UDK 811.172’373.46:811.111
811.172’373.46:004.7.056
Izvorni znanstveni rad
Rukopis primljen 19. VI. 2023.
Prihvaćen za tisak 27. X. 2023.
https://doi.org/10.31724/rihjj.49.2.12

Sigita Rackevičienė
Mykolas Romeris University, Faculty of Human and Social Studies, Institute of Humanities
Ateities st. 20, LT-08303 Vilnius, Lithuania
orcid.org/0000-0001-5794-0296
sigita.rackeviciene@mruni.eu

Andrius Utka
Vytautas Magnus University, Institute of Digital Resources and Interdisciplinary Research
V. Putvinskio st. 23-216, LT-44243 Kaunas, Lithuania
orcid.org/0000-0001-5212-4310
andrius.utka@vdu.lt

Agnė Bielinskienė
Vytautas Magnus University, Institute of Digital Resources and Interdisciplinary Research
V. Putvinskio st. 23-216,LT-44243 Kaunas, Lithuania
orcid.org/0000-0002-9209-2605
agne.bielinskiene@vdu.lt

Liudmila Mockienė
Mykolas Romeris University, Faculty of Human and Social Studies, Institute of Humanities
Ateities st. 20, LT-08303 Vilnius, Lithuania
orcid.org/0000-0001-7153-7276
liudmila@mruni.eu

LITHUANIAN-ENGLISH CYBERSECURITY TERMBASE: PRINCIPLES OF DATA COLLECTION AND STRUCTURING

The aim of the paper is to present compilation and structuring principles, scope and development possibilities of the bilingual Lithuanian-English cybersecurity termbase. The paper discusses different approaches to terminology management, the best practices of which have been used to collect cybersecurity terminology and compile the termbase. Data collection has been mainly based on semasiological and corpus-driven approaches involving creation of deep learning systems trained to extract terminology from the cybersecurity corpora. To achieve systematicity and comprehensiveness of the dataset, the onomasiological and corpus-based approaches have also been incorporated in the data collection process. The termbase design decisions (its macrostructure and microstructure) have been based on onomasiological principles, while term variation has been handled by applying the descriptive approach. The termbase has been developed in the open-source cloud-based terminological management platform *Terminologue*. To ensure interoperability, the termbase has been exported into the TBX format and deposited into the CLARIN-LT repository. The paper also discusses possibilities of publishing terminological data as linguistic linked open data and linking it with other terminological resources and cybersecurity ontologies. The termbase is expected to be useful for cybersecurity specialists, translators, terminographers, lexicographers and the general public, as well as to contribute to the development of the Lithuanian cybersecurity terminology.

1. Introduction

The cybersecurity (CS) domain has gained special relevance in the current public and private life marked by an ever-growing role of global connectivity, cloud services and challenges to ensure the security of sensitive data on all levels: state, business, and individual. Cyber awareness and cyber hygiene have become indispensable not only for governmental institutions and companies, but also for every internet user. Consequently, the need to understand and use terminology of this domain has increased considerably. The CS domain is particularly dynamic. New concepts are constantly developed, and their designations are primarily created in English. The quick pace of its development makes it particularly difficult for other languages not to lag behind; therefore, English terminology prevails in CS communication between experts. However, the national terminology is necessary for communicating the domain-specific knowledge to lay people and communities, facilitating the understanding of CS threats, as well as raising cyber awareness and media literacy. Simultaneously, the national terminology is mandatory for national legislation, documentation of executive bodies, educational materials, etc.

The Lithuanian CS terminology is still evolving. The CS concepts often have several Lithuanian designations which are used interchangeably and create confusion for non-experts. Many concepts still lack Lithuanian designations or, if they do exist, they are not widely recognised and are often replaced by semi-localised or non-localised English terms. Therefore, collection, management and dissemination of the national CS terminology is particularly relevant for diverse user groups.

The paper presents the Lithuanian-English cybersecurity termbase (below referred to as CS termbase) compiled by the researchers of two universities: Mykolas Romeris University and Vytautas Magnus University. The following aspects of the termbase development are discussed:

- approaches to terminological data collection and development of the term index;
- structuring the terminological data and developing the macrostructure and the microstructure of the termbase;
- the scope of the terminological data included in the termbase;
- approaches to handling term variation in the termbase;
- interoperability and possibilities to transform the termbase into linguistic linked open data (LLOD).

The termbase has been compiled in the open-source cloud-based terminological management platform *Terminologue* designed and administered by Gaois research group at the Dublin City University and is freely available online.[1] The termbase is continuously edited, expanded and modified in response to the newly obtained data and user needs. The latest version of the termbase in the TermBase eXchange (TBX) format is deposited in the CLARIN-LT repository (Utka et al. 2023).

2. Approaches to terminology management

Terminology management is multi-dimensional as it involves dealing with conceptual, linguistic, and pragmatic aspects of terminology. This section discusses the main approaches to terminology management and presents those that have been used for data collection and representation in the CS termbase.

[1] Lithuanian-English Cybersecurity Termbase / Lietuvių-anglų kalbų kibernetinio saugumo terminų bazė https://www.terminologue.org/csterms/.

2.1. Onomasiological and semasiological approaches

In the workflow of terminology management, the traditional onomasiological (also called knowledge-driven or concept-based) approach oriented primarily towards concepts and their characteristics is usually combined with the semasiological (also called lexicon-driven or word-based) approach, the starting point of which is terms and their relations in texts. The best practices of both approaches are applied to collect, analyse and structure the datasets (see L'Homme 2004, 2018, 2020; Warburton 2015; UNESCO 2005). The application of the semasiological approach facilitates resolution of problematic issues within the onomasiological approach, such as the lack of appropriate definitions of concepts, and allows considering various linguistic properties and functions of terms which provide important information on both the terms and the concepts they designate (L'Homme 2004, 2020). Thus, application of both approaches contributes to a more comprehensive terminological data collection, analysis, and representation.

2.2. Prescriptive and descriptive approaches

Handling terminology variation in natural language also requires researchers to choose a particular approach (descriptive vs. prescriptive/normative) or a combination of both (Warburton 2015; Bielinskienė et al. 2015; UNESCO 2005; Zeller 2005).

The prescriptive approach seeks to develop the standardised terminology and its work constitutes "an agreement by users to adopt a term for common and repeated use in given circumstances" (UNESCO 2005: 11). Therefore, it is mostly concerned with the quality of terminology and the choice of the most appropriate term for a specific concept designation based on various criteria which may include precision and clarity, systematicity, user-friendliness, derivability, absence of inappropriate connotations, and compliance with language norms (see Gaivenis 2002: 36−49; UNESCO 2005: 11). The principle of univocity and avoidance of synonymy and polysemy are emphasised as they allow achieving unambiguity and precision in specialised communication (L'Homme 2020: 10−11; Sandrini 2014). Onomasiological principles are most relevant in this approach as the choice of the most appropriate term requires concept analysis, synonym ranking, and crafting of definitions (Warburton 2015: 650).

Meanwhile, the descriptive approach observes and records the usage of terms in various discourses, emergence of a variety of terminological designations and their usage without making any "value judgement about them" (Warburton 2015: 650). The descriptive approach grounds its methodology on semasiological principles and aims to record term usage trends in natural language. As terms are an integral part of the language lexicon, polysemy and synonymy are considered natural phenomena in terminology.

Although the two approaches have a different focus on terminology management, they are both closely interconnected. The descriptive approach provides important information for prescriptive terminology and standardisation of terms. Data on the usage of terms, their frequency and distribution, as well as comparative quantitative analyses of synonymous terms allow establishing the dominant term variants in various discourses and in the language in general. Thus, the information provided by the descriptive approach can become a scientific foundation for application of the prescriptive approach and standardisation of terminology. It helps to establish the most appropriate terminology to be used in a specific discourse and to provide recommendations to terminographers for compiling higher-quality terminological resources.

2.3. Corpus-based and corpus-driven approaches

Whichever approach is applied to terminological analysis, corpora are usually used as the main resource as they enable handling huge amounts of everchanging-data and provide a lot of information on terminology of specialised domains (Marcinkevičienė 2000). Corpus analysis methods comprise corpus-based, corpus-driven, or mixed approaches (Tognini-Bonelli 2001). The corpus-based approach "uses corpus evidence mainly as a repository of examples to expound, test or exemplify given theoretical statement;" meanwhile, in the corpus-driven approach, "the theoretical statement can only be formulated in the presence of corpus evidence and is fully accountable to it" (Tognini-Bonelli 2001: 10−11).

In the terminology work based on the onomasiological and prescriptive approaches, corpus-based methods are often applied. They enable researchers to exemplify the usage of terms with various context examples as well as to search for term definitions. Meanwhile within the semasiological and descriptive frameworks, corpus-driven methods are of primary importance. Corpora are used as the main source to extract the existing terminology, collect statistical information on the term usage in natural language, its meaning, frequency, contextual environment, and systemic relations with other terms in knowledge-rich contexts (Meyer 2001; Kovalevskaitė et al. 2015). Corpora are also used for automatic terminology extraction that enables efficient recognition of terms in texts and thus facilitates the processes of developing and updating termbase indexes, as well as serves as translation support or as a basis for indexing document collections (Heylen and De Hertog 2015).

2.4. Approaches applied in the compilation of the CS termbase

While compiling the CS termbase, a combination of the above-discussed approaches has been applied. Data collection has been mainly based on the semasiological and corpus-driven methodology though onomasiological and corpus-based methods have also been incorporated. The design of the termbase (its macrostructure and microstructure) has been based on onomasiological principles, while the terminology variation has been handled applying the descriptive approach with a minimal application of some prescriptive aspects. Automatic terminology extraction methods based on deep learning systems have been applied for data collection from corpora. The training corpora with manually labelled CS terms have been compiled to train the neural networks to perform the term extraction tasks (see the pilot study on Lithuanian CS term extraction in Rokas et al. 2000; bilingual CS data extraction methodology in Rackevičienė et al. 2021; the results of the most recent study are being prepared for publication).

The sections below present the workflow and the chosen approaches and methods applied to data collection, structuring, and representation. The main work on the compilation of the termbase has been done by linguists who have background in terminology, lexicography, and corpus and computational linguistics.

However, the support of a domain expert (cybersecurity researcher at Mykolas Romeris University) has been indispensable and has contributed to all main stages of the project.

3. Collection of terminological data and development of the term index for the CS termbase

This section presents the main approaches and methods which have been used for the collection of the CS terminological data from corpora compiled for the purposes of the project and other sources selected by the domain expert. Subsequently, the development of the term index for the termbase is described.

3.1. Collection of data from corpora

The term collection methodology has been mainly based on the semasiological (lexicon- and corpus-driven) approach. Two types of corpora have been compiled for collection of the terminological data: a parallel CS corpus composed of original texts in English and their translations into Lithuanian (1.4 m words), and a comparable CS corpus composed of original texts in English and Lithuanian (4 m words). The combination of the two corpus types, namely parallel and comparable, has allowed to have a much wider variety of text types and discourses which would not be possible with parallel texts only as parallel texts of this domain are very limited.

The parallel corpus has been composed mainly of the EU legislative acts and related documents (regulations, directives, communications, recommendations, etc.) produced by the main EU institutions: the European Parliament, the Council of the European Union, the European Commission, etc. and translated by the translation departments of these institutions. Both English and Lithuanian versions of the EU legislative acts have been extracted from the EUR-Lex database which contains the EU legal acts and other public documents in all EU languages. Meanwhile, the comparable corpus contains much more diverse texts produced in various discourses: legislative, administrative, informative, academic, and media. The legislative, administrative, and informative texts encompass legislative documents, reports, guidelines, etc. produced by national and international legislative bodies, cybersecurity agencies, and other institutions responsible for cybersecurity policy and its implementation. The academic papers include scientific articles, textbooks, as well as MA and PhD theses. The media texts consist of both mass and specialised media articles on CS topics. Thus, the texts for the comparable corpus have been extracted from various sources: national and international legal and administrative document repositories, academic databases, and media portals (see a comprehensive presentation of the corpora in Utka et al. 2022).

Based on the compiled corpora, two small-scale training corpora have been created with manually labelled CS terms for training deep learning systems: a parallel training corpus (102,583 words) and a comparable training corpus (231,061 words) (Rackevičienė and Utka 2023). The training corpora have been used in multiple experiments for development of the most efficient neural network models for term extraction from the English and Lithuanian datasets. A number of monolingual and multilingual Transformer-based models have been tested in the experiments (such as BERT, DistilBERT, LitLatBERT, RoBERTa, and ELECTRA). The experiments have shown that the best results could be obtained by pretraining the Multilingual BERT model with cybersecurity data achieving the best score of 81.7%. Most frequent automatically extracted terms have been included in the CS termbase.

The whole workflow of the CS terminology collection is presented in Figure 1.

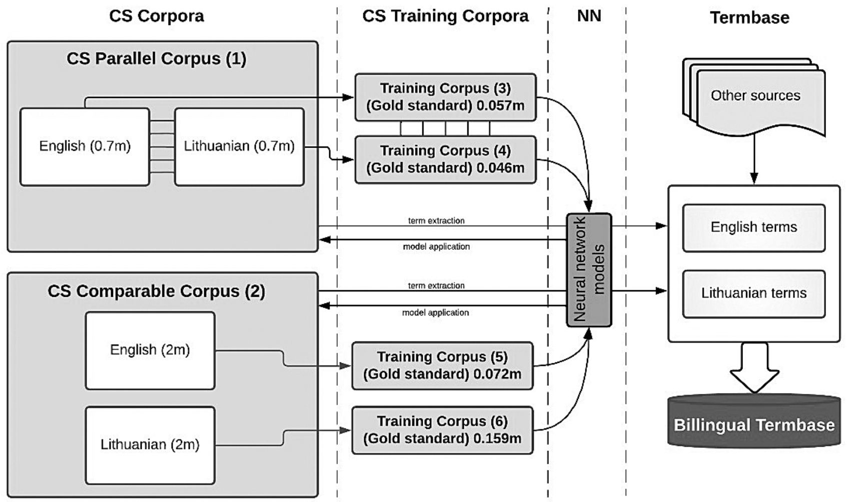


Figure 1. Workflow of the CS terminology collection

Corpora have also been used for semi-automatic extraction of knowledge-rich contexts used for concept definitions, as well as term usage in context examples.

3.2. Collection of data from other sources

To include all important terms of the domain, the onomasiological (knowledge-driven) approach has simultaneously been applied for collection of the data. It has involved the analysis of terminology of international and national institutions that work on CS issues, and term-and-definition lists provided in the documents of these institutions (e.g., glossaries of the European Union Agency for Cybersecurity and the USA National Institute of Standards and Technology, legal acts of the EU and the Republic of Lithuania, ISO standards, etc.). Terms and definitions provided in these sources have been examined and the selected terms and their definitions have been extracted and systematised by the domain expert.

3.3. Development of the term index

The term index has been developed based on the lists of terms extracted from the corpora and terms listed by the domain expert. The lists have been contrasted to establish repetitive and non-repetitive terms, and final lists of Lithuanian and English terms have been drafted. These final lists have been used for term selection and development of the term index for the termbase. Alignment of the Lithuanian-English terms has been performed by the neural network models specially developed for this purpose, as well as by the manual search for equivalents in the parallel and comparable CS corpora. For selection purposes the terms have been categorised according to two criteria: frequency of terms in the corpora and characteristics of the concepts they represent.

3.3.1. Frequency of the terms

In order to establish the trends in the usage of the CS terminology, the terms used in the corpora have been grouped according to their frequencies into five categories and labelled accordingly: very frequent (not less than 1,000 hits in the corpora), frequent (not less than 200 hits), fairly common (not less than 20 hits), rare (not less than 4 hits), and very rare (1 to 3 hits). Terms, which have been added by the expert, but do not occur in the corpora, have been ascribed to the last category.

3.3.2. Conceptual characteristics of the terms

To have a term index, which represents the whole conceptual structure of the domain, the existing CS ontologies and conceptual models have been examined (Thinyane and Christine 2020; Syed and Zhong 2018; Jia et al. 2018). Based on the collected information and the semantic analysis of the terms, a conceptual CS model has been developed by the project researchers: the terminologists in cooperation with the cybersecurity domain expert (see Figure 2).

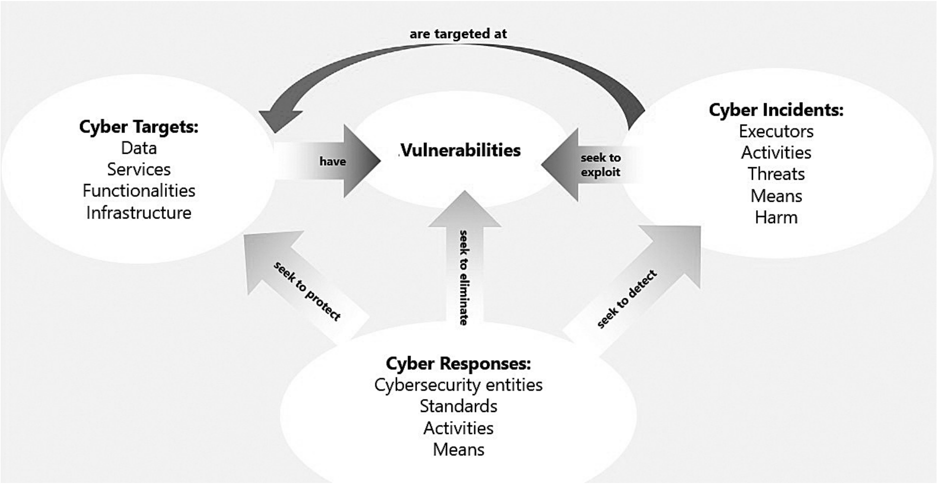


Figure 2. The conceptual model of the CS domain

The conceptual model delineates the main CS subdomains, each of which comprises a set of several concept groups representing classes of both material and non-material entities, as well as processes inherent to the cybersecurity domain:

- Cyber Responses comprise concepts that represent animate entities (CS practitioners and organisations responsible for maintaining cybersecurity of information resources and infrastructure, referred to as CS entities in the model), processes (activities performed by organisations) and inanimate concrete and abstract entities (cybersecurity means and standards).
- Cyber Incidents also include concepts that represent animate entities (executors of malicious cyber activities), processes (malicious activities themselves or their potential, referred to as Activities and Threats in the model) and inanimate entities (means used for performing malicious activities and harm caused by them).
- Vulnerabilities comprise concepts that refer to inanimate entities which cover various kinds of weaknesses in security systems.
- Cyber Targets include concepts that refer to inanimate entities which serve as the primary assets protected by cybersecurity systems.

In addition to presenting the classes of entities and processes within the cybersecurity domain, the model reflects the major relationships among the subdomains, indicating the objectives of the activities that take part in the domain. The subdomain of Cyber Responses, that represents the core mission of the CS domain, is interrelated with all other subdomains. Each of these relations signifies a particular objective directed towards a specific subdomain (‘seek to detect’, ‘seek to eliminate’, ‘seek to protect’). Activities outlined under the Cyber Incidents are interrelated with the Vulnerabilities (through the relation ‘seek to exploit’) and the Cyber Targets (through the relation ‘are targeted at’). Moreover, the subdomain of Cyber Targets is closely related with the Vulnerabilities though the relationship ‘have’, indicating the reasons for protection needs.

The developed conceptual model has enabled to select terms for the termbase more systematically so that they represent all main CS subdomains. The model has also been used for development of the termbase macrostructure (see section 4.2).

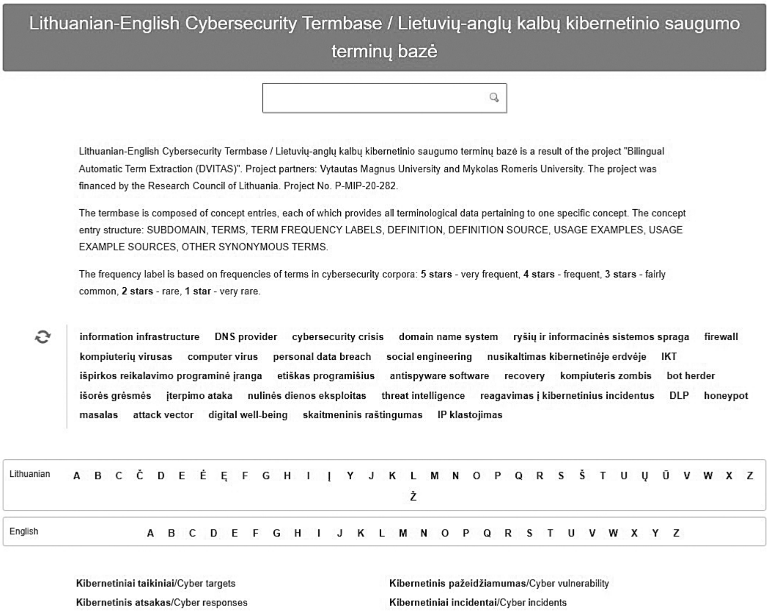
4. Compilation of the termbase

This section presents the *Terminologue* platform used for compilation of the termbase, representation of the terminological data in the termbase, and solutions for handling term variation.

4.1. The *Terminologue* platform and term query possibilities

The *Terminologue* platform is a cloud-based instance of the open-source *Terminologue* software that provides various functionalities for terminology management and termbase development[2]. *Terminologue* is localised in 14 languages including Lithuanian. The interface provides the following search possibilities: random search, alphabetical search, and search according to the subdomains of the termbase (see Figure 3).

[2] https://www.terminologue.org/



Lithuanian-English Cybersecurity Termbase / Lietuvių-anglų kalbų kibernetinio saugumo terminų bazė

Lithuanian-English Cybersecurity Termbase / Lietuvių-anglų kalbų kibernetinio saugumo terminų bazė is a result of the project "Bilingual Automatic Term Extraction (DVITAS)". Project partners: Vytautas Magnus University and Mykolas Romeris University. The project was financed by the Research Council of Lithuania. Project No. P-MIP-20-282.

The termbase is composed of concept entries, each of which provides all terminological data pertaining to one specific concept. The concept entry structure: SUBDOMAIN, TERMS, TERM FREQUENCY LABELS, DEFINITION, DEFINITION SOURCE, USAGE EXAMPLES, USAGE EXAMPLE SOURCES, OTHER SYNONYMOUS TERMS.

The frequency label is based on frequencies of terms in cybersecurity corpora: **5 stars** - very frequent, **4 stars** - frequent, **3 stars** - fairly common, **2 stars** - rare, **1 star** - very rare.

information infrastructure DNS provider cybersecurity crisis domain name system ryšių ir informacinės sistemos spraga firewall kompiuterių virusas computer virus personal data breach social engineering nusikaltimas kibernetinėje erdvėje IKT išpirkos reikalavimo programinė įranga etiškas programišius antispyware software recovery kompiuteris zombis bot herder išorės grėsmės įterpimo ataka nulinės dienos eksploitas threat intelligence reagavimas į kibernetinius incidentus DLP honeypot masalas attack vector digital well-being skaitmeninis raštingumas IP klastojimas

Lithuanian A B C Č D E Ė Ę F G H I Į Y J K L M N O P Q R S Š T U Ų Ū V W X Z Ž

English A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Kibernetiniai taikiniai/Cyber targets **Kibernetinis pažeidžiamumas**/Cyber vulnerability

Kibernetinis atsakas/Cyber responses **Kibernetiniai incidentai**/Cyber incidents

Figure 3. Interface of the CS termbase

4.2. The macrostructure of the termbase

Compilation of the termbase has started with the development of its macrostructure representing categorisation of CS concepts and their terminological designations into thematic groups and subgroups. Categorisation of the concepts has been based on the conceptual framework developed for the project purposes (see Figure 2). Following the framework, the concepts have been classified into four main thematic groups which have been subdivided into smaller subgroups that represent the subdomains of the CS domain: Cyber Targets (data, services, functionalities, infrastructure), Vulnerabilities, Cyber Incidents (executors, attacks, threats, means, harm), and Cyber Responses (CS entities, standards, activities, means) (see Figure 2). The subdomain of Vulnerabilities currently contains the smallest number of concepts and is not subdivided further. However, when more data is collected, this subdomain might also be divided into subgroups according to the sources and types of vulnerabilities based on the existing typologies (e.g., Kizza 2020).

4.3. The microstructure of the termbase

The microstructure of the termbase is based on the concept-oriented approach that enables to organise the terminological data around the concepts that the data pertains to. Data categories are stored on two main levels: the concept level and the language level. The concept level is composed of the language-independent data categories (the concept ID and the subdomain the concept belongs to), while the language level is composed of the linguistic data categories stored on several sublevels: Terms, Definitions, Examples, and Other synonymous terms (for more on this sublevel see subsection 4.5). Figure 4 presents the structure of each concept entry.

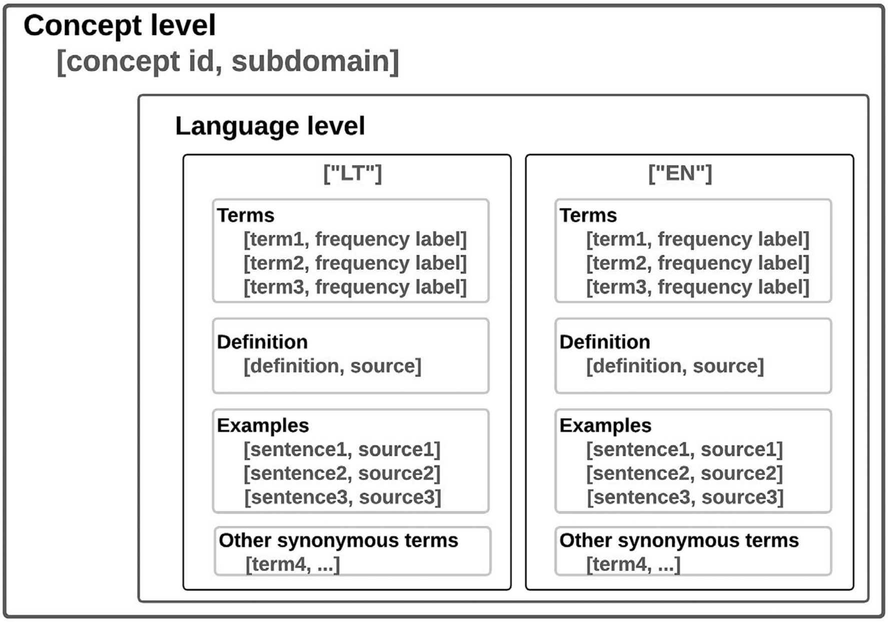


Figure 4. The microstructure of a concept entry

4.4. The scope of the termbase

Presently, the CS termbase consists of 234 concepts that are represented by 1,009 terms, as well as 468 definitions and 761 examples in English and Lithuanian. A more detailed scope of the concepts and terms across all subdomains is presented in Table 1 below.

Table 1. Scope of the CS termbase

Termbase divisions representing CS subdomains	Concepts	Terms (including other synonymous terms)		
		Lithuanian	English	Total
Cyber Targets	33	70 (8)	53 (1)	123 (9)
Cyber Responses	88	188 (39)	170 (31)	358 (70)
Cyber Vulnerability	9	22 (8)	12 (2)	34 (10)
Cyber Incidents	104	302 (102)	192 (31)	492 (133)
All subdomains	234	582 (157)	427 (65)	1,009 (222)

4.5. Handling term variation

The characteristic feature of the CS terminology is the usage of competing synonymous terms, which is especially evident in the Lithuanian corpora. As the data presented in Table 1 indicates, 234 concepts included in the termbase are designated by 582 Lithuanian terms and 427 English terms. The following concepts have most terminological designations detected in the Lithuanian corpora: ‘DDoS attack’ (14), ‘spam’ (11), ‘DoS attack’ (9), ‘botnet’ (9), and ‘spamming’ (8).

Variation of terms is a natural phenomenon in such rapidly evolving domains as cybersecurity. In dynamic domains, “competing terms are used in parallel until unambiguous terms are gradually established, either through a natural selection process or by conscious standardization” (Schmitz 2015). The synonymous CS terms differ in various aspects such as: (1) the origin (native/foreign/hybrid), (2) the denotation type (primary/figurative), (3) the length and explicitness (the number of constituents and use/disuse of abbreviations), and (4) the lexical structure, e.g.,

(1) Lithuanian designations of different origin:

‘spam’

native: *brukalas;*

foreign: *spamas;*

hybrid: *„spam“ laiškai.*

(2) Lithuanian designations of different denotation type:

‘brute force attack’

figurative (calque of the English term): *brutalios jėgos ataka*;

primary (descriptive): *slaptažodžių parinkimo ataka.*

(3) Lithuanian designations of different length and explicitness (different number of constituents and including/not including abbreviations):

‘malware’

2 constituents (one of which is an abbreviation): *kenkimo PĮ*;

3 constituents (no abbreviations included): *kenkimo programinė įranga*

(4) Lithuanian designations of different lexical structure:

‘biometric authentification’

terms with heads which differ in their suffixes:

suffix -*avimas: biometrinis autentifikavimas;*

suffix -*acija: biometrinė autentifikacija*.

‘zero-day vulnerability’

terms with different heads:

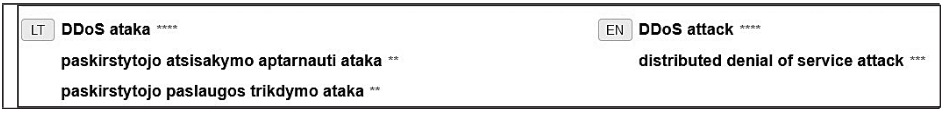
the head *pažeidžiamumas: nulinės dienos pažeidžiamumas;*

the head *spraga: nulinės dienos spraga.*

In the CS termbase, term variation has been handled by using mainly the descriptive approach with the aim of presenting the variety of synonymous terms detected in the corpora. However, some preferences have been observed, namely, the term frequency and term compliance to the basic language norms discussed in the subsections below.

4.5.1. Frequency of synonymous terms

Synonymous terms have been categorised according to their frequencies. The most frequent synonyms (up to 3) have been presented on the Term level with frequency labels that indicate their frequency categories (see subsection 3.3.1 and Figure 5). Other terms have been presented on the Other synonymous terms level.



LT **DDoS ataka** ****
paskirstytojo atsisakymo aptarnauti ataka **
paskirstytojo paslaugos trikdymo ataka **

EN **DDoS attack** ****
distributed denial of service attack ***

Figure 5. Terms and their frequency labels

4.5.2. Compliance to the language norms

Compliance of terms to the basic Lithuanian language norms has been ensured. Unlocalised English terms used in their original form in Lithuanian texts have not been included in the termbase, e.g., *fake news, phishing, rootkit* (usually written in quotation marks). English terms whose localisation is incomplete have been included but are presented on the "Other synonymous terms" level, e.g., *phishingas, phishing'as.* However, some other terms whose compliance to the language norms could also be questionable have been included on the Term level based on their frequency in the corpora which is the same or even higher than the frequency of the synonymous terms which comply with the language norms, e.g., multiword terms with indefinite adjectives (*aktyvi ataka* 'active attack', *pasyvi ataka* 'passive attack') used instead of definitive adjectives required by the language norms (*aktyvioji ataka, pasyvioji ataka*) and hybrid multi-word terms that include English abbreviations or even phrases (*MitM ataka* 'MitM attack', *"Man in the Middle" ataka* 'Man-in-the-Middle attack').

This strategy has been pursued due to the chosen descriptive approach and an attempt to record the whole variety of term usage in authentic texts which can be applied to further term categorisation according to the requirements relevant for standardisation.

5. Further steps: enhancing interoperability by conversion of the termbase into LLOD

In order to ensure interoperability, the CS termbase was exported into the XML TermBase eXchange (TBX) format and deposited in the CLARIN-LT repository (Utka et al. 2023). TBX is a universally acknowledged format for facilitating storage and sharing terminological data as it allows importing TBX termbases into different terminology management systems or computer-assisted translation tools, as well as converting the format into structured formats of different complexity (CSV, RDF, SQLITE, etc.) for diverse usage.

As the last step in the workflow, we are also considering the conversion of the CS termbase into linguistic linked open data (LLOD). LLOD termbases have many advantages: they are linked to the global LLOD network and to other termbases and are reusable, searchable, interoperable, and discoverable across the Semantic Web. Thus, representing any lexical database as linked data is seen as a good practice (Chiarcos et al. 2013; Bosque-Gil et al. 2016; Di Buono et al. 2020; Rodriguez-Doncel et al. 2015). Specifically, in the cybersecurity domain, the Unified Cybersecurity Ontology (UCO) serves as the knowledge core (Syed et al. 2016) and it potentially could become the main linking target for the CS termbase.

An important step in the conversion process of a termbase into linked data is choosing appropriate and comprehensive modelling scheme. The OntoLex-Lemon Model "provides a core vocabulary (OntoLex) to represent *linguistic information* related to ontology and vocabulary elements" (W3C 2019). The model has several extensions (modules), among which the Lexicography module (lexicolog) is commonly used for modelling lexicographic data as linked data (W3C 2019). While the module has all necessary tools for representing traditional dictionaries, it lacks vocabulary for representing specific terminology data, i.e. terminological definitions, different type sources, term notes, or reliability codes. Besides, OntoLex-Lemon has a limited vocabulary for expressing information about word vectors or corpus-based lexicography (such as frequency, lemmas, or collocations).

The problem has been recently addressed by the Ontology-Lexicon community group, which has started developing a Terminology module[3] and a module for frequency, attestation, and corpus information OntoLex-FrAC (Chiarcos et al. 2022). The new modules are particularly relevant to the CS termbase as they are intended to handle specific terminological information, as well as frequency information, links to corpora, and authentic examples. Thus, the new representation capabilities of the new modules are seen as an important future step towards further enhancing the interoperability of the CS termbase.

6. Final remarks

The work on the CS termbase confirms that the combination of the semasiological and onomasiological approaches to terminology of a specific domain is indispensable to developing a comprehensive termbase, which would represent the conceptual framework of the domain and include the main terms with the relevant explanatory and usage information (definitions, frequency labels, and contextual examples). The support of a CS expert is important in all stages of the workflow: collection of texts for corpora compilation, validation of terminologically annotated datasets, development of a conceptual domain model and systematic term index, as well as collection of definitions and their formulation.

The CS termbase compilation process proves that the descriptive approach has to precede the prescriptive one. The descriptive approach helps to reveal a whole variety of the term usage in different discourses and types of authentic texts and provides a lot of information important for further work on term management. The categorisation of Lithuanian terms according to prescriptive requirements and provision of prescriptive recommendations requires another project encompassing an analysis of different target groups' needs, Lithuanian standardisation traditions and norms, as well as further cooperation with CS experts.

[3] https://www.w3.org/community/ontolex/wiki/Terminology.

Despite having to solve numerous issues of compatibility and improvement of conversion tools, we believe that conversion of the CS termbase into LLOD is worth pursuing in order to enhance its accessibility, interoperability, and reusability.

Acknowledgment

This research was conducted within the framework of the COST Action CA18209 “European network for Web-centred linguistic data science (NexusLinguarum)”, a use case on cybersecurity terminology.

References

Bielinskienė, Agnė; Boizou, Loïc; Grigonytė, Gintarė; Kovalevskaitė, Jolanta; Rimkutė, Erika; Utka, Andrius. 2015. *Lietuvių kalbos terminų automatinis atpažinimas ir apibrėžimas.* Monography. Vytautas Magnus University. Kaunas. https://portalcris.vdu.lt/server/api/core/bitstreams/559abb58-b5d1-4fa5-a907-45af03a3feb3/content (Accessed 21 November 2022).

Bosque-Gil, Julia; Gracia, Jorge; Gómez-Pérez, Asunción. 2016. Linked data in lexicography. *Kernerman Dictionary News* 24. 19–24. https://lexicala.com/wp-content/uploads/2021/03/kdn24_2016.pdf (Accessed 21 November 2022).

Chiarcos, Christian; Moran, Steven; Mendes, Pablo N.; Nordhoff, Sebastian; Littauer, Richard. 2013. Building a Linked Open Data cloud of linguistic resources: Motivations and developments. *The People's Web Meets NLP*. Springer. Berlin, Heidelberg. 315–348. doi: 10.1007/978-3-642-35085-6_12.

Chiarcos, Christian; Apostol, Elena-Simona; Kabashi, Besim; Truică, Ciprian-Octavian. 2022. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. *Proceedings of the 29th International Conference on Computational Linguistics.* International Committee on Computational Linguistics. Gyeongju. Republic of Korea. 4018–4027. https://aclanthology.org/2022.coling-1.353.pdf (Accessed 21 November 2022).

Di Buono, Maria Pia; Cimiano, Philipp; Elahi, Mohammad Fazleh; Grimm, Frank. 2020. Terme-à-LLOD: Simplifying the conversion and hosting of terminological resources as linked data. *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020).* Eds. Ionov, Maxim; McCrae, John P.; Chiarcos, Christian; Declerck, Thierry; Bosque-Gil, Julia; Gracia, Jorge. European Language Resources Association (ELRA). Paris. 28–35.

Gaivenis, Kazimieras. 2002. *Lietuvių terminologija: teorijos ir tvarkybos metmenys*. Lietuvių kalbos instituto leidykla. Vilnius.

Heylen, Kris; De Hertog, Dirk. 2015. Automatic term extraction. *Handbook of Terminology. Volume 1.* Eds. Kockaert, Hendrik J.; Steurs, Frieda. John Benjamins. Amsterdam.

Jia, Yan; Qi, Yulu; Shang, Huaijun; Jiang, Rong; Li, Aiping. 2018. A practical approach to constructing a knowledge graph for cybersecurity. *Engineering* 4/1. 53–60.

Kizza, Joseph Migga. 2020. Introduction to computer network vulnerabilities. *Guide to Computer Network Security. Texts in Computer Science*. Springer. Cham. 87–103. doi.org/10.1007/978-3-030-38141-7_4.

Kovalevskaitė, Jolanta; Bielinskienė, Agnė. 2015. Deskriptyviosios terminologijos tyrimų rezultatų panaudojimas terminologijos praktikoje. *Terminologija* 22. 149–168. http://lki.lt/wp-content/uploads/2017/06/terrminologija_22_maketas-ilovepdf-compressed-ilovepdf-compressed.pdf (Accessed 21 November 2022).

L'Homme, Marie-Claude. 2004. Lexico-semantic approach to the structuring of terminology. *Proceedings of CompuTerm 2004. 3rd International Workshop on Computational Terminology.* 7–14. COLING. Geneva, Switzerland. https://aclanthology.org/W04-1801.pdf (Accessed 21 November 2022).

L'Homme, Marie-Claude. 2018. Maintaining the balance between knowledge and the lexicon in terminology: A methodology based on frame Semantics. *Lexicography ASIALEX* 4. 3–21. doi.org/10.1007/s40607-018-0034-1.

L'Homme, Marie-Claude. 2020. *Lexical semantics for terminology. An introduction*. John Benjamins publishing Company. Amsterdam.

Marcinkevičienė, Rūta. 2000. Terminografija ir tekstynas. *Terminologija* 6. 5–22. https://etalpykla.lituanistikadb.lt/fedora/objects/LT-LDB-0001:J.04~2000~1367177232040/datastreams/DS.002.0.01.ARTIC/content (Accessed 21 November 2022).

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. *Recent Advances in Computational Terminology*. Eds. Bourigault, Didier; Jaquemin, Christian; L'Homme, Marie-Claude. John Benjamins. Amsterdam. 279–302.

Rackevičienė, Sigita; Utka, Andrius, Mockienė, Liudmila; Rokas, Aivaras. 2021. Methodological framework for the development of an English-Lithuanian Cybersecurity Termbase. *Kalbų studijos (Studies about Languages)* 39. 85–92. doi.org/10.5755/j01.sal.1.39.29156.

Rackevičienė, Sigita; Utka, Andrius. 2023. Developing Training Corpora for Automatic Extraction of Cybersecurity Terminology. *Terminologica. TOTh 2022. Terminologie & Ontologie: Théories et Applications: Actes de la conférence, 2 & 3 juin 2022*. Université Savoie Mont Blanc. Chambéry. 75–95.

Rodriguez-Doncel, Victor; Santos, Cristiana; Casanovas, Pompeu; Gómez-Pérez, Asunción; Gracia, Jorge. 2015. A linked data terminology for copyright based on OntoLexlemon. *AI Approaches to the Complexity of Legal Systems*. Springer. Cham. 410−423.

Rokas, Aivaras; Rackevičienė, Sigita; Utka, Andrius. 2020. Automatic extraction of Lithuanian cybersecurity terms using deep learning approaches. *Human Language Technologies – The Baltic Perspective (Baltic HLT 2020).* Eds. Utka, Andrius; Vaičenonienė, Jurgita; Kovalevskaitė, Jolanta; Kalinauskaitė, Danguolė. IOS Press. Amsterdam, Berlin, Washington. 39–46. doi.org/10.3233/FAIA200600.

Sandrini, Peter. 2014. Multinational legal terminology in a paper dictionary? Ed. Aodha, Máirtín Mac. *Legal lexicography: a comparative perspective.* Ashgate Publishing. Surrey, Burlington. 141−152. https://www2.uibk.ac.at/downloads/trans/publik/legalpaper.pdf (Accessed 21 November 2022).

Schmitz, Klaus-Dirk. 2015. Terms in texts and the challenge for terminology management. *Terminologija* 22. 15–18. http://lki.lt/terminologija-22 (Accessed 21 November 2022).

Syed, Zareen; Padia, Ankur; Finin, Tim; Mathews, Lisa; Joshi, Anupam. 2016. UCO: A Unified Cybersecurity Ontology. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence Artificial Intelligence for Cyber Security: Technical Report WS-16-03.* https://www.researchgate.net/publication/287195565_UCO_A_Unified_Cybersecurity_Ontology (Accessed 10 October 2023).

Syed, Romilla; Zhong, Haonan. 2018. Cybersecurity vulnerability management: an ontology-based conceptual model. *AMCIS 2018 Proceedings*. 6. https://aisel.aisnet.org/amcis2018/Semantics/Presentations/6 (Accessed 21 November 2022).

Thinyane, Mamello; Christine, Debora. 2020. *The SMART Citizen Cyber Resilience Ontology (SC2RO).* 2020. United Nations University Institute in Macau. Macau, China. https://i.unu.edu/media/cs.unu.edu/page/4822/SC2RO_overview_0.1.pdf (Accessed 25 November 2022).

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins.

UNESCO. 2005. *Guidelines for terminology policies: formulating and implementing terminology policy in language communities*. United Nations Educational, Scientific and Cultural Organization. Paris. https://unesdoc.unesco.org/ark:/48223/pf0000140765 (Accessed 21 November 2022).

Utka, Andrius; Rackevičienė, Sigita; Mockienė, Liudmila; Rokas, Aivaras; Laurinaitis, Marius; Bielinskienė, Agnė. 2022. Building of parallel and comparable cybersecurity corpora for bilingual terminology extraction. *Selected Papers from the CLARIN Annual Conference 2021.* Eds. Monachini, Monica; Eskevich, Maria. Linköping University Electronic Press. Linköping. 126−138. doi.org/10.3384/ecp18912.

Utka, Andrius; Rackevičienė, Sigita; Bielinskienė, Agnė; Laurinaitis, Marius; Mockienė, Liudmila; Rokas, Aivaras. 2023. *Lithuanian-English Cybersecurity Termbase v.0.1.* CLARIN-LT digital library in the Republic of Lithuania. http://hdl.handle.net/20.500.11821/55 (Accessed 14 April 2023).

W3C. 2019. *The OntoLex Lemon Lexicography Module.* Final Community Group Report 17 September 2019. https://www.w3.org/2019/09/lexicog/ (Accessed 8 December 2022).

Warburton, Kara. 2015. Terminology management. *Routledge Encyclopedia of Translation Technology.* Ed. Sin-wai, Chan. Routledge. London and New York.

Zeller, Irma. 2005. *Automatinis terminų atpažinimas ir apdorojimas*. Doctoral dissertation. Faculty of Humanities, Vytautas Magnus University. Kaunas.

Litavsko-engleska terminološka baza kibernetičke sigurnosti: načela strukturiranja i prikupljanja podataka

Sažetak

Cilj je rada predstaviti načela sastavljanja dvojezične litavsko-engleske terminološke baze kibernetičke sigurnosti, opseg terminoloških podataka uključenih u terminološku bazu i mogućnosti njezina daljnjega razvoja. U radu se raspravlja o različitim pristupima upravljanju terminologijom, od kojih su najbolje prakse korištene za prikupljanje terminologije kibernetičke sigurnosti i sastavljanje baze pojmova. Prikupljanje podataka uglavnom se temelji na semasiološkim pristupima i pristupima vođenim korpusom koji uključuju stvaranje sustava dubokoga učenja osposobljenih za izlučivanje terminologije iz korpusa kibernetičke sigurnosti. Kako bi se postigla sustavnost i sveobuhvatnost skupa podataka, u proces prikupljanja podataka ugrađeni su onomasiološki i korpusni pristupi. Odluke o oblikovanju pojmovne baze (njezine makrostrukture i mikrostrukture) temeljene su na onomasiološkim načelima, dok je terminološka varijacija riješena primjenom deskriptivnoga pristupa. Terminološka baza razvijena je u otvorenoj platformi za upravljanje terminologijom *Terminologue*. Kako bi se osigurala interoperabilnost, baza pojmova pretvorena je u TBX format i pohranjena u repozitorij CLARIN-LT. U radu se također raspravlja o mogućnostima objavljivanja terminoloških podataka kao jezičnih povezanih podataka i njihova povezivanja s drugim resursima/ontologijama kibernetičke sigurnosti. Očekuje se da će izrađena baza pojmova biti korisna stručnjacima za kibernetičku sigurnost, prevoditeljima i široj javnosti, kao i da će doprinijeti razvoju terminologije kibernetičke sigurnosti u Litvi.

Keywords: cybersecurity, termbase, terminology management, termbase structure, LLOD.
Ključne riječi: kibernetička sigurnost, terminološka baza, upravljanje terminologijom, struktura terminološke baze, LLOD.