



Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe

D2.6 Report on Lynx acquired corpora

PROJECT ACRONYM	Lynx
PROJECT TITLE	Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe
GRANT AGREEMENT	H2020-780602
FUNDING SCHEME	ICT-14-2017 - Innovation Action (IA)
STARTING DATE (DURATION)	01/12/2017 (36 months)
PROJECT WEBSITE	http://lynx-project.eu
COORDINATOR	Elena Montiel-Ponsoda (UPM)
RESPONSIBLE AUTHORS	Ēriks Ajausks (TILDE), Christian Sageder (openlaws), Andis Lagzdīņš (TILDE), Víctor Rodríguez-Doncel (UPM)
CONTRIBUTORS	Roberts Rozis (TILDE), Rinalds Vīksna (TILDE), Matīss Rikters (TILDE)
REVIEWERS	Víctor Rodríguez-Doncel (UPM), Maria Khvalchik (SWC)
VERSION STATUS	V1.0 Final
NATURE	Report
DISSEMINATION LEVEL	Public
DOCUMENT DOI	10.5281/zenodo.3692591
DATE	29/02/2020 (M27)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780602

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
01	1 st draft of the document	25/01/2020	Ēriks Ajausks (TILDE)
02	Contributions to 2 nd part	05/02/2020	Ēriks Ajausks (TILDE), Roberts Rozis (TILDE), Andis Lagzdīņš (TILDE), Rinalds Vīksna (TILDE), Matīss Rikters (TILDE)
03	Contributions to 1 st part	10/02/2020	Christian Sageder (openlaws), Elena Montiel-Ponsoda (UPM)
04	Contributions to 1 st part and general review	13/02/2020	Víctor Rodríguez-Doncel (UPM)
05	Review	14/02/2020	Maria Khvalchik (SWC)
06	Harvesters table	26/02/2020	Víctor Rodríguez-Doncel (UPM)
07	Review of harvesters table	27/02/2020	Christian Sageder (openlaws)
1.0	Final review	27/02/2020	Christian Sageder (openlaws), Elena Montiel-Ponsoda (UPM)

DISCLAIMER

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content. Neither the Lynx consortium as a whole, nor a certain party of the Lynx consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

ACRONYMS LIST

CKAN – Web based management system for the storage and distribution of data

CSV – Comma separated values (file format)

DOC, DOCX – Filenames extension for document files

HTML – Hypertext Markup Language

IPR – Intellectual Property Rights

JPG/JPEG – Joint Photographic (Experts) Group (compression for digital images)

JSON – Java Script Object Notation

MT – Machine Translation

NMT – Neural Machine Translation

OCR – Optical Character Recognition

PDF – Portable Document Format

RDF – Resource Description Framework

SKOS – Simple Knowledge Organization System

SMT – Statistical machine translator

TBX – TermBase eXchange

TMX – Translation Memory eXchange (file format)

TSV – Tab-separated values

TXT – A filename extension for text files

XLS, XLSX – Filenames extension for Microsoft Excel sheet format

XML - Extended Markup Language

XSL – Extensible Stylesheet Language

XSLT –Extensible Stylesheet Language Transformations (for transforming XML)

TABLE OF CONTENTS

LIST OF TABLES.....	4
LIST OF FIGURES.....	5
EXECUTIVE SUMMARY.....	6
1 INTRODUCTION.....	7
2 INDEXED LEGISLATION CORPORA.....	8
2.1 CONTRACT CORPORA.....	8
2.2 STANDARDS, RECOMMENDED PRACTICES, REGULATIONS, AND RESOLUTIONS.....	9
2.3 LABOR LAW CORPORA.....	9
2.4 GENERAL LEGAL CORPORA.....	13
2.5 CORPORA INDEXING METHOD.....	14
2.6 LEGAL KNOWLEDGE GRAPH ONTOLOGY.....	15
2.7 HARVESTED DOCUMENTS.....	16
3 CREATED TRANSLATION CORPORA.....	18
3.1 CORPORA CREATION WORKFLOW.....	18
3.1.1 General parallel corpora creation workflow.....	18
3.1.2 Corpora creation from web crawled data.....	18
3.2 LIST OF TRANSLATION CORPORA.....	20
3.3 LIST OF CORPORA FOR MACHINE TRANSLATION TRAINING.....	23
CONCLUSIONS.....	27
ANNEX 1. MAIN EU LABOR LAW LEGISLATION CORPORA.....	28
REFERENCES.....	29

LIST OF TABLES

Table 1. Legislation and case law from openlaws.com	13
Table 2. Harvested data sources.....	17
Table 2. Resources used to build corpora	22
Table 3. Segment count in LYNX_Geothermal	23
Table 4. Corpora used for EN-NL NMT systems	24
Table 5 Corpora used for EN-ES NMT systems.....	25
Table 6. Corpora used for EN-DE NMT systems	25
Table 7. Corpora used for NL-DE NMT systems.....	26

LIST OF FIGURES

Figure 1. Web scrapper for collective agreements in Spain.....	12
Figure 2. From HTML/PDF tables to TXT tables in the Lynx Document.	12
Figure 3. Collective agreements extracted in Spain, per labour law authority.....	13
Figure 4. Example of NIF annotated LynxDocument with parts.....	15
Figure 5. General language resource processing workflow used for Lynx corpora creation.....	18
Figure 6. Web-crawled data processing workflow	19
Figure 7. Parallel corpus processing workflow	19

EXECUTIVE SUMMARY

This deliverable summarizes the final work on acquired corpora (as part of WP2) within the context of the Lynx project. The aim of this task is to provide a description of the corpora collection methods, and the resulting collected corpora by Lynx partners around the different use cases. There are three business cases for each corpus. The first case is related to Compliance Assurance Services for Contracts, the second is related to Compliance Assurance Services in Oil & Gas and Energy, and the third Business Case is about Compliance Assurance Services in Labor Law. This document serves as reference material for the corpora collected to cover the needs of the three business cases, and for the first steps in the method followed to index that corpora. Furthermore, the document describes the corpora preparation workflow to be used in the training of Neural MT engines for specific languages and domains. Finally, this document reports on the term extraction process on the compiled corpora.

1 INTRODUCTION

This report summarizes the work done in Task 2.4 “Indexing of corpora” and Task 2.5 “Translation corpora creation” in WP2 and is a follow up of D2.3 ‘Intermediate report on Lynx acquired corpora’. T2.4 includes work done by openlaws, UPM, SWC, and DFKI and is scoped to provide a compilation of data to cover the needs of the different Lynx business cases. In the respective section “Indexed legislation corpora”, partners provide information about the corpora gathered throughout the project. The report summarizes the work done on data compilation and preparation. The detailed description and corpora interlinkage with services was included in deliverable D3.4.

The second part of the report focuses on translation corpora creation, with specific focus on the identification of language resources in public online sources and data repositories in the legal domain, including parallel data, monolingual data, glossaries, etc. It provides the workflow of web-crawling, automated extraction, and other data collection methods applied for the Lynx business cases. A significant part of the report consists of the description of language resource processing (i.e. aligning, and re-formatting), which is a crucial part of translation corpora creation for developing MT engines.

2 INDEXED LEGISLATION CORPORA

In this section we describe the process of acquisition and indexing of textual resources to support the three business cases defined in the project, that is, Business Case 1 “Compliance Assurance Services for Contracts”, Business Case 2 “Compliance Assurance Services in Oil & Gas and Energy”, and Business Case 3 “Compliance Assurance Services in Labor Law”.

By *indexing* of textual resources we mean the curation, storage and organization of the acquired textual resources. Importantly, this requires that documents are transformed to the agreed Lynx document format and that they be made available to the services developed in WP3. These descriptors are listed in the data-models of the Lynx project¹, and legal-specific standards such as ELI. We note that this definition of indexing implies (and goes beyond) the one used in the term “document indexer” (e.g. Elasticsearch, Solr), in that documents are organized in a way that they can be efficiently accessed.

In the following subsections we describe which corpora were acquired by several partners to cover the needs specified by the business cases, and the type and number of documents compiled. The format of the original data is also specified, and, whenever performed. After that, we also describe the general legislation corpora acquired directly through the openlaws.com platform for different legislation. Finally, we refer to the method followed to index the compiled resources.

2.1 CONTRACT CORPORA

The contract corpora collected in Lynx are intended to fulfil the needs of Business Case 1 “Compliance Assurance Services for Contracts”. For this use case only the Austrian and European Legal corpora (legislation and case law) is harvested. The pilot showcase will analyze contracts and extract the necessary information from them in order to be compliant with the present legislation in Austria.

The legal corpora is a subset of the National and Federal law and the Jurisdiction of Austria, mainly legislation which is below the index number “2 ZIVIL- UND STRAFRECHT” (2 CIVIL AND CRIMINAL LAW), and here the following index numbers.

- 20 PRIVATRECHT ALLGEMEIN (PRIVATE LAW GENERAL)
- 21 HANDELS- UND WERTPAPIERRECHT (COMMERCIAL AND SECURITIES LAW)
- 26 GEWERBLICHER RECHTSSCHUTZ (INDUSTRIAL PROPERTY RIGHTS)

For the case law all decisions were taken which refer to legislation within this index number. This possible as this information is already available within openlaws.com.

The legislation / case law is available in openlaws.com internal document format and is converted into the Lynx defined format by a tool.

The format for contracts is mainly PDF, either already from an electronic version of the contract or a scan and OCR of the contract or MS-Word. Harvesting incl. OCR of contracts is done within the Business Case 1 and out of the scope of WP2. Only a small sample set of contract template is provided for testing and training purposes of the WP3 services. The contracts are passed from openlaws.com to Lynx for analysis, but are not stored within Lynx.

¹<http://lynx-project.eu/data2/data-models>

2.2 STANDARDS, RECOMMENDED PRACTICES, REGULATIONS, AND RESOLUTIONS

In Business Case 2, Compliance Assurance Services in Oil & Gas and Energy, led by DNV-GL, the assembled corpus consists of four types of documents: (a) Standards (b) Recommended Practices, (c) Regulations, and (d) Government resolutions. Documents of types (a) and (b) are produced by DNV-GL experts, based on several decades of experience, and have been pre-categorized, being only the “Energy” and “Oil and Gas” categories relevant for this project. Documents of type (c) are published by the EU Publications Office (as part of EurLex) and local official legislation publishers (in this case, we consider the Dutch Mining Act). Documents of type (d) are published by the relevant local authorities, in this case, the Dutch Ministry of Economic Affairs and Climate Policy.

Standards and Recommended Practices are accessible from the DNV-GL website², all in English language, and they are produced by the different departments of DNV-GL. In terms of Regulations, Directives 2009/28/EC and 2007/2/EC have been considered, as well as the Dutch mining act³. Resolutions are harvested from the Dutch Ministry of Economic Affairs and Climate Policy⁴. The listed documents all come as PDF files, in a variety of layouts and types of content (text, pictures, mathematical formulae, diagrams, and tables), which makes parsing text a daunting task.

For sources where only a small number of documents are available and whose document format differs between the single documents were manually transformed into the Lynx format. In a further step, a GUI tool will be made available to support this manual transformation.

2.3 LABOR LAW CORPORA

This section describes the corpora acquired in relation to Labor Law. The Labor Law corpora collected in Lynx are intended to fulfil the needs of Business Case 3, Compliance Assurance Services in Labor Law. The resulting pilot will showcase the access to interlinked relevant legal information in the labor law sector across multiple orders, jurisdictions, and languages. CUATRECASAS, the Spanish law firm participating in this project, leads the development of this pilot.

According to CUATRECASAS, the main types of documents in the labor law domain are (i) legislation at EU level and Member State level; (ii) case law (judgements related to labor law in the different jurisdictions; (iii) collective bargaining agreements (official documents, agreed upon between unions and business, that determine the conditions of work for a specific sector and function at the same level as ordinary laws) and (iv) employment contracts (standard binding contracts between workers and companies), and customs (other legal documents with special features).

To accomplish the purpose of accessing labor law information across jurisdictions, we created the following collections, as explained below:

- EU labor law legislation and case law
- Spanish labor law legislation, case law and collective bargaining agreements
- Austrian labor law legislation, case law and collective bargaining agreements
- German labor law legislation

The choice of these jurisdiction has been made considering the languages of the project and the interest of CUATRECASAS.

²<https://rules.dnvgl.com/servicedocuments/dnvgl/#!/home>

³<https://zoek.officielebekendmakingen.nl/stb-2002-542.html>

⁴<https://www.sodm.nl>

EU labor law legislation and case law

EU legislation, as described in section 2.4 ‘General legal corpora’ is taken directly from openlaws.com. Within openlaws.com, a set of regulations has been identified as relevant for labor law and is publicly available⁵. The list of main EU labor law legislation and case law is available in Annex 1. In addition, more than 100 decisions of the Court of Justice have been identified, which are also available in the collection of openlaws.com.

Austrian labor law legislation, case law, and collective bargaining agreements

Austrian legislation related to labor law also has been taken from openlaws.com. In total, more than 200 Regulations have been identified as relevant for labor law.

Austrian collective bargaining agreements have been collected from the website of the Austrian Chamber of Commerce (Wirtschaftskammern Österreichs), which organizes them into seven different sectors:

- Commerce and Craftworks
- Commerce
- Industry
- Information and Consulting
- Tourism and Entertainment
- Transport and Traffic
- Bank and Insurance

From each of these sectors, we have selected the most representative topics that contain the desired information. Thus, we have discarded outdated documents and other types of data that are not required at this phase of the project: salary tables, summaries, and additional information. As a result, a corpus of 50 PDF files in German language has been created.

Spanish labor law legislation, case law and collective bargaining agreements

General information

Spanish labor law is published, as any other piece of legislation, by the official gazetteer BOE (Boletín Oficial del Estado). Its harvesting is described in the next section. Case law is not offered or free.

Collective bargaining agreements are documents signed between the employer and representatives of the employees. They are the result of an extensive negotiation process between the parties regarding topics such as wages, hours, and terms and conditions of employment.

There are different types of agreements in Spain, besides the most important norm, which the «Estatuto»:

- «Estatuto de los trabajadores». This is the most general norm (last version, 2015), and it prevails when no other rule is determined by «convenios» below.
- «Convenios sectoriales estatales y nacionales». Negotiated by trade unions and companies representing a sector, they have a nationwide and sectoral scope. They are published by BOE.
- «Convenios sectoriales autonómicos.». As before, but with a scope limited to an autonomous community.
- «Convenios sectoriales provinciales.». Sectoral agreements with a province as scope, but this time negotiated directly by representatives of the employees.
- «Convenios sectoriales interprovinciales.». Signed by trade unions, they affect several provinces.
- «Convenios sectoriales locales o comarcales.». They have the minimum geographical scope.

⁵ <https://openlaws.com/public-folder-categories/4c44ce46-6ebf-4ef0-aeef-29f0e1427b48>

- «Convenios colectivos de empresa». The scope is limited to one company (with more than 6 employees).

As of 2020, there are about 4500 «company collective agreements» (affecting ~1M employees), about state 1500 «sectoral collective agreements» (affecting ~10M employees), the rest being another 1500 collective agreements affecting another 10M employees.

Collective bargaining agreements are published by different authorities at national, regional (Autonomous Communities) and province level, depending on their geographical scope –a total of 58 authorities publishing agreements. Besides the official gazettes, the Spanish Ministry for Work publishes a website⁶ which provides easier access to the documents –it is a web portal for search, but it does not host the document themselves. This website is known as REGCON, Register for Collective Agreements from the Ministry of Labor, Migration, and Social Affairs.

Thus, there is not a single format nor a single publisher (but 58) and the proper ingestion (structured and with good metadata) of these documents is expensive and error prone if not done carefully. Yet, a bulk extraction was made with the sole intention of feeding terminology extraction services. The total number of documents harvested in Spanish language sum up 6888 (those in Catalan, Basque and Galician) were discarded for NLP purposes.

Technical description of the harvesters

Dedicated harvesters were put in place to collect and adequately format documents into the Lynx document structure. The harvesters were integrated into one application, whose workflow is defined in the next figure. This application is described in detail in Bautista (2020).

In the first step a search is made with a very simple HTTP message.

```
curl -X POST https://expinterweb.empleo.gob.es/regcon/pub/consultaPublicaEstatat -d autoridadLaboral=1
```

The HTML is then analyzed and links are followed to extract, page by a page. The exploration of this pages is relatively costful, as cookies have to be properly handled and the system has to be deceived. The output of this task is a full list of document links (links whose domain always falls within the 58 authorities), which becomes thus a clean result (see 2nd step in the figure below) stored as a CSV file.

In the third step, iterates the list obtained in the step before, and dedicated parsers intervene to collect, parse and transform documents from their various formats into the one of LynxDocument. Metadata is also collected in this step, and slightly enriched: the raw NACE (Eurostat, 2009) code (CNAE in Spanish) is looked up and transformed (from number to textual description).

⁶ <https://expinterweb.empleo.gob.es/regcon/pub/consultaPublicaEstatat>

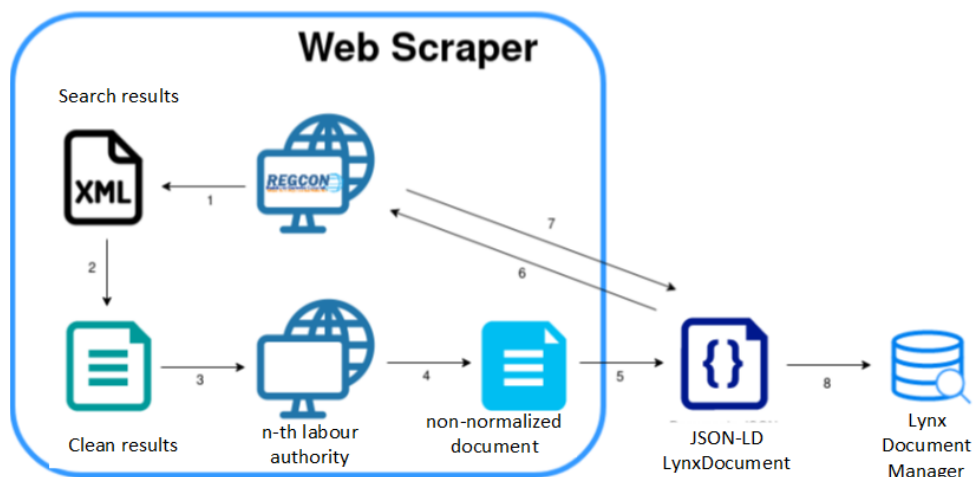


Figure 1. Web scraper for collective agreements in Spain

Whereas adapting the metadata is a simple task, processing the text and trying to maintain the structure is a costly process. In this project, a first-level structuring approach was followed, and an attempt to maintain information in tables (very frequent in these documents) was made. An example of strings searched in the plain-text only cases follows:

```

{"Capítulo", "Capítol", "CAPÍTOL", "CAPITOL", "CAPÍTULO", "CAPITULO", "Artículo", "Article", "ARTICLE", "ARTICULO",
"ARTÍCULO", "Art.", "ART.", "Anexo", "Annex", "ANEXO", "ANNEX", "Sección"};

```

An example of transformed table is shown in the next figure.

Subgrupo	Sueldo base		C. grupo		Total	
	Mensual	Anual	Mensual	Anual	Mensual	Anual
A1	1.957,20	27.400,80	248,85	2.986,20	2.206,05	30.387,00
A2	1.816,01	25.424,14	248,85	2.986,20	2.064,86	28.410,34
A3	1.698,37	23.777,18	248,85	2.986,20	1.947,22	26.763,38
B1	1.644,89	23.028,46	221,94	2.663,28	1.866,83	25.691,74
C1	1.417,16	19.840,24	158,01	1.896,12	1.575,17	21.736,36
C2	1.290,34	18.064,76	158,01	1.896,12	1.448,35	19.960,88
D1	1.193,31	16.706,34	129,85	1.558,20	1.323,16	18.264,54
D2	1.153,25	16.145,50	129,85	1.558,20	1.283,10	17.703,70
E	1.102,97	15.441,58	126,98	1.523,76	1.229,95	16.965,34

Subgrupo	Sueldo base		C. grupo		Total	
	Mensual	Anual	Mensual	Anual	Mensual	Anual
A1	1.957,20	27.400,80	248,85	2.986,20	2.206,05	30.387,00
A2	1.816,01	25.424,14	248,85	2.986,20	2.064,86	28.410,34
A3	1.698,37	23.777,18	248,85	2.986,20	1.947,22	26.763,38
B1	1.644,89	23.028,46	221,94	2.663,28	1.866,83	25.691,74
C1	1.417,16	19.840,24	158,01	1.896,12	1.575,17	21.736,36
C2	1.290,34	18.064,76	158,01	1.896,12	1.448,35	19.960,88
D1	1.193,31	16.706,34	129,85	1.558,20	1.323,16	18.264,54
D2	1.153,25	16.145,50	129,85	1.558,20	1.283,10	17.703,70
E	1.102,97	15.441,58	126,98	1.523,76	1.229,95	16.965,34

Figure 2. From HTML/PDF tables to TXT tables in the Lynx Document.

In the last step, documents are ingested in the document manager. The total number of documents fully structured in the LynxDocument structure follows.

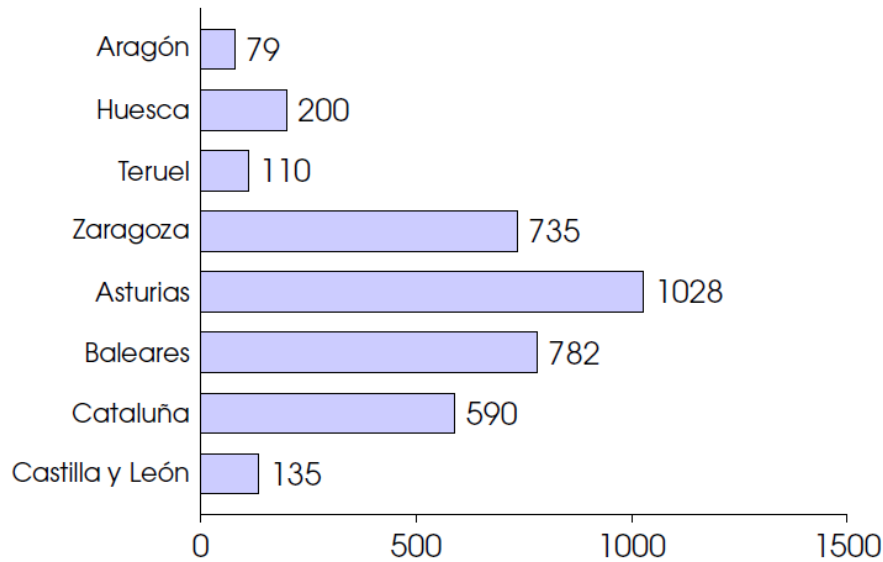


Figure 3. Collective agreements extracted in Spain, per labour law authority

2.4 GENERAL LEGAL CORPORA

EU, Austria and Germany legislation and case law

General legal corpora from the EU, Austria and Germany have been made available directly from the openlaws.com platform. The openlaws platform is a SaaS which helps to find legal information more easily, organizes it according to the user desires and shares it with others. The user can create a network of legislation, case law, legal literature and legal experts – both on a national and a European level. Table 1 summarizes the legislation and case law compiled for such jurisdiction.

	Legislation	Case law
Austria	Federal law National law	Constitutional Court (VfGH) Administrative High Court (VwGH) Supreme Court (OGH) Higher Regional Courts (OLG) District Courts (LG) Obersten Patent- und Markensenats (OPMS) Federal Administrative Court (BVwG) National Administrative Court (LVwG)
EURLex	Sector 0: Consolidated texts R - Regulations L - Directives	Sector 6: EU case law CJ: Court of Justice - Judgment TJ: General Court - Judgment
Germany	Federal law	Not available

Table 1. Legislation and case law from openlaws.com

The corpora in openlaws.com are already split into text fragments at the article level for legislation, even if the target source, e.g., EUR-Lex, does not provide this granularity level. This is necessary for further annotations and linking. In addition, these fragments are already linked to other text fragments and/or decisions at the national and European level. This information is either extracted by a human or during the load process from the original source to openlaws.com.

openlaws's website also has historic and future data, but this is not used within Lynx. For the time being, Lynx only works with the current version (as of Spring 2020).

Spanish legislation and case law

In Spain, legislation is published in the Official Gazette website or BOE⁷. These documents are harvested – directly from the website with a web scrapper whose code is in the Lynx software repos⁸. The last update of the harvesters dates back as of January 2020.

The documents are available in XML and PDF formats. Within the official website, there is a section⁹ devoted to labor law which has been used to guide the downloading them. Only documents related to the labor law business case have been harvested.

Case law is published in Spain by CENDOJ¹⁰ (*Centro de Documentación Judicial del Consejo General del Poder Judicial*), the National Center for Judicial Documentation of Spain. Documents cannot be downloaded in bulk, and harvesting is not possible due to legal restrictions. Here the search was restricted to case law issued by the Spanish Supreme Court, using keywords such as “laboral” (labor) and “empleo” (employment).

Due to the legal restrictions, a small set of documents was downloaded for research purposes (e.g. characterization of documents, design of data models) but not published

2.5 CORPORA INDEXING METHOD

This section offers a summary of the methods followed to index the corpora compiled for the three business cases defined in Lynx. In order to process the compiled textual resources using the services that make up the Lynx Platform and to extract knowledge that will aid in the first steps of the compliance process for a new project, documents have to be put into an agreed standardized format. This allows for all services to have access to it and for all facilities for indexing and rendering to be shareable among use cases.

For business case 1 “**Compliance Assurance Services for Contracts**” the contracts themselves will not be indexed. The text itself is extracted from the pdf with Apache Tika. The text corpus of the contracts is annotated to train the different services, but will not be part of the Lynx Legal Knowledge Graph. The main idea is to extract information out of the contracts using the Lynx services.

For business case 2 “**Compliance Assurance Services in Oil & Gas and Energy**”, this has been achieved through a custom-made python script using the library PDFMiner.six¹¹ and a set of hand-crafted regular expressions that help identify titles and subtitles. While this approach implies quite a bit of manual work, it is suitable for this dataset in which most documents come in the same format. Specifically, it is necessary to define the area of the documents in which the text is found (excluding headers and footers, for example), as well as combinations of regular expressions and font types that determine that a given piece of text is a title or a subtitle. With this in hand, the whole document can be parsed and divided into sections. We note that this approach ignores all images and considers all text that does not fit the aforementioned regular expressions as equal. This means that text in figure legends, tables, and normal paragraphs is combined. However, since there are always separations between two pieces of text, all

⁷ <http://www.boe.es>

⁸ https://gitlab.com/superlynx/upm_dcm/-/tree/master/lkg

⁹ <https://www.boe.es/legislacion/codigos/codigo.php?id=93&modo=1¬a=0&tab=2>

¹⁰ <http://www.poderjudicial.es/>

¹¹ <https://github.com/pdfminer/pdfminer.six>

services developed in tasks 3.2 and 3.3 work as expected, except for Summarization, which might require adjustments for this behavior.

As for **Business Case 3 “Compliance Assurance Services in Labor Law”**, a service called Document Structure Extractor (StrEx) has been created. The first version of this service is described in D3.1. This service is able to handle multiple types of documents (legislation, case law...) and can extract the different sections in them. For specific sources, such as the Spanish official state gazette, customized extractors have been implemented to recognize sections (further information on this specific service can be found in deliverable 3.1). Regular expressions frequently appearing in certain types of documents have also been identified to improve the detection of their respective structure. Additionally, a generic algorithm has been designed to extract the basic structure of other kinds of documents, such as contracts or judgments, relying on common divisions such as sections or articles.

The **General Legal Corpora** for Austria, Germany, EURLex, is provided by openlaws.com. To extract the content out of openlaws.com, a web service that is described in Deliverable 3.4 has been developed that allows to receive the legislation / decision / collection of documents via the openlaws.com REST API, and to provide it in the Lynx defined format either in RDF or JSON. In the current stage the legislation contains sub parts on an article level, so each individual article can be referenced. For case law there are no sub parts, the decision is provided as one text block at all.

The tool will also be used to feed these documents directly into the Lynx document manager on a regular (e.g., daily) basis. Details of the tool are described in D3.4.

2.6 LEGAL KNOWLEDGE GRAPH ONTOLOGY

All collected corpora have been converted into the Lynx data model which is online available at <http://lynx-project.eu/doc/lkg/>

Document annotations are done in the Natural Language Processing Interchange Format (NIF). This format allows the annotation of linked data and the usage of external ontologies. The data (documents) of the project are going to be converted into NIF, annotated using the semantic annotation services, and stored in the Document Manager and Legal Knowledge Graph.

A basic example of a NIF annotated LynxDocument with its parts is shown in Figure 4:

```
{
  "@context": "http://lynx-project.eu/doc/jsonld/lynxdocument.json",
  "@id": "http://lynx-project.eu/doc/samples/doc006",
  "@type": ["lkg:LynxDocument", "nif:Context"],
  "text": "Art.1 This is the fourth Lynx document. Art.2 This is the fourth Lynx document. Art 2.1. Empty.",
  "parts": [
    {
      "@id": "http://lynx-project.eu/doc/samples/doc006#offset_0_40",
      "@type": ["lkg:LynxDocumentPart", "nif:OffsetBasedString"],
      "offset_ini": "0",
      "offset_end": "39",
      "title": "Art. 1"
    },
    {
      "@id": "http://lynx-project.eu/doc/samples/doc006#offset_41_94",
      "@type": ["lkg:LynxDocumentPart", "nif:OffsetBasedString"],
      "offset_ini": "40",
      "offset_end": "94",
      "title": "Art. 2"
    }
  ]
}
```

Figure 4. Example of NIF annotated LynxDocument with parts

2.7 HARVESTED DOCUMENTS

Table 2 provides an overview of the different data sources, the harvesters and the harvested documents. This table will be updated and detailed in D2.8 [M36]

Jurisdiction	URI	Source Data Format	ELI implemented	Type of docs	Estimate of Volume	Domain	Sub-Domain	Harvester	Harvester is public?	Content availability online
ES	boe.es	PDF,HTML,XML	yes	Legislation	approx. 100,000	law	State and Autonomous Community legislation	UPM	yes	yes
ES	sede.asturias.es	XML, PDF	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
ES, CAT	caib.es	HTML, RDF	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
ES, CAT	https://dogc.gencat.cat	RDF, TURTLE, XML, HTML, PDF	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
ES	http://boa.aragon.es	XML, HTML, JSON, PDF	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
ES	bocyl.jcyl.es	XML	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	10672 ¹² 13	law	federal	openlaws	no ¹⁴	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	1289	law	Burgenland	openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	318	law	Kärnten	openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	1054	law	Niederösterreich	openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	876	law	Oberösterreich	openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	1144	law	Salzburg	openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	1098	law	Steiermark	openlaws	no	yes

¹² Number of laws, not documents. Within RIS 1 article is 1 document

¹³ Laws which are currently in force

¹⁴ Only the converter from openlaws.com to a Lynx-Document

AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	573	law	Tirol	openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	779	law	Voralberg	openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	467	law	Wien	openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	22619	VfgH		openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	116077	VwgH		openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	134473	Justiz		openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	132869	BVwG		openlaws	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	21554	LVwG		openlaws	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	Level 1	Legislation		Legal acts	Directives	openlaws	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	Level 1	Legislation		Legal acts	Regulations	openlaws	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	ECLI	jurisprudence		Judgment	Court of Justice	openlaws	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	ECLI	jurisprudence		Judgment	General Court	openlaws	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	-	Legislation		Consolidated texts	Directives	openlaws	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	-	Legislation		Consolidated texts	Regulations	openlaws	no	yes
DE	https://www.gesetze-im-internet.de	PDF,HTML,XML, EPUB	No	Legislation	6497		Federal	openlaws	no	yes
DE	http://www.rechtsprechung-im-internet.de	PDF,HTML,XML		jurisprudence	51625			openlaws	no	yes

Table 2. Harvested data sources

3 CREATED TRANSLATION CORPORA

3.1 CORPORA CREATION WORKFLOW

3.1.1 General parallel corpora creation workflow

This section contains information about the creation workflow for parallel corpora. This workflow depends on resource type, file format, content of source data, and other factors. In general, the process starts with acquiring original data, processing data into two parallel Moses files¹⁵ (two plain text files with aligned sentences in utf-8 encoding), converting it to final delivery format, and performing quality evaluation (see Figure 2). Data processing includes Cleaning, Aligning, Language identification, Converting, Anonymization, Filtering, Evaluation, and other steps. If problems with data are found at any of these steps, adaptation of some previous steps and reprocessing is required.

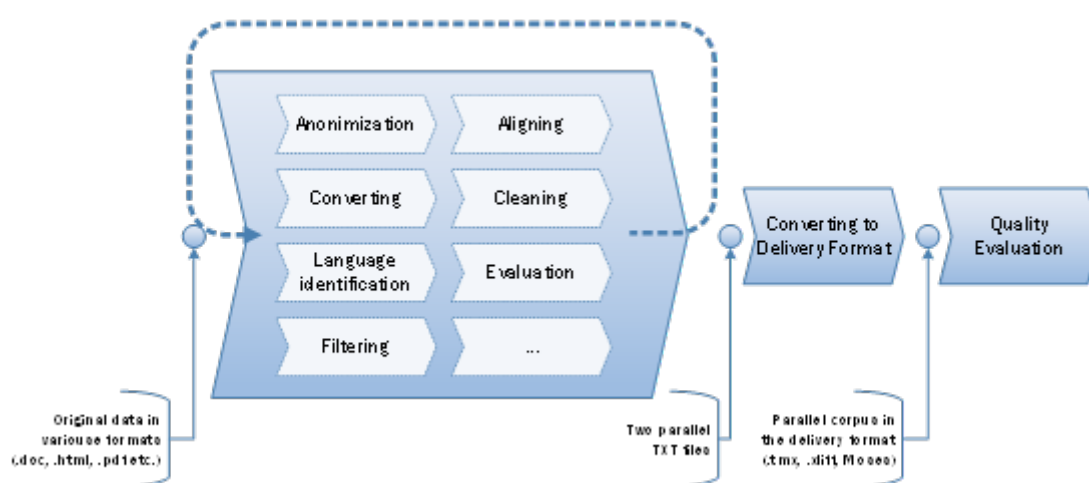


Figure 5. General language resource processing workflow used for Lynx corpora creation

Original data are in many different file formats (such as .doc, .docx, .pdf, .xls, .rtf, .odt, etc.), web content (.html files, lists of links, etc.), translation memory and other localization file formats (.tmx, .xliff, .ttx, etc.), and database dumps (.json, .xml, .sql, etc.). Delivery data is prepared in the agreed format as TMX (Translation Memory eXchange) format, which is standard in localization, or Moses format used by Moses SMT and other MT engines. Quality evaluation is done to ensure that the delivery data meets required quality requirements and does not contain translator or automatic processing errors.

3.1.2 Corpora creation from web crawled data

Web crawling that was used to compile corpus for all business cases. Depending on partners resources, this process was applied to further compile corpora. The first step in this regard was the identification of useful websites. Useful websites are the ones that contain parallel or at least comparable content in two languages. This resource identification was performed by the use case partners and yielded a list of URLs to be crawled and processed. Web-crawling data processing workflow is displayed in Figure 3.

¹⁵ <http://www.statmt.org/moses/>

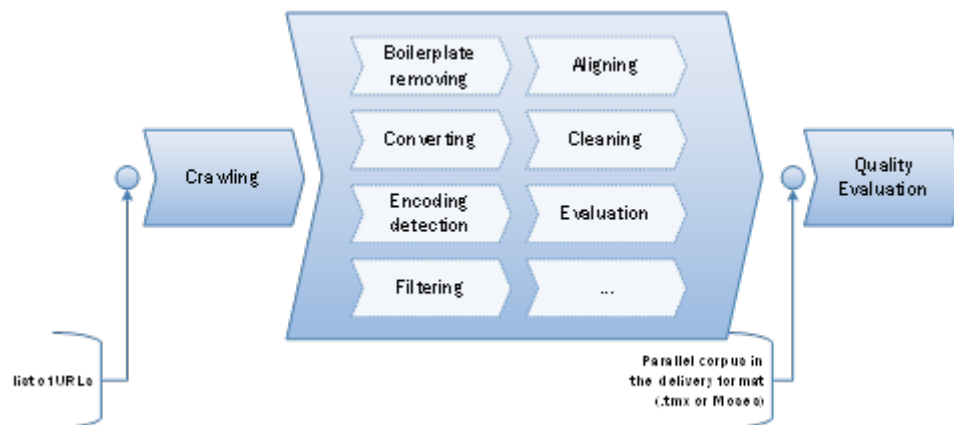


Figure 6. Web-crawled data processing workflow

In general, corpora creation starts by crawling specified URLs and extracting more links from the same domain. Collected links are retrieved, contents are stored, and additional links may be extracted and collected in a loop, until no new links can be found. Web crawling yields .html, .pdf, .doc, and other file types, and for every file type, the appropriate tool are used for text extraction in correct encoding (usually utf-8), boilerplate (ads, headers, footers, etc.) removal, and metadata (title, keywords, author, publisher, etc.) extraction. Further processing is used for language identification, text segmentation, duplicate document removal, and anonymization (see first part of Figure 4).

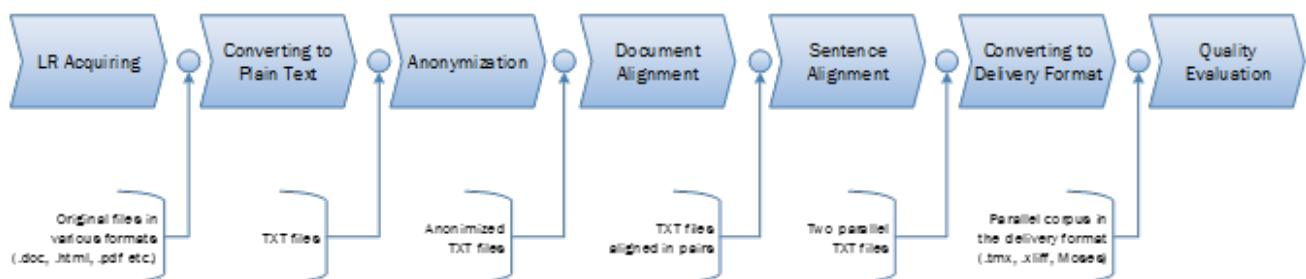


Figure 7. Parallel corpus processing workflow

Processed documents are then aligned at the document level using various heuristics, i.e., similar filenames (for .doc, .pdf, etc.), links pointing to the same document in another language (for .html), and content similarity (pictures, numbers, structure, etc.). This produces a list of paired files that should contain the same text in different languages, and each file pair is then further processed to find matching segments (sentences) in both languages. Sentence alignment is done using either Microsoft Bilingual sentence aligner or HunAlign and produces plaintext files containing matching sentences. Matching sentences may further be simply concatenated into a single Moses file or exported into the required output format. As a post-processing step, several filters may be applied with the purpose of favoring document pairs and aligned segments that are most useful (e.g., having good maximum/minimum length of segment, length ratio of segments, language filters, etc.) for training MT engines. The output file is then evaluated by taking random samples of segments and giving them to a human evaluator. Figure 4 summarizes this process.

3.2 LIST OF TRANSLATION CORPORA

Further in the workflow, various resources (documents and websites) were provided by Lynx partners and later processed to create our geothermal energy related corpora. Table 2 provides examples of processed documents.

ID	Category	Title	Link	Document type	Language	Tool used
1	Directive	DIRECTIVE 2009/28/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL	https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009L0028&from=EN	PDF	DE EN ES NL	General workflow
2	Directive	Directive 2007/2/EC as regards interoperability of spatial data sets and services	https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=LEGISSUM:l28195&from=EN	HTML	DE EN ES NL	Web crawling
3	Act	Dutch Mining Act	https://zoek.officielebekendmakingen.nl/stb-2002-542.html	HTML	NL	Monolingual Web crawling
4	Website	NLOG	https://www.nlog.nl/en	HTML	EN NL	Web crawling
5	Website	NLOG	https://www.nlog.nl/geothermie	HTML	EN NL	Web crawling
6	Website	NLOG	https://www.nlog.nl/en/geothermal-energy	HTML	EN NL	Web crawling
7	Report	EBN Annual report (see Dutch translation below)	https://www.ebn.nl/wp-content/uploads/2017/06/Focus-on-Energy-2017.pdf	PDF	EN	General workflow
8	Report	EBN Jaarverslag	http://cdn.instantmagazine.com/upload/11408/web_-_2018_04_24_-_ebn_-_jaarverslag_2017.8bcde623ed6b.pdf	PDF	NL	General workflow
9	Website	DAGO - Dutch Association Geothermal Operators	https://www.dago.nu/nl/geothermie	HTML	NL	Web crawling
10	Website	DAGO - Dutch Association Geothermal Operators	https://www.dago.nu/en/geothermie	HTML	EN	Web crawling
11	Website	Thermogis - Geothermal mapping	https://www.thermogis.nl/	HTML	NL	Web crawling

12	Website	Thermogis - Geothermal mapping	https://www.thermogis.nl/en	HTML	EN	Web crawling
13	Article	Gebruik van warmte uit bodem in 5 jaar verdubbeld	https://www.cbs.nl/nl-nl/nieuws/2015/41/gebruik-van-warmte-uit-bodem-in-5-jaar-verdubbeld	HTML	NL	Web crawling
14	Article	Use geothermal heat doubled in the past 5 years	https://www.cbs.nl/en-gb/news/2015/41/use-geothermal-heat-doubled-in-the-past-5-years	HTML	EN	Web crawling
15	Website	Hoewerktaardwarmte.nl	Hoewerktaardwarmte.nl	HTML	NL	Monolingual Web crawling
16	Website	WHAT IS DEEP GEOTHERMAL ENERGY	https://vito.be/en/vito-insights/deep-geothermal/what-deep-geothermal-energy	HTML	EN	Web crawling
17	Website	WAT IS DIEPE GEOTHERMIE?	https://vito.be/nl/diepe-geothermie/wat-diepe-geothermie	HTML	NL	Web crawling
18	Website	DEEP GEOTHERMAL ENERGY IN FLANDERS	https://vito.be/en/vito-insights/deep-geothermal/deep-geothermal-energy-flanders	HTML	EN	Web crawling
19	Website	GEOTHERMIE IN VLAANDEREN	https://vito.be/nl/diepe-geothermie/geothermie-vlaanderen	HTML	NL	Web crawling
20	Website	VITO REVEALS GEOTHERMAL ENERGY POTENTIAL IN THE BORDER REGION	https://vito.be/en/vito-reveals-geothermal-energy-potential-border-region	HTML	EN	Web crawling
21	Website	GEOTHERMISCH POTENTIEEL IN GRENSREGIO	https://vito.be/nl/geothermisch-potentieel-grensregio	HTML	NL	Web crawling
22	Website	GEOTHERMAL ENERGY BOOSTS EMPLOYMENT	https://vito.be/en/geothermal-energy-boosts-employment	HTML	EN	Web crawling
23	Website	GEOTHERMIE BOOST WERKGELEGENHEID	https://vito.be/nl/geothermie-boost-werkgelegenheid	HTML	NL	Web crawling
24	Website	DEEP GEOTHERMAL ENERGY IN THE KEMPEN: WHAT'S NEXT?	https://vito.be/en/news/deep-geothermal-energy-kempen-what%E2%80%99s-next	HTML	EN	Web crawling

25	Website	DIEPE GEOTHERMIE IN DE KEMPEN: WHAT'S NEXT?	https://vito.be/nl/nieuws/diepe-geothermie-de-kempen-what%E2%80%99s-next	HTML	NL	Web crawling
26	Website	DEVELOPMENT & TESTING OF INNOVATIVE GEOPHYSICAL METHODS FOR EVALUATION GEOTHERMAL POTENTIAL	https://vito.be/en/news/development-testing-innovative-geophysical-methods-evaluation-geothermal-potential	HTML	EN	Web crawling
27	Website	ONTWIKKELEN & TESTEN VAN INNOVATIEVE GEOFYSISCHE METHODES VOOR EVALUATIE GEOTHERMISCH POTENTIEEL	https://vito.be/nl/nieuws/ontwikkelen-testen-van-innovatieve-geofysische-methodes-voor-evaluatie-geothermisch	HTML	NL	Web crawling
28	Website	VITO BRINGS GEOTHERMAL HEAT TO THE SURFACE	https://vito.be/en/news/vito-brings-geothermal-heat-surface	HTML	EN	Web crawling
29	Website	VITO BRENGT AARDWARMTE AAN DE OPPERVLAKTE	https://vito.be/nl/nieuws/vito-brengt-aardwarmte-aan-de-oppervlakte	HTML	NL	Web crawling
30	Website	BALMATT ENERGY PLANT	https://vito.be/en/vito-insights/deep-geothermal/balmatt-energy-plant	HTML	EN	Web crawling
31	Website	BALMATT-SITE	https://vito.be/nl/diepe-geothermie/balmatt-site	HTML	NL	Web crawling
32	Website	GEOTHERMAL ENERGY SOURCE	https://vito.be/en/news/geothermal-energy-source	HTML	EN	Web crawling
33	Website	AARDWARMTE BRON VAN ENERGIE	https://vito.be/nl/nieuws/aardwarmte-bron-van-energie	HTML	NL	Web crawling

Table 3. Resources used to build corpora

The given resources were examined and aggregated for convenient processing by domain. In this way, resources #4, #5, and #6 were processed together, as were #7- #8, #9-#10, #11-#12, #13-#14, and #16-#33, to produce bilingual corpora. The processed corpora (LYNX_Geothermal) was uploaded to <https://www.letsmt.eu> for further use.

	Mono	de	en	es	nl
de	1,939		261	824	854
en	10,982	261		976	9,745
es	2,100	824	976		300

nl	13,066	854	9,745	300
----	--------	-----	-------	-----

Table 4. Segment count in LYNX_Geothermal

Uploaded corpora size statistics are given in Table 4. Most segments belong to the EN-NL language pair, as most resources were Dutch geothermal-energy-related websites. Some DE-EN and EN-ES segments were extracted from resources #1 and #2.

Resources #3 and #15 were found to be unilingual; therefore, only NL content was extracted and used to build the NL monolingual corpus. The monolingual corpus contains 2,167 segments and, similarly to the bilingual resources it was uploaded to <https://www.letsmt.eu> for further use.

3.3 LIST OF CORPORA FOR MACHINE TRANSLATION TRAINING

Table 5, Table 6, Table 7 and Table 8 show which public or proprietary corpora were used to train our NMT systems. We used the corpora filtering workflow of Pinnis (2018) to remove most of the lower-quality data from these corpora before training the NMT systems. The following issues are addressed by various filters:

1. Source-source or target-target entries in parallel data. (Equal source/target entries are filtered out).
2. Sentence splitting issues. (Segments with more than 1000 symbols or more than 400 tokens are filtered out; the numerical thresholds can be adjusted for each individual training task.)
3. Data corruption through optical character recognition (OCR), e.g., when processing PDF documents. (Segments containing tokens with >50 symbols are filtered out.)
4. Redundancy issues. (Duplicate entries are filtered out).
5. Partial translation (also sentence splitting) issues. (Entries where the length ratio between the source and target segments is too small (e.g., <0.3) are filtered out.)
6. Foreign language data issues. (Entries containing letters from neither source nor target languages are filtered out)
7. Sentence misalignment issues. (Sentences failing a cross-lingual alignment test using c-eval (Zariņa et al., 2015) are filtered out.)
8. Incorrect language filtering using an automatic language detection tool (Shuyo, 2010).
9. Low content overlap filtering using the cross-lingual alignment tool MPAligner (Pinnis, 2013).
10. Digit mismatch filtering. (This showed to be effective in identifying parallel corpora sentence segmentation issues).

Table 4 lists the corpora used for the EN-NL NMT systems with a total size of 41 639 229 parallel sentences.

Corpus name
LYNX_Geothermal
DGT-TM
EUBookShop
DCEP
EESC
Europarl v7
TAUS - Legal

JRC-Acquis (v.3.0)
EMA
OPUS - EMEA
TAUS - Information Technology
RAPID
Tatoeba
OPUS - ECB
TAUS - Business
Global Voices Parallel Corpus
TAUS - Electronics
European Ombudsman
EUROSTAT Combined Nomenclature
OPUS - European Constitution
TAUS - Manufacturing
EUROSTAT PRODCOM
JRC-Names
Geo Names
Europe's Languages in the Digital Age
Regions
TAUS - Misc

Table 5. Corpora used for EN-NL NMT systems

Table 5 lists the corpora used for the EN-ES NMT systems with total size of 81 176 632 parallel sentences.

Corpus name
DGT-TM
MultiUN
DCEP
Europarl v7
TAUS - Legal
JRC-Acquis (v.3.0)
European Ombudsman
OPUS - European Constitution
United Nations Parallel Corpus
TAUS - Information Technology
EUBookShop
EESC
EMA
OPUS - EMEA
RAPID
Global Voices Parallel Corpus
Tatoeba
TAUS - Telecommunications
TAUS - Electronics
OPUS - ECB
TAUS - Automotive

TAUS - Misc
TAUS - Business
TAUS - Financial
TAUS - Manufacturing

Table 6 Corpora used for EN-ES NMT systems

Table 6 lists the corpora used for the EN-DE NMT systems with total size of 87 542 066 parallel sentences.

Corpus name
OPUS - Open Subtitles
TAUS - Information Technology
DGT-TM
DCEP
EESC
Europarl v7
RAPID
EMA
JRC-Acquis (v.3.0)
Tatoeba
TAUS - Electronics
MultiUN
OPUS - ECB
TAUS - Business
Libre Office
Global Voices Parallel Corpus
German-English Parallel Corpus de-news
TAUS - Manufacturing
TAUS - Misc
OPUS - European Constitution
JRC-Names
EUROSTAT PRODCOM
ECDC-TM
Geo Names
ParaCrawl parallel corpus
Common Crawl parallel corpus
News Commentary Corpus

Table 7. Corpora used for EN-DE NMT systems

Table 7 lists the corpora used for the NL-DE NMT systems with total size of 15 463 248 parallel sentences.

Corpus name
DGT-TM
DCEP
OPUS - Open Subtitles
Europarl v7

JRC-Acquis (v.3.0)
OPUS - EMEA
RAPID 2016
EMA 2016 UNIQUE
TED Talks 2018
WIT ³ - Web Inventory of Transcribed and Translated Talks.
OPUS - ECB
OPUS - KDE4
Tatoeba.org 2017
Tatoeba.org 2018
Tatoeba.org
RAPID
European Ombudsman
EUROSTAT Combined Nomenclature
OECD iLibrary Summaries
OPUS - European Constitution
JRC-Names
EUROSTAT PRODCOM 2014
ECDC-TM
DG EAC TM
EUROSTAT NACE v2
LYNX Geothermal
Geo Names
Regions

Table 8. Corpora used for NL-DE NMT systems

CONCLUSIONS

This report summarizes the work done in Task 2.4 “Indexing of corpora” and Task 2.5 “Translation corpora creation” under WP2 of the Lynx project. Firstly, the corpora collection approach, the analysis of data format, and the results of the terminology extraction have been explained. Then, the corpora creation workflow for Lynx acquired corpora has been described, including the workflow of parallel corpora and corpora creation from the web crawled data. Finally, the list of translation corpora and the one for machine translation used for training have been presented.

As for terminology preparation for MT, it must be noted that a domain-specific bilingual terminology is a resource often used to control lexical quality of domain-specific MT systems. However, not all term collections that have been prepared by terminologists, translators, or automatic means can be directly used (or should be used) in MT. This is because of ambiguity of the terms in the term collections, which, in turn, is due to insufficient morphological, syntactic, or semantic information that describes each term in the term collections. Furthermore, MT systems are source-to-target systems that in typical scenarios (i.e., if we ignore multi-way NMT systems and other multi-task neural network-based systems) translate from one source language to one target language. Therefore, term collections for MT systems have to be prepared as bilingual collections that define term pairs. However, we know that a term entry in a term database may consist of a term in multiple languages and even multiple term variants, e.g., different variants for different registers (e.g., neutral, technical, slang, etc.), types (e.g., abbreviation, initialism, short form, full form, etc.), etc. This means that before integrating terms into an MT system, each term collection has to be transformed and pre-processed into a bilingual term collection that consists of term pairs.

ANNEX 1. MAIN EU LABOR LAW LEGISLATION CORPORA

- Council Directive 1999/70/EC of 28 June 1999 concerning the framework agreement on fixed-term work concluded by ETUC, UNICE, and CEEP
- Council Directive 2001/86/EC of 8 October 2001 supplementing the Statute for a European company with regard to the involvement of employees
- Council Directive 2010/18/EU of 8 March 2010 implementing the revised Framework Agreement on parental leave concluded by BUSINESSEUROPE, UEAPME, CEEP, and ETUC and repealing Directive 96/34/EC (Text with EEA relevance)
- Council Directive 89/391/EEC of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work
- Council Directive 91/533/EEC of 14 October 1991 on an employer's obligation to inform employees of the conditions applicable to the contract or employment relationship
- Council Directive 92/104/EEC of 3 December 1992 on the minimum requirements for improving the safety and health protection of workers in surface and underground mineral-extracting industries (twelfth individual Directive within the meaning of Article 16 (1) of Directive 89/391/EEC)
- Council Directive 92/85/EEC of 19 October 1992 on the introduction of measures to encourage improvements in the safety and health at work of pregnant workers and workers who have recently given birth or are breastfeeding (tenth individual Directive within the meaning of Article 16 (1) of Directive 89/391/EEC)
- Council Directive 94/33/EC of 22 June 1994 on the protection of young people at work
- Directive 2003/41/EC of the European Parliament and of the Council of 3 June 2003 on the activities and supervision of institutions for occupational retirement provision
- Directive 2003/88/EC of the European Parliament and of the Council of 4 November 2003 concerning certain aspects of the organization of working time
- Directive 2008/94/EC of the European Parliament and of the Council of 22 October 2008 on the protection of employees in the event of the insolvency of their employer (Codified version) (Text with EEA relevance)
- Directive 96/71/EC of the European Parliament and of the Council of 16 December 1996 concerning the posting of workers in the framework of the provision of services.

REFERENCES

- Bautista Salinero, J. (2020). Extracción y Normalización de Convenios Colectivos. Trabajo fin de Grado, Universidad Politécnica de Madrid. To appear in <http://oa.upm.es>
- Eurostat. (2009) Nace Rev. 2 Metadata. URL: https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC
- Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* (pp. 562-570).
- Pinnis, M. (2018). Tilde's Parallel Corpus Filtering Methods for WMT 2018. In *Proceedings of the Third Conference on Machine Translation* (pp. 952–958).
- Shuyo, N. (2010). Language detection library for java. Retrieved Jul, 7, 2016.
- Zariņa, I., Nikiforovs, P., Skadiņš, R. (2015). Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.