RESPECTUS PHILOLOGICUS

Respectus Philologicus
ISSN: 2335-2388
vincas.grigas@leidykla.vu.lt
Vilniaus Universitetas
Lituania

Distribution of Terms Across Genres in the Annotated Lithuanian Cybersecurity Corpus

Rackevi#ien#, Sigita; Utka, Andrius; Bielinskien#, Agn#; Rokas, Aivaras

Distribution of Terms Across Genres in the Annotated Lithuanian Cybersecurity Corpus
Respectus Philologicus, vol. 41, núm. 46, 2022
Vilniaus Universitetas, Lituania
Disponible en: https://www.redalyc.org/articulo.oa?id=694473588002
DOI: https://doi.org/10.15388/RESPECTUS.2022.41.46.105

Linguistic research

Distribution of Terms Across Genres in the Annotated Lithuanian Cybersecurity Corpus

Sigita Rackevičienė sigita.rackeviciene@mruni.eu
Mykolas Romeris University, Lituania
https://orcid.org/0000-0001-5794-0296
Andrius Utka andrius.utka@vdu.lt
Vytautas Magnus University, Lituania
https://orcid.org/0000-0001-5212-4310
Agnė Bielinskienė agne.bielinskiene@vdu.lt
Vytautas Magnus University, Lituania
https://orcid.org/0000-0002-9209-2605
Aivaras Rokas aivaras.rokas@vdu.lt
Vytautas Magnus University, Lituania
https://orcid.org/0000-0003-3602-3872

Respectus Philologicus, vol. 41, núm. 46, 2022

Vilniaus Universitetas, Lituania

Aprobación: 10 Enero 2022

DOI: https://doi.org/10.15388/RESPECTUS.2022.41.46.105

Redalyc: https://www.redalyc.org/articulo.oa?id=694473588002

Abstract: The paper provides results of the frequential distribution analysis of cybersecurity terms used in the Lithuanian cybersecurity corpus composed of texts of different genres. The research focuses on the following aspects: overall distribution of cybersecurity terms (their density and diversity) across genres, distribution of English and English-Lithuanian terms and their usage patterns in Lithuanian sentences, and, finally, the most frequent cybersecurity terms and their thematic groups in each genre. The research was performed in several stages: compilation of a cybersecurity corpus and its subdivision into genre-specific subcorpora, manual annotation of cybersecurity terms, automatic lemmatisation of annotated terms and, finally, quantitative analysis of the distribution of the terms across the subcorpora. The results reveal the similarities and differences of the use of cybersecurity terminology across genres which are important to consider to get a complete picture of terminology usage trends in this domain.

Keywords: cybersecurity domain, corpus annotation, terminology annotation, lemmatisation, distribution analysis.

Introduction

The cybersecurity (CS) domain has gained special relevance in the current public and private life, becoming more and more digitalised due to the ever-growing role of information technologies and pandemic challenges. Security of online activities and protection of sensitive data has become indispensable for every internet user; consequently, the need to understand and use terminology denoting rapidly changing phenomena of this domain has increased considerably. Lithuanian cybersecurity terminology is still evolving: many cybersecurity concepts lack Lithuanian designations, or their Lithuanian designations exist but are not widely used and well-known. Such concepts are often denoted by

English or English-Lithuanian terms, which are used in various patterns in Lithuanian texts. Thus, the research on cybersecurity terminology in Lithuanian texts is believed to give insights on their usage and contribute to their management and dissemination.

The paper presents results of the frequential distribution analysis of terminology across genre-specific subcorpora in a Lithuanian cybersecurity corpus compiled and annotated for purposes of the research. The main aim of the paper is to discover tendencies of the usage of cybersecurity terms in Lithuanian texts of different genres. The research focuses on the following issues:

- overall distribution of cybersecurity terms (their density and diversity);
- distribution of English and hybrid (English-Lithuanian) terms and their usage patterns in Lithuanian sentences;
- establishment of most frequent cybersecurity terms and their main thematic groups in each genre.

In order to compile a dataset for the distribution analysis of cybersecurity terms across genre-specific texts, the following tasks were accomplished: 1) compilation of a corpus composed of texts of different genres used in specialised and popular discourses; 2) manual annotation of cybersecurity terms in the corpus texts and 3) automatic lemmatisation of the annotated units. The annotated data enabled the quantitative analysis which allowed drawing conclusions on the usage of terminology in the Lithuanian texts of the cybersecurity domain. The annotated corpus also has the added value: it can be used as a training and validation dataset for the development of automatic terminology extraction methods, which are future research objectives of the authors.

1. Related work

So far, the research on Lithuanian cybersecurity terminology has focused on the terms used in EU documents: Stunžinas analysed Lithuanian terms with the constituent „kibernetinis“ (‘cyber’) in EU documents and compared them with their synonymous variants in online texts (Stunžinas, 2017); Rackevičienė and Mockienė investigated English terms that include the lexical item “cyber”, and their Lithuanian counterparts used respectively in the English and Lithuanian versions of EU documents (Rackevičienė, Mockienė, 2020). Thus, the corpus-driven distribution analysis presented in this paper will complement the research on Lithuanian cybersecurity terminology by indicating its usage tendencies across different genres of texts published in Lithuania.

The presented distribution analysis is based on the results of manual terminology annotation, which is a specific type of language data annotation widely applied in projects on the development of automatic term extraction (ATE) methods. A corpus with manually annotated terms (gold standard corpus) was used for the development of tools for automatic extraction of Lithuanian education and science terminology

(Bielinskienė et al., 2015); numerous ATE research projects for other languages are reported in Bada et al. (2010); Schumann, Fischer (2016); Hätty et al. (2017), etc.

Term recognition, and subsequently its annotation and extraction, are based on two basic qualities of a term: unithood, which refers to the degree of stability of syntagmatic combinations and termhood, which refers to the degree to which a stable lexical unit is related to some domain-specific concepts (Kageura, Umino, 1996; Nakagawa, 2001; Hätty, Schulte im Walde, 2018). The first criterium is relevant only to term candidates that are multi-word expressions, while the second is relevant to term candidates of all forms.

In addition, in term recognition processes, it is important to consider that "a term candidate can be associated to a domain to different degrees" (Hätty, Schulte im Walde, 2018). Roelcke (1999) groups terms into four layers: intra-subject terminology specific to the focus domain, inter-subject terminology specific both to the focus domain and other domains, extra-subject terminology not specific to the focus domain but used within it and non-subject terminology, which is shared across all specific domains (Roelcke, 1999 as cited in Hätty et al., 2017). This classification is presented in Figure 1.

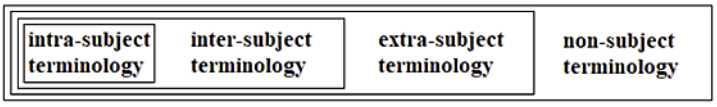


Fig. 1.
Layers of terminology according to Roelcke (1999), translated by Hätty et al. (2017)

Thus, terminology annotation involves several tasks: firstly, determining which lexical units function as terms in a text, and secondly, determining which domain terms belongs to. In addition, terms may be categorised further according to their conceptual characteristics. Each terminology annotation project develops its annotation scheme and term candidate evaluation criteria. These criteria vary from detailed and strict to loose and liberal and depend on the project aims and approach to the notion of termhood (Hätty et al., 2017).

2. Stages and methodology of the research

The research presented in the paper was performed in several stages, each of which is described below.

In the **first stage**, texts on cybersecurity issues written in different genres and used in specialised and popular discourses were collected, and a cybersecurity (CS) corpus was compiled.

In the **second stage**, the cybersecurity terms used in the corpus were annotated manually by four annotators using the annotation software developed for the purposes of the project – *QuickTag*.

The terminology annotation scheme and guidelines were based on linguistic and conceptual annotation criteria. The linguistic criteria

determined the grammatical categories of lexical units which had to be annotated: it was decided to limit annotation to terms which are nouns, noun phrases, initialisms that function as nouns or are parts of noun phrases. Meanwhile, the conceptual criteria determined the main tagset, which comprised the following categories (c.f. Roelcke, 1999, as cited in Hätty et al., 2017):

- Intra-subject terminology – the terminology of the cybersecurity domain;
- Inter-subject terminology – the terminology used in cybersecurity and other closely related domains.

The distinction between intra- and inter-subject terminologies was introduced to analyse what domains are mostly related to and dependent on cybersecurity. However, it was very difficult to achieve inter-annotator agreement on this distinction, and in our subsequent quantitative analyses, terms tagged as intra- and inter-subject terminology were combined.

In addition to the main tagset, terms were tagged with additional attributes, which allowed indicating special cases of term usage (e.g., terms used in abbreviated forms; complex terms interrupted by other words). A special attribute was added to terminological units consisting of a combination of English and Lithuanian words, e.g., *botnet tinklas* 'botnet network', *DDoS ataka* 'DDoS attack'. In the paper, such terminological multi-word units are referred to as English-Lithuanian hybrids (c.f. multi-word hybrids in Mockienė (2016); hybrid complex terms in Wiese (2018)).

The annotation software tool *QuickTag* allowed attaching tags from the main tagset (special tags for Lithuanian and English terms) and additional attributes to the annotated data. It also allowed tagging of nested terms, i.e., terms nested in more complex terms. Finally, *QuickTag* extracted lists of annotated units to an *MS Excel* spreadsheet file with statistical metadata for analysis purposes.

In the **third stage** of the research, the tagged terms were grouped according to the genres of texts they appeared in and then automatically lemmatised.

Lemmatisation was performed using the morphological analyser developed under the project *Semantika.lt*. [1] The analyser is one of the two most used Lithuanian morphological analysers (the second being *Lemuoklis* [2]); it was chosen due to its higher lemmatisation precision determined by Kapočiūtė-Dzikienė et al. (2017).

In the paper, the lemmatised terms are referred to as unique terms, i.e., a unique term is the main form of a term that generalises all its grammatical forms.

In the **final (fourth) stage**, the tagged terms and unique terms of each genre were quantitatively analysed and compared using *MS Excel* software functions of data sorting and analysis. The analysis focused on the following parameters: density and diversity of terms (cf. the studies on terminological density by Ferraresi (2019), lexical density and diversity

by Nasseri, Thompson (2021)) and prevalence of terms based on their frequency and distribution (c.f. Sinclair, 1991; Biber, 2002). The results obtained in different genres were juxtaposed to determine the similarities and differences of the term usage.

In the sections below, the structure of the compiled corpus and the results of the quantitative analysis of the annotated data are presented.

3. Structure of the cybersecurity corpus

The cybersecurity domain is highly heterogenous and includes various types of texts used in specialised and popular discourses. In order to represent the diversity of lexis usage in the cybersecurity domain, texts from four genres were selected for the research. Thus, the corpus includes four subcorpora: three of the subcorpora contain texts of genres specific to specialised discourses (legal texts, expert texts and academic texts), and one subcorpus contains texts of a popular discourse (media texts) (c.f. Wall, 2007). The size of the whole corpus is 135,667 words; the sources of the corpus cover the period 2011–2021.

Subcorpus of legal texts includes legally-binding documents on cybersecurity: The Cybersecurity Law passed by the Parliament of the Republic of Lithuania (Seimas) and resolutions issued by the Government of the Republic of Lithuania. The resolutions deal with the following issues: approval of the cybersecurity strategy, plans on management of cyber incidents, organizational and technical cybersecurity requirements for critical information infrastructure and state information resources.

Subcorpus of expert texts comprises texts produced by cybersecurity practitioners: reports, information bulletins and recommendations by the National Cyber Security Centre under the Ministry of the National Defence of the Republic of Lithuania and recommendations by "Microsoft" company. These texts contain an analysis of the cybersecurity situation in Lithuania (cyber resilience of various devices and software, cyber incidents that occurred in Lithuania) and recommendations on how to protect computers and data from cyber-attacks.

Subcorpus of academic texts includes educational and scientific texts: textbooks and theses on the investigation and management of cyber incidents.

Subcorpus of media texts comprises popular and specialised texts on various cybersecurity issues: articles in the mass media portals *15min.lt* and *Delfi.lt*, articles in more specialised portals *technologijos.lt* and *sprendimaiverslui.lt* and articles in the special issue of the news portal *Apžvalga* dedicated to cybersecurity (*Kibernetinio saugumo apžvalga*).

The percent proportions of the subcorpora, and the size in words of each subcorpus are provided in Figure 2 and Figure 3.

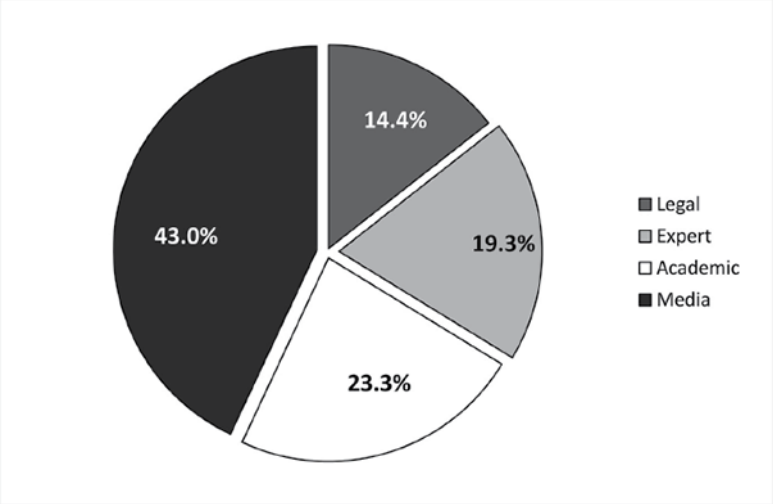


Fig. 2.
Composition of the CS corpus: percent proportions of subcorpora

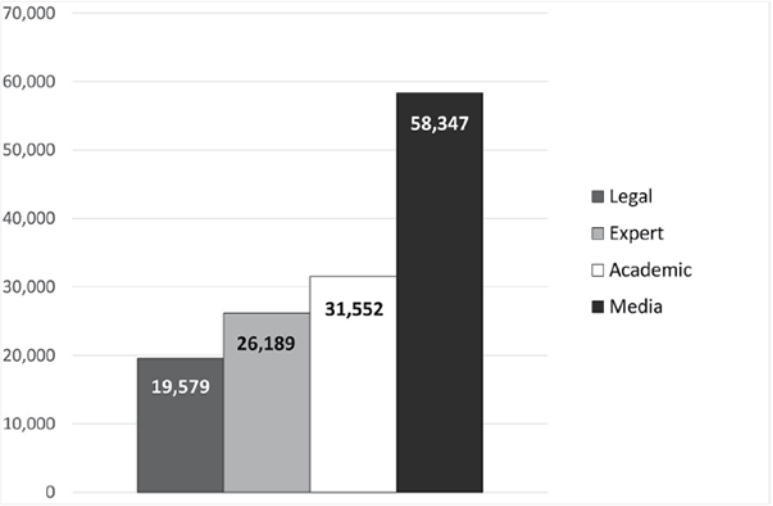


Fig. 3.
Composition of the CS corpus: size in words of each subcorpus

The distribution of different genres in the corpus was mainly determined by the accessibility of CS texts; therefore, media articles, which are most accessible, constitute the most significant part of the corpus (43.0%). The subcorpora containing texts of specialised discourses (legal texts, expert texts, academic texts) are considerably smaller, but their size was sufficient for comparative analysis of the data.

4. Distribution analysis of cybersecurity terms across genre-specific subcorpora

The terminological data of the compiled corpus was manually annotated and automatically lemmatised according to the methodology described in Section 2 above. The compiled dataset allowed to perform distribution analysis, the results of which are presented in this section.

4.1. Overall distribution of cybersecurity terms

The overall distribution analysis focuses on the two parameters: density and diversity of cybersecurity terms across the genre-specific subcorpora. The measurement of terminology density is based on the number of all annotated cybersecurity terms, while the measurement of terminology diversity is on the number of unique (lemmatised) cybersecurity terms.

The quantitative analysis of the data has revealed that the annotated corpus contains 8,813 annotated cybersecurity terms out of which 2,579 terms are unique. As the genre-specific subcorpora differ in size, the relative frequencies of terms per 1,000 words have been calculated to compare the density and diversity of terminology in the subcorpora. The following formulae have been used for the calculations (cf. Biber et al., 2002):

- Density of terminology:

$$\text{Relative frequency} = \frac{\text{Number of all terms in a subcorpus}}{\text{Number of words in a subcorpus}} \times 1{,}000.$$

- Diversity of terminology:

$$\text{Relative frequency} = \frac{\text{Number of unique terms in a subcorpus}}{\text{Number of words in a subcorpus}} \times 1{,}000.$$

Relative frequencies of all annotated terms and unique terms in the genre-specific subcorpora are provided in Fig. 4.

Figure 4 shows that the density of CS terms in the whole corpus equals 64.99 terms per 1,000 words, while diversity equals to 19.01 unique terms per 1,000 words.

The density of CS terms is the highest in the expert and legal texts, while the second position according to this measure is taken by the academic texts. Meanwhile, the media texts have the lowest density of all genres. The density counts confirm the overall tendencies of terminology usage: popular discourse texts (media texts), the addressee of which is the general public, are terminologically less dense than specialised discourse texts, which are much more specialized and targeted mostly at experts and professionals.

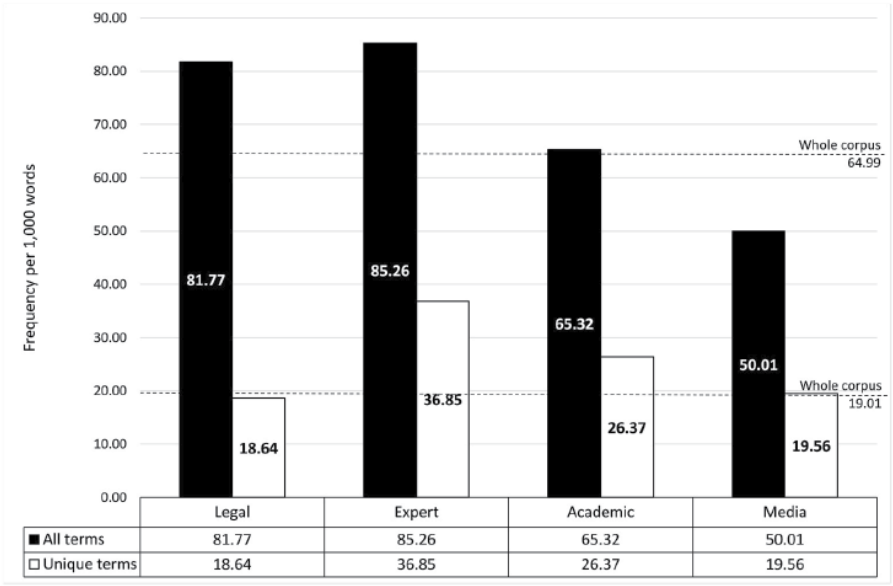


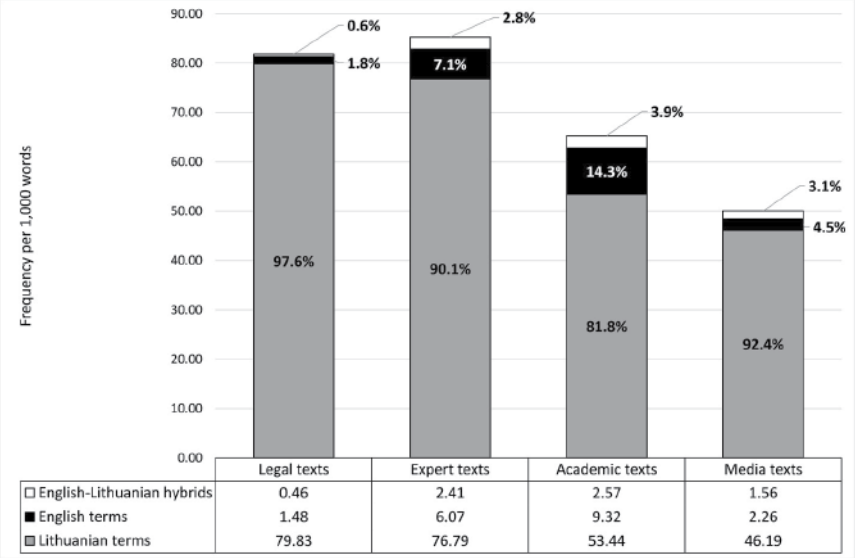
Fig. 4.
Density and diversity of CS terms across subcorpora

The distribution according to the diversity measure is different. The diversity of cybersecurity terms is highest in the expert texts; they are followed by the academic texts, and the lowest diversity of cybersecurity terms was detected in the legal and media texts. The diversity counts reveal the particular usage of cybersecurity terms in the legal texts: even though the legal texts contain a high number of cybersecurity terms, their diversity is very low, as terms are used repetitively. It could be explained by the nature of legal acts on cybersecurity: most of them describe general issues related to cybersecurity strategic planning and requirements and do not contain extensive descriptions of technical cybersecurity details, which would require more diverse terminology.

The analysis shows that, as it could be expected, the expert texts produced by cybersecurity practitioners are the most valuable for terminology extraction as their terminological density and diversity are the highest among the investigated text genres.

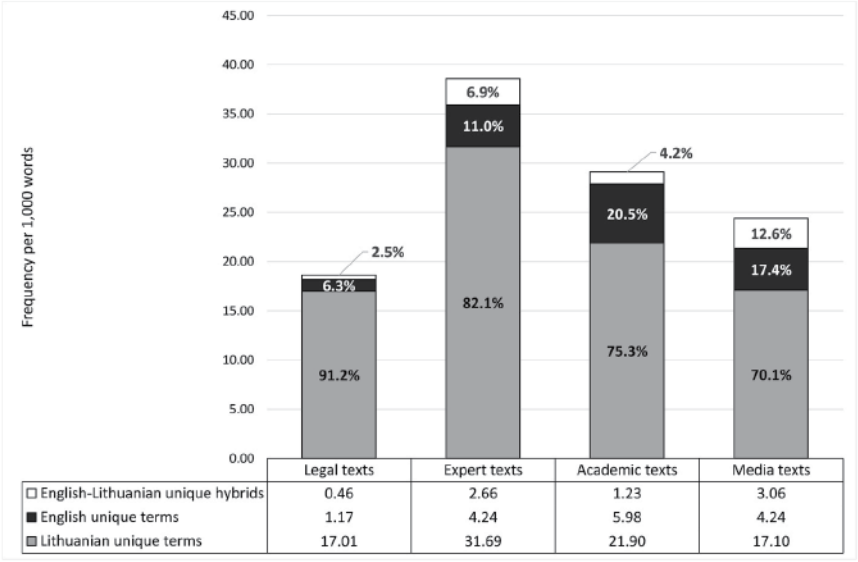
4.2. Distribution and usage patterns of English and English-Lithuanian terms

Lithuanian terms and English terms used in the CS corpus were tagged with separate tags during the annotation. English-Lithuanian hybrids (combinations of English and Lithuanian words) were tagged as Lithuanian terms with an additional attribute indicating that they are English-Lithuanian hybrids. Based on the tagged and lemmatised data, the proportions of the tagged and unique Lithuanian terms, English terms and English-Lithuanian hybrids were calculated (see Fig. 5 and Fig. 6).



	Legal texts	Expert texts	Academic texts	Media texts
English-Lithuanian hybrids	0.46	2.41	2.57	1.56
English terms	1.48	6.07	9.32	2.26
Lithuanian terms	79.83	76.79	53.44	46.19

Fig. 5.
Density of Lithuanian, English and English-Lithuanian terms across subcorpora



	Legal texts	Expert texts	Academic texts	Media texts
English-Lithuanian unique hybrids	0.46	2.66	1.23	3.06
English unique terms	1.17	4.24	5.98	4.24
Lithuanian unique terms	17.01	31.69	21.90	17.10

Fig. 6.
Diversity of Lithuanian, English and English-Lithuanian terms across subcorpora

The data presented in Figures 5 and 6 reveal that English and English-Lithuanian terms are present in all genres included in the CS corpus. The highest density of English and English-Lithuanian terms was detected in the academic texts. The second position is taken by the expert texts and the third position by the media texts. The legal texts are the "cleanest", as only 2.4% of all tagged terms in them are English or English-Lithuanian hybrids. The diversity of English and English-Lithuanian terms is the highest in the media texts. They are followed by the academic texts and by the expert texts. The legal texts have the lowest diversity of English and English-Lithuanian terms (8.8%). The counts show that legal texts stand out among the texts of other genres. This might be explained

by several reasons: firstly, the higher requirements to the language used in legal acts in Lithuania (they should contain only standardised terminology) and, secondly, contents of legal acts, most of which do not contain technical details that would require specific terminology; namely technical cybersecurity terminology usually contains unlocalised elements.

English terms are used in various ways in the CS corpus. These ways may be grouped into two main patterns (leaving aside rare cases): usage of the English terms in bracketed insertions in Lithuanian sentences and usage of English terms as integral parts of Lithuanian syntactic structures.

The first usage pattern (usage of English terms in bracketed insertions) is most frequent; it was detected in all genres represented in the corpus. The English terms used according to this pattern are inserted in brackets with the shortening *angl*. 'English' (in some cases, this shortening is missing), e.g.:

žalingo kodo programinė įranga (angl. malware),
tikslingos atakos (angl. advanced persistent threat),
elektroninių paslaugų trikdymo (angl. Denial of Service, DoS) atakos,
finansiniai trojos arkliai (angl. financial trojans)

The second usage pattern (usage of English terms as integral parts of Lithuanian sentences) is not as frequent as the previous pattern: it was detected in three of the genres represented in the corpus (expert, academic and media), most of the cases of this usage pattern were present in the media texts. This usage has the following main subpatterns:

- English terms used in the original form with or without quotation marks [3] : *„fake news“, „phishing“, „rootkit“, ransomware,* DLL, SSH, APT.
- English terms used in the semi-localised form (with Lithuanian endings) with or without quotation marks: *phishingas, phishingo, botneto, botnetą, „botnetas“, botus, bot'us, „botai“, „botų“, „botus“, „botais“.*

There are some clear tendencies in the usage of English terms belonging to the first subpattern: the English terms, which are nouns/noun phrases, are mostly used in quotation marks, while the terms which are initialisms are used without them:

> *„Fake news“ tapo viena didesnių 2017 m. problemų... 'Fake news have become one of bigger problems in 2017'. Jeigu prastai parašysite programą ir paliksite atviras galimybes jai pačiai atlikti DLL paiešką, tuomet gali kilti sunkumų.* 'If the software is poorly coded and the possibilities to perform DLL search on its own are left open, problems can be faced'.

The English terms belonging to the second subpattern are used with Lithuanian endings. In most cases, the endings are added directly, in one case – with an apostrophe. However, these terms still seem foreign in Lithuanian sentences as their localisation is limited to the added case ending; therefore, they are often written in quotation marks that indicate that they are taken from the vocabulary of a foreign language. The English

terms belonging to this subpattern were detected mostly in the media texts:

> *O taip pat šiomis dienomis stebime neįtikėtinus kiekius phishingo... 'And we're also tracking incredible amounts of phishing these days.' Galime tikėtis daugiau „botų", kenksmingų melagingų naujienų, DDoS atakų ir naujų išpirkos reikalaujančių kenkėjų. 'We can expect more bots, malicious fake news, DDoS attacks, and new ransomware.'*

In addition to the discussed two main usage patterns of English terms, a considerable number of cases of English-Lithuanian hybrid terms was detected. In most cases, such hybrid terms have the following structure: an English word/phrase/initialism + a Lithuanian noun/noun phrase. A Lithuanian constituent designates the generic concept, while an English attribute specifies it, e.g.:

botnet tinklas 'botnet network', man-in-the-middle kibernetinės atakos 'man-in-the-middle cyber attacks', code cave metodas 'code cave method', DDoS ataka 'DDoS attack', C2 karas 'C2 war'.

In some cases, the English attribute is written in quotation marks:

„Brute force" atakos 'Brute force attacks', „open-rdp" nuotolinė prieiga 'open-rdp remote access', „botnet" tinklas 'botnet network'.

Such hybrid formations are present in all genres represented in the corpus. As was mentioned above, they were tagged separately, and therefore, were included in the quantitative density and diversity analysis. The highest density of such formations was detected in academic and media texts, while the highest diversity was in media and expert texts.

The analysis of English and English-Lithuanian terms reveals that such terms are used for designation of various types of cyber-attacks and specific technical concepts referring to computer software that may be affected by cyber-attacks or used to complete them. The reasons for their usage may be lack of Lithuanian designations, unawareness of their existence or attempts to make information clearer as existing Lithuanian designations are still not widespread and not well-known. The latter reason is evident in the usage of English terms in bracketed insertions, which follow Lithuanian terms: such cases indicate that Lithuanian terms are still not well-known and, therefore, the authors of the texts add original English terms to make the information clearer.

4.3. The most frequent cybersecurity terms across genre-specific subcorpora

The annotated data also allowed us to determine which cybersecurity terms dominate in each subcorpus. On the basis of the compiled dataset, TOP 10 lists of the most frequent terms in every subcorpus have been generated (see Tables 1–2).

Table 1
Lithuanian TOP 10 CS terms in legal and expert texts

Terms in legal texts	Rel. freq.	Terms in expert texts	Rel. freq.
kibernetinis incidentas 'cyber incident'	9.91	*slaptažodis* 'password'	3.51
kibernetinis saugumas 'cybersecurity'	7.51	*kibernetinis incidentas* 'cyber incident'	2.41
RIS 'communication and information system'	3.88	*antivirusinė programa* 'antivirus program'	2.14
ypatingos svarbos informacinė infrastruktūra 'critical information infrastructure'	3.42	*interneto svetainė* 'website'	1.79
kibernetinių incidentų valdymas 'cyber incident management'	2.86	*kenkimo PĮ* 'malicious software'	1.37
kibernetinio saugumo subjektas 'cyber security subject'	2.50	*kibernetinis saugumas* 'cybersecurity'	1.37
RIS naudotojas 'user of communication and information system'	1.84	*pažeidžiamumas* 'vulnerability'	1.34
ypatingos svarbos informacinės infrastruktūros valdytojas 'critical information infrastructure manager'	1.63	*piktavalis* 'hacker'	1.18
valstybės informacinis išteklius 'state information resource'	1.53	operacinė sistema 'operating system'	1.03
slaptažodis 'password'	1.43	*IP adresas* 'IP address'	0.95

Table 2.
Lithuanian terms TOP 10 CS terms in academic and media texts

Terms in academic texts	Rel. freq.	Terms in media texts	Rel. freq.
incidentas 'incident'	3.83	*kibernetinis saugumas* 'cybersecurity'	3.10
įkaltis 'evidence'	3.20	*dezinformacija* 'dezinformtion'	2.01
NEE 'crime in the electronic space'	2.57	*kibernetinė ataka* 'cyber attack'	1.27
kibernetinis incidentas 'cyber incident'	1.39	*ataka* 'attack'	1.17
elektroninis nusikaltimas 'electronic crime'	1.27	programišius 'hacker'	1.13
kibernetinis saugumas 'cybersecurity'	0.98	*DI (dirbtinis intelektas)* 'artificial intelligence'	0.99
nusikaltimas elektroninėje erdvėje 'crime in the electronical space'	0.79	*kibernetinė erdvė* 'cyberspace'	0.84
informacinis karas 'information warfare'	0.70	*kibernetinė grėsmė* 'cyber threat'	0.74
nusikaltimas 'crime'	0.67	*kibernetinis incidentas* 'cyber incidence'	0.70
elektroninis įkaltis 'electronic evidence'	0.63	*propaganda* 'propaganda'	0.70

The frequency values in TOP 10 lists show that the dominating terms in the legal texts are most repetitive: their frequency values are higher than the frequency values of the terms in respective positions in TOP 10 lists of other subcorpora.

Only two terms occur in the TOP 10 lists of all subcorpora: *kibernetinis saugumas* ('cybersecurity'), *kibernetinis incidentas* ('cybersecurity incident'). In addition, *slaptažodis* ('password') occurs in the lists of two subcorpora: the expert texts and the legal texts. However, the relative frequencies of these terms are very different: frequencies of *kibernetinis saugumas* vary from 7.51 in the legal texts to 0.98 in the academic texts; frequencies of *kibernetinis incidentas* vary from 9.91 in the legal texts to 0.70 in the media texts; frequencies of *slaptažodis* are 3.51 in the expert texts and 1.43 in the legal texts. Moreover, the synonymous terms denoting a hacker are used in the expert and media texts: *piktavalis* and *programišius*. Their frequencies are similar: 1.18 and 1.13 respectively. Other terms are non-repetitive and not synonymous across the genres. Their frequency depends on the dominating topics in each subcorpus.

In the legal texts, the dominating terms mostly designate the concepts referring to the objects of the Republic of Lithuania which have to be protected from cyber attacks: communication and information systems, in particular those constituting the critical information infrastructure

and storing the state information resources. Related to them are the terms denoting general concepts of a cyber incident and cyber incident management.

In the expert texts, most terms in the TOP10 list designate the concepts referring to malware and protection measures against it. The list also includes general IT terms (e.g., terms denoting an operating system, an IP address), which are an important layer of terminology in this subcorpus as it deals mainly with technical recommendations on computer software protection against cyber threats.

The most frequent terms in the academic texts reveal the main topics in this subcorpus – cybercrimes and their investigation methodology: six of the most frequent terms denote the concepts of a crime committed in cyberspace and digital evidence.

The TOP 10 list of the media texts includes cybersecurity terms denoting the general concepts of a cyber-attack, a cyber threat and a cyber incident. In addition, the list contains the terms designating the concepts referring to the phenomena of disinformation and propaganda.

Several TOP10 lists include full terms and their abbreviated forms: in the academic texts, a crime in the cyberspace is designated even by four terms (*nusikaltimas elektroninėje erdvėje . elektroninis nusikaltimas – nusikaltimas – NEE*), a cyber incident – by two terms (*kibernetinis incidentas – incidentas*); in the media texts, a cyber-attack – by two terms (*kibernetinė ataka – ataka*). Such cases indicate that in coherent texts dominating terms are often used in abbreviated forms when their meanings are clear from the context.

English terms and English-Lithuanian terms are not present in TOP 10 lists as their occurrence frequencies are much lower than frequencies of Lithuanian terms. However, as the analysis in the section above indicates, they constitute an important lexical layer of Lithuanian cybersecurity texts. English terms and English-Lithuanian hybrids mostly designate technical concepts referring to types of cyber-attacks and malware used to complete them. The most frequent English terms in the whole corpus include: *trojan horse, exploit, backdoor, botnet, phishing, denial of service, fake news, APT* (advanced persistent threat). The most frequent English-Lithuanian hybrids are *DDoS ataka* 'DDoS attack' and „*botnet. tinklas* 'botnet net'.

Conclusion

The conducted distribution analysis allows drawing the following conclusions:

1. The cybersecurity domain is highly heterogenous and encompasses various types of texts used in specialised and popular discourses. Specialised discourse texts are of different genres: legal texts which encompass legally binding documents on cybersecurity, expert texts which comprise texts produced by cybersecurity practitioners and academic texts written by

cybersecurity researchers. Numerous popular discourse texts on various cybersecurity issues may be found in media portals which may be grouped further into mass media texts and specialised media texts. They are targeted at different groups of readers (common readers and readers especially interested in cybersecurity or domains related to it) and, therefore, differ in the degree of popularisation.

2. The investigated expert texts are the most valuable for terminology extraction as their terminology density and diversity are the highest among the investigated genres. The lowest density of cybersecurity terms was established in the media texts; their terminological diversity is also rather low. The legal texts stand out among other genres: their terminological density is rather high, while their terminological diversity is low. This could be explained by the nature of the legal acts on cybersecurity: most legal acts describe general issues related to cybersecurity strategic planning and requirements and do not contain extensive descriptions of technical cybersecurity details, which would require diverse terminology.
3. English and English-Lithuanian terms are present in all genres included in the corpus. Their lowest density and diversity were established in the legal texts. This might be explained by several reasons: firstly, higher standards for the language used in legal acts in Lithuania and, secondly, contents of legal acts, most of which do not contain technical details that would require specific terminology.

   There are two main patterns of usage of English terms in the CS corpus: their usage in bracketed insertions which follow Lithuanian terms (e.g. žalingo kodo programinė įranga (angl. malware)) and their usage (in original or semi-localised form) as integral parts of the Lithuanian syntactic structures (e.g. „fake news“, APT, phishingas, „botai“). In addition, English terminological units are used in hybrid multi-word terms composed of English and Lithuanian constituents (e.g. botnet tinklas ‘botnet network’). Bracketed insertions and hybrid formations were detected with similar frequencies in all genres represented in the corpus, while original and semi-localised English terms were mostly present in the media texts.

   The reasons for the usage of English and English-Lithuanian terms may be the lack of Lithuanian designations, unawareness of their existence or attempts to make information clearer as existing Lithuanian designations are still not widespread and not well-known.
4. Only some basic terms occur in the TOP 10 lists of all genre-specific subcorpora. Most terms prevailing in the subcorpora are specific and reflect the dominating topics in each subcorpus. English terms and English-Lithuanian terms

are not present in TOP 10 lists as their frequency is much lower than the frequency of Lithuanian terms. They mostly designate concepts referring to types of cyber-attacks and malware used to complete them.

All in all, the distribution analysis reveals that it is important to investigate terminology across genres to get a full picture of its usage trends, as each genre has its own characteristics. Legal acts contain the "cleanest" terminology; however, it is not that diverse. The terminology of cybersecurity practitioners and researchers is often very specific, constituting professional jargon. Meanwhile, media texts contain a wide diversity of terminological formations, reflecting the evolution of cybersecurity terminology and attempts to create clearest and/or most attractive Lithuanian equivalents of English terms. The term usage trends show that Lithuanian cybersecurity terminology is still very young, inconsistent, often containing gaps filled with original English terms. Therefore, their collection, research and management are especially important for their further development.

Acknowledgments

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

Bada, M., Eckert, M., Palmer, M., Hunter, L., 2010. An overview of the CRAFT concept annotation guidelines. Proceedings of the Fourth Linguistic Annotation Workshop, pp. 207–211. Available at: https://aclanthology.org/W10-1833.pdf [Accessed 4 October 2021].

Biber, D., Conrad, S., Leech, G., 2002. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Bielinskienė, A., Boizou, L., Grigonytė, G., Kovalevskaitė, J., Rimkutė, E., Utka, A., 2015. Lietuvių kalbos terminų automatinis atpažinimas ir apibrėžimas [Automatic extraction and definition of Lithuanian terms]. Kaunas: Vytauto Didžiojo universitetas. Available at: https://www.lvb.lt/primo-explore/fulldisplay?vid=ELABA&docid=ELABAPDB8767813&context=L [Accessed 4 January 2022]. [In Lithuanian].

Ferraresi, A., 2019. How specialized (or popularized)? Terminological density as a clue to text specialization in the domain of food safety. Lingue Linguaggi, 29, pp. 17–39. Available at: https://www.researchgate.net/publication/333973359_HOW_SPECIALIZED_OR_POPULARIZED_Terminolo [Accessed 18 October 2021]. https://doi.org/10.1285/i22390359v29p17.

Hätty, A., Schulte im Walde, S. S., 2018. Fine-grained termhood prediction for German compound terms using neural networks. Proceedings of the Joint Workshop on Lin guistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pp. 62–73. Available at: https://aclanthology.org/W18-4909.pdf [Accessed 3 September 2021].

Hätty, A., Tannert, S., Heid, U., 2017. Creating a gold standard corpus for terminological annotation from online forum data. Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017). Available at: https://aclanthology.org/volumes/W17-70/ [Accessed 3 September 2021].

Kageura, K., Umino, B., 1996. Methods for automatic term recognition. *Papers of the National Center for Science Information Systems*, pp. 1–22.

Kapočiūtė-Dzikienė, J., Rimkutė, E., Boizou, L., 2017. A comparison of Lithuanian morphological analyzers. In: Ekštein, K., Matoušek, V., eds. Text, Speech, and Dialogue. TSD 2017. *Lecture Notes in Computer Science*, 10415. Springer, Cham. https://doi.org/10.1007/978-3-319-64206-2_6.

Mockienė, L., 2016. Formation of terminology of constitutional law in English, Lithuanian and Russian. Ph. D. Vilnius: Mykolas Romeris University. Available at: https://vb.mruni.eu/object/elaba:15561061/ [Accessed 4 January 2022].

Nakagawa, H., 2001. Experimental evaluation of ranking and selection methods in term extraction. In: Bourigault, D., Jacquemin, Ch, L'Homme, M-C., eds. *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins Publishing.

Nasseri, M., Thompson, P., 2021. Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. Assessing Writing, 47, 100511. Available at: https://www.sciencedirect.com/science/article/pii/S1075293520300726 [Accessed 18 October 2021]. https://doi.org/10.1016/j.asw.2020.100511.

Rackevičienė, S., Mockienė, L., 2020. Cyber law terminology as a new lexical field in legal discourse. *International Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique*. Springer Nature 33(3), pp. 673–687. https://doi.org/10.1007/s11196-020-09690-0.

Roelcke, T., 1999. *Fachsprachen. Grundlagen der Germanistik*. Erich Schmidt Verlag.

Schumann, A. K., Fischer, S., 2016. Compasses, magnets, water microscopes: annotation of terminology in a diachronic corpus of scientific texts. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 3578–3585. Available at: https://aclanthology.org/L16-1568.pdf [Accessed 3 September 2021].

Sinclair, J., 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stunžinas, R., 2017. Europos Sąjungos kibernetinio saugumo terminai su dėmeniu kibernetinis, (-ė): reikšmės, kilmė, sinonimija ir variantai. Terminologija [European Union Cybersecurity Terms with Compound kibernetinis, (-ė) (cyber(-): Meanings, Origin, Synonymy and Variations. Terminology], 24, pp. 145–163. Available at: http://lki.lt/wp-content/uploads/2018/03/Terrminologija_24_maketas2.pdf [Accessed 4 January 2022]. [In Lithuanian].

Wall, D. S., 2007. *Cybercrime: The transformation of crime in the information age*. USA: Wiley.

Wiese, I., 2018. Medical language. In: Humbley, J., Budin, G., Laurén, Ch. *Languages for special purposes: an international handbook*. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110228014.

Notes

1 https://semantika.lt/Help/Info/Solutions
2 https://klc.vdu.lt/anotatorius/
3 In Lithuanian, quotation marks are applied in a different way than in English: the opening mark is written at the level of commas, while the closing mark is at the level of apostrophes („x“).