Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction

Andrius Utka
Vytautas Magnus University
Kaunas, Lithuania
andrius.utka@vdu.lt

Sigita Rackevičienė
Mykolas Romeris University
Vilnius, Lithuania
sigita.rackeviciene@mruni.eu

Liudmila Mockienė
Mykolas Romeris University
Vilnius, Lithuania
liudmila@mruni.eu

Aivaras Rokas
Vytautas Magnus University
Kaunas, Lithuania
aivaras.rokas@vdu.lt

Marius Laurinaitis
Mykolas Romeris University
Vilnius, Lithuania
laurinaitis@mruni.eu

Agnė Bielinskienė
Vytautas Magnus University
Kaunas, Lithuania
agne.bielinskiene@vdu.lt

Abstract

The paper aims at presenting English-Lithuanian corpora for bilingual term extraction (BiTE) in the cybersecurity domain within the framework of the project DVITAS. It is argued that a system of parallel, comparable, and training corpora for BiTE is particularly useful for less-resourced languages, as it allows efficiently to combine strengths and avoid weaknesses of comparable and parallel resources. A special focus is given to the availability of sources in the cybersecurity domain and issues related to copyright-protected publications, as well as the data curation performed for building the corpora and depositing them to CLARIN-LT repository.

1 Introduction

The model of combining several types of corpora has been chosen for the bilingual terminology extraction project DVITAS.[1] The aim of the project is to develop a methodology for automatic extraction of English and Lithuanian terms of a specialised domain from parallel and comparable corpora, as well as to create a publicly available bilingual termbase. Cybersecurity (CS) terminology has been chosen as a specialised domain for the project because of its particular relevance in today's digitalised world in which cybersecurity awareness and cyber hygiene skills are indispensable for every Internet user. The compiled termbase is believed to be valuable both for specialists of the domain and the general public, as well as drafters of legal and administrative documents, and translators.

The project aims at employing current deep learning terminology extraction methods. In 2020, the project team (Rokas et al., 2020) completed a pilot study on semi-supervised automatic extraction of Lithuanian CS terms from a Lithuanian monolingual corpus. A small-scale manually annotated dataset (66,706 word corpus with 1,258 annotated cybersecurity terms) was used as a training data. The pilot study was performed in several stages: firstly, various baseline LSTM and GRU networks were tested using the Adam optimiser and FastText embeddings; secondly, each of the best baseline LSTM and GRU networks were tested with various optimisers; and finally, the best model was compared with a model that has been trained using multilingual BERT embeddings (Rokas et al., 2020). The latter approach proved to be the most efficient: Bidirectional Long Short-Term Memory model (Bi-LSTM) using multilingual Bidirectional Encoder Representations from Transformers (BERT) embeddings reached F1 score of 78.6%.

The methodology used in the pilot study will be modified and tested on different configurations of neural networks taking into account the methods applied in related research. In studies by other scholars,

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

[1] https://klc.vdu.lt/dvitas/en

Andrius Utka, Sigita Rackevičiene, Liudmila Mockiene, Aivaras Rokas, Marius Laurinaitis and Agne Bielinskiene 2022. Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction. *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 126–138. DOI: https://doi.org/10.3384/9789179294441

who applied neural networks for term extraction as sequence labelling task and used larger annotated datasets, higher F1 score was achieved: e.g., Kucza et al. used a dataset with 78,567 annotated terms and with Bi-LSTM reached F1 score of 86.73% (Kucza et al., 2018). Other studies on sequence labeling tasks with multilingual BERT embeddings show that reduction of the number of languages to three in BERT models may help to achieve higher results compared with the ones achieved with multilingual BERT (Ulčar and Robnik-Šikonja, 2020). As deep neural network models achieve higher and higher F1 scores, they reveal their performance and effectiveness in the tasks they were trained for and prove their spot as one of the peak state-of-the-art approaches to terminology extraction.

Thus, we believe that more comprehensive training and testing data obtained from larger bilingual corpora will allow to improve the preliminary results. Precisely, that is the goal of the present paper - to present the motivation behind the idea of creating such a resource, as well as to present solutions to encountered problems and challenges.

2 Related Research

Bilingual/multilingual term extraction, which is widely used for terminographic purposes, is performed by using two types of corpora - parallel and comparable. These two types of corpora are distinguished by the nature of texts that are used to build them. A parallel corpus (bilingual or multilingual) is the one "that contains source texts and their translations", whereas "a comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness", which means that it should include "the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period" (McEnery and Xiao, 2007).

Term extraction from parallel corpora has been already applied for several decades (Kupiec, 1993). It is considered to be relatively easy, at least from a technical point of view, as in a parallel corpus, which typically consists of aligned sentences, source and target terms appear in the same aligned pair of sentences. Parallel corpora are particularly useful for translation studies as their analysis provides insights into various equivalence issues. They are also extensively used to develop machine translation (MT) systems and computer-assisted translation (CAT) tools like translation memories (TM) (McEnery and Xiao, 2007). Moreover, "specialized parallel corpora can be especially useful in domain-specific translation research" (McEnery and Xiao, 2007).

Lately, the importance of comparable data have been increasing, as more and more papers have appeared on term extraction from comparable corpora (Vintar, 2010; Delpech et al., 2012; Gornostay et al., 2012; Aker et al., 2013; Chu et al., 2016). Notably, since 2008 a number of valuable research papers on the usage of comparable corpora for term extraction have been published in Proceedings of the Workshop on Building and Using Comparable Corpora (BUCC)[2]

As the extraction of data from comparable corpora is not that straightforward and accurate as from parallel corpora, scholars have applied a variety of data extraction methods or combinations thereof. For instance, Steingrímsson et al. suggested combining three different approaches for effective bitext extraction from comparable corpora, namely combining crosslingual information extraction (CLIR), contextualised embeddings, and word alignments; this method is particularly useful for low-resourced scenarios (Steingrímsson et al., 2021). Sanjanasri et al. used Apache Spark framework for mining bilingual word pairs from a comparable corpus (Sanjanasri et al., 2021). Vintar et al. suggested applying intersections of word embeddings for mining semantic relations from comparable corpora (Vintar et al., 2020). Huidrom et al. proposed using the web as a source for building a comparable corpus for a less-resourced language pair using the heuristic approach based on sentence-length information and a bilingual dictionary when such is available (Huidrom et al., 2021). Terryn et al. presented a new approach to monolingual and multilingual term annotation and automatic term extraction based on the gold standard (Terryn et al., 2020).

Researchers indicate several important advantages of using comparable data. First, term extraction from comparable corpora provides valuable terminological data as these data reflect the usage of termi-

[2]Workshop on Building and Using Comparable Corpora (BUCC) - `https://aclanthology.org/venues/bucc/`.

nology in original languages which is much more natural than the usage of terminology in translations that are inevitably influenced by source languages. McEnery et al. also highlight that "specialised comparable corpora are particularly helpful for highly domain-specific translation tasks" (McEnery and Xiao, 2007). Another important advantage of using comparable data is the possibility to include data sources of a much larger variety, as comparable data is not limited to translated resources, which might be scarce or lack diversity, especially in cases when both the original language and the translation language are not English (Alonso et al., 2012; Delpech et al., 2012; Goeuriot et al., 2009; Morin and Prochasson, 2011; Morin et al., 2011; Rivera et al., 2013; Terryn et al., 2020). Thus, building and using comparable corpora for under-resourced languages next to parallel corpora could be very important for the analysis of such languages. And finally, comparable corpora are less expensive to build than parallel corpora as text alignment is not needed for their compilation.

Therefore, some scholars have introduced the idea of combining comparable and parallel corpora to benefit from the advantages provided by both (McEnery and Xiao, 2007; Bernardini, 2011; Morin and Prochasson, 2011; Biel, 2016; Giampieri, 2018), yet some researchers concentrate solely on comparable corpora (Steyaert and Rigouts Terryn, 2019; Vintar et al., 2020).

Thus, there is enough evidence to assume that for an efficient bilingual terminology extraction for English and Lithuanian languages, we need to build a resource consisting of parallel and comparable corpora.

3 Cybersecurity Domain and Availability of the Sources

The analysis of the cybersecurity sources revealed that this domain is highly heterogeneous and encompasses diverse types of information accumulated in various discourses. Ideally, the cybersecurity corpora should be representative of the whole cybersecurity domain and its constituent genres of texts produced in various discourses.

Wall in his study on cybercrime distinguishes four main discourses relevant to the CS domain: legislative/administrative discourse, academic discourse, expert discourse and popular, emotional or layperson's discourse (Wall, 2007). Similarly, for our corpora we distinguished legal, administrative-informative, academic, and media discourses. We ascribed expert texts written by cybersecurity practitioners to the administrative-informative discourse. The sources of these discourses were investigated and assessed for compilation of the corpora. Two most important criteria of source assessment were their suitability for compilation of a comparable corpus and a parallel corpus and their availability.

Most sources were suitable for compilation of the comparable corpus, which consists of the original texts in English and Lithuanian. Meanwhile, the sources suitable for the parallel corpus (English original texts and their translations into Lithuanian) were much more sparse. More detailed description of suitability of the sources for the parallel and comparable corpora is given in Subsection 4.1.1.

Though there were numerous sources suitable for corpora compilation, not all of them were freely available. Documents produced by national and international legislative and administrative bodies are commonly accessible without any restrictions. Meanwhile, the access to academic publications is often restricted. Most relevant academic sources are published by major publishing companies and protected by intellectual property rights. As we had to ensure proper usage of these texts, we examined the legal framework related to copyright protection and text and data mining (TDM) activities, as well as possibilities to acquire permissions to reuse relevant copyright-protected publications for corpora compilation, data extraction and storage in CLARIN-LT repository.

For a long time TDM activities have faced conservative intellectual property protection and strict restrictions (small-scale use, no possibility to develop derivative products, etc.) on the usage of legally protected sources. This situation has hampered big data projects which are necessary for development of various AI applications. Therefore, numerous studies have appeared discussing the situation and necessary changes in legal frameworks (Rosati, 2018; Sag, 2019; Flynn et al., 2020).

In the US, the notion of fair use of copyrighted works has been questioned and reinterpreted in law courts, which have ruled that copying of copyright-protected works for TDM research purposes satisfies fair use criteria and is not an infringement. The lawsuits concerned the Google Book Search Project (the

cases Authors Guild v. GOOGLE, 2015; Authors Guild v. HaithiTrust, 2014) [3].

In the UK, a TDM exception was included in the statutory amendments to copyright law which came into effect in 2014. Since 2014 Copyrights, Designs and Patents Act (Article 29A) has allowed performing TDM activities provided that the TDM practitioner has lawful access to the resource and that TDM activities are performed for non-commercial purposes [4].

Until recently, the EU did not have a uniform legislation regarding copyright protection related to TDM. Thus, TDM activities were subject to national copyright legislation and exceptions applied to copyright protection. E.g. in the Lithuanian copyright law (Law on Copyright and Related Rights of the Republic of Lithuania, last amended in 2015) there were no exceptions to copyright protection concerning TDM activities; Article 22 on reproduction of the copyright-protected work for purposes of teaching or scientific research did not respond to modern needs of TDM activities as it emphasized the purpose of illustration, short works, and short extracts [5].

In 2019, The Directive on Copyright and Related Rights in the Digital Single Market (2019) was adopted with the aim to uniform and modernise the EU copyright protection, adapting it to the implementation of new technologies. Article 3 of the Directive "Text and data mining for the purposes of scientific research" states that "Member States shall provide for an exception to the rights provided for in <...> of this Directive for reproductions and extractions made by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access." The Article also states that "Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results."[6]

The Directive had to be transposed to national legislations in two years; in case of delays, the Directive had to be applied directly. In December, 2021, the Lithuanian Law on Copyright and Related Rights amended according to the Directive's provisions was submitted to the legislature for adoption.

However, the EU Directive has already provoked criticism and calls for further development. The major problem remains sharing research datasets. The new Directive exempts researchers performing TDM activities from the obligation to obtain authorisation from rightholders of texts; however, "corpora compiled on the basis of such exceptions cannot be freely shared, which in a long run may have serious consequences for Open Science and the functioning of research infrastructures such as CLARIN ERIC" (Kamocki et al., 2019). Thus, even if research activities are freed from requirement to obtain permission from rightholders, "knowledge transfer, citizen science and user innovation may paradoxically become more difficult, as they require sharing of data between various groups of stakeholders" (Kamocki et al., 2019). In order to prevent this, "it is important to remember that even if certain research activities are exempted from the rules of copyright, proper licensing is still necessary to efficiently and widely share the fruits of researchers' work" (Kamocki et al., 2019).

In our work, the above-discussed legal problems became the reality. While working on collection of necessary sources, we selected 20 books on cybersecurity written by researchers and practitioners of the field and published by various publishing houses. As all books were copyright-protected, we contacted the publishing houses inside and outside the EU (8 in all) in order to request a permission to use the books for the compilation of our corpora and the storage in the CLARIN's repository.

We have found out that almost all publishing houses ask to fill permission request forms or use per-

[3]Prof. William T. Fisher III. Authors Guild v. GOOGLE, Inc. https://opencasebook.org/casebooks/493-copyright/resources/9.2.5-authors-guild-v-google-inc/; Michael Risch. Authors Guild, Inc. v. HaithiTrust https://opencasebook.org/casebooks/409-an-open-internet-law-casebook/resources/6.3.1-authors-guild-inc-v-hathitrust/

[4]Copyrights, Designs and Patents Act https://www.legislation.gov.uk/ukpga/1988/48/contents; CLARIN Legal Information Platform https://www.clarin.eu/content/clic-text-and-data-mining-tdm-exceptions-uk-and-france

[5]Law on Copyright and Related Rights of the Republic of Lithuania (English translation) https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/5f13b560b2b511e59010bea026bdb259?jfwid=32wf6i76

[6]Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC http://data.europa.eu/eli/dir/2019/790/oj

mission request systems. The request forms are meant mostly for commercial requests to republish the content owned by publishers in other publications. Some publishing houses have special request forms for educational purposes or for authors who want to reuse certain content (e.g. images, charts, tables) in their academic dissertations/theses. However, none of the forms were suitable for data mining and sharing; therefore, we had to include extensive comments explaining the specificity of our case.

As our application procedure was the same as the one of commercial requests, we had to indicate exact number of pages we want to reuse; the reuse of the whole book was not possible. We also had to indicate the title of the publication in which the content would be reused, the publisher, the format of publication, print run/number of expected users/download forecast, and other details relevant to print and online publications. We had to assure that we would be able to make the material secure, so that it would be password-protected against illegal copying/downloading/distribution. In most cases the permission to reuse the material could be obtained for one edition of the publication or a maximum period of one year.

In the comment slots (where they were available) we explained that the texts of the books would be reused as datasets for a research project on machine learning and terminology extraction, the processed texts would be stored in CLARIN-LT repository and the access to them will be restricted to academic users only via authentication service with university logins.

Despite our detailed explanations of our research aims in the forms, as well as correspondence with the publishers, none of the publishing houses granted us free reuse of the requested book extracts. The charges for an extract ranged from 200 to 5,000 Eur. The permission to use texts of one of the books was rejected because the rights were held by the author, not by the publishing house. As we did not have funds allocated for this in our budget, none of the copyrighted books were included in our corpora.

Our experience reveals that publishers do not have special permission request options for reuse of texts as datasets for TDM activities for scientific purposes. In addition, the publishers are not familiar with CLARIN infrastructure, its policy, aims and functions. Therefore, corpora stored in CLARIN repositories have to comply with the same requirements as commercial publications.

Thus, our corpora do not include copyright-protected books on cybersecurity which would be very important for our terminology extraction research. We had to rely on the inclusion of rather large bulk of publicly available media texts into the comparable corpus. In order to ensure its use for scientific purposes, we provided access to the corpus only to academic users of CLARIN-LT repository.

4 Corpora System for Bilingual Terminology Extraction

Five CS corpora have been compiled for this project: a parallel corpus of English texts and their Lithuanian translations (approx. 1.4 million words), a comparable corpus composed of two subcorpora: original English texts and original Lithuanian texts (approx. 4 million words), and three training (gold standard) corpora (approx. 0.1 million words each). The system of corpora and a flowchart of BiTE is presented in Figure 1.

Two of the corpora, namely *English-Lithuanian Parallel CS Corpus*[7] and *English-Lithuanian Comparable CS Corpus*[8], have been deposited to CLARIN-LT repository[9]. The parallel corpus is accessible under the CLARIN public licence (PUB), while the comparable corpus under the CLARIN academic licence (ACA).

The next subsections will present two important aspects of these two resources, namely data curation and composition.

4.1 Data Curation

Data curation is a very important and time-consuming activity, which ensures the quality and endurance of any dataset. The process of data curation typically involves the following steps: 1) discovering data sources; 2) acquiring textual data; 3) cleaning, deduplicating and transforming of extracted data, and 4)

[7] https://clarin.vdu.lt/xmlui/handle/20.500.11821/46
[8] https://clarin.vdu.lt/xmlui/handle/20.500.11821/47
[9] https://clarin.vdu.lt/xmlui/

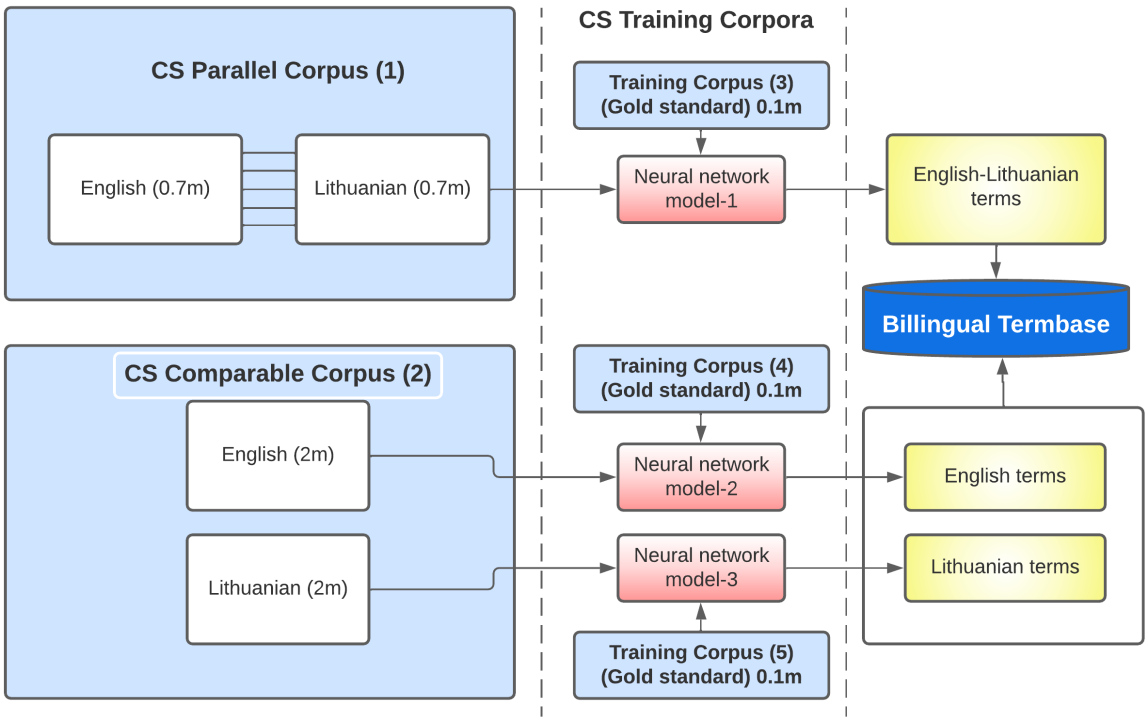

Figure 1. Corpora system for BiTE

integrating the data with other data sources. In the following subsections we will present the data curation steps, which are relevant for our project.

4.1.1 Discovering Data Sources

Presently, all the data for any corpora come from the web, however, texts from different discourses come in different quantities, formats and pose different challenges for a researcher.

First and most obvious source of textual information for our purposes is legal documents on cybersecurity, such as cybersecurity strategies, laws, government resolutions, minister orders, etc. Official national and EU legally binding and non-binding documents are commonly accessible without any restrictions. The documents of these categories can be acquired for both comparable and parallel corpora for both languages (see Table 4 and Table 3).

The second source of information is texts produced by CS experts (practitioners at national and international cybersecurity agencies and other institutions), containing reports, recommendations, information bulletins, guidelines, are also freely available; however, most of them are suitable only for the comparable corpus, as only a handful of them have been translated.

The third source of information is academic research publications on the cybersecurity topic. However, as we have shown in Section 3, access to academic publications is often restricted. Besides, research books and papers written in English and Lithuanian are seldom translated, thus, again acquired academic publications are only suitable for the comparable corpus.

The fourth source of information is media articles. This is by far the most voluminous source of information, as textual information on cybersecurity can be easily acquired by scraping various news portals ranging from general to specialised. Media articles can only be used for the comparable corpus, as only very rarely one can find genuine translations with alignable sentences.

As it could be expected, the volume of information on cybersecurity for English and Lithuanian lan-

guages differ across different sources. As mentioned previously, for the parallel corpus it is almost impossible to acquire original and translated academic texts and it is likewise difficult to find translated informative texts or media articles. Therefore the corpus relies mainly on EU English documents translated into Lithuanian.

The situation with the data for the comparable corpus is somewhat better: we could find data from all four discourses for both languages, but, clearly, Lithuanian sources cannot ensure the same diversity and be of a comparable size to global sources of the English language. Thus, the obvious solution was, firstly, to acquire as much as possible of Lithuanian data on cybersecurity, and then try to construct the similar structure for English.

4.1.2 Acquiring Textual Data

The data files containing relevant textual information have been acquired from the web by a variety of methods:

- using custom developed scrapers, benefiting from *Selenium WebDriver Beautiful Soup* modules for Python and targeting general portals (e.g. BBC for English and Delfi for Lithuanian), specific cybersecurity news portals (e.g. Bleeping Computer) and official EU portals (e.g. EUR-LEX); the method produces clean plain text files;
- manual downloading of PDF files, where possible (e.g. enisa[10] portal);
- manual downloading of web-pages, where scraping was not practical;
- downloading of scientific works in PDF from our home universities' databases (e.g. master theses and doctoral dissertations);

Once the data files have been downloaded, the textual data need to be extracted from PDF, MS Word, or HTML files.

4.1.3 Cleaning, Deduplicating, and Transforming

The acquired data is not always intact:

- textual data extracted from PDF files often contain various problems with line breaking, extra spacing, extra tabs, processing pictures and tables, footnotes interfering with the main text, list of references, text in another language, etc.;
- scraped files are usually in a better shape, however, one can frequently download duplicates of the same text or extra information from a website; besides, dynamically loaded web pages may obscure the full data and introduce a plethora of web scraping issues.

The above mentioned problems are difficult to fix automatically, as cluttered files differ depending on a source. We had to use semi-automatic or even manual find-and-replace routines for cleaning the files. The deduplication process was alleviated by employing a custom fuzzy matching algorithm to the scraped data, which was then followed by a semi-manual checking.

After the files have been cleaned, they have to be transformed into the final form. For the parallel corpus it's semi-automatic alignment on the sentence level. We chose to use *LF Aligner*, which is well-suited for EU official documents (Varga et al., 2005). The resulting files are translation memory exchange (TMX) files.

In addition, English and Lithuanian texts of the comparable corpus have been morphologically annotated. For the English language we have used a large trained pipeline from the spaCy library[11]. The resulting files are in a vertical tabulated format that marks "word", "lemma", "universal POS", and "fine grained POS"[12] (see Table 1). Due to it's minimal structural complexity, the chosen format is easily transformable into any other format.

[10] `https://www.enisa.europa.eu/`
[11] `https://spacy.io/models`
[12] `https://github.com/explosion/spaCy`

A	a	DET	DT
wave	wave	NOUN	NN
of	of	ADP	IN
criticism	criticism	NOUN	NN
was	be	AUX	VBD
launched	launch	VERB	VBN
from	from	ADP	IN
the	the	DET	DT
privacy	privacy	NOUN	NN
supporters	supporter	NOUN	NNS
.	.	PUNCT	.
<s>	<s>		

Table 1. An example of an annotated sentence in the English CS comparable corpus

Likewise, the Lithuanian files have been morphologically annotated with a Lithuanian tagger from SEMANTIKA-2 project[13], where the information of morphological analysis is presented as "word", "lemma", and "msd tag"[14] (see Table 2).

Ekspertai	ekspertas	Ncmpnn-
vieningai	vieningai	Rgp
teigė	teigti	Vgma3—n–ni-
,	,	Tc
kad	kad	Cg
reiškinys	reiškinys	Ncmsnn-
pavojingas	pavojingas	Agpmsnn
.	.	Tp
<s>	<s>	Xh

Table 2. An example of an annotated sentence in the Lithuanian CS comparable corpus

4.1.4 Integrating with Other Data Sources

In our case the integration with other data sources has involved the storage and sharing of the compiled corpora on the CLARIN-LT repository. CLARIN's DSpace-based depositing service (Mišutka et al., 2015) is conveniently built and as a depositor you only will be required to consider the following steps:

- description of the data resource;
- supplying the resource with relevant metadata;
- choosing appropriate format acknowledged by the research community;
- choosing of an appropriates licence;
- archiving of the data.

4.2 Composition of English-Lithuanian CS Corpora

4.2.1 English-Lithuanian Parallel CS Corpus

The parallel corpus includes the EU legal acts and other documents from the time period of 2010-2020. There are 80 files in English and Lithuanian aligned on the sentence level in the corpus. The total size is 1.4m words (EN - 773,373; LT - 633,942). The number of unique words (*types*) is 12,171 for English,

[13] https://semantika.lt/

[14] https://github.com/Semantika2/Morfologiniu-zymeliu-standartas

and 31,558 for Lithuanian. The corpus contains 35,415 aligned segments. The documents are extracted from the EUR-LEX database and other EU institutional repositories (see Table 3).

Document categories	**Subcategories**	**Proportion**
Legally binding (secondary legislation)	Regulations of the European Parliament and of the Council; Directives of the European Parliament and of the Council; Decisions of the European Parliament and of the Council	60%
Official non-binding	Communications of the European Commission; Reports of the European Commission; Recommendations of the European Commission; Opinions of the Committees of the EU; Briefing papers of the Court of Auditors	40%

Table 3. Structure of the parallel corpus (2010-2020)

4.2.2 English-Lithuanian Comparable CS Corpus

The CS comparable corpus compiled for the project includes texts from the time period of 2010-2021 (except for a few important documents from an earlier period). There are 1,708 files in English and 2,567 in Lithuanian. The total size of the corpus is 4m words (EN - 2,000,586; LT - 2,000,343). The number of unique words (*types*) is 37,565 for English, and 101,076 for Lithuanian. Text categories, subcategories and their proportions within the corpus are presented in Table 4.

Text categories	**Subcategories**	**EN**	**LT**
Academic	Scientific articles, monographs, MA and PhD theses, textbooks	19%	30%
Administrative-informative	Reports and recommendations of Cybersecurity Centres; booklets and posters	8%	11%
Legal	CS strategies, laws, government resolutions, ministry orders	18%	4%
Media	Mass media articles, specialised media articles	55%	55%

Table 4. Structure of the comparable corpus (2010-2021)

When compiling a comparable corpus it is very important to ensure similar sampling procedures for compared languages, as the goal is to compare how a particular domain is reflected in two distinct languages. As mentioned earlier, in the case of English-Lithuanian CS comparable corpus, it was difficult to attain the ideal balance of text categories (see Figure 2). Nevertheless, we have achieved, that the media part and other parts (legal, administrative-informative and academic) if taken together, would be equal for both languages (55% and 45%). Thus, we have attained the balance between two parts of the corpus, one of which is more popularised, while the other is more specialised.

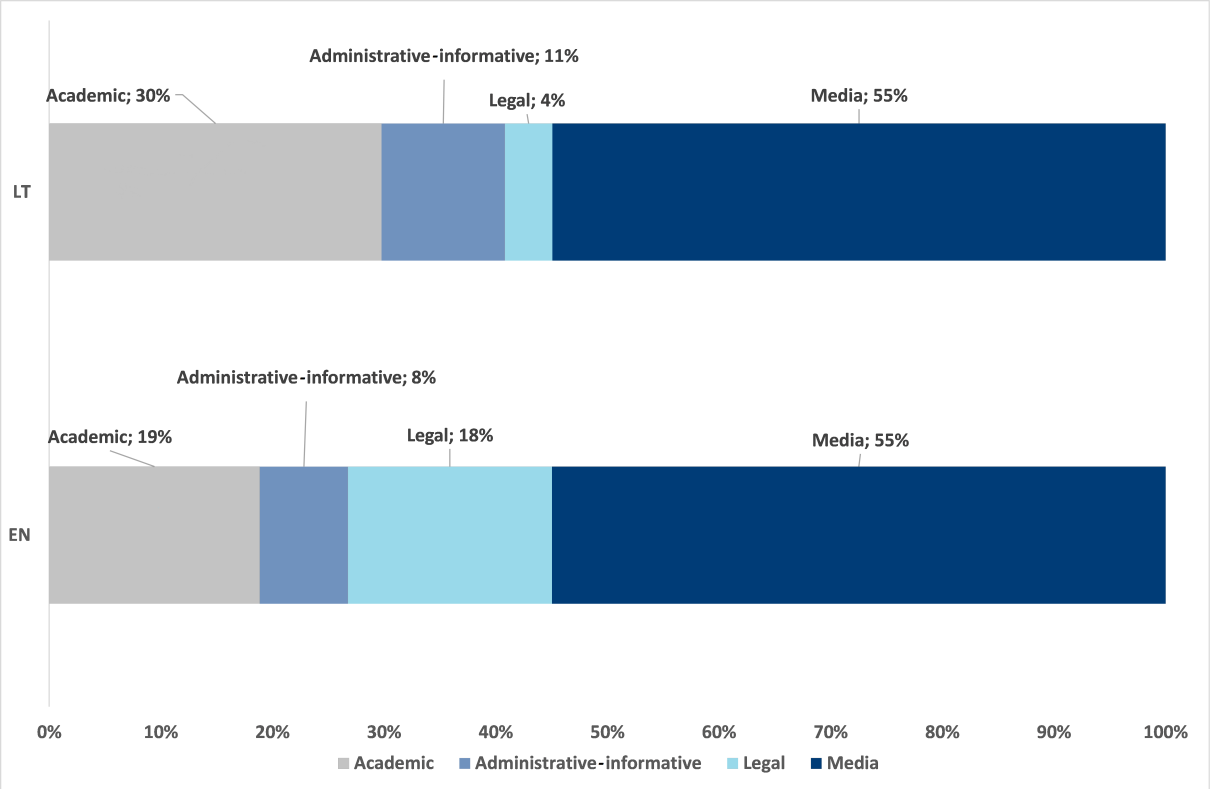

Figure 2. Subcorpora proportions in the English-Lithuanian CS comparable corpus

Ten most frequent lemmas of English and Lithuanian common nouns suggest that the two subcorpora in the comparable corpus are comparable, as 7 out of 10 nouns are at the top 10 in the both corpora:

1.	security	14,825	saugumas	'security'	23,110
2.	information	7,724	duomuo	'datum'	14,357
3.	attack	7,338	sistema	'system'	11,910
4.	system	6,384	informacija	'information'	10,815
5.	datum	6,175	tinklas	'network'	7,817
6.	cybersecurity	6,073	internetas	'internet'	7,061
7.	threat	4,613	vartotojas	'user'	6,760
8.	network	4,536	valstybė	'state'	6,479
9.	service	4,129	programa	'program'	6,169
10.	user	3,855	ataka	'attack'	6,121

4.2.3 Training corpora

In order to train neural networks to perform BiTE, training corpora (gold standard) have been compiled. They have been composed of the same text categories as the main corpora. The comparable training corpora contain legal texts (legislative acts and government resolutions), administrative-informative texts (reports and recommendations by CS experts), academic publications (theses and textbooks), and media articles. Parallel training corpus is composed of the most important EU legal acts and other documents on cybersecurity issues.

The corpora are being manually annotated by tagging three categories of terminological data: terms of the CS domain, terms related to the CS domain, as well as proper names relevant to the CS domain. Four terminology researchers are working on annotation of the training corpora in constant cooperation with a cybersecurity expert who consults and validates the annotation results (see more in (Rackevičienė

et al., 2021)).

5 Concluding Remarks

The analysis of cybersecurity sources revealed that this domain is highly heterogeneous and encompasses diverse types of information accumulated in various discourses. However, availability of some sources is limited. The limitations mainly concern the scientific publications, most of which are copyright-protected. Their reuse for TDM activities and storage in research data repositories involves tackling complex (and not adapted to these aims) permission request procedures and often are charged by rightholders. Though CLARIN constantly raises legal issues related to storing and sharing language resources, further steps are evidently needed by the research community to foster the development of Open Science.

Acquisition of quality data and its curation proved to be a challenging task due to its dynamic nature, necessitating manual reviewing and removing clutter or fixing incomplete elements of the data. The data curation is time-consuming and never ending process, which, nevertheless, needs to be continued in order to ensure quality and longevity of a resource.

Despite the fact that we could not include all planned sources to our corpora, the compiled parallel and comparable corpora contain reasonable variation, as they represent the cybersecurity domain in four different discourses in international and national settings. Thus, we believe that the corpora will provide sufficient data for the future research: deep learning-based terminology extraction, terminology analyses, as well as compilation of a bilingual cybersecurity termbase.

Acknowledgements

The research is carried out under the project "Bilingual Automatic Terminology Extraction" funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282). The project is also included as a use case in COST action "European Network for Web-Centred Linguistic Data Science" (CA18209).

References

Aker, A., Paramita, M. L., and Gaizauskas, R. J. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 402–411. The Association for Computer Linguistics.

Alonso, A., Blancafort, H., de Groc, C., Million, C., and Williams, G. 2012. Metricc: Harnessing comparable corpora for multilingual lexicon development. In *15th EURALEX International Congress*, pages 389–403.

Bernardini, S. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS - A Journal of Professional Communication*, 26.

Biel, Ł. 2016. Mixed corpus design for researching the eurolect: A genre-based comparable-parallel corpus in the pl eurolect project. *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora.*

Chu, C., Dabre, R., and Kurohashi, S. 2016. Parallel sentence extraction from comparable corpora with neural network features. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Delpech, E., Daille, B., Morin, E., and Lemaire, C. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In Kay, M. and Boitet, C., editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 745–762. Indian Institute of Technology Bombay.

Flynn, S., Geiger, C., Quintais, P. J., Margoni, T., Sag, M., Guibault, L., and Carroll, M. W. 2020. Implementing user rights for research in the field of artificial intelligence: A call for international action. *European Intellectual Property Review*, 42(7):393–398.

Giampieri, P. 2018. Online parallel and comparable corpora for legal translations. *Altre Modernità*, 20:237–252.

Goeuriot, L., Morin, E., and Daille, B. 2009. Compilation of specialized comparable corpora in french and japanese. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC 2020)*, pages 55–63.

Gornostay, T., Ramm, A., Heid, U., Morin, E., Harastani, R., and Planas, E. 2012. Terminology extraction from comparable corpora for latvian. In Tavast, A., Muischnek, K., and Koit, M., editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012*, volume 247 of *Frontiers in Artificial Intelligence and Applications*, pages 66–73. IOS Press.

Huidrom, R., Lepage, Y., and Khomdram, K. 2021. Em corpus: a comparable corpus for a less-resourced language pair manipuri-english. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 60–67.

Kamocki, P., Ketzan, E., Wildgans, J., and Witt, A. 2019. New exceptions for text and data mining and their possible impact on the clarin infrastructure. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, Linköping Electronic Conference Proceedings, pages 66–71. Linköping University Electronic Press, Linköpings universitet.

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., and Stüker, S. 2018. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In Yegnanarayana, B., editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2072–2076. ISCA.

Kupiec, J. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 17–22.

McEnery, A. and Xiao, Z. 2007. Parallel and comparable corpora: What is happening? In Anderman, G. and Rogers, M., editors, *Incorporating Corpora: The Linguist and the Translator*, pages 18–31. Multilingual Matters.

Mišutka, J., Kamran, A., Košarko, O., Josífko, M., Ramasamy, L., Straňák, P., and Hajič, J. 2015. Linguistic digital repository based on DSpace 5.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Morin, E. and Prochasson, E. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (BUCC 2011)*, pages 27–34.

Morin, E., Hazem, A., and Saldarriaga, S. P. 2011. Bilingual lexicon extraction from comparable corpora as metasearch. In *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (BUCC 2011)*, pages 35–43.

Rackevičienė, S., Utka, A., Mockienė, L., and Rokas, A. 2021. Methodological framework for the development of an english-lithuanian cybersecurity termbase. *Studies about Languages/Kalbų studijos*, 39:85–92.

Rivera, O. M., Mitkov, R., and Pastor, G. C. 2013. A flexible framework for collocation retrieval and translation from parallel and comparable corpora. *Multiword Units in Machine Translation and Translation Technology*.

Rokas, A., Rackevičienė, S., and Utka, A. 2020. Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In Utka, A., Vaičenonienė, J., Kovalevskaitė, J., and Kalinauskaitė, D., editors, *Human language technologies - the Baltic perspective: proceedings of the 9th international conference, Baltic HLT, Kaunas, Vytautas Magnus University, Lithuania, 22-23 September 2020*, pages 39–46. IOS Press.

Rosati, E. 2018. *The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market : technical aspects*. European Parliament.

Sag, M. 2019. The new legal landscape for text mining and machine learning. *Journal of the Copyright Society of the USA*, 66.

Sanjanasri, Menon, V. K., Kp, S., and Wolk, K. 2021. Mining bilingual word pairs from comparable corpus using apache spark framework. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 2–7.

Steingrímsson, S., Lohar, P., Loftsson, H., and Way, A. 2021. Effective bitext extraction from comparable corpora using a combination of three different approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17.

Steyaert, K. and Rigouts Terryn, A. 2019. Multilingual term extraction from comparable corpora: Informativeness of monolingual term extraction features. In Sharoff, S., Zweigenbaum, P., and Rapp, R., editors, *Proceedings of the 12th Workshop on Building and Using Comparable Corpora (BUCC 2019)*, pages 16–25, Varna, Bulgaria, September.

Terryn, A. R., Hoste, V., and Lefever, E. 2020. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, 54(2):385–418.

Ulčar, M. and Robnik-Šikonja, M. 2020. Finest bert and crosloengual bert: less is more in multilingual models. *arXiv e-prints*, June.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.

Vintar, Š., Grčić Simeunović, L., Martinc, M., Pollak, S., and Stepišnik, U. 2020. Mining semantic relations from comparable corpora through intersections of word embeddings. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora (BUCC 2020)*, pages 29–34, Marseille, France, May. European Language Resources Association.

Vintar, Š. 2010. Bilingual term recognition revisited. *Terminology*, 16:141–158.

Wall, D. S. 2007. *Cybercrime: The Transformation of Crime in the Information Age*. Polity, Cambridge, UK.