Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 100–108
Marseille, 20-25 June 2022

Introducing the CURLICAT Corpora: Seven-language Domain Specific Annotated Corpora from Curated Sources

Tamás Váradi[1], Bence Nyéki[1], Svetla Koeva[2], Marko Tadić[3], Vanja Štefanec[3], Maciej Ogrodniczuk[4], Bartłomiej Nitoń[4], Piotr Pęzik[5], Verginica Barbu Mititelu[6], Elena Irimia[6], Maria Mitrofan[6],Vasile Păiș[6], Dan Tufiș[6], Radovan Garabík[7], Simon Krek[8], Andraž Repar[8]

[1]Hungarian Research Centre for Linguistics, Budapest, Hungary, {varadi.tamas,nyeki.bence}@nytud.hu
[2]Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia, Bulgaria, svetla@dcl.bas.bg
[3]University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia,
{marko.tadic,vanja.stefanec}@ffzg.hr
[4]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland,
{maciej.ogrodniczuk,bartlomiej.niton}@ipipan.waw.pl
[5]University of Łódź, Poland, pezik@uni.lodz.pl
[6]RACAI, Bucharest, Romania, {vergi,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@ijs.si, andraz.repar@cjvt.si

Abstract

This article presents the current outcomes of the CURLICAT CEF Telecom project, which aims to collect and deeply annotate a set of large corpora from selected domains. The CURLICAT corpus includes 7 monolingual corpora (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian) containing selected samples from respective national corpora. These corpora are automatically tokenized, lemmatized and morphologically analysed and the named entities annotated. The annotations are uniformly provided for each language specific corpus while the common metadata schema is harmonised across the languages. Additionally, the corpora are annotated for IATE terms in all languages. The file format is CoNLL-U Plus format, containing the ten columns specific to the CoNLL-U format and three extra columns specific to our corpora as defined by Várádi et al. (2020). The CURLICAT corpora represent a rich and valuable source not just for training NMT models, but also for further studies and developments in machine learning, cross-lingual terminological data extraction and classification.

Keywords: national corpora, comparable corpora, domain corpora

1. Introduction

The present paper introduces the CURLICAT corpora and related resources compiled in the CEF Telecom Action[1] of the same name. The CEF Telecom project Curated Multilingual Language Resources for CEF.AT (CURLICAT)[2] aims to enhance the eTranslation system[3] developed by the European Commission through supplying seven large corpora consisting of subsets of national and/or reference corpora 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 section 3, whereas section 4 introduces the additional sources of data since it was expected that there will not be enough data available from the national corpora to meet the envisaged sizes of data for each domain. The format and annotation, as well as the metadata of the multilingual corpora are described in sections 5 and 6 respectively. In the section 7, future processing steps are described: IATE[4] annotation, the issues of anonymisation and IPR protection. Some conclusions are given in section 8.

2. Rationale and Objectives

The CURLICAT CEF Telecom project is pursued as a contribution to the ultimate goal of breaking down linguistic barriers to the creation of the Digital Single Market[5] in Europe, one pillar of which will be multilingual digital service infrastructures (such as the Online Dispute Resolution[6], the e-justice platform[7] or Europeana[8]). The eTranslation system, itself a digital service infrastructure, is a building block that will help to make these infrastructures become truly multilingual. The eTranslation system faces the daunting task of supplying quality machine translation (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 CURLICAT corpus meets the above requirements on several counts. It supplies the monolingual texts from several selected domains. The choice of domains is justified for training the MT systems that will support relevant DSIs. However,

[1] https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom
[2] http://curlicat.eu
[3] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
[4] https://iate.europa.eu/home

[5] https://ec.europa.eu/digital-single-market/en
[6] https://ec.europa.eu/consumers/odr/main/?event=main.home2.show
[7] https://e-justice.europa.eu/home.do?action=home
[8] https://www.europeana.eu/portal/en

the training material collected so far either by the MT development unit at the DGT or through several language resources collection campaigns (e.g. ELRC[9] and ELG[10]) did not consist of data from national corpora particularly because, surprising as it may be, parts of national corpora are not automatically available to the EC, hence, the CURLICAT corpora represent new material, which will be available through the ELRC-SHARE[11] repository.

The list of domains was not uniformly defined in advance since in different national and/or reference corpora the domain classifications use different criteria and therefore yield different categories. However, the overall list of domains could be composed of the following wider categories: science, culture, health, nature, politics, education, social issues, economy and finance.

From another perspective, the collection of corpora for selected domains may be seen as a large comparable corpus across seven languages. In addition to the standard lemmatization and morphosyntactic analysis plus named entities, all corpora will be annotated for the IATE terms in respective languages.

3. Composition of the Corpora

The corpora cover domain specific subsets from the following national and/or reference corpora:

- Bulgarian National Corpus,
- Croatian National Corpus,
- Hungarian Gigaword Corpus (HGC), an extended new edition of the Hungarian National Corpus,
- Polish Open Science Metadata Corpus (POSMAC)
- CoRoLa, the reference corpus of the contemporary Romanian,
- Slovak National Corpus,
- Gigafida, a reference corpus of the Slovenian..

The planned size of each individual language corpus was at least 2 million sentences consisting of at least 20 million words. This means at least 5 million words and at least 500,000 sentences per domain if we take into account the four main domains (science, culture, health, economy and finance). However, there are some other domains covered by the collected language specific sub-corpora (politics, nature, education).

In this section we present in detail only the first versions of the seven corpora, which are exctracted from the national and/or reference corpora. Some quantitative information on that process is presented in Table 1, where the column headings are the language codes. These numbers do not represent the final statistics since parts of the corpora are still undergoing cleanup and additional resources may be added before the end of the project.

language	bg	hr	hu	pl[12]	ro	sk	sl
documents	6036	31k	630	127k	26k	5.5k	2340
sentences [M]	1.68	1.04	3.37	2.67	3.56	1.01	1.72
tokens [M]	22.81	20.77	69.36	65.01	95.10	13.02	36.76

Table 1: Basic information about the first versions of corpora as subsets of the national and/or reference corpora.

3.1 Details of the Bulgarian Corpus

The Bulgarian corpus consists of 6,036 documents with 22,809,225 tokens overall. The main source for the CURLICAT data is the Bulgarian National Corpus. However, to ensure enough copyright free documents we identified several new sources: library of scientific texts (books and PHD theses) and several other websites providing texts from required thematic domains. The linguistic annotation in the Corpus is divided into: (i) general annotation (tokenisation and sentence splitting), and (ii) detailed annotation. The detailed annotation includes morphosyntactic tagging (POS tagging and rich MSD annotation), and lemmatisation. The annotation of Bulgarian texts is further extended with noun phrases, and named entities.

The Bulgarian CURLICAT corpus is annotated with several annotation modules integrated in a pipeline (Koeva et al. 2020). 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, Genov 2011). The data is dependency parsed with NLP-Cube[13].

The accuracy of tagging with the BGLPC is 96.58%, and of lemmatization 98.31%.[14] The accuracy reported for the universal dependency parsing is as follows: 96.36 Morpho (measuring morphological attributes), 98.53 UPOS (Universal Part of Speech: measuring the correct prediction of the universal part of speech), 92.47 UAS (Unlabeled Attachment Score: measuring the linking to the correct word), and 88.93 LAS (Labelled Attachment Score: measuring both linking to another word and correctly predicting the link's label) (Boroş at al. 2018: 178).

3.2 Details of the Croatian Corpus

The Croatian corpus consists of 31,076 documents and 20,770,220 tokens overall. The corpus is composed of the subset of documents from the Croatian National Corpus[15] (Tadić, 2009) using the selection criteria such as publication source and topic coverage in accordance with the targeted domains. Application of these criteria resulted

[9] https://elrc-share.eu

[10] https://european-language-grid.eu

[11] https://elrc-share.eu

[12] The current version of the corpora includes only titles and abstracts (or their translations) in the Polish language. Full texts will be added as part of further work.

[13] https://opensource.adobe.com/NLP-Cube/index.html

[14] http://dcl.bas.bg/en/webservices/

[15] https://hnk.ffzg.unizg.hr

in the majority of documents coming from the domain of culture. The documents are selected mostly from the culture pages of several national and regional daily newspapers and from the entire issues of a specialised bi-weekly for culture. The data for other domains will be taken from additional sources (see Section 4).

The corpus was sentence-split and tokenized using a fork[16] of ReLDI tokenizer tool[17], adapted for use within the MARCELL project. Morphological and syntactic annotation was performed using the UDPipe tool (Straka et al., 2016) with pretrained v2.5 models for Croatian, while the NER was done using CLASSLA-StanfordNLP v1.0 standard Croatian NER models[18].

For these models (although with a different tokenization tool) the following scores are reported: 95.3 F1 score for lemmatization, 90.4 for morphosyntax, and 78.1 for dependency parsing (LAS). The annotation of the IATE terms by way of matching these terms with SWE/MWEs in the corpus is in progress.

3.3 Details of the Hungarian Corpus

The Hungarian corpus representing the subset of the Hungarian national corpus contains 630 documents[19] with 69,358,132 tokens. Apart from retrieving samples from the HNC, we also contacted several text providers in order to collect enough data in the domains of science, economy and culture.

The data was analysed with the e-magyar text processing system[20] (Váradi et al., 2018; Indig et al., 2019). The accuracy scores in Table 2 were originally reported in the latest paper on the updates of the e-magyar system (Simon et al., 2020).

The system was enhanced with detokenization functionality (precisely for the requirements of the previous MARCELL CEF-project since we use the data format from that project) to provide SpaceAfter=No annotation indicating no whitespace between two tokens in the original text. The corpus does not include dependency annotation, but it does contain noun phrase annotation. Additional scripts were created for extracting the necessary metadata, for converting the data to CoNLL-U Plus format, and for the annotation of IATE terms in the text. Furthermore, a new NER module was created based on the fine-tuning of huBERT (Nemeskey, 2020; Nemeskey, 2021), a neural language model that has achieved state-of-the-art results (approximately 99% of accuracy) in this task. The NER annotation is of special importance as it provides input data to the anonymisation module that is currently under development.

3.4 Details of the Polish Corpus

The current version of the Polish Open Science Metadata Corpus[21] contains over 127k documents with more than 65M tokens.

For the purposes of the project 318,088 scientific publications were acquired over the programmatic interface endpoints provided by the Library of Science platform[22]. They were mostly articles and scientific studies and less frequently reviews from 45 disciplines and 8 fields of science published by more than 400 different publishers in more than 1000 scientific journals. The data was initially imported at the metadata level into a relational database using a collector tool[23]. Although the minimum size of the data (in tokens) to be delivered is almost reached by simply including titles and abstracts (in Polish) of the collected texts, efforts were made to extract sentences from the full text PDF documents to meet threshold requirements for the number of sentences . From over 19k of full texts available with CC-BY and CC-BY-SA licences, we obtained 48M additional tokens and over 1.8M sentences.

Corpus texts were automatically annotated with state-of-the-art NLP tools for Polish within the collector framework:

- tokenized and morphologically analysed with Morfeusz2 morphological analyser (Kieraś and Woliński 2017)
- disambiguated with Concraft-pl tagger (Waszczuk, 2012) used to obtain segmentation, LEMMA and XPOS, Concraft-pl reaches 97.06% lemmatization accuracy
- named entity recognition with Liner2 (Marcińczuk et al. 2013), which reaches 0.81 F1 Final score (Wawer and Małek, 2018)
- dependency-parsed with COMBO parser (Rybak and Wróblewska 2018, Wróblewska and Rybak 2019), used to obtain UPOS (98.56 F1 score), FEATS (94.63 F1 score), HEAD and DEPREL (89.37 CLAS F1 Score) values

3.5 Details of the Romanian Corpus

The Romanian corpus contains 26k files, comprising over 95 million tokens. All the texts were obtained from the CoRoLa corpus (Tufiş et al., 2019).

The texts were extracted from the original corpus format and converted into TXT files. Each file has multiple levels of annotation: first, the texts were tokenized, lemmatized, morphologically annotated and dependency parsed using the UDPipe tool (Straka et al., 2016). Then, nominal phrases were identified using a component[24] extracted from the Tokenizing, Tagging and Lemmatizing (TTL) text processing platform developed at RACAI (Ion, 2007). Named entities were identified using a tool (Păiş et al., 2021b) trained on the LegalNERo corpus (Păiş et al., 2021a). IATE terms were identified using an internal tool (Coman et al., 2019), initially developed in the context of the MARCELL project, for annotating the Romanian sub-corpus (Tufiş et al., 2020).

The Romanian BLARK tools were recently evaluated in (Păiş et al., 2021a). According to this evaluation, the model used for annotating the Romanian corpus offers 99.88% F1 for tokenization, 95.91% lemma accuracy, 97.15% UPOS accuracy and 96.24% XPOS accuracy

[16] https://github.com/zzl-ffzg/reldi-tokeniser

[17] https://github.com/clarinsi/reldi-tokeniser

[18] http://hdl.handle.net/11356/1322

[19] The low number of documents is due to composition of the selection, which is dominated by entire books.

[20] http://e-magyar.hu

[21] http://clip.ipipan.waw.pl/POSMAC

[22] https://bibliotekanauki.pl/

[23] http://git.nlp.ipipan.waw.pl/Marcell/collector

[24] https://github.com/racai-ai/RomanianTTLChunker

(considering MSD tags), and 84.35% UAS (for dependency parsing), when evaluated on the RRT corpus. The anonymization system used in the Romanian CURLICAT corpus (Păiș et al., 2021b) is based on a NER system (Păiș et al, 2021c) which has an overall macro F1=86.84%, with individual F1 scores 98.86% for person entities, 80.89% for organizations and 76.01% for locations (these being the entities considered for anonymization). However, the anonymization system further uses textual markers, such as words written with first letter uppercase inside a sentence to further anonymize potentially unidentified entities, thus increasing the anonymization accuracy above the NER system's performance.

Annotation of IATE terms is performed using a new tool which should provide better results than the previously used tool in the MARCELL project, but there is currently no formal evaluation available. The new tool was already integrated into the RELATE portal and a paper describing it is currently under preparation.

The pipeline was integrated in the RELATE platform (Păiș et al., 2019) in order to allow high-speed parallel processing (Păiș et al., 2020) of the entire corpus. Finally, the resulting documents were exported in the CURLICAT specific format.

3.6 Details of the Slovak Corpus

The Slovak corpus contains 5,570 documents with 13 million tokens. The data is obtained as the subset from the Slovak National Corpus[25] (Garabík and Šimková 2012) corpora: a redistributable subset of the *prim-9.0* corpus, the corpus of Slovak Wikipedia and Necykopédia[26] *wiki-2018-03* and the *od-justice-1.0* corpus of court rulings[27]. The data has been converted from the original corpus format, deduplicated by Onion (Pomikálek 2011; Benko 2019), tokenized, lemmatized and morphologically annotated with the Slovak MorphoDita model (Garabík and Bobeková, 2021) and dependency parsed with UDPipe (Straka et al., 2016). Named entity recognition is annotated with the NameTag 1 recognizer (Straková et al. 2014), trained on the Slovak named entity annotated corpus (Garabík, 2021).

The NLP pipeline of Slovak starts with and depends on lemmatization and full MSD tagging, additional steps depend on the text being already lemmatized and tagged. We are using a Slovak version of MorphoDita, trained on manually lemmatized and annotated corpus[28] r-mak-6.0. The analyser reaches 93.5% accuracy (all the accuracies are measured on a mixed genre and domain general language corpus) on the lemma+tag combination; 96.9% on the lemma+POS, and 98.2% on lemmas. Neglecting differences in case, the accuracy rises to 94.0% for the lemma+tag, and 99.0% for the lemmas only. These numbers include words not present in the morphological database (out-of-vocabulary) that are lemmatized by a statistical guesser; if we limit ourselves to known words, the lemma+tag accuracy will be 94.8%, lemma+POS 97.8% and lemmas 99.1%.

Though the available data reach the target size easily, it is extremely unbalanced with respect to style, genre and domain. Therefore we performed an additional data collection task, where we identified data sources with permissive licences and collected additional documents. This includes notably a subset of the Greenie online library[29] (a portal providing access to freely accessible, though not always redistributable e-books, mostly in Slovak) and scientific journals published within the Open Science framework (unfortunately, there is no central repository of Slovak Open Science content), focusing on domains of interest.

3.7 Details of the Slovenian Corpus

The Slovenian corpus contains 2340 documents with 36 million tokens. The data was obtained from the Gigafida 2.0, which is a reference corpus of written Slovene. It comprises daily news, magazines, a selection of web texts (a certain portion of which covers news texts as well), and different types of publications (fiction, school books, and non-fiction). The texts have been selected and automatically processed with the aim of creating a corpus that represents a sample of modern standard Slovene and can be used for research in linguistics and other branches of the humanities, for compiling modern dictionaries, grammars, and learning materials, as well as for developing language technologies for Slovene. To identify those documents that can be classified into one of the domains of interest for CEF DSIs, namely culture, education, economy, finance, health, politics, we carried out a review of the documents included in the Gigafida 2.0 corpus. We were able to identify a large number of documents which allowed us to almost meet the minimum size of the data (in tokens) to be delivered, while the remainder will be procured from new text providers.

The selected samples were tokenized with the Slovenian tokenizer Obeliks4j (Grcar et al., 2012), and lemmatized, tagged and dependency parsed with a fork[30] of the StanfordNLP parser (Peng et al., 2018) trained on the ssj500k training corpus (Krek et al., 2017), which provides state-of-the-art performance[31] for the Slovene language with an accuracy of 97.06% for morphosyntactic tagging and 99.63% for lemmatization. Additional scripts have been created to extract metadata and annotate IATE terms.

4. Additional Sources of Data

Since for some domains the national and/or reference corpora in some languages couldn't provide enough data to reach the planned size of domain corpora, either for the limited number of samples or for the lack of IPR clearance, alternative sources of data had to be identified in the second step.

For each of the corpora the partners identified relevant sources of data. It should be noted that this project task is still running, so we can't present the exact figures yet, but the expected size for each domain in each language is one of the project requirements and we will certainly meet them.

[25] https://korpus.juls.savba.sk

[26] https://necyklopedia.org

[27] https://korpus.sk/OpenData.html

[28] https://korpus.sk/ver_r(2d)mak.html

[29] https://greenie.elist.sk/

[30] https://github.com/clarinsi/classla-stanfordnlp

[31] https://github.com/clarinsi/babushka-bench

For the data from additional sources, the whole chain of processing steps, described in Section 3, will have to be applied. Additional pre-processing steps are needed for the original texts files available in the formats that require them, such as text extraction from PDF or HTML.

5. Format and Annotation

The corpora use the CoNLL-U Plus format. 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 3 columns are specific to our corpora.

The columns are separated by a TAB character. There are the following columns (the detailed description of the first ten CoNLL-U columns, as well as the internal format of the file can be found at the Universal Dependencies site[32] and in (Váradi et al. 2020)):

ID FORM LEMMA UDPOS XPOS FEATS HEAD DEPREL DEPS MISC CURLICAT:NE CURLICAT:NP CURLICAT:IATE

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[33]
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. CURLICAT:NE: the BIO (Beginning-In-Out) format annotation of the current token, O if it is not part of a named entity
12. CURLICAT:NP: the BIO format annotation of the current token, O if it is not part of a noun phrase
13. CURLICAT:IATE: the annotation of a IATE term by the language-independent code if it is (part of) a IATE 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-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). The complete text of the respective sentence is included as the text attribute.

5.1 Accuracy of Lemmatisation and MSD

In this subsection we present the cumulative accuracy of lemmatisation and MSD-tagging for comparison between languages. Although different combinations of processing tools were used for different languages, it could be noted that the performance over all languages is comparable.

Language	Tagging	Accuracy
Bulgarian	lemma	0.983
	MSD	0.966
Croatian	lemma	0.953
	MSD	0.904
Hungarian	lemma	0.963
	MSD	0.963
Polish	lemma	0.971
	MSD	0.946
Romanian	lemma	0.959
	MSD	0.962
Slovak	lemma	0.982
	MSD	0.935
Slovenian	lemma	0.996
	MSD	0.971

Table 2: Accuracy of lemmatization and MSD tagging

6. Metadata

Principles of metadata encoding from CEF-project MARCELL (Váradi et al. 2020) are to be followed also in the current endeavour of creating a common metadata annotation schema. Metadata is, therefore, seen as a collection of information classified as *obligatory* (all partners have to provide it), *optional* (the field can be missing or containing an empty value in some language corpora), or *local* (annotation specific for a given language corpus, included for fidelity to the original source annotation).

Obligatory metadata fields in the future CURLICAT schema, that can be easily provided for all language corpora, are the following:

- *identifier:* is a short string uniquely identifying the document in its language corpora;
- *language:* the ISO 639-1 codes of the specific represented languages;
- *date:* the date for the creation of the document in the original source/national corpora, in ISO 8601 format;
- *title:* the human readable title of the source document, in the original language, e.g the title

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

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

of the book, chapter, paper, newspaper article etc. based on which the document was created;

- *type:* further specifies the type of the source document, in English e.g. book, chapter, paper, newspaper article, blogpost, etc.
- *source:* the name of the organization that published the source document, be it a Journal, Publishing House, Blog, Website, etc., in the original language;
- *domain:* the domain covered in the document, in English, selected from the predefined list of CURLICAT domains and based on the domain metadata fields in the source corpora;
- *no_of*: no_of_sentences, no_of_words, no_of_punctuation, no_of_tokens: the total number of sentences, words, punctuation marks and tokens (words + punctuation marks) in the document.

Some examples of *optional* metadata fields that are taken into consideration are:

- *keywords:* contains several keywords related to the content of the document;
- *url:* is the original individual address the document was accessed at, if applicable);
- *style:* the literary style of the text in the document, selected from a predefined list: imaginative, memoirs, administrative, law, journalistic, etc;
- *author:* the name/s of the person/s that created the text in the source document;
- *subdomain:* a further classification of the documents into narrower categories, e.g. scientific fields for the Science domain, or cultural fields for the Culture domain;
- *issn_isbn_eisbn*: the International Standard Serial Number or International Standard Book Number of the source document.

Some local fields that we included come from the scientific publications descriptions that comprise the Polish corpus: *title in English, abstract in English, issue volume, issue number, page range, full text license, reviewers*, etc.

Such elaborated metadata schema will allow easy selection of relevant subcorpora, using metadata value as a criterion, thus facilitating the training of different in-domain language models.

7. Future Processing Steps

7.1 Terminology Annotation

The corpora are currently being annotated for IATE terms. In the latest version of IATE database, that CURLICAT intends to use for annotation, we found the following number of terms for the respective languages[34]: Bulgarian 52,836, Croatian 33,932, Hungarian 64,780, Polish 93,003, Romanian 66,106, Slovak 60,685, Slovenian 71,921. 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.

7.2 Anonymisation

The General Data Protection Regulation (GDPR)[35] contains provisions and requirements for handling personal data of individuals, whatever their nationality or place of residence, and applies to any enterprise established in the European Economic Area (EEA). Furthermore, it considers that "the protection of natural persons in relation to the processing of personal data is a fundamental right". Different concepts need to be considered, such as: right to be forgotten, privacy by design (Spiekermann, 2012), transparency (full information is provided to individuals) and accountability (demonstrable responsibility for using personal data) (Goddard, 2017). Article 4, paragraph 5, of the GDPR, defines "pseudonymisation" as the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information. A distinction must be made between "local" pseudonymization, where a pseudonym is used within a single text to replace certain personal information, and "global" pseudonymization, where the same identifier is used in multiple texts to replace the same personal information (Francopoulo and Schaub, 2020). In the case of "global" pseudonymization, some re-identification may still be possible by combining different information from different contexts.

For the Romanian language, we considered the “local” pseudonymization approach, since this is more resilient to de-anonymization attacks. Furthermore, in order to allow more natural language processing algorithms to take advantage of the corpus, we decided to keep suffixes, specific to Romanian named entities, as part of the pseudonym being used. For this reason, we used the format _#TYPE#ID_suffix, where TYPE is the named entity type being replaced (PER,LOC,ORG,UNK), ID is the local identifier associated with the pseudonym and suffix is the corresponding suffix (if present). For example, “Maria” becomes “_#PER#1_”, while “Mariei” becomes “_#PER#1_ei”. Experiments (Păiș et al., 2021c) have shown that implementation of this anonymization scheme in the raw text is handled well by the UDPipe annotation tool (it is treated as a single token, lemmatization is performed well given the suffix, the token is correctly identified as proper noun, etc.).

For the remaining languages, we will follow a general approach where we will reuse an existing named-entity recognition (NER) model and adapt it to an anonymisation setting. The NER models are language-specific and based on the local versions of the BERT language model (e.g., huBERT (Nemeskey, 2020; Nemeskey, 2021) for Hungarian or SloBERTa (Ulčar and Robnik-Šikonja, 2021) for Slovenian). The identified entities will be replaced by placeholders of the same gender, grammatical categories (case, number) and the same inflectional paradigm.

While the anonymisation models will be developed for all languages involved in the project, anonymisation will not be applied to the entire corpus, but only where required by the text providers as a condition for including their documents in the corpus. For example, anonymisation will

[34] https://iate.europa.eu/download-iate [accessed on 2022-01-17]

[35] Council Regulation 2016/679, 2016 O.J. (L119) (EU) 1 https://eur-lex.europa.eu/eli/reg/2016/679/oj

not be performed for the Polish language documents, since they consist only of scientific texts and do not contain sensitive or personal data, and no anonymisation will be performed also for Croatian texts in the domains of science, economy and finance, for the same reason.

7.3 Intellectual Property Rights (IPR) Clearance

We selected documents for the Bulgarian corpus with licences such as: Universal Public Domain Dedication (CC0 1.0); Attribution 4.0 International (CC BY 4.0) and Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Thus, we are avoiding copyrighted material, which might limit the use of our dataset only for academic purposes.

We have chosen documents from government websites (ministries, government agencies, municipalities, European Union bodies and institutions), academic websites and their repositories granting access to full-text journal articles, other research articles, monographs, books, dissertations; ResearchGate repository with the applicable rights, the Bulgarian Portal for Open Science repository, media and NGOs (blogs, media, political bodies) websites which grant access to texts which are released under Creative Commons licences[36].

The current Croatian data are selected from daily and bi-weekly newspapers culture pages and for some of these documents the IPR are cleared, and for the rest the negotiations with publishers are in process. For other relevant domains the sources of data will be from the open access journals with permissive licences, such as CC-BY or CC-BY-SA.

All metadata in the Polish Library of Science (including titles and abstracts) are available under the CC0 licence. Full texts are available with various licences but only texts available with CC-BY and CC-BY-SA licences will be extracted to maximise the use of the resulting resource.

As the data selected for CURLICAT delivery was extracted from the Romanian national reference corpus CoRoLa which was IPR cleared for search, but not for distribution, we had to contact the text providers with new agreement proposals, asking for their permission to distribute the selected document. We selected the major data providers and sent them updated agreement forms. Out of the 62 sent letters, we received 28 positive responses. At the end of this, a number of 3,042 documents were IPR-cleared and together with the IPR-free documents they cover the necessary delivery data in the CURLICAT project. However, we continue the campaign for IPR clearance of the entire CoRoLa corpus with respect to distribution.

The documents in the main Slovak National Corpus *prim-9.0* are annotated for their licensing status, and the redistributable subset of the corpus has been selected for the CURLICAT corpus. The exact licences vary, but only those allowing redistribution have been chosen. The size of the subset is 1.7 million tokens, or 93 thousand sentences. Other corpora of the Slovak National Corpus selected for the projects are the corpus of courtroom proceedings *od-justice-1.0* (exempt from copyright protection) of 1.3 billion tokens, 40 million sentences (after deduplication; the corpus contains many similar or identical sentences, thanks to its nature); the corpus of Slovak wikipedia texts *wiki-2019-08* (47 million tokens, 4 million sentences) and the *Necyklopédia* subset of corpus *wiki-2018-03* (1 million tokens, 70 thousand sentences), both under the CC BY-SA 3.0 licence. Additional texts collected from the Greenie library and from scientific publications adhering to the Open Science principles are covered predominantly under various variants of the CreativeCommons licence. Additionally, text acquisition for the Slovak National Corpus is an ongoing process where the text providers are asked to release their work under open licensing terms, if possible.

Most of the Slovenian data was selected from the Gigafida 2.0 national corpus, the text provision agreement of which provides that 10% of the corpus can be shared under the Creative Commons Attribution-ShareAlike 4.0 (CC-BY-SA 4.0) licence. New licensing agreements, preferably via the Gigafida 2.0 pipeline, will be concluded with additional text providers to fulfil the requirements of the project.

8. Conclusions

We have described the composition and processing of the first version of domain dependent monolingual corpora in seven EU-official yet moderately under-resourced languages: Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian. These corpora are the first major result of the running CEF project CURLICAT. The metadata and the annotations are uniformly provided for each language specific corpus. The annotations follow the CoNLL-U Plus format with three additional specific columns as defined in the MARCELL project (Váradi et al. 2020). Beside the standard morphosyntactic analysis (lemmatization and PoS/MSD-tagging), named entity, the corpus is enriched with the annotations of IATE terms for some languages while the same processing for the rest of the languages is under way.

We strongly believe that this highly enriched set of domain dependent corpora 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).

9. Acknowledgements

The work reported here was supported by the European Commission in the CEF Telecom Programme (Action No: 2019-EU-IA-0034 CURLICAT).

10. Bibliographical References

Benko, V. (2019). Deduplication in Large Web Corpora. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, Mannheim, pp. 17–22.

[36] CC BY-SA 4.0: https://creativecommons.org/licenses/by-sa/4.0/deed.en, CC BY-NC-ND 4.0: https://creativecommons.org/licenses/by-nc-nd/4.0/, CC BY-NC 2.0: https://creativecommons.org/licenses/by-nc/2.0/, CC BY-NC-SA 2.5 BG, CC BY 3.0: https://creativecommons.org/licenses/by/3.0/, CC BY-SA 3.0 IGO: https://creativecommons.org/licenses/by-sa/3.0/igo/.

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.

Francopoulo, G., & Schaub, L.-P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. In *Proceedings of the Workshop on Legal and Ethical Issues (Legal2020)*, pp. 9–14.

Garabík, R. (2021). Rozpoznávanie pomenovaných entít v slovenčine – webové rozhranie. In: *Slovenská reč*,. 86(3): 402 – 405.

Garabík, R., Bobeková, K. (2021). Lematizácia, morfologická anotácia a dezambiguácia slovenského textu – webové rozhranie. In: *Slovenská reč*, 86(1):. 104–109.

Garabík, R., Karčová, A., Šimková, M., Brída, R., Žáková, A. (2016). *Skloňovanie podstatných mien v slovenčine s korpusovými príkladmi*. Vydavateľstvo Mikula, Bratislava.

Garabík, R., Šimková, M. (2012). The Slovak National Corpus and its Corpus Linguistic Resources. In *Prace Filologiczne*, vol. LXIII. Warszawa 2012. Wydział polonistyki Uniwersytetu Warszawskiego, pp. 109–119.

Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): European Regulation that has a Global Impact. *International Journal of Market Research*, 59(6), pp. 703–705.

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

Koeva S., Obreshkov N., Yalamov M. (2020). Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. In *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.

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.

Nemeskey, D. M. (2020). *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Eötvös Loránd University.

Nemeskey, D. M. (2021). Introducing huBERT. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, pages 3–14, Szeged.

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.

Păiș, V., Ion, R., & Tufiș, D. (2020). A Processing Platform Relating Data and Tools for Romanian Language. In *Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020)*, pp. 81–88.

Păiș, V., Mitrofan, M., Gasan, C. L., Ianov, A., Ghiță, C., Coneschi, V. S., & Onuț, A. (2021a). *Romanian Named Entity Recognition in the Legal domain (LegalNERo)* [Dataset]. Zenodo, doi:10.5281/zenodo.4772094.

Păiș, V., Mitrofan, M., Gasan, C. L., Coneschi, V., & Ianov, A. (2021b). Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop (NLLP 2021)*, pp. 9–18.

Păiș, V., Irimia, E., Ion, R., Tufiș, D., Mitrofan, M., Barbu Mititelu, V., Avram, A.M., Curea, E. (2021c). Romanian text anonymization experiments from the CURLICAT project. In *Proceedings of the International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR 2021)*.

Păiș, V., Ion, R., Avram, A.M, Mitrofan, M., Tufiș, D. (2021d). In-depth evaluation of Romanian natural language processing pipelines. In the Romanian Journal of Information Science and Technology (ROMJIST). vol. 24, no. 4, pp. 384--401, 2021a

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.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech republic.

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.

Simon, E., Indig, B., Kalivoda, Á., Mittelholcz, I., Sass, B., Vadász, N. (2020). Újabb fejlemények az e-magyar háza táján. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020)*, pages 29–42, Szeged.

Spiekermann, S. (2012). The Challenges of Privacy by Design. *Commun. ACM*, 55(7), pp. 38–40.

Steinberger, R., Pouliquen, B., Widiger,A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC 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 8th Language Resources and Evaluation Conference (LREC 2012)*, 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 the 10th Language Resources and Evaluation Conference (LREC 2016)*.

Straková, J., Straka, M. and Hajič, J. (2014). 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 (ACL 2014)*, pp. 13-18, Association for Computational Linguistics.

Tadić, M., (2009) New Version of the Croatian National Corpus. In Hlaváčková D., Horák A., Osolsobě K., Rychlý P. (eds.) *After Half a Century of Slavonic Natural Language Processing*, pp. 221–228, Tribun EU, Brno.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Păiş, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019). Little Strokes Fell Great Oaks. Creating Corola, The Reference Corpus of Contemporary Romanian. *Revue Roumaine de Linguistique*, no./issue 3, pp. 227–240.

Tufiş, D., Mitrofan, M., Păiş, V., Ion, R., Coman, A. (2020). Collection and Annotation of the Romanian Legal Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, pp. 2773–2777.

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.

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiş, V., Tufiş, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., & Brank, J. (2020). The MARCELL Legislative Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 3761–3768.

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 Recent Advances in Natural Language Processing conference (RANLP 2005)*, 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.

Wawer A., Małek E. (2018). *Results of the PolEval 2018 Shared Task: Named Entity Recognition*. In: M. Ogrodniczuk and Ł. Kobyliński (eds.) Proceedings of the PolEval 2018 Workshop, pp. 53–62. Institute of Computer Science, Polish Academy of Sciences.

Wróblewska A., Rybak P. (2019). *Dependency parsing of Polish*. Poznań Studies in Contemporary Linguistics, 55(2):305–337, 2019.