MARCELL

Agreement number: INEA/CEF/ICT/A2017/1565710
Action No: 2017-EU-IA-0136

Deliverable 2:
Domain Specific Classification of National Legislative Corpora

Version 1.0
31/03/2021

Document information

Activity:	Activity 2: Domain Specific Classification of National Legislative Corpora
Deliverable number:	D2
Deliverable title:	Domain Specific Classification and Multilingual Clustering of National Legislative Corpora
Indicative submission date:	31/03/2021
Actual submission date of deliverable:	31/03/2021
Main Author(s):	Tamás Váradi, Nikola Obreshkov, Martin Yalamov, Marko Tadić, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraž Repar
Contributors:	Tinko Tinchev,, Bence Nyéki, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Bálint Sass, Vanja Štefanec
Version:	1.0

Version history

Version	Date	Status	Name of the responsible partner
0.1	27/02/2021	Initial version	IBL
0.9	30/03/2021	Completed	IBL
1.0	31/03/2021	Approved	ILR

Executive summary

This Deliverable presents the main outcomes of Activity 2 aiming to annotate the collected documents organized in seven large-scale monolingual corpora of national legislation with the EuroVoc descriptors and IATE terms; to classify the documents into coherent groups where each document is linked to correct EuroVoc top-level domains and to provide multilingual clusterization resulting in clusters containing similar documents from the national legislation of the seven EU countries and aligned at the EuroVoc top-level Domains. The resulting dataset – the MARCELL Legislative Corpus – includes 7 monolingual sub-corpora (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian) containing the total body of respective national legislative documents. Besides the standard morphosyntactic analysis plus named entity and dependency and noun phrase annotation, the corpora are 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 five extra columns specific to our corpora. The seven large-scale monolingual corpora of national legislation are classified and organized by EuroVoc top-level domains in thematically related sets of documents. A multilingual clusterization is provided based on a linguistic analysis of each document and on the extraction of language-specific and translation-equivalent features used for a vector-space representation of the documents. The result is multilingual clusters of similar national legislation documents, each of which is aligned to EuroVoc top-level domains. The classification and the multilingual clusterization are integrated within the dockets' infrastructure in order to guarantee the reusability and sustainability of the project outcomes.

The annotated, classified and clusterized corpora represent a rich and valuable source for further studies and developments in machine translation, machine learning, cross-lingual terminological data extraction and classification.

Table of Contents

1. Description and Location of the MARCELL Legislative Corpus 5
2. Annotation of National Legislation with IATE Terms and EuroVoc descriptors 6
1.1 The Bulgarian annotation 6
1.2 The Croatian annotation 6
1.3 The Hungarian annotation 7
1.4 The Polish annotation 8
1.5 The Romanian annotation 8
1.6 The Slovak annotation 9
1.7 The Slovenian annotation 9
2. Format and Annotation Conventions 9
3. Metadata 11
3. Document Classification: Seven Monolingual Corpora Linked by EuroVoc Classes 13
3.1 The Classification of Bulgarian Documents 13
3.2 The Classification of Croatian documents 14
3.3 The Classification of Hungarian documents 14
3.4 The Classification of Polish documents 15
3.5 The Classification of Romanian documents 16
3.6 The Classification of Slovak documents 17
3.7 The Classification of Slovenian documents 17
4. Multilingual Clusterization: Aligning at the EuroVoc Top-level Domains 18
5. Bibliographical references 19

1. Description and Location of the MARCELL Legislative Corpus

The present deliverable introduces the MARCELL Legislative Corpus – a collection of all effective national legislation from Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia in seven large-scale linguistically processed monolingual subcorpora, classified and organized by EuroVoc top-level domains in thematically related sets of documents and grouped by means of the multilingual clusterization. According to the legislation in all these countries, such texts are free of any intellectual property restrictions.

The subcorpora have been deposited in ELRC-SHARE using "Provide a URL" method. The data are available at:

- MARCELL BULGARIAN SUBCORPUS AT ELRC-SHARE
- MARCELL CROATIAN SUBCORPUS AT ELRC-SHARE
- MARCELL HUNGARIAN SUBCORPUS AT ELRC-SHARE
- MARCELL POLISH SUBCORPUS AT ELRC-SHARE
- MARCELL ROMANIAN SUBCORPUS AT ELRC-SHARE
- MARCELL SLOVAK SUBCORPUS AT ELRC-SHARE
- MARCELL SLOVENIAN SUBCORPUS AT ELRC-SHARE

Table 1 presents the basic statistics of each of the subcorpora; their detailed description follows below.

Language	bg	hr	hu	pl	ro	sk	sl
Documents [k]	29	33.2	26	27.49	163.23	13	25
Paragraphs [k]	1313	9287	84	1589	NA	2346	7514
Sentences [k]	2731	9625	718	2145	4434	2520	8722
Tokens [M]	61	63	31	46	412	33	128
Raw size [MiB]	612	624	300	330	2662	224	802
Time span	1946–2021	1990-2019	1991–2021	2000–2021	1990–2021	1993–2021	1974–2021

Table 1: Basic information about the subcorpora.

In the v2 version, all the subcorpora add another column to the CONLLU+ files (EUROVOCMT) and an EuroVoc field to the document metadata.

2. Annotation of National Legislation with IATE Terms and EuroVoc descriptors

The annotation of the national legislation documents with EuroVoc descriptors and IATE terms serves the purpose of classification and clusterization of multilingual documents. The annotation is based on the tagging and lemmatization obtained under Activity 1. The single and multiword term detection and annotation rely on techniques already available for different languages.

1.1 The Bulgarian annotation

The MARCELL Bulgarian subcorpus consists of 29,648 documents (at the end of March 2021), which are classified into fifteen types: Administrative court; Agreements; Amendments: Legislative acts; Compacts; Conventions; Decrees; Decrees of the Council of Ministers; Decisions of the Central Election Committee; Decisions of the Constitutional Court; Decisions of the Council of Ministers; Guidelines; Instructions; Laws (Acts); Memorandums; Resolutions. The corpus is a selection from a larger legal domain dataset and contains universally binding legal acts. The time span of the documents is 1946–2020.

The data has been retrieved from the Bulgarian State Gazette (http://dv.parliament.bg), the Bulgarian government official journal, publishing documents from the official institutions such as the government, the National Assembly of Bulgaria, the Constitutional Court, etc. A C++ based NLP Pipeline for Bulgarian, constructed such as to answer the requirements of the project for autonomy and sustainability, is continuously feeding the Bulgarian corpus with newly issued legislative documents. Data is extracted from a single web source and further transformed. The transformation phase makes changes to the data format, filters document types, organises data in structures, and accumulates data with metadata and linguistic information. The annotation modules of the pipeline integrate a sentence splitter, a tokenizer, a part-of-speech tagger, a lemmatizer, a UDs parser, a named entity recogniser, a noun phrase parser, a EuroVoc descriptor annotator and an IATE term annotator. The sentence splitter, the tokenizer, the part-of-speech tagger and the lemmatizer are organised in a chain: Bulgarian Language Processing Chain – BGLPC (Koeva and Genov, 2011). The data is dependency parsed with NLP-Cube (Boroș et al., 2018). An additional annotation tool was developed to annotate IATE terms and EuroVoc descriptors within the corpus (Koeva et al. 2020). A preliminary disambiguation excluding IATE terms which are expected to be inappropriate in every context was performed. Additional annotation of the EuroVoc MTs corresponding to annotated EuroVoc descriptors was added to ensure a better correspondence between the individual parts of the multilingual corpus.

1.2 The Croatian annotation

The Croatian corpus consists of 33,245 documents that represent the national legislation from 1990 until October 2019. The corpus is composed of legally binding acts (laws, regulations, decisions, orders, etc.) and internally binding acts (ordinances, recommendations, etc.). There are 12 different text types with ordinances (11,354), decisions (7,626) and laws (3,637) 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)[1], which has, as a part of its mission, the securing of online accessibility to all Croatian legal documentation, we received the final data set from their database in October 2019 and we are presenting the figures of that current state.

Regarding the copyright issues, at the webpage of RDD[2] there is a statement: "Information for reuse on the website of the Central Catalogue of Official Documents of the Republic of Croatia of the Central Office for the Development of the Digital Society is available to users without restrictions and for free use with Open license. The Open licence shall allow the user to use any information to which it relates, including the spatial and temporal unlimited, free of charge, not exclusive and personal right to use the information subject to the licence. The open licence relates both to the content and structure of the dataset in question representing public sector information, as well as to metadata relating to the information concerned."

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 (Padró et al., 2014): paragraphs and sentences are split, tokens are identified and morphologically and syntactically annotated. An annotation tool is being developed to annotate IATE terms and EuroVoc descriptors within the corpus by the way of matching these terms with SWE/MWEs in the corpus. The corpus overall size is almost 10.3 M sentences and around 102 M tokens.

1.3 The Hungarian annotation

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. There are 11 different text types in the corpus covering different kinds of legal texts: law, regulation, decree, etc. The documents were published in the period between 1991 and 2019.

The data was analysed with the e-magyar text processing system[3] (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. Additional scripts were created for extracting the necessary metadata, for converting to CoNLL-U Plus format, for annotating IATE terms and EuroVoc descriptors in the text, as well as for classifying the documents into top-level EuroVoc domains. EuroVoc MT codes corresponding to the EuroVoc descriptors were also added to the annotation.

The raw data is 31.2M tokens, the analysed corpus is 2.9GB in CoNLL-U Plus format.

According to Para 1, (4) of the Law LXXVI of 1999 The laws and other legal instruments of the state, the verdicts of the court, the decrees of the authorities, the announcements and files of the official bodies and official standards are outside the scope of the present law. The

1 http://rdd.gov.hr

2 http://www.digured.hr/Uvjeti-koristenja (translated by EU Council Presidency Translator, https://hr.presidencymt.eu)

3 http://github.com/dlt-rilmta/emtsv

data was downloaded from kozlonyok.hu, where it is explicitly stated that “anyone can directly download editions of the journal without registration.”

1.4 The Polish annotation

The Polish corpus contains 27485 documents of 21 types representing universally binding legal acts (law, regulation, 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 2000–2021 and the set covers only documents in effect.

The data were retrieved from Dziennik Ustaw[4] and Monitor Polski[5], 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 and HTML files (unfortunately not all documents are available in HTML format)to textual format , tokenized and morphologically analysed with Morfeusz[6] (Kieraś and Woliński 2017), disambiguated with Concraft2 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 EuroVocannotation.

According to the Polish law, pursuant to Article 4(1) of the Act of 4 February 1994 on copyright and related rights, normative acts and their official drafts are not subject to copyright6 and as such are in the public domain.

1.5 The Romanian annotation

The Romanian corpus contains 163,236 files, which represent the body of national legislation ranging from 1990 to 2021. This corpus includes mainly: governmental decisions, ministerial orders, decisions, decrees and laws. All the texts were obtained via crawling from the public Romanian legislative portal[7]. We have not distinguished between in force and "out of force" laws because it is difficult to do this automatically and there is no external resource to use to distinguish between them. The texts were extracted from the original HTML format and converted into TXT files. Each file has multiple levels of annotation: firstly the texts were tokenized, 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 NER tool developed at RACAI (Păiș, 2019), nominal phrases were identified also with TTL, while IATE terms and EuroVoc descriptors were identified using an internal tool (Coman et al., 2019). All processing tools were integrated into an end-to-end pipeline RELATE[8] (Păiș et al., 2019).

4 http://dziennikustaw.gov.pl/

5 http://monitorpolski.gov.pl/

6 http://sejm.gov.pl/Sejm9.nsf/komunikat.xsp?documentId=9B9295F193BFCDC6C12584A9004530E7

7 http://legislatie.just.ro/

8 http://racai.relate.ro

1.6 The Slovak annotation

The Slovak corpus contains 13 thousand documents (33 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[9] 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).

Copyright issues in Slovak Republic are regulated by the law nr. 185/2015 Z. z., act of 1 July 2015 ("Copyright Act"). Pursuant to Section 5, paragraph b, a text of legislation or a decision of public authority are not subject to copyright.

The second uploaded version of the Slovak subcorpus added texts up to 1 December 2020. The lemmatization and morphosyntactic description has been improved by adding 11 thousand manually proofread paradigms to the morphological database and statistical guesser for out-of-dictionary words has been improved by adding heuristic filtering of reconstructed lemmas. The Named Entity Recognizer has been substantially improved by using the full MorphoDiTa model (Straková et al. 2014; Garabík 2021).

1.7 The Slovenian annotation

The Slovenian corpus contains 25 thousand documents (802[10] MB in size, 128 M tokens), ranging from 1974 to 2018. The data was obtained from the Slovenian Open Data Portal[11]. 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 (Logar et al., 2012), and lemmatized, tagged and dependency parsed with a fork[12] 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. The legislation is published in the Slovenian Open Data Portal under the CC-BY 4.0 license.

2. Format and Annotation Conventions

The MARCELL Legislative Corpus delivers the data in two formats: the CoNLL-U Plus format and XML files. Each language specific subcorpus 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 5 columns are specific to our corpora.

9 https://www.slov-lex.sk

10 The v1 version of the corpus reported a size of 5GB which was the size of the original JSON file available at the Open Data portal. The format of that file changed and reporting the size of the raw text files is a more relevant piece of information. The size of the raw text data of v1 corpus was 634 MB.

11 https://podatki.gov.si/

12 https://github.com/clarinsi/classla-stanfordnlp

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[13]):

1. **ID:** Word index, integer starting at 1 for each new sentence; may be a range for multiword 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[14]
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 if it is part of a name entity ('O' otherwise; obligatory types PER, LOC, ORG; facultative types TIME, DATE)
12. **MARCELL:NP:** the BIO format annotation of the current token if it is part of a noun phrase ('O' otherwise)
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 EuroVoc descriptor level IDs if it is a term ('_' otherwise)
15. **MARCELL:EUROVOCMT:** the EuroVoc MT level IDs if there is a term related with the MT ('_' otherwise)

Each document in the corpora is uniquely identified by its identifier constructed in the form XX-ID, where XX is the language code and ID is a unique identifier within one language corpus, derived from 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-ID-p2s1 marks the first sentence in the second paragraph of the document ID in the XX corpus). Complete text of the respective sentence is included as the text attribute.

XML format follows best practices in XML representation of corpus data. Individual files are enclosed within a <text> XML tag; documents are denoted by the <doc> tag, paragraphs and sentences are marked with <p> and <s> tags, respectively. Document metadata, paragraph and sentence identifiers are marked as attributes of the corresponding XML tags; the attribute names and their values are identical to those of the CoNLL-U format. Individual

13 https://universaldependencies.org/format.html

14 https://universaldependencies.org/u/pos/index.html

tokens are enclosed in <token> tags, with the token annotation marked by XML tags named identically to the CoNLL-U columns (for compatibility reasons converted to lowercase and the colon : is replaced by underscore _).

3. 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. Alternatively, it may also be a hash code generated from the URL of the source document.
- **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 (semicolon separated if more than one)
- **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
- **status** is the legal status of the document, e.g. in effect, repealed, etc.
- **eurovoc** contains up to six EuroVoc Top Level Domains (for some languages with weights) obtained by document classification.

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. These keys can be either obligatory (each document must contain this annotation), facultative (this annotation key can be missing in some language corpora – which is not the same as containing an empty 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 facultative 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		×	×				x
status				×			
eurovoc	×	x	x	x	x	x	x

Table 2. Available annotation data in source archives.

Obligatory annotation is:

- **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
- **eurovoc** – ID(s) of the EuroVoc top-level domain(s) optionally accompanied with a weight and resulting from either automatic or manual classification

Facultative annotation is:

- **url** – address the individual document has been accessed at
- **keywords** – several keywords separated by commas or pipe characters, in the original language
- **topic** – human readable topic of the document contents, in the original language
- **eurovoc** titles – ID(s) of the EuroVoc top-level domain(s) extracted from the documents' titles
- **parent** – link to the parent document.

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.

3. Document Classification: Seven Monolingual Corpora Linked by EuroVoc Classes

Classifying the national legislation documents into EuroVoc classes serves the purpose of compiling multilingual domain-specific corpora corresponding to EuroVoc top-level domains. Only a few of the national legislations in the seven countries have been (manually) classified so far according to the EuroVoc Thesaurus (Croatian and Slovenian). However, the manual classification is limited roughly to ⅔ of the documents in Slovenian. Different partners implemented different approaches to classification. The initial task was to categorise national legislation documents with the JRC EuroVoc indexer software – JEX (Steinberger et al., 2012). The high number of categories used by JEX Indexer, combined with a very unevenly balanced training set, is a big challenge for a multi-label categorisation task and even bigger for a one-label classification task. The JRC EuroVoc indexer JEX was used both for a multi-label classification of national legislation documents according to the multilingual EuroVoc thesaurus and as a training and testing source for a deep-learning classification.

The result of the classification is seven large-scale monolingual corpora of national legislation organized by EuroVoc top-level domains in thematically related sets of documents (linked by EuroVoc classes). The classification is made available as a part of the pipe-lines of all languages in order to guarantee sustainability and reusability.

3.1 The Classification of Bulgarian Documents

We made several experiments with: a) JEX indexer; b) Neural models (built with TensorFlow[15] and Keras) using JEX dataset or an unbalanced training dataset for Bulgarian annotated with IATE terms and EuroVoc descriptors; c) Statistical method measuring the domain-specific IATE terms and EuroVoc descriptors within the documents; d) Classification predictive modeling based on document title. The Statistical classifier was combined with term extraction, grouping of the primary document and its secondments and the Classification predictive modeling based on document titles and the results of the experiments were reported in Obreshkov et al. (2020).

It is well known that the quality of a training dataset depends not only on the amount of the annotated documents but also on the relatively equal number of documents classified to each class and on how the documents were chosen to allow clear differentiation between the classes. Meeting these conditions predetermined to a great extent the preference for simple methods based on the distribution of characteristic words.

The Statistical classifier groups legislative documents containing Eurovoc terms related to one Top Level Domain. In addition, IATE pointers to Eurovoc Micro Thesauruses or Top Level Domains are taken into account if a particular term is not presented in EuroVoc. The Statistical classifier is designed to work as a multi-label classifier providing confidence measures for the correctness of assigned classes. It relies on data pre-processing, which is part of the Bulgarian Language Processing Chain: tokenization, PoS Tagging and Lemmatization and EuroVoc descriptors and IATE terms annotations following the priority of the longest match and the first match. The minimum and maximum number of labels as well as the confidence threshold can be set as parameters (in our case 1, 6, and 0,12).

[15] https://www.tensorflow.org

The second classification method (Classification predictive modeling based on document titles) complements the Statistical classifier. It relies on parsing the titles of legislative documents which have a specific structure: with the legislative category into which the document fits at the beginning, and the differentiators for the document describing its content at the end. The differentiators are processed both as bag-of-words neglecting word order and detecting common phrases which are treated as single units and as ordered sequences. Data pre-processing with the Bulgarian Language Processing Chain is also applied. The Classification predictive modeling based on titles of documents results in a single label assignment to one of the EuroVoc Top Level Domains.

Two manually annotated datasets were developed for the evaluation: the first one comprises 517 documents which are distributed relatively equally among the 21 EuroVoc Top Level Domains; the second one consists of 667 documents and an unbalanced number of documents are classified to up to three Top Level Domains.

The Statistical modeling and the Classification predictive modeling are integrated in the Bulgarian pipe-line. It was proven that the JEX Indexer can be easily integrated as well. The Neural model with the best evaluation results is also one of the outcomes of the project. Our results are: for JEX indexer with the Unbalanced dataset (F1 0.41%, precision 0.42% and recall 0.40%); for Statistical modeling with the Unbalanced dataset (F1 0.49%, precision 0.41% and recall 0.60%); for Statistical modeling and Modeling on titles with the Unbalanced dataset (F1 0.48%, precision 0.35% and recall 0.54%).

3.2 The Classification of Croatian documents

The EuroVoc top-level domains are manually assigned for all Croatian legislative documents. Some documents are classified to more than one label and the number of labels per document is up to 14.

3.3 The Classification of Hungarian documents

The Hungarian EuroVoc classification tool makes use of a FastText classifier (Joulin et al., 2016). The classifier was trained on the Hungarian subcorpora of the ACQUIS (Steinberger et al., 2006) and OPOCE (Publications Office of the European Union) multilingual corpora[16], a sample of 1200 documents from the Hungarian MARCELL legislative subcorpus, which were manually classified into EuroVoc top-level domains, as well as the documents from the Croatian MARCELL legislative subcorpus that had been annotated with EuroVoc top-level domains. Thus, a bilingual classifier was trained although it was only aimed at classifying the texts of the Hungarian MARCELL legislative subcorpus. The EuroVoc descriptors assigned to the documents of the ACQUIS and OPOCE subcorpora were converted into EuroVoc top-level domains using the EuroVoc thesaurus hierarchy.[17]

For training the classifier, pre-trained word and character n-gram vectors were exploited (Bojanowski et al., 2017). These vectors were trained on a dataset consisting of the following resources apart from the subcorpora mentioned above: the remaining texts of the

[16] The subcorpora can be downloaded from https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer

[17] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

Hungarian MARCELL legislative subcorpus that were not manually annotated, the Wikipedia subcorpus of the Hungarian Webcorpus 2.0[18] (Nemeskey, 2020), the Croatian Translations of the Acquis Communautaire[19] and the Croatian Newscrawl and Wikipedia subcorpora from 2016 of the Leipzig Corpora Collection[20] (Goldhahn et al., 2012).

The performance of the classifier was evaluated twice. Firstly, 10-fold cross validation was done on the dataset collected for training the classifier. In this case, an average precision of 82% and a recall of 80% were achieved. Secondly, cross-validation was done by splitting the 1200 manually annotated documents from the Hungarian MARCELL legislative subcorpus into 10 parts and then in each step adding 9 of these 10 parts to the other Hungarian and Croatian documents for training. Then the model performance was evaluated on the remaining annotated texts from the Hungarian MARCELL legislative subcorpus. In this case, the model performance could be observed only on the documents represented in the Hungarian MARCELL legislative subcorpus. The average results are the following: precision of 62%, recall of 59%. The results given above were achieved by predicting maximum 5 labels to each document with probabilities larger than 0.35.

When the whole Hungarian MARCELL legislative subcorpus was classified into top-level EuroVoc domains, at least one label was assigned to each document: when there was no label with a probability greater than 0.35, the label with the highest probability was chosen.

The Hungarian pipeline that includes the classifier can be downloaded as an archived docker image.[21]

3.4 The Classification of Polish documents

The Polish TLD Classifier implements two approaches to identifying EuroVoc domain depending on the availability of auxiliary metadata use in the process of classification:

- For documents scrapped from ISAP[22] we use topic keywords available for documents provided by this source;
- For documents with missing or unknown keywords we use text-based classification.

Keywords based TLD annotation

For documents described with ISAP keywords we use a vector-based similarity algorithm with embeddings created with the Language-Agnostic BERT Sentence Embeddings language model.[23]

In this approach a vector representation is first created for each ISAP keyword and each EuroVoc descriptor (we use Polish and English labels of descriptors).

[18] https://hlt.bme.hu/en/resources/webcorpus2

[19] http://meta-share.ffzg.hr/repository/browse/croatian-translations-of-acquis/547866326c1811e28a985ef2e4e6c59e6758e8d15e7a445e9471e185a758b50c/

[20] https://wortschatz.uni-leipzig.de/en/download/

[21] http://corpus.nytud.hu/marcell-share/docker/

[22] https://isap.sejm.gov.pl/isap.nsf/ByYear.xsp

[23] https://ai.googleblog.com/2020/08/language-agnostic-bert-sentence.html

Using such representations we can find EuroVoc descriptors most similar to specific ISAP keywords and then determine their EuroVoc domains. Six domains with top similarity scores between keyword and descriptor are selected. The algorithm uses a minimum similarity threshold of 0.5 which means that in some cases the number of finally selected EV domains can be smaller than six.

In the document header section, the TLD classification weights obtained with this method are marked as "sim" because the scores represent embeddings similarity, ex.:

keywords = prywatyzacja | stocznie morskie

eurovoc = 12/1.0 36/0.7894 48/0.7441 52/0.7185 56/0.7158 40/0.712

tld_score = sim

Title based TLD annotation

For documents without ISAP keywords or when similarity of all most similar domains is 0.5 or below we use a title-based multilabel classified based on the HerBERT model[24]. Each output of the neural network represents one EuroVoc domain and the probability that the title of a document points to a given domain. As for the keyword-based classifier, only probabilities higher than 0.5 are taken into account, when all of them are 0.5 or lower the highest scored domain is taken. Thus this approach can return from 1 to 6 selected EV domains. We use documents with the best similarity scores from keywords approach as a training dataset in this approach. The titles of such annotated documents and their selected domains are then used to fine-tune the HerBERT model.

Documents classified with this approach have TLD classification weights marked as "prob" because the scores represent probability that document title point to specific domain, ex.:

 # title = Komunikat Ministra Infrastruktury i Rozwoju z dnia 14 stycznia 2014 r. w sprawie zmienionej listy projektów indywidualnych dla Programu Operacyjnego Innowacyjna Gospodarka, 2007-2013

 # eurovoc = 16/0.8367

 # tld_score = prob

3.5 The Classification of Romanian documents

The EuroVoc classification model for Romanian language was trained using the FastText tool and exploiting pre-trained word embeddings (Păiș and Tufiș, 2018) based on the CoRoLa (Mititelu et al., 2019) corpus. For training we used the Romanian part of the ACQUIS (Steinberger et al., 2006) and OPOCE (Publications Office of the European Union) multilingual corpora, which were initially used for training the JEX tool.

We performed parameter tuning using grid search and cross-validation on 10 splits of the training corpora. Our results indicate a 6% increase in F1 score compared to the JEX tool for EuroVoc ID classification (our results are F1 53.53%, precision 50.93% and recall 56.41%). The top-level domain classification of the documents, as needed within the MARCELL project, achieved an F1 score of 70.80% (with precision 64.90% and recall 77.89%). We also performed an evaluation on the intermediate MT labels, part of the EuroVoc hierarchy which provided an F1 score of 61.83%.

[24] https://huggingface.co/allegro/herbert-base-cased

The model is served using a changed version of FastText supporting serving trained models. This implementation can be downloaded from our github repository[25]. The model can be tested within the RELATE platform[26] and allows specification of maximum number of labels and the minimum threshold. The pre-trained model can also be downloaded in BIN[27] or VEC[28] format.

3.6 The Classification of Slovak documents

We made several experiments with: a) JEX indexer; b) Statistical method measuring the domain-specific IATE terms and EuroVoc descriptors within the documents (adapted from the Bulgarian annotation); c) FastText classifier (adapted from the Romanian annotation) based on word embeddings trained on CommonCrawl corpus (Grave et al. 2018).

The Statistical classifier groups legislative documents containing Eurovoc terms related to one Top Level Domain. In addition, IATE pointers to Eurovoc Micro Thesauruses or Top Level Domains are taken into account if a particular term is not presented in EuroVoc. The Statistical classifier is designed to work as a multi-label classifier providing confidence measures for the correctness of assigned classes. It relies on data pre-processing, which is part of the Slovak Language Processing Chain: Tokenization, PoS Tagging and Mophosyntactic Description, Lemmatization and EuroVoc descriptors and IATE terms annotations following the priority of the longest match and the first match. The minimum and maximum number of labels as well as the confidence threshold can be set as parameters (in our case 0, 6, and 0.1).

The JEX indexer has F1 49.42%, precision 47.05%, recall 52.04%. The FastText classifier achieves F1 54.26%, precision 51.67% and recall 57.13% for EuroVoc ID classification.

Both JEX and the Statistical classifier are fully integrated into the Slovak processing pipeline, selectable during Docker build stage; the FastText classifier is implemented offline, external to the pipeline.

3.7 The Classification of Slovenian documents

We made experiments with: a) JEX indexer; b) Statistical method measuring the domain-specific IATE terms and EuroVoc descriptors within the documents (adapted from the Bulgarian annotation).

The Statistical classifier groups legislative documents containing EuroVoc terms related to one Top Level Domain. In addition, IATE pointers to EurovVoc Micro Thesauruses or Top Level Domains are taken into account if a particular term is not presented in EuroVoc. The statistical classifier is designed to work as a multi-label classifier providing confidence measures for the correctness of assigned classes. It relies on data pre-processing, which is part of the Slovenian Language Processing Chain: Tokenization, PoS Tagging and Mophosyntactic Description, Lemmatization and EuroVoc descriptors and IATE terms

25 https://github.com/racai-ai/ServerFastText

26 https://relate.racai.ro/index.php?path=eurovoc/classify

27 https://relate.racai.ro/resources/EUROVOC/eurovoc.ro.bin

28 https://relate.racai.ro/resources/EUROVOC/eurovoc.ro.vec

annotations following the priority of the longest match and the first match. The minimum and maximum number of labels as well as the confidence threshold can be set as parameters (in our case 0, 6, and 0.1).

We have compared the performance of the classifiers against the manually annotated Slovenian documents. The default JEX configuration assigned correct labels to 3801 documents out of the total 9652 documents with manually assigned EuroVoc descriptors, while the statistical method was able to assigned correct labels to 4398 documents.

The Statistical classifier are fully integrated into the Slovenian processing pipeline, which checks whether a manually assigned label exists. If a manually assigned label is not present, the pipeline runs the classifier to automatically assign the label.

4. Multilingual Clusterization: Aligning at the EuroVoc Top-level Domains

The multilingual clusterization is based on a linguistic analysis of each document that allows extracting language-specific and translation-equivalent features used for a vector-space representation of the documents. The goal is to align the MARCELL multilingual corpora at the top level EuroVoc domains.

A system was designed to automate the clusterization of MARCELL comparable corpora in CoNLL-U Plus format. The system operates in four consequent modules: preprocessing, transformation, documents' vector representation and clusterization, which are controlled via a single configuration file. Among other technical details the configuration file specifies: the languages, the resources for the multilingual annotation, the documents that will be transformed and clustered, and the desired number of clusters, which ensures the reusability of the system for different languages and with different resources.

In our experiments the comparable corpora are transferred into a monolingual one, then they are represented by different methods and clustered by different algorithms. For the purpose of the monolingual transformation the system supports four types of resources: the EuroVoc descriptors, the IATE terms, a multilingual lexicon of multiword terms and a multilingual lexicon of single words. During the preprocessing module the resources are converted into strings of lemmas. During the transformation module the documents are converted into strings of lemmas from the available resources. To ensure the correct transformation a lemma represented in the different resources is associated only once according to the following priority: EuroVoc, IATE, Multiword Term Lexicon and Single Word Lexicon, while to select among the overlapping entries in a particular resource the longest match is applied. Documents' vector representation and the documents' clusterization are performed with the scikit-learn library[29], which offers different algorithms for clusterization: K-Means, Gaussian Mixture Models, etc. In order to improve the scalability for a large number of clusters the MiniBatchKMeans[30] version is used. The ratio of monolingual versus multilingual clusters, as well as the Silhouette Index, the Davies–Bouldin Index and the Calinski–Harabasz Index, are calculated for the evaluation of the clusterization results.

[29] https://scikit-learn.org/stable/

[30] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html

Complementary, for Bulgarian the distribution of primary acts and their amendments among clusters is calculated. Every cluster is associated with up to six EuroVoc top-level domains along with their relevance weights.

The result is multilingual clusters of comparable national legislation documents, each cluster aligned to EuroVoc top-level domains or the MARCELL multilingual corpora aligned at the top level EuroVoc domains. The multilingual clustering is integrated within the dockers' infrastructure in order to guarantee the reusability and sustainability of the project outcomes. Each update of the multilingual corpora restarts the clusterization docker.

The results of the clusterization and the alignment of the MARCELL multilingual corpora at the top level EuroVoc domains are visualised with a user interface.

5. Bibliographical references

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. arXiv preprint arXiv:1607.04606.

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 Modeling, 0(1): 41–63.

Garabík, R. (2021). Rozpoznávanie pomenovaných entít v slovenčine (demo). In: Slovenská reč. In print.

Goldhahn, D., Eckart, T., Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In Proceedings of the 8th International Language Resources and Evaluation (LREC'12).

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.

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.

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. arXiv preprint arXiv:1607.01759.

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

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.

Koeva, S., Obreshkov, N., Yalamov, M. (2021). Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 6988–6994.

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

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.

Obreshkov, N., M. Yalamov, Sv. Koeva. (2020). Categorisation of Bulgarian Legislative Documents. Proceedings of the Fourth International Conference “Computational Linguistics in Bulgaria” (CLIB 2020), Sofia: Institute for Bulgarian Language, pp. 53–-62.

Mititelu, V., Tufiș, D., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019) Little Strokes Fell Great Oaks. Creating CoRoLa, The Reference Corpus of Contemporary Romanian. In Revue Roumaine de linguistique, No./Issue 3.

Nemeskey, D. M. (2020). Natural Language Processing methods for Language Modeling. PhD thesis. Eötvös Loránd University.

Padró, L., Agić, Ž., Carreras, X., Fortuna, B., García-Cuesta, E., Li Zhixing, Štajner, T., Tadić, M. (2014) Language Processing Infrastructure in the XLike Project. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014), pp 3811-3816.

Păiș, V., Tufiș, D. (2018) Computing distributed representations of words using the CoRoLa corpus. Proceedings of The Romanian Academy Series A, Vol. 19, No. 2, pp. 403-409.

Păiș, V. (2019). Contributions to semantic processing of texts; Identification of entities and relations between textual units; Case study on Romanian language. (PhD Thesis) Romanian Academy, Bucharest.

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.

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.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiș, D., Varga, D. (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+languages. In Proceedings of The 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.

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.

Straková, J., Straka, M., Hajič, J.: Open-source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association forComputational Linguistics, Baltimore, Maryland. 2014. pp. 13 – 18.

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.

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.