Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3761–3768
Marseille, 11–16 May 2020

The MARCELL Legislative Corpus

Tamás Váradi[1], Svetla Koeva[2], Martin Yalamov[2], Marko Tadić[3], Bálint Sass[1], Bartłomiej Nitoń[4], Maciej Ogrodniczuk[4], Piotr Pęzik[5], Verginica Barbu Mititelu[6], Radu Ion[6], Elena Irimia[6], Maria Mitrofan[6], Vasile Păiș[6], Dan Tufiș[6], Radovan Garabík[7], Simon Krek[8], Andraž Repar[8], Matjaž Rihtar[8], Janez Brank[8]

[1]Research Institute for Linguistics, Budapest, Hungary, {varadi.tamas,sass.balint}@nytud.hu
[2]IBL, BAS, Sofia, Bulgaria, {svetla,martin}@dcl.bas.bg
[3]University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia, marko.tadic@ffzg.hr
[4]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland,
bartek.niton@gmail.com, maciej.ogrodniczuk@ipipan.waw.pl
[5]University of Łódź, Poland, pezik@uni.lodz.pl
[6]RACAI, Bucharest, Romania, {vergi,radu,elena,maria,vasile,tufis}@racai.ro
[7]Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia,
garabik@kassiopeia.juls.savba.sk
[8]IJS, Ljubljana, Slovenia, {simon.krek,matjaz.rihtar,janez.brank}@ijs.si, andraz.repar@cjvt.si

Abstract

This article presents the current outcomes of the MARCELL CEF Telecom project aiming to collect and deeply annotate a large comparable corpus of legal documents. The MARCELL corpus includes 7 monolingual sub-corpora (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian) containing the total body of respective national legislative documents. These sub-corpora are automatically sentence split, tokenized, lemmatized and morphologically and syntactically annotated. The monolingual sub-corpora are complemented by a thematically related parallel corpus (Croatian-English). The metadata and the annotations are uniformly provided for each language specific sub-corpus. Besides the standard morphosyntactic analysis plus named entity and dependency and/or noun phrase annotation, the corpus is enriched with the IATE and EuroVoc labels. The file format is CoNLL-U Plus Format, containing the ten columns specific to the CoNLL-U format and four extra columns specific to our corpora. The MARCELL corpora represent a rich and valuable source for further studies and developments in machine learning, cross-lingual terminological data extraction and classification.

Keywords: law corpus, comparable corpus, under-resourced languages

1. Introduction

The present paper introduces the MARCELL corpus and related resources compiled in the CEF Telecom[1] Action of the same name. The CEF Telecom project Multilingual Resources for CEF.AT in the Legal Domain (MARCELL)[2] aims to enhance the eTranslation system[3] developed by the European Commission through supplying seven large scale corpora consisting of national legislative documents effective in Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia.

The structure of the paper is as follows. We describe the rationale and objectives of the work in section 2. The composition of the corpus in each of the seven languages is presented in 3, whereas section 4 introduces the Croatian–English parallel corpus, also created in the project. The format and annotation, as well as the metadata of the multilingual corpora are described in sections 5 and 6 respectively. In some languages the corpus is already enriched with annotation of terminology in IATE[4] and EuroVoc[5] as described in section 7, the annotation work is still in progress for the rest of the languages. The multilingual corpus will be subdivided into several sub-domains corresponding to top-level categories of the EuroVoc system. This work is briefly described in section 8. The issue of sustainability with the aspects it involves is addressed in section 9 before some conclusions are given in section 10.

2. Rationale and Objectives

The MARCELL CEF Telecom Action is pursued with the ultimate goal of breaking down linguistic barriers to the creation of the Digital Single Market[6] in Europe, one pillar of which will be multilingual digital service infrastructures (such as the Online Dispute Resolution[7], the e-justice platform[8] or Europeana[9]). The eTranslation system, itself a digital service infrastructure, is a building block that will help to make these infrastructures become multilingual. The eTranslation system faces the daunting task of supplying quality MT service in all domains of relevance to the growing number of digital service infrastructures and for all the official languages of the EU.

As is well known, one bottleneck to MT is the scarcity of quality data, which means primarily parallel texts, but recently monolingual data has been usefully employed through the technique of back translation (Sennrich et al., 2015). Preferably, the data should cover specific domains relevant for the fields of application. The MARCELL

[1]https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom
[2] http://marcell-project.eu
[3]https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
[4] https://iate.europa.eu/home
[5] https://eur-lex.europa.eu/browse/eurovoc.html
[6] https://ec.europa.eu/digital-single-market/en
[7]https://ec.europa.eu/consumers/odr/main/?event=main.home2.show
[8] https://e-justice.europa.eu/home.do?action=home
[9] https://www.europeana.eu/portal/en

corpus fills the above requirements on several counts. It supplies the total body of national legislative documents that are in effect in seven member states of the EU. The choice of domain may require justification in view of the fact that the existing eTranslation system was trained on legislative parallel documents. However, the training material consisted of EU legislation (Steinberger et al., 2006) and, surprising as it may be, national legislation is not automatically available to the EC, hence, the MARCELL corpus represents new material. Apart from its rather marked stylistic features, the legal domain is notoriously heterogeneous in terms of content. Therefore, as an innovative feature, the documents in seven monolingual corpora will be classified in terms of the EuroVoc domains (such as politics, economics, trade, education and communication, science, etc.). The classification will thus yield twenty one thematic sub-corpora in each language. From another perspective, the cross-lingual mapping may be seen as twenty one comparable corpora across seven languages. In addition to the standard lemmatization and morphosyntactic analysis plus named entity and dependency annotation, the whole corpus will be enriched with the IATE and EuroVoc terminology set.

3. Composition of the Corpus

The corpus gathers all effective national legislation from Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia in seven large-scale linguistically processed monolingual sub-corpora of national legislation. According to the legislation in all these countries, such texts are free of any intellectual property restrictions. Some quantitative information on the seven sub-corpora are presented in Table 1, where the column heads are the language codes.

3.1 Details of the Bulgarian Corpus

The Bulgarian corpus consists of 25,283 documents (at the beginning of November 2019) which are classified into eleven types: Administrative court; Agreements; Amendments, Legislative acts; Conventions; Decrees; Decrees of the Council of Ministers; Guidelines; Instructions; Laws (Acts); Memorandums; Resolutions. The corpus is a selection from a larger legal domain dataset (113,427 documents distributed in 52 types) and contains universally binding legal acts. The time span of the documents is 1946-2019.

The data has been retrieved from the Bulgarian State Gazette (http://dv.parliament.bg), the Bulgarian government official journal, publishing documents from official institutions like government, National Assembly of Bulgaria, Constitutional Court, etc. The data were extracted from the original HTML format, filtered by document type, tokenized, sentence split, tagged and lemmatized with a fine-grained version of the Bulgarian Language Processing Chain (Koeva and Genov, 2011). The data were dependency parsed with NLP-Cube[10]. The named entities (persons, locations, organisations and other) and specific types of noun phrases were annotated with a rule-based annotation tool (Koeva and Dimitrova, 2015). An annotation tool was developed to annotate IATE terms and EuroVoc descriptors within the corpus.

3.2 Details of the Croatian Corpus

The Croatian corpus consists of 33,559 documents that represent the national legislation from 1990 until today. The corpus is composed of legally binding acts (laws, regulations, decisions, orders, etc.) and internally binding acts (ordinances, recommendations, etc.). There are 12 different texts types with ordinances (11,367), decisions (7,708) and laws (3,789) as three most frequent text types. In collaboration with the Central State Office for the Development of the Digital Society of the Republic of Croatia (RDD)[11], which has, as a part of its mission, the duty to ensure online accessibility to all Croatian legal documentation, we received the data from their database in October 2019 and we are presenting the figures of that current state. The data were delivered in a proprietary XML format that had to be converted into a CoNLL-U Plus format and the relevant accompanying metadata were extracted from the RDD database.

The corpus was analysed with the Croatian Language Web Services (hrWS, see at META-SHARE[12]): sentences are split, tokens are identified and morphologically and syntactically annotated. The annotation of the IATE terms and EuroVoc descriptors by way of matching these terms with SWE/MWEs in the corpus is in progress. The corpus overall size is almost 9.6 M sentences and around 63 M tokens.

3.3 Details of the Hungarian Corpus

The Hungarian corpus representing the Hungarian national legislation contains 26,821 documents retrieved from PDF files of the official gazette Magyar Közlöny, which is freely available online for download[13]. There are 11 different text types in the corpus covering different kinds of legal texts: laws, regulations, decrees, etc.

The data was analysed with the e-magyar text processing system[14] (Váradi et al., 2018; Indig et al., 2019). The system was enhanced with detokenization functionality (precisely for the requirements of the MARCELL project) to provide SpaceAfter=No annotation indicating no whitespace between two tokens in the original text. The corpus does not contain dependency annotation, but it does contain noun phrase annotation. Additional scripts were created for extracting the necessary metadata, for converting to CoNLL-U Plus format, and for the annotation of IATE terms and EuroVoc descriptors in the text.

The raw data is 31.2 M tokens and 302 MB in size. The analysed corpus is 2.9 GB in CoNLL-U Plus format.

3.4 Details of the Polish Corpus

The Polish corpus contains 22,341 documents of 19 types representing universally binding legal acts (law, regulation,

[10] https://opensource.adobe.com/NLP-Cube/index.html
[11] http://rdd.gov.hr
[12] http://meta-share.ffzg.hr
[13] http://kozlonyok.hu/
[14] http://e-magyar.hu

etc.) or binding internal acts (such as resolutions of the Sejm, Senate and some state administration bodies, e.g. the Council of Ministers). The time span of the documents is 1992-2020 and they amount to 272 MB and 38 M tokens in raw data (values represent only documents considered being in effect).

The data were retrieved from Dziennik Ustaw[15] and Monitor Polski[16], the official and publicly available sources of Polish law, publishing Acts of Parliament, Regulations of the Ministers, uniform acts and amendments. The data was converted from editable PDF files to textual format (unfortunately an XML version of those documents was unavailable), tokenized and morphologically analysed with Morfeusz2 (Kieraś and Woliński, 2017), disambiguated with Concraft-pl tagger (Waszczuk, 2012), named entity recognition with Liner2 (Marcińczuk et al., 2018) and dependency-parsed with COMBO (Rybak and Wróblewska, 2018). Additional scripts were created (and used) for IATE terms and EuroVoc descriptors annotation.

3.5 Details of the Romanian Corpus

The Romanian corpus contains 144,131 files containing 4,300,131 sentences, which represent the body of national legislation ranging from 1881 to 2018. This corpus includes mainly governmental decisions, ministerial orders, decisions, decrees and laws. All the texts were obtained via crawling from the Romanian legislative portal[17]. We have not distinguished between legal documents that are in effect and those that are not because it is difficult to distinguish them automatically in the absence of any external resource to use for the process. The texts were extracted from the original HTML format and converted into TXT files (more than 2.6 GB). Each file has multiple levels of annotation: firstly the texts were tokenized (more than 375 M), lemmatized and morphologically annotated using the Tokenizing, Tagging and Lemmatizing (TTL) text processing platform developed at RACAI (Ion, 2007), then dependency parsed with NLP-Cube (Boroș et al., 2018), named entities were identified using a tool developed at RACAI (Păiș et al., 2019), nominal phrases were identified also with TTL, while IATE terms and EuroVoc descriptors were identified using an internal tool (Coman et al., 2019).

3.6 Details of the Slovak Corpus

The Slovak corpus contains 13,600 documents (32 M tokens) of legally binding acts starting from the year 1993 (following minor orthography reform in 1991, but it also coincides with the independence of Slovakia). The data is obtained from the *Slov-Lex legislative and information portal* archive[18] of the acts approved by the Slovak Parliament. The data has been converted from the original HTML format, filtered by date and document length, tokenized, lemmatized and morphologically annotated with the Slovak MorphoDita model (Garabík and Šimková, 2012) and dependency parsed with UDPipe (Straka et al., 2016).

3.7 Details of the Slovenian Corpus

The Slovenian corpus contains 21,556 documents (5 GB in size, 127 M tokens), ranging from 1974 to 2018. The data was obtained from the Slovenian *Open Data Portal*[19]. The original file type is JSON, which contains individual documents in HTML format. The data in the corpus was extracted from the HTML documents, tokenized with the Slovenian tokenizer Obeliks4j (Grčar et al., 2012), and lemmatized, tagged and dependency parsed with a fork[20] of the StanfordNLP parser (Peng et al., 2018) trained on ssj500k training corpus (Krek et al., 2017). Additional scripts have been created to extract metadata and annotate IATE terms and EuroVoc descriptions.

language	bg	hr	hu	pl	ro	sk	sl
documents [k]	25	34	26	22	144	13	22
sentences [k]	3281	9592	962	1754	4300	2473	7647
tokens [M]	45	63	31	38	375	32	127
raw size [MiB]	1080	N/A[21]	302	272	2600	180	5000
time span	1946 2019	1990 2019	1974 2019	1992 2020	1881 2018	1993 2019	1974 2018

Table 1: Basic information about the sub-corpora.

4. The Croatian-English Parallel Corpus

Since Croatia joined the EU in 2013 only, the role of Croatian as one of the EU official languages lacks six to nine years of systematic accumulation of translation memories (TMs) during the translation process in different EU bodies which other languages in the project had. In order to overcome this situation, an additional task agreed upon was to build the Croatian-English Parallel Corpus of Croatian National Legislation, with texts from 1990 to 2019, that has been translated into English[22] with a size of at least 1,800 documents.

Unlike the situation with the monolingual Croatian legal documents, the English translations were received only in PDF, which was produced starting with the versions from late 1990s. Consequently, we had to deal with the text extraction from different PDF varieties and sources that diminished the quality of automatic extraction. The extracted texts and originals were converted into a plain TXT format which was processed for sentence splitting and

[15] http://dziennikustaw.gov.pl/
[16] http://monitorpolski.gov.pl/
[17] http://legislatie.just.ro/
[18] https://www.slov-lex.sk
[19] https://podatki.gov.si/
[20] https://github.com/clarinsi/classla-stanfordnlp
[21] Not applicable because data is received in marked up format.
[22] Contrary to popular misconceptions, English remains an official language of the European Union even after the withdrawal of the United Kingdom from the EU.

aligned with LF-aligner[23] open source tool that uses the HunAlign in the background (Varga et al., 2005). Alignments of all 1,816 documents were manually inspected and corrected in order to produce the high quality aligned TMX files and thus a reliable parallel corpus that can be used for noiseless training of NMT systems. The corpus size is 396,984 TUs with 14.4 and 17.7 M tokens in Croatian and English respectively

5. Format and Annotation

The corpora use the CoNLL-U Plus format. Each language specific sub-corpus observes the same format, which was deliberately modelled after the CoNLL-U format by including several additional columns. The first ten (1 to 10) columns keep their CoNLL-U values, while the following 4 columns are specific to our corpora.

The columns are separated by a TAB character. There are the following columns (the detailed description of the CoNLL-U columns, as well as the internal format of the file can be found at the Universal Dependencies site[24]):

ID FORM LEMMA UDPOS XPOS FEATS HEAD DEPREL DEPS MISC NER NP IATE EuroVoc

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes
2. FORM: Word form (including punctuation)
3. LEMMA: Lemma
4. UPOS: Universal part-of-speech tag[25]
5. XPOS: Language-specific part-of-speech tag (morpho-syntactic description)
6. FEATS: List of morphological features
7. HEAD: Head of the current word (its ID or zero)
8. DEPREL: Universal dependency relation to the HEAD
9. DEPS: Enhanced dependency graph (optional)
10. MISC: Other information; e.g. missing white space between the token and the following one
11. MARCELL:NE: the BIO format annotation of the current token, O if it is not part of a named entity
12. MARCELL:NP: the BIO format annotation of the current token, O if it is not part of a noun phrase
13. MARCELL:IATE: the annotation of a IATE term by the language-independent code if it is (part of) a IATE term ('_' otherwise)
14. MARCELL:EuroVoc: the upper level domain label in the EuroVoc thesaurus if it is a term ('_' otherwise)

Unless mentioned otherwise, the underscore (_) is used to denote unspecified values in all fields.

Each document in the corpora is uniquely identified by its identifier constructed in the form XX-legal-ID, where XX is the language code and ID is a unique identifier within one language corpus, derived from the document identification number (e.g. by replacing characters disallowed in CoNLL-U format). Paragraphs and sentences are numbered (starting from 1) and assigned each a unique identifier as well (e.g. XX-legal-ID-p2s1 marks the first sentence in the second paragraph of the document ID in the XX corpus). The complete text of the respective sentence is included as the text attribute.

6. Metadata

Data for each of the languages come from a separate source, often developed as a government supported access to the legal system of the particular country; all these systems were developed independently and offer widely diverging modes of access and data annotation (document metadata). Nevertheless, some common (and obvious) annotation items can be extracted and used as a base of common annotation schema.

Table 2 below captures existing (or trivially obtainable) important metadata in source archives that serve as a base for the annotation of documents in the corpora. Note that these keys and values need not be presented directly in the source documents, but can be unambiguously extracted or derived from other metadata (e.g. date can be obtained from file name or transformed from native-language date description):

- *identifier* is a short string uniquely identifying the document within one language (one archive); usually it is the official legal act number, often including the year of publication and a chronologically assigned number
- *date* is either the date when the document was created, or the date when the legal act went into effect (if both are present, the most relevant one is selected)
- *title* is an informative, usually official name of the document
- *type* further specifies the legal type of the document, e.g. regulation, law, announcement, legally binding decision, etc.
- *issuer* is the organization issuing (publishing) the documents
- *keywords* contain several keywords related to the content of the document
- *url* is the original individual address the document was accessed at, in case the documents are available separately, each at its own URL (not if the whole legal body was obtained as one big archive)
- *topic* roughly specifies the subject of the document

Annotation of the documents in the corpora is based on the source metadata, but transforms or adds several annotation keys that are constructed during corpora compilation (in particular, the original raw values are checked, cleaned up and unified, e.g. by normalizing capitalization or automatically fixing common spelling mistakes in the original metadata). These keys can be either *obligatory* (each document must contain this annotation), *optional* (this annotation key can be missing in some language corpora – which is not the same as containing an empty

[23] https://sourceforge.net/projects/aligner/

[24] https://universaldependencies.org/format.html

[25] https://universaldependencies.org/u/pos/index.html

value), or *local* (annotation specific for a given language corpus, containing less important information, e.g. included for completeness to capture data for the original source annotation, or less accurate data, etc.). Obligatory and optional keys are harmonized across all the language specific corpora.

language	bg	hr	hu	pl	ro	sk	sl
identifier	×	×	×	×	×	×	×
date	×	×	×	×	×	×	×
title	×	×	×	×	×	×	×
type	×	×	×	×	×	×	×
issuer	×		×	×	×		
keywords		×		×			
url	×	×		×	×		
topic		×	×				

Table 2. Available annotation data in source archives.

Obligatory annotation items are as follows:

- *id* – unique identifier of the document within all the corpora, following CoNLL-U conventions
- *date* – date of the document, in ISO 8601 format, with accuracy given by source metadata (at least the year)
- *title* – human-readable title (name) of the document, in the original language
- *type* – legal type of the document, in the original language
- *entype* – legal type of the document, in English (harmonized across the languages)

Optional annotation items are as follows:

- *url* – address the individual document has been accessed at
- *keywords* – several keywords separated by commas, in the original language
- *topic* – human-readable topic of the document contents, in the original language

Local annotation follows this convention in key naming – key name without a language prefix means the value is either language-agnostic, or in the original language; key name prefixed by *en* means the value is in English.

7. Terminology Annotation

The Bulgarian, Hungarian, Romanian and Slovak corpora were annotated with IATE terms (45,592 terms in Bulgarian, 51,957 in Hungarian, 56,228 in Romanian, and 46,399 in Slovak are available at the language specific sections of IATE) and EuroVoc descriptors. Single-word and multiword terms within the documents were annotated if their lemma and part-of-speech coincide with the lemma and the part-of-speech of an IATE term or an EuroVoc descriptor. For example, BG инструктор; RO monitor (EN Instructor): IATE ID: 1394636; EuroVoc field – 3206 (Education and Communications). Such annotation has some drawbacks because no disambiguation could be performed with respect to IATE terms and EuroVoc descriptors. For Croatian, Slovak, Slovenian and Polish the annotation work is still in progress.

8. Future Work

8.1 Thematic Document Linking and Clustering

To facilitate the identification of topical clusters in the comparable corpus we have used two document classification and linking techniques. The first approach relies on the JEX dataset (Steinberger et al., 2013), which contains samples of legal documents in 22 European languages annotated with EuroVoc terms. A convolutional neural network supervised classifier is first trained on the original JEX dataset and subsequently applied to fixed-length portions of the documents from the MARCELL corpus. The aggregated document-level EuroVoc labels obtained by the classifier for each document can be used as cross-lingual topical descriptors of their content.

The second approach is meant to generate unlabelled links between relevant sections of documents in the different languages. This is motivated by the fact that the length of legal documents may vary significantly and some of the largest documents may include topical sections which are related to subsections of documents in other languages. As a first step, we use the models provided by Schwenk and Douze (2017) to compute language agnostic embeddings (LASER) for every sentence in the multilingual corpus. Next, the resulting sentence vectors are indexed using the FAISS vector search library (Johnson et al., 2017) and an interlingual distance matrix of all sentences is computed. The similarity measure between documents in different languages is then calculated as an overlap coefficient of similar sentences which they contain.

Table 3 illustrates the outcome of document similarity measurement between a sample of Slovak and Polish documents. Eventually, all MARCELL sub-corpora will undergo this clustering and thematic document linking approach.

#	**Slovak**	**Polish**	**Distance**
1.	Toto nariadenie nadobúda účinnosť dňom 1. januára 1952.	Rozporządzenie wchodzi w życie z dniem 1 stycznia 2015 r.	0.091
2.	Poplatky sú príjmom štátnej pokladnice v rámci rozpočtu Ministerstva financií.	Opłaty stanowią dochód budżetu państwa.	0.092
3.	Inak platia primerane ustanovenia odsekov 2 a 4.	Przepisy ust. 2 i 3 stosuje się odpowiednio.	0.113

4.	Plavecká knižka sa vydáva zásadne na dobu päť rokov.	Książeczkę żeglarską wystawia się na okres 10 lat.	0.084
5.	Úrad pre normalizáciu má najmä tieto úlohy:	Do zadań formacji należy:	0.149

Table 3: Similarity distance between documents.

8.2 Semantic Micro-Alignment

The legislative text genre is highly structured and formalised. Semantic micro-alignment will be used to segment the documents into smaller units to make use of the technology employed in cross lingual semantic alignment.

First, we will make assessment of segmentation options for all seven languages: sentence splitting, comparing size of paragraphs, identification of higher-level segments in documents (e.g. articles), (manual) evaluation of comparability of text segments on micro-level.

Then, we will provide micro-alignments of semantically equal or related segments of text. In general, the task will align textual segments on the sub-document level. The technology used within the task was developed within FP7 projects XLike[26] and XLime[27] focused on cross-lingual knowledge-extraction. In particular, for this task we will use the systems and components from Wikifier (Brank et al., 2017), XLing (Rupnik et al., 2016) and EventRegistry (Leban et al., 2014) all dealing with statistical and semantic cross-lingual annotations and alignments.

The result of the task will be the aligned multilingual comparable corpus and a component in the tool chain integrated from pre-existing components, operating across all target languages for semantic alignment of sub-document text segments.

9. Sustainability

Sustainability of the project involves two aspects: continual feeding of the repository with new incoming data and ensuring time-resistance of the processing pipelines against the OS updates and other changes between hosts and environments.

For sustainability in the data collection, we opted to leave the individual crawlers out of the language processing chains. Their complexity and implementation depend on the data structuring at each source provider, the access rights granted to the project partner, the format of the published documents, the possible necessary conversions into raw texts, the rate of data updates, among others. The new data may be sent to a partner by owners based on a contractual agreement, or may be periodically (e.g. monthly) downloaded by partners from some open-access sites. Irrespective of the data acquisition procedure, the new text data shall be archived and sent (by each partner) to the single-access point language processing platform, where the corresponding language dependent processing flow will be activated.

The second aspect of sustainability refers to containerisation of the language specific processing flows. Members of the consortium provided 7 language specific pipelines that will be "dockerized" and assembled into a single-access point environment. By using Docker[28] containers and embedding all the necessary runtime libraries, independence of any uncontrolled external updates at OS level is achieved. The single access point will receive an archive with text documents and their language ID. The contents of the archives will be transferred to the specific "dockerized" language processing chain. Each of the language processing flows has the same input-output behaviour: they receive a collection of text documents in the specific language and output a collection of processed documents. In case of improvements to processing pipelines for certain languages, consortium members have the ability to provide an updated container which will replace the previous one, without interfering with other processing flows. Furthermore, the use of containerization enables scalability of processing resources with the number of new documents, by instantiating as many containers as needed to allow for efficient parallel processing.

After the language specific processing the documents are archived and sent to the next processing hub for the multilingual clustering and comparable documents semantic alignment phase.

The output of these processing services, together with the raw data, will be sent to the ELRC-Share, the repository of language resources developed and maintained by ELRC[29], where the seven sub-corpora and the Croatian-English parallel corpus have already been uploaded.

10. Conclusions

We have described the composition and processing of a large comparable corpus in seven EU-official yet under-resourced languages: Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian, containing the total body of national legislative documents. This corpus is the major result of the running CEF-project MARCELL. The metadata and the annotations are uniformly provided for each language specific sub-corpus. The annotations follow the CoNLL-U Plus format with four additional MARCELL-specific columns. Beside the standard morphosyntactic analysis (lemmatization and PoS/MSD-tagging), named entity and dependency and/or noun phrase annotation, the corpus is enriched with the IATE and EuroVoc labels for some languages and the same processing for the rest of the languages is under way.

An additional result of the MARCELL project is the Croatian-English parallel corpus of at least 1800 documents of national legislation translated into English and delivered in TMX format.

[26] http://xlike.org
[27] http://xlime.eu
[28] https://www.docker.com
[29] https://www.elrc-share.eu/

We strongly believe that this highly enriched corpus will represent a valuable basic language resource for different kinds of linguistic research, starting with more traditional (e.g. contrastive linguistic issues) up to more contemporary ones (e.g. cross-lingual legal terminology extraction, cross-lingual entity mapping or neural machine translation training).

11. Acknowledgements

The work reported here was supported by the European Commission in the CEF Telecom Programme (Action No: 2017-EU-IA-0136). We wish to thank the following colleagues for their valuable work in the project: Tsvetana Dimitrova, Dimitar Georgiev, Valeri Kostov, Nikola Obreshkov, Valentina Stefanova, Tinko Tinchev, Božo Bekavac, Matea Filko, Daniela Katunar, Ivana Simeon, Krešimir Šojat, Vanja Štefanec, Dávid Halász, Balázs Indig, Ágnes Kalivoda, Noémi Vadász and Gregor Leban.

12. Bibliographical References

Brank, J., Leban, G., Grobelnik, M. (2017). Annotating Documents with Relevant Wikipedia Concepts. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*.

Boroș, T., Dumitrescu, Ș.D., Burtica, R. (2018). NLP-Cube: End-to-End Raw Text Processing With Neural Networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics. pp. 171–179.

Coman, A., Mitrofan, M., Tufiș, D. (2019). Automatic identification and classification of legal terms in Romanian law texts. In *Proceedings of the International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR 2019)*, pp. 3–12.

Garabík, R., Šimková, M. (2012). Slovak Morphosyntactic Tagset. *Journal of Language Modelling*, 0(1): 41–63.

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., Makrai, M. (2019). One format to rule them all – The emtsv pipeline for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop*, pp. 155–165, Florence, Italy.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian.* (PhD Thesis) Romanian Academy, Bucharest.

Johnson, J., Douze, M., Jégou, H. (2017). *Billion-scale similarity search with GPUs*. arXiv preprint arXiv:1702.08734.

Kieraś, W., Woliński, M. (2017). Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, XCVII(1):75–83.

Koeva, S.,. Dimitrova, T. (2014). Rule-based Person Named Entity Recognition for Bulgarian. Slavic Languages in the Perspective of Formal Grammar. In: *Proceedings of FDSL 10.5*, Series Linguistic International, vol. 37, pp. 121–139, Peter Lang.

Koeva, S., Genov, A. (2011). Bulgarian Language Processing Chain. In *Proceeding of the Workshop on the Integration of Multilingual Resources and Tools in Web Applications*, Hamburg.

Krek, S. et al. (2017). *Training corpus ssj500k 2.0. Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1165.

Leban, G., Fortuna, B., Brank, J., Grobelnik, M. (2014). Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 107–110.

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede.

Marcińczuk, M., Kocoń, J., Gawor, M. (2018). Recognition of Named Entities for Polish-Comparison of Deep Learning and Conditional Random Fields Approaches. In M. Ogrodniczuk and Ł. Kobyliński (eds.): *Proceedings of the PolEval 2018 Workshop*, pp. 63–73, Institute of Computer Science, Polish Academy of Science, Warszawa.

Păiș, V., Tufiș, D., Ion, R. (2019). Integration of Romanian NLP tools into the RELATE platform. In *Proceedings of the International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR 2019)*, pp 181–192.

Peng, Q., Dozat, T., Zhang, Y., Manning, C.D. (2018). Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 160–170.

Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., Grobelnik, M. (2016). News across languages – cross-lingual document similarity and event tracking. *Journal of Artificial Intelligence Research* 55:283–316.

Rybak, P., Wróblewska, A. (2018). Semi-Supervised Neural System for Tagging, Parsing and Lemmatization. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 45–54. Association for Computational Linguistics.

Schwenk, H., Douze, M. (2017). Learning Joint Multilingual Sentence Representations with Neural Machine Translation, In *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

Sennrich, R., Haddow, B., Birch, A. (2015). *Improving Neural Machine Translation Models with Monolingual Data*. CoRR, https://arxiv.org/abs/1511.06709.

Steinberger, R., Pouliquen, B., Widiger,A., Ignat, C., Erjavec, T., Tufiș, D., Varga, D. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th LREC Conference*, Genoa, Italy, 22-28 May, 2006, pp. 2142–2147.

Steinberger, R., Ebrahim, M., Turchi, M. (2012). JRC EuroVoc Indexer JEX-A freely available multi-label categorisation tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 798–805.

Straka, M., Hajič, J., Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of LREC 2016*.

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B. (2018). E-magyar – A Digital Language

Processing System. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1307–1312, Miyazaki, Japan.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Nemeth, L., Tron, V. (2005). Parallel corpora for medium density languages. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, & N. Nikolov (Eds.), *Proceedings of the RANLP 2005 (Recent Advances in Natural Language Processing)*, pp. 590–596, Borovets, Bulgaria.

Waszczuk, J. (2012). Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pp. 2789–2804, Mumbai, India.