian-English Multilingual Corpus	Czech

Universal Dependencies 1.0	Multiple languages
Universal Dependencies 1.1	Multiple languages
Universal Dependencies 1.2	Multiple languages
Universal Dependencies 1.3	Multiple languages
Universal Dependencies 1.4	Multiple languages
Universal Dependencies 2.0	Multiple languages
Universal Dependencies 2.0 alpha (obsolete)	Multiple languages
Universal Dependencies 2.0 β€“ CoNLL 2017 Shared Task Development and Test Data	Multiple languages
Universal Dependencies 2.1	Multiple languages
Universal Dependencies 2.2	Multiple languages
Universal Dependencies 2.3	Multiple languages
Universal Dependencies 2.4	Multiple languages
Universal Dependencies 2.5	Multiple languages
Universal Dependencies 2.6	Multiple languages
Universal Dependencies 2.7	Multiple languages
Universal Derivations v0.5	Multiple languages
Universal Derivations v1.0	Multiple languages
Urdu Monolingual Corpus	Urdu
VALLEX 2.5	Czech
VALLEX 3.0	Czech
Video699: lecture recordings and lecture materials	English | Czech
VPS-30-En	English
VPS-GradeUp (2016-10-10)	English
Vystadial 2013 - Czech data	Czech
Vystadial 2013 - English data	English
Vystadial 2016 - Czech data	Czech
W2C - Web to Corpus β€“ Corpora	Multiple languages
WMT 13 Test Set	Multiple languages
WMT 2011 Testing Set	English | Slovak | Czech
WMT16 APE Shared Task Data	German | English
WMT16 APE Shared Task Data - Reference sentences	German
WMT16 Quality Estimation Shared Task Training and Development Data	German | English
WMT16 Tuning Shared Task Models (Czech-to-English)	English | Czech
WMT16 Tuning Shared Task Models (English-to-Czech)	English | Czech
WMT17 De-En APE Shared Task Data	German | English
WMT17 En-De APE Shared Task Data	German | English
WMT17 Quality Estimation Shared Task Training and Development Data	German | English
WMT17 Quality Estimation Shared Test Data	German | English
WMT18 APE Shared Task: En-DE NMT Train and Dev Data	German | English
WMT18 Quality Estimation Shared Task Test Data	German | English | Latvian | Czech
WMT18 Quality Estimation Shared Task Training and Development Data	German | English | Latvian | Czech
Word representations for multiple languages	English | Latin | Hungarian | German | Spanish | Czech
WordSim353-cs: Evaluation Dataset for Lexical Similarity and Relatedness, based on WordSim353	English | Czech

Table 9: LRs from the LINDAT-CLARIAH-CZ repository

A.E. LREC Shared LRs (71 LRs)

This annex provides an overview of the LRs ingested into ELG from the LREC Shared LRs.

Resource name	Language(s)
A Tweet Dataset Annotated in Four Emotion Dimensions	English
AcadOnto	English
Adimen-SUMO v2.6	English
Amazigh annotated corpus with POS tags	Standard Moroccan Tamazight
An Amazigh lexicon with POS tags	Standard Moroccan Tamazight
Arabic Dialects Dataset	Arabic
Arabic Sentiment Lexicon	Arabic | Egyptian Arabic
AuCoPro - Splitting	Dutch | Afrikaans
Bengali Lemmatization Dataset	Bengali
Classical Chinese Evaluation Dataset (Twenty-Five Histories)	Chinese
CLIPv2.1	Portuguese
Collected Movie Reviews in Serbian	Serbian
CoreSC/ART corpus	English
Cornell eRulemaking Corpus -- CDCP	English
CorpusDRF	French
Database of Lexical Simplification Errors	English
Dataset of Nuanced Assertions on Controversial Issues (NAoCI dataset)	English
Datasets for classification experiments IS-pros	English
Datasets for modernising historical Basque words	Basque
Diachronic Ontologies from People's Daily	Chinese
DiLAF African languages-French dictionaries	French | Songhai languages | Hausa | Kanuri | Tamashek
Dot type corpus with experts	English
Dutch sentiment lexicon	Dutch
Emotion Movie Transcript Corpus	English
English-Malayalam-MorphGenerated Forms	English | Malayalam
EN-Hi : Humor Detection Code-mixed texts	English | Hindi
filesRo.zip	French | Turkish | Portuguese | Italian | Spanish | Romanian
German Lexemes Annotated with Senses and Substitutions (GLASS)	German
GOST Lexicon	English
HAI Alice-corpus	English
HiEve	English
Hungarian Verb Clusters	Hungarian
IITP:ReviewSentimentDataset	Hindi
Katakana-English Scientific Terms Lexicon	Japanese | English
Khresmoi Query Translation Test Data for the Medical Domain version 1.0	German | French | English | Czech
Konstanz Resource of Questions (KRoQ)	Modern Greek | German | French | Spanish
Korean L2 Unknown Words - Labeled Dataset	Korean
Lexical Resources for English- Malayalam	English | Malayalam
LIdioms	Russian | English | Italian | German | Portuguese
LuxId	Luxembourgish

MCDTB	Chinese
Metaphors for Economic Inequality in English, Farsi, Spanish, and Russian	Persian | Spanish | English | Russian
metaTED	English
MotaMot French-Khmer Pivot Database	French | Khmer
MSimlex999_Polish	Polish
Multilingual corpora with coreferential annotation of person entities	Portuguese | Galician | Spanish
NL2KB	English
NoSta-D: German NER Dataset Train/Dev	German
OSS Online Communication Messages	English
OSSMETER Threaded Corpus	English
PACE Corpus	German | English
Package for Consistent Evaluation of Morphological Segmentation	Finnish
POS Tagged Dialectal Arabic Data	Arabic
Possession identification in blogs	English
Priberam Compressive Summarization Corpus	Portuguese
Relational Noun Lexicon	English
RTE+SR 1-7	English
Sclera2cornetto	Dutch
Semantic verb classes for English, Polish, and Croatian	English | Polish | Croatian
SemRelData	German | English | Russian
SerbMR-2C	Serbian
SerbMR-3C	Serbian
Swedish Literary Corpus	Swedish
TAP-DLND 1.0 : A Corpus for Document Level Novelty Detection	English
The Epic Epigraph Graph	English
The N2 Corpus	English
Translation errors	Portuguese
TSix	English
Urdu Summary Corpus	Urdu
Webis-CLS-10	English | French
wiki_zh_ja_corpus	Japanese | Chinese

Table 10: LRs from the ELRA-SHARE-LRs

A.F. Zenodo (73 LRs)

This annex provides an overview of the LRs ingested into ELG from Zenodo.

Resource name	Language(s)
A Computational Theory for the Emergence of Grammatical Categories in Cortical Dynamics	English
A Corpus of Modern Burmese	Burmese
A Kam Lexicon	English | Kam
A part-of-speech (POS) lexicon of Classical Tibetan for NLP	Tibetan
Australian Federal Legislation - Principal acts in force	English
BioPortal Snapshot 30.03.2017	English
BNa13EN13	English
Bootstrapped Lexicon of English Verbal Polarity Shifters	English

Bootstrapped Lexicon of German Verbal Polarity Shifters	German
BuzzFeed-Webis Fake News Corpus 2016	English
cante100 Metadata	Spanish
cante2midi Metadata	Spanish
canteFAN Metadata	Spanish
Catalan United Nations v1.0 test set	Catalan
Character mentions in the German novel "Corpus Delicti" by Juli Zeh and annotations	English
Corpus Linguistic Analysis of the BMSatire Descriptions corpus	English
corpusCOFLA Metadata	Spanish
COVID-19 Open Research Dataset (CORD-19)	English
Dataset de Lectura distante y visualizaciГ³n de textos en ArqueologГa y disciplinas afines	Spanish
Datasets from `Discovering and analysing lexical variation in social media text'	English
Duhumbi Lexicon	Chug
Duhumbi Phonology - Syllabic and Word Stress	Chug
Fear in Akkadian Texts	Akkadian
French Word Sense Disambiguation with Princeton WordNet Identifiers	English | French
FTa13EN11	English
Gold standard corpus, ontologies, and Entity-Quality ontology annotations for evolutionary phenotypes	English
Greek Old Testament Toponyms	Modern Greek (1453-)
Hansard Speeches and Sentiment V1.0	English
Hansard Speeches and Sentiment V1.0.1	English
Hypernym-LIBre: A free Web-based Corpus for Hypernym Detection [PoS and Dep-parsed Version]	English
Hypernym-LIBre: A free Web-based corpus from Hypernym Detection [Hearst Pattern extractions from Hypernym-LIBre]	English
Hypertension - Florida Annotated Corpus for Translational Science (FACTS), Vital Sign Ontology Annotations	English
Inflected lexicon of Russian Nouns in IPA notation	Russian
Italian Lexical Simplification Benchmark	Italian
IT-VaLex: Index Thomisticus Valency Lexicon	Latin
JULIELab/MEmoLon (Data)	Multiple languages
Komnzo lexicon	Uncoded languages | English
Korean ideophonic reduplicatives sorted by their vowel harmony patterns	Korean
Lexicon	Modern Greek | English
Lexicon of English Verbal Polarity Shifters	English
Lexicon of Place Names in the Alsatian Dialects	French | Swiss German
Lexicon of Polarity Shifting Directions	English
Lexicon_V2	Modern Greek | English
Linguatec Tolosa Treebank	Occitan (post 1500)
Lubrang Brokpa Lexicon - basic word list	Brokpake
MeSCCon - Medical Spanish Chemical compound, drug and medication Name Lexicon (unfiltered version)	Spanish
MeSDiCon - Medical Spanish Disease and symptom name Collection lexicon (unfiltered initial version)	Spanish
Middle Dutch syllabified words	Dutch
New Testament Toponyms	Modern Greek | English
Non-Homeric hexameter	Modern Greek

OntoCorp	English
Pere lexicon	Mbre
Pere lexicon of flora and fauna	Mbre
Polarity Shifter Resources	German | English
Polynesian Segmented Data	Multiple languages
Prosopographical Database of Judeans in Babylonia (outside Yahudu and the Murašû Archive)	Judeo-Persian
Romance Verbal Inflection Dataset 2.0	Romance languages
Salvaging the Internet Hate Machine: Using the discourse of extremist online subcultures to identify emergent extreme speech	English
Sentiment analysis of tech media articles using VADER package and co-occurrence analysis	English
Sentiment analysis of tech media articles using VADER package and co-occurrence analysis (01.2016-04.2019)	English
Sharvard Corpus	Spanish
Shenzhen address corpus (part)	Chinese
SIMPITIKI corpus for simplification in Italian	Italian
Simple Italian sentences ranked by readability	Italian
SpokenChichewaCorpus	English | Nyanja
STa13EN11	English
The Consonant Challenge Corpus	English
The Rodrigo corpus	English
Twitter historical dataset: March 21, 2006 (first tweet) to July 31, 2009 (3 years, 1.5 billion tweets)	Japanese | French | Portuguese | English | German | Spanish
VnEmoLex: A Vietnamese emotion lexicon for sentiment intensity analysis	Vietnamese
Webis Comparative Web Search Questions Corpus 2020	English
Words or terms? Sanskrit annotated dataset	Sanskrit
XML_corpus	German

Table 11: LRs from Zenodo

A.G. Conceptual Mapping Between LINDAT-CLARIAH-CZ and ELG

This annex provides the conceptual mapping tables used in the conversion of LINDAT-CLARIAH-CZ's metadata into ELG's metadata (mappings from LINDAT-CLARIAH-CZ into ELG and vice-versa).

Mapping from ELG to LINDAT-CLARIAH-CZ

MetatadataRecord

Element	data type	CV	Attribute	data type of attribute	CV	Mandatoriness	Cardinality	LINDAT	comment for LINDAT
metadataRecordIdentifier	string		metadataRecordIdentifierScheme	CV	doi; handle; ...	M	1		pre-fixed value that we will overwrite
metadataCreationDate	date					M	1		the first day the record will be imported into ELG
metadataLastDateUpdated	date					M	1		for metadata records that are changed
metadataCurator	person					M	1		someone from LINDAT?
compliesWith	fixed	ELG-SHARE				M	1		

sourceOf-Metada-taRecord	string			M when applicable	1	CLARIN/LINDAT or whatever they pre-fer
sourceMetad ataRecord	metada-taRecord			M when applicable	1	the handle id

LRCommon

Element	data type	CV	Attribute	data type of attribute	CV	Mandatori-ness	Cardi-nality	LINDAT	comment for LINDAT
entityType	fixed	Languag-eRe-source				M	1		
resource-Name	multilin-gual					M	1m	title	
resource-ShortName	multilin-gual					R	1m		
description	multilin-gual					M	1m	descrip-tion	
LRIdentifier	string		LRIdentifi-erScheme	CV	doi; handle; ...	R when ap-plicable	multi-ple		LINDAT has only one identifier; better to use this for metadata rec-ord identifier (with the relation hasMetadata)
logo	URI					R	1		
version	string					M	1		undefined?
additional-Info.landingP age	landingP age (URI) / email (string)					M	multi-ple	con-tactPer-son	map to addition-alInfo.email but also to con-tactPerson; add also metada-taRecord handle as additional-Info.landingPage?
keyword	multilin-gual					M	multi-ple	subject	only subject and this is not manda-tory;
domain	multilin-gual					R	multi-ple		
subject	multilin-gual					R	multi-ple	subject	
resourcePro-vider	person / organiza-tion /group					R	multi-ple	publisher	
publica-tionDate	date					R	1	date is-sued	
resourceCre-ator	person / organiza-tion /group					R	multi-ple	author	resourceCrea-tor/person always
fundingPro-ject	project					R when ap-plicable	multi-ple	fund-ingPro-ject	funding org, code, project name, type (eu, own, na-tional, other), openaire id (on eu funding; where applicable)
intendedAp-plication	CV		LTarea (OMTD)			R	multi-ple		
isDocument-edBy	docu-ment					R	multi-ple		
isDocument-edBy	docu-ment					R	multi-ple	isRefer-encedBy	url but we also need a title
replaces	related LR					R	multi-ple	replaces	handle id; we also need the title

LRSubclass	ToolService / Corpus / LCR / LD					M	1		

Corpus

Element	data type	CV	Attribute	data type of attribute	CV	Mandatoriness	Cardinality	LINDAT	comment for LINDAT
lrType	fixed	corpus				M	1		
corpusSubclass	CV	raw; annotated; annotations				M	1		undefined? or guess if they have some information on annotation?
CorpusMediaPart	CorpusTextPart / CorpusAudioPart / CorpusVideoPart / CorpusImagePart / CorpusTextNumericalPart					M	multiple		to be automatically created from mediatype
datasetDistribution	DatasetDistribution					M	multiple		to be automatically created
personalDataIncluded	boolean					M	1		no by default for open licences?
sensitiveDataIncluded	boolean					M	1		no by default for open licences?
anonymized	boolean					M when applicable	1		

CorpusMediaPart

	Element	data type	CV	Mandatoriness	Cardinality	LINDAT	comment for LINDAT
CorpusTextPart	mediaType	fixed	text	M	1	mediaType	
	lingualityType	CV	monolingual; bilingual; multilingual	M	1		can be computed
	multilingualityType	CV	parallel; comparable; ...	M when applicable	1		undefined?
	language	CV	bcp47	M	multiple	language	map from BCP47
	languageVariety	string		R	multiple		
	modalityType	CV	spokenLanguage; bodyGesture; ...	R when applicable	multiple		
	textType	string (+ id + scheme)		R when applicable	multiple		
	TextGenre	string (+ id + scheme)		R when applicable	multiple		
CorpusAudioPart	mediaType	fixed	audio	M	1	mediaType	
	lingualityType	CV	monolingual; bilingual; multilingual	M	1		can be computed
	multilingualityType	CV	parallel; comparable; ...	M when applicable	1		undefined?
	language	CV	bcp47	M	multiple	language	map from BCP47
	languageVariety	string		R	multiple		
	modalityType	CV	spokenLanguage;	R when applicable	multiple		

			bodyGesture; ...				
	AudioGenre	string (+ id + scheme)		R when applicable	multiple		
	SpeechGenre	string (+ id + scheme)		R when applicable	multiple		
CorpusVideoPart	mediaType	fixed	video	M	1	mediaType	
	lingualityType	CV	monolingual; bilingual; multilingual	M	1		can be computed
	multilingualityType	CV	parallel; comparable; ...	M when applicable	1		undefined?
	language	CV	bcp47	M	multiple	language	
	languageVariety	string		R	multiple		
	modalityType	CV	spokenLanguage; bodyGesture; ...	R when applicable	multiple		
	VideoGenre	string (+ id + scheme)		R when applicable	multiple		
	typeOfVideoContent	multilingual		M	multiple		undefined?
CorpusImagePart	mediaType	fixed	image	M	1	mediaType	
	lingualityType	CV	monolingual; bilingual; multilingual	M	1		can be computed
	multilingualityType	CV	parallel; comparable; ...	M when applicable	1		undefined?
	language	CV	bcp47	M	multiple	language	
	languageVariety	string		R	multiple		
	modalityType	CV	spokenLanguage; bodyGesture; ...	R when applicable	multiple		
	ImageGenre	string (+ id + scheme)		R when applicable	multiple		
	typeOfImageContent	multilingual		M	multiple		undefined?

Dataset Distribution

	Element	data type	CV	Mandatoriness	Cardinality	LINDAT	comment for LINDAT		Element	data type
CorpusTextPart	mediaType	fixed	text	M	1	mediaType		CorpusTextPart	mediaType	fixed
	lingualityType	CV	monolingual; bilingual; multilingual	M	1		can be computed		lingualityType	CV
	multilingualityType	CV	parallel; comparable; ...	M when applicable	1		undefined?		multilingualityType	CV
	language	CV	bcp47	M	multiple	language	map from BCP47		language	CV
	languageVariety	string		R	multiple				languageVariety	string
	modalityType	CV	spokenLanguage; bodyGesture; ...	R when applicable	multiple				modalityType	CV
	textType	string (+ id + scheme)		R when applicable	multiple				textType	string (+ id + scheme)

	TextGenre	string (+ id + scheme)		R when applicable	multiple				TextGenre	string (+ id + scheme)
CorpusAudioPart	mediaType	fixed	audio	M	1	mediaType		CorpusAudioPart	mediaType	fixed
	lingualityType	CV	monolingual; bilingual; multilingual	M	1		can be computed		lingualityType	CV
	multilingualityType	CV	parallel; comparable; ...	M when applicable	1		undefined?		multilingualityType	CV
	language	CV	bcp47	M	multiple	language	map from BCP47		language	CV
	languageVariety	string		R	multiple				languageVariety	string
	modalityType	CV	spokenLanguage; bodyGesture; ...	R when applicable	multiple				modalityType	CV
	AudioGenre	string (+ id + scheme)		R when applicable	multiple				AudioGenre	string (+ id + scheme)
	SpeechGenre	string (+ id + scheme)		R when applicable	multiple				SpeechGenre	string (+ id + scheme)
CorpusVideoPart	mediaType	fixed	video	M	1	mediaType		CorpusVideoPart	mediaType	fixed
	lingualityType	CV	monolingual; bilingual; multilingual	M	1		can be computed		lingualityType	CV
	multilingualityType	CV	parallel; comparable; ...	M when applicable	1		undefined?		multilingualityType	CV
	language	CV	bcp47	M	multiple	language			language	CV
	languageVariety	string		R	multiple				languageVariety	string
	modalityType	CV	spokenLanguage; bodyGesture; ...	R when applicable	multiple				modalityType	CV
	VideoGenre	string (+ id + scheme)		R when applicable	multiple				VideoGenre	string (+ id + scheme)
	typeOfVideoContent	multilingual		M	multiple		undefined?		typeOfVideoContent	multilingual
CorpusImagePart	mediaType	fixed	image	M	1	mediaType		CorpusImagePart	mediaType	fixed

LexicalConceptualResource

Element	data type	CV	Attribute	data type of attribute	CV	Mandatoriness	Cardinality	LINDAT	comment for LINDAT
lrType	fixed	lexicalConceptualResource				M	1		
lcrSubclass	CV	terminology; ontology; ...				M	1	detailed type	
enconding-Level	CV	morphology; syntax; ..				M	multiple		
ContentType	CV	transcription; example; ...				R	multiple		
LexicalConceptualResource-MediaPart	LexicalConceptualResourceTextPart / LexicalConceptualResourceAudioPart / LexicalConceptualResourceVi-					M	multiple		to be automatically created from mediatype

	deoPart / Lexical-ConceptualRe-sourceImagePart								
datasetDis-tribution	DatasetDistribution					M	multiple		to be auto-matically cre-ated
person-alDataIn-cluded	boolean					M	1		
sensitive-DataIn-cluded	boolean					M	1		
anony-mized	boolean					M when applica-ble	1		
<various relations>	related LR					R	multiple		

LexicalConceptualResourMediaPart

	Element	data type	CV	Attribute	data type of attrib-ute	CV	Manda-toriness	Cardinal-ity	LINDAT	comment for LINDAT
LexicalCon-ceptualRe-sourceText Part	mediaType	fixed	text				M	1		
	linguali-tyType	CV	monolingual; bi-lingual; multilin-gual				M	1		can be com-puted
	multilin-gualityType	CV	parallel; compa-rable; ...				R when applica-ble	1		
	language	CV	bcp47				M	multiple		
	language-Variety	string					R	multiple		
	metalan-guage	CV	bcp47				M	multiple		
	modali-tyType	CV	spoken-Language; bodyGesture; ...				R when applica-ble	multiple		
LexicalCon-ceptualRe-sourceAudi oPart	mediaType	fixed	audio				M	1		
	linguali-tyType	CV	monolingual; bi-lingual; multilin-gual				M	1		can be com-puted
	multilin-gualityType	CV	parallel; compa-rable; ...				M when applica-ble	1		
	language	CV	bcp47				M	multiple		
	language-Variety	string					R	multiple		
	metalan-guage	CV	bcp47				M	multiple		
	modali-tyType	CV	spoken-Language; bodyGesture; ...				R when applica-ble	multiple		
LexicalCon-ceptualRe-sourceVi-deoPart	mediaType	fixed	video				M	1		
	linguali-tyType	CV	monolingual; bi-lingual; multilin-gual				M	1		can be com-puted
	multilin-gualityType	CV	parallel; compa-rable; ...				M when applica-ble	1		

	language	CV	bcp47				M	multiple		
	language-Variety	string					R	multiple		
	metalan-guage	CV	bcp47				M	multiple		
	modali-tyType	CV	spoken-Language; bodyGesture; ...				R when applica-ble	multiple		
	typeOfVide oContent	multi-lingual					M	multiple		
LexicalCon-ceptualRe-sourceImag ePart	mediaType	fixed	image				M	1		
	linguali-tyType	CV	monolingual; bi-lingual; multilin-gual				M	1		can be com-puted
	multilin-gualityType	CV	parallel; compa-rable; ...				M when applica-ble	1		
	language	CV	bcp47				M	multiple		
	language-Variety	string					R	multiple		
	metalan-guage	CV	bcp47				M	multiple		
	modali-tyType	CV	spoken-Language; bodyGesture; ...				R when applica-ble	multiple		
	typeOfImag eContent	multi-lingual					M	multiple		

ToolService

Element	data type	CV	Mandatori-ness	Cardinality	LINDAT	comment for LINDAT
lrType	fixed	toolService	M	1	type	
function	CV / string	function (OMTD)	M	multiple		
SoftwareDistri-bution	SoftwareDistribu-tion		M	multiple		to be automatically created
languageDe-pendent	boolean		M	1	language dependent	
inputConten-tResource	pro-cessingResource		M	multiple		
outputResource	pro-cessingResource		R when appli-cable	multiple		
formalism	multilingual		R	1m		
method	CV	method (OMTD)	R	1		
implementa-tionLanguage	string		R	1		
trl	CV	trl1; trl2; ...	R	1		
evaluated	boolean		M	1		
parameter	parameter		R when appli-cable	multiple		

SoftwareDistribution

Element	data type	CV	Attribute	data type of attribute	CV	Mandatori-ness	Cardi-nality	LINDAT	comment for LINDAT
SoftwareDistribu-tionForm	CV	down-loadable; dockerImage; ...				M	1		
executionLoca-tion	URI					M when ap-plicable	1		
downloadLoca-tion	URI					M when ap-plicable	1		
dockerDownload-Location	URI					M when ap-plicable	1		

serviceAdapterDownloadLocation	URI					M when applicable	1		
accessLocation	URI					M when applicable	1		
demoLocation	URI					M when applicable	1	demoURI	
isDescribedBy	document					R	1		
additionalHwRequirements	string					R when applicable	1		
webServiceType	CV	REST; ...				R when applicable	1		
licenceTerms	Licence					M	multiple	licenseURL	we need to map to our values & see what we'll do with those we don't have
cost.amount & cost.currency	float & CV	euro; ...				R when applicable	1		

ProcessingResource

Element	data type	CV	Attribute	data type of attribute	CV	Mandatoriness	Cardinality	LINDAT	comment for LINDAT
processingResourceType	CV					M	1		
samplesLocation	URI					R	1		
language	CV					M when applicable	multiple	languageISO	map iso-3 to bcp47
mediaType	CV	text; audio; ...				R when applicable	1		
dataFormat	CV	dataformat (OMTD)				R when applicable	multiple		
characterEncoding	CV	utf-8; ...				R when applicable	multiple		
annotationType	CV	annotationType (OMTD)				R when applicable	multiple		

Mapping from LINDAT-CLARIAH-CZ to ELG

field_name	repeatable	required	type_bind	note	repeatable	required	ELG field	issues	issues
author	true	true	all	family name, given	y	n	resourceCreator/person/surname, givenName		
date issued	false	true	all	YYYY(-MM(-DD))	n	n	publicationDate		
description	false	true	all	Free text	n	y	description		
language iso	true	false	toolService	ISO 639; 3 letter language codes	y	uc	inputContentResource/language & outputResource/language	bcp47 - we already have the mapping from 3 to 2-letter codes	
language iso	true	true	corpus,lexicalConceptualResource,languageDescription		y	y	language	bcp47 - we already have the mapping from 3 to 2-letter codes	need to know the mediaType of the resource to add it at the right place

publisher	true	true	all	Name of organization, usually a University-faculty-department	y	n	resourceProvider/organization/organizationName		
relation isreferencedby	true	false	all	url of paper(s) talking about the resource; not very common	y	n	isDocumentedIn, isDescribedIn, isCitedBy	document (title + id, which can be a url)	decide on only one relation => isCitedBy; if only URL, what do we put on title?
relation isreplacedby	true	false	all	handle of next version	y	n	(reverse of replaces)	LR (title + id)	
relation replaces	true	false	all	handle of previous version	y	n	replaces	LR (title + id)	
source uri	false	false	all	Usually a project url;			?	this is the source for LINDAT; difference with pubisher?; if it's a project, it could go to the fundingProject	
subject	true	true	all	subject keywords	y	n	keyword		
title	false	true	all		y (multil.)	y	resourceName		
type	false	true	all	corpus/languageDescription/lexicalConceptualResource/toolService	n	y	lrType	map to new values from the ontology	problem with inconsistent use (models)
contact person	true	true	all	surname; given name; email; organization they belong to	y	y (additionalInfo)	additionalInfo.email + contact.person	we can also add metadataRecord handle as additionalInfo.landingPage	
demo uri	false	false	all	url where the resource can be explored or tried out	y	n	softwareDistribution.demoLocation; dataDistribution.demoLocation		
embargo termslift	false	false	all	End of embargo - date; very rarely used	n	n	availabilityEndDate		
size info	true	false	corpus,languageDescription,lexicalConceptualResource	Values (sizeUnits) from metashare v2; if not provided mapped to number of bitstream in the metashare mapping.	y	n	size (amount & sizeUnit) on distributionXFeature	mapping of values of sizeUnit	need to know the mediaType of the resource to add it at the right place
sponsor	true	false	all	funding org, code, project name, type (eu, own, national, other), openaire id	y	n	fundingProject	we also have full metadata records for eu projects; see how they can be mapped (code?)	

				(on eu funding; where applicable)					
detailed type	false	true	languageDescription	grammar, other	n	y	LanguageDescriptionSubclass	grammar, MLModel, NGramModel	ngram corpora have become ngram models
detailed type	false	true	lexicalConceptualResource	wordList, computationalLexicon, ontology, wordnet, thesaurus, framenet, terminilogicalResource, machineReadableDictionary, lexicon, other	n	y	lcrSubclass	map to ontology values	
detailed type	false	true	toolService	tool, service, platform, suiteOfTools, infrastructure, architecture, nlpDevelopmentEnvironment, other			-		
media type	false	true	corpus,lexicalConceptualResource	text, audio, video, image	n	y	mediaType	same values but not always used in LINDAT, e.g., https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-188?show=full	
media type	false	true	languageDescription	text, video, image	n	y	mediaType	same values	
language dependent	false	true	toolService	true/false	n	y	languageDependent		
license url	false	true	all having a bitstream	URL of license text; limited to licenses known by the repository; but users might ask for new license to be added; each license has a PUB/ACA/RES label attached to it	y	y	licenceTerms/name & URL	all licences from SPDX + licences specific to LRs (ELRA, MS) + commercial licences	resources without licences? need to know source licences for mappings to SPDX to avoid duplicates; decide what to do with new licences when added in LINDAT