

English- Lithuanian Language Resources for the Cybersecurity Domain

Sigita Rackevičienė
Mykolas Romeris
University

CLARIN



Five bilingual language resources for the cybersecurity domain are deposited in CLARIN-LT repository, two more are being prepared for the depositing.



English-Lithuanian Parallel Cybersecurity Corpus - DVITAS



(Vytautas Magnus University; Mykolas Romeris university / 2022-02-05)

Author(s):

Utkā, Andrius ; Rackevičienė, Sigita ; Rokas, Aivaras ; Bielinskienė, Agnė ; Mockienė, Liudmila ; Laurinaitis, Marius

Šis įrašas turi 3 bylas (6.56 MB).

Publicly Available

English-Lithuanian Comparable Cybersecurity Corpus - DVITAS



(Vytautas Magnus university; Mykolas Romeris university / 2022-02-05)

Author(s):

Utkā, Andrius ; Rackevičienė, Sigita ; Rokas, Aivaras ; Bielinskienė, Agnė ; Mockienė, Liudmila ; Laurinaitis, Marius

Šis įrašas turi 12 bylas (66.34 MB).

Academic Use

English-Lithuanian Parallel Cybersecurity Corpus - DVITAS v2.0



(Mykolas Romeris university; Vytautas Magnus university / 2024-12-31)

Author(s):

Mickevič, Jolanta ; Utkā, Andrius ; Rackevičienė, Sigita ; Rokas, Aivaras ; Bielinskienė, Agnė ; Mockienė, Liudmila ; Laurinaitis, Marius

Šis įrašas turi 3 bylas (9.1 MB).

Publicly Available

Lithuanian-English Cybersecurity Termbase v.0.1



(Vytautas Magnus University; Mykolas Romeris University / 2023-04-13)

Author(s):

Utkā, Andrius ; Rackevičienė, Sigita ; Bielinskienė, Agnė ; Laurinaitis, Marius ; Mockienė, Liudmila ; Rokas, Aivaras

Šis įrašas turi 3 bylas (916.04 KB).

Publicly Available

Survey Data on Preferences of Lithuanian Cybersecurity Terminology



(Mykolas Romeris University; Vytautas Magnus University / 2024-10-04)

Author(s):

Rackevičienė, Sigita ; Utkā, Andrius

Šis įrašas turi 3 bylas (39.98 KB).

Publicly Available

Language resources for the Cybersecurity (CS) domain are being created as an integral part of the project and its post-project activities on CS terminology.



VYTAUTAS
MAGNUS
UNIVERSITY
MCMXXII



Mykolas Romeris
University

CLARIN-LT



Nexus
Linguarum

Why Cybersecurity (CS)?

Relevance – global digital connectivity
& cloud-based services

Security – challenges of securing
sensitive data

Dynamics – rapid evolution of the
domain

Standardisation – need to harmonise
Lithuanian terminology



Application of Language Resources (LR):

Technology

- Development of deep learning systems for automatic term extraction

Terminology Management

- Creation of a bilingual termbase

Terminology Research

- Analysis of conceptual, linguistic & pragmatic aspects
- Analysis of sociolinguistic aspects

Parallel and comparable corpora

Corpora in CLARIN-LT repository

English-Lithuanian comparable and parallel corpora are openly available in the CLARIN-LT repository:

- Comparable corpus: plain texts and morphologically annotated texts
- Parallel corpora: aligned TMX files (the 1st version and the updated 2nd version)



Corpus

CLARIN-LT

English-Lithuanian Parallel Cybersecurity Corpus - DVITAS

(Vytautas Magnus University; Mykolas Romeris university / 2022-02-05)

Author(s):
Utkā, Andrius ; Rackevičienė, Sigita ; Rokas, Aivaras ; Bielinskienė, Agnė ; Mockienė, Liudmila ; Laurinaitis, Marius

Šis įrašas turi 3 bylas (6.56 MB).

Publicly Available

Corpus

CLARIN-LT

English-Lithuanian Comparable Cybersecurity Corpus - DVITAS

(Vytautas Magnus university; Mykolas Romeris university / 2022-02-05)

Author(s):
Utkā, Andrius ; Rackevičienė, Sigita ; Rokas, Aivaras ; Bielinskienė, Agnė ; Mockienė, Liudmila ; Laurinaitis, Marius

Šis įrašas turi 12 bylas (66.34 MB).

Academic Use

Corpus

CLARIN-LT

English-Lithuanian Parallel Cybersecurity Corpus - DVITAS v2.0

(Mykolas Romeris university; Vytautas Magnus university / 2024-12-31)

Author(s):
Mickevič, Jolanta ; Utkā, Andrius ; Rackevičienė, Sigita ; Rokas, Aivaras ; Bielinskienė, Agnė ; Mockienė, Liudmila ; Laurinaitis, Marius

Šis įrašas turi 3 bylas (9.1 MB).

Publicly Available

Corpora on *Sketch Engine* platform



DASHBOARD

KS_LT PALYGINAMASIS **CORPUS INFO** **MANAGE CORPUS**

- Word Sketch**
Collocations and word combinations
- Word Sketch Difference**
Compare collocations of two words
- Thesaurus**
Synonyms and similar words
- Concordance**
Examples of use in context
- Parallel Concordance**
Translation search
- Wordlist**
Frequency list
- N-grams**
Multiword expressions (MWEs)
- Keywords**
Terminology extraction
- Trends**
Diachronic analysis, neologisms
- Text type analysis**
Statistics of the whole corpus

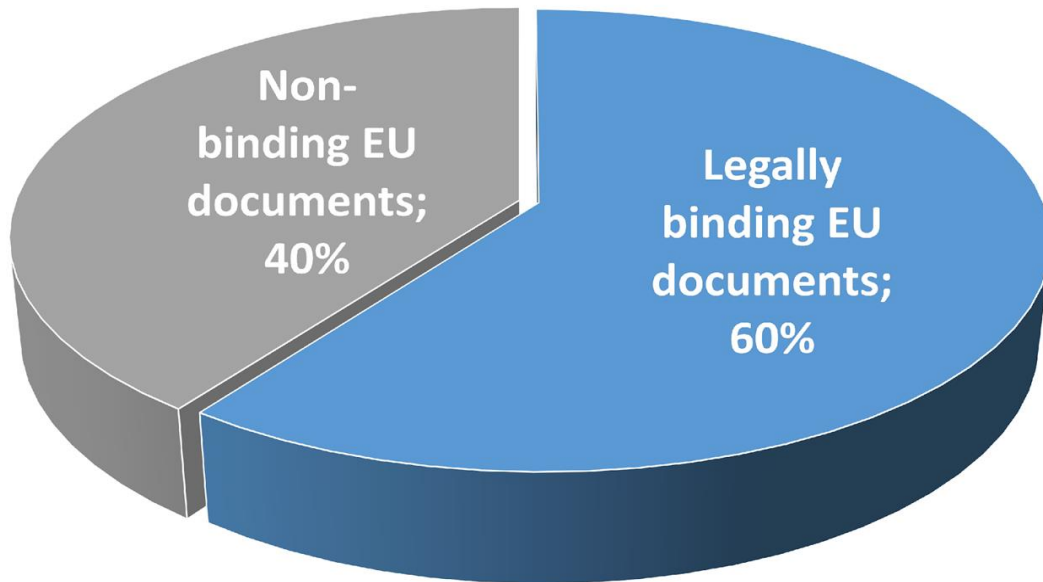
- KS_LT palyginamasis
- KS lygiagretusis, Lithuanian
- KS lygiagretusis#2, Lithuanian
- KS_EN palyginamasis
- KS lygiagretusis, English
- KS lygiagretusis#2, English

Texts types in corpora

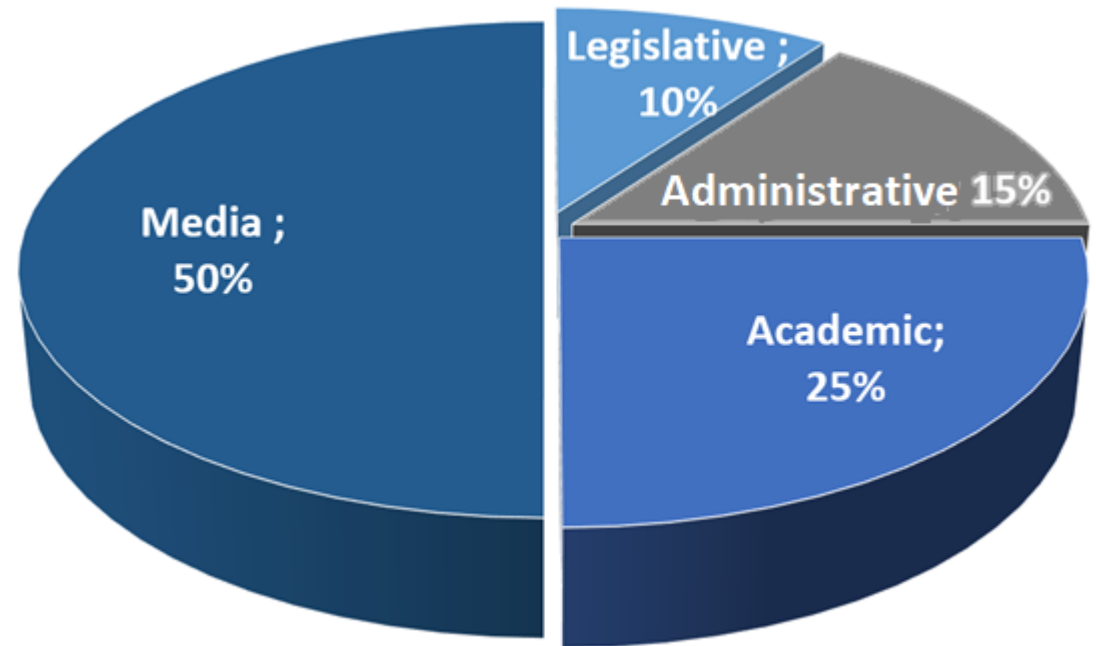
EN-LT Parallel corpus (Version 2)

EN-LT Comparable corpus

Period: 2006-2022

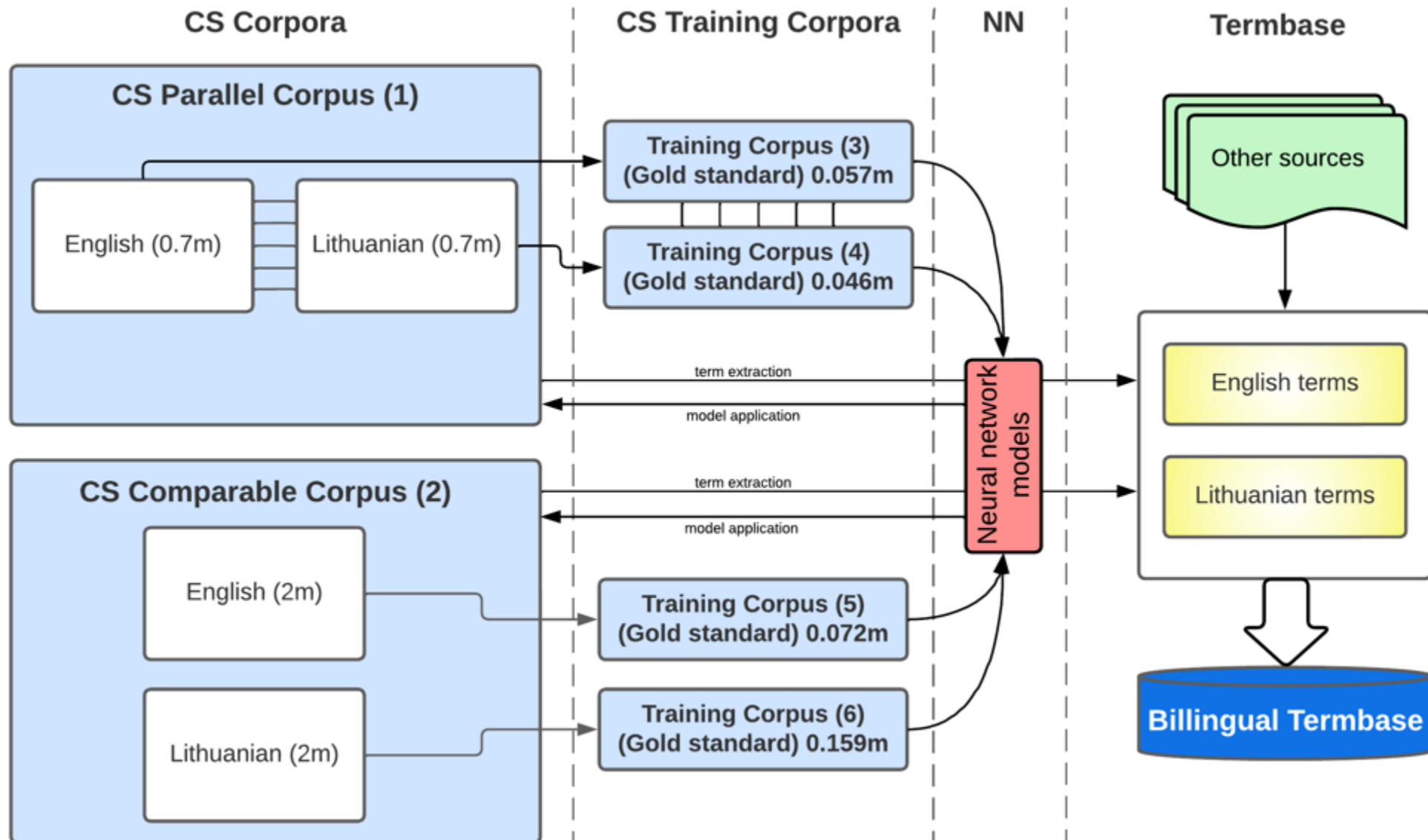


2m words



4m words

CYBERSECURITY RESOURCE SYSTEM



Bilingual terminology database (lexical-conceptual resource)

Bilingual CS Termbase Available Online:

CLARIN-LT repository – in TBX format for machine learning and AI development

Terminologue platform (administered by Dublin City University) – user-friendly format for terminology users: translators, editors, CS students and professionals, and other Internet users

English-Lithuanian CS termbase in the CLARIN-LT repository in TBX (termbase exchange) format used for representing and exchanging terminological data in a structured XML format.

LexicalConceptualResource

CLARIN-LT


Lithuanian-English Cybersecurity Termbase v.0.1

(Vytautas Magnus University; Mykolas Romeris University / 2023-04-13)

Author(s):
Utkā, Andrius ; Rackevičienė, Sigita ; Bielinskienė, Agnė ; Laurinaitis, Marius ; Mockienė, Liudmila ; Rokas, Aivaras

Šis įrašas turi 3 bylas (916.04 KB).

Publicly Available



```
<termEntry id="eid-61">
  <descrip type="subjectField">Kibernetiniai
incidentai » Veiklos</descrip>
  <langSet xml:lang="lt">
    <ntig>
      <termGrp>
        <term>kibernetinė ataka</term>
        <termNote
type="normativeAuthorization">*****</termNote>
      </termGrp>
    </ntig>
    <ntig>
      <termGrp>
        <term>kibernetinis išpuolis</term>
        <termNote
type="normativeAuthorization">****</termNote>
      </termGrp>
    </ntig>
```

English-Lithuanian CS termbase on Terminologue platform in a web-based user-oriented format

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

EditingAdministrationConfiguration

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 project was included as a use case in the COST Action CA18209 "European network for Web-centred linguistic data science (NexusLinguarum)".

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.

DDoS attack paskirstytojo paslaugos trikdymo ataka cyber incident management data centre hacker cybersecurity certification informacijos saugumo incidentas cross-site scripting attack botnet turinio iškraipymo ataka DNS hijacking tapatybės ir prieigos valdymas „Surask klaidą“ programa dictionary attack zero-day bug rizikos mažinimas asimetrinė kriptografija kriptografinis raktas duomenų vagystė hash value cybersecurity hygiene DLP trolis netikra paskyra IIoT Computer Emergency Response Team backup ransomware attack cyberspace intelligence 1 cyberwar

Lithuanian

A B C Č D E Ė Ė 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

Kibernetinė gynyba ir atsakas/Cyber defence and response
Kibernetiniai taikiniai/Cyber targets

Kibernetinis pažeidžiamumas/Cyber vulnerabilities
Kibernetiniai incidentai/Cyber incidents

Macrostructure of the termbase

► Kibernetinė gynyba ir atsakas/Cyber defence and response

► Kibernetiniai taikiniai/Cyber targets

Kibernetinis pažeidžiamumas/Cyber vulnerabilities

► Kibernetiniai incidentai/Cyber incidents

▼ Kibernetiniai incidentai/Cyber incidents

Atakos/Attacks

Grėsmės/Threats

Priemonės/Means

Vykdytojai/Executors

Žala/Harm

▼ Kibernetiniai taikiniai/Cyber assets

Duomenys/Data

Funkcionalumai/Functionalities

Infrastruktūros/Infrastructures

Paslaugos/Services

Programinė įranga/Software

Techninė įranga/Hardware

▼ Kibernetinė gynyba ir atsakas/Cyber defence and response

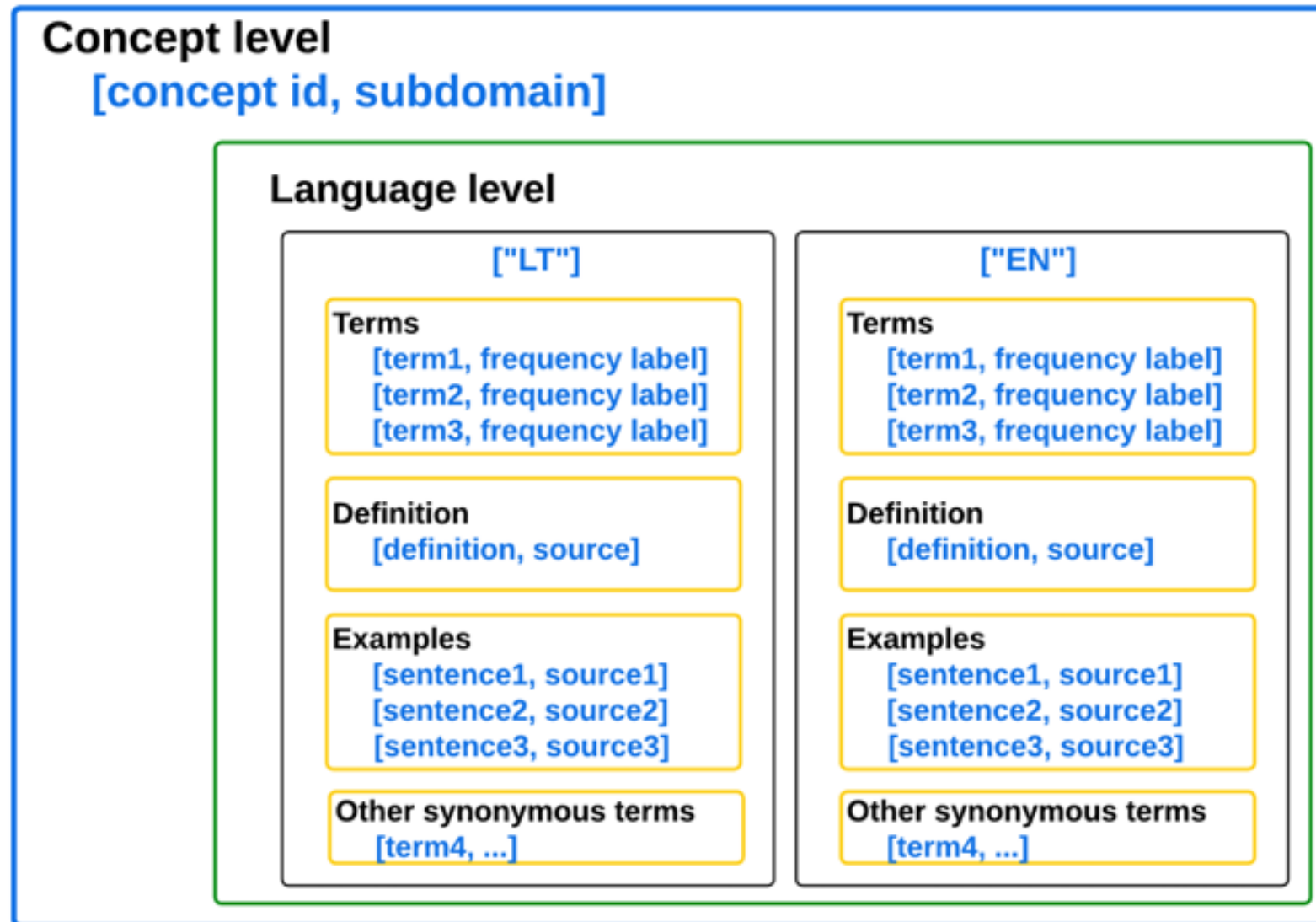
Priemonės/Means

Standartai/Standards

Subjektai/Cybersecurity entities

Veiklos/Activities

Microstructure of a termbase (structure of a terminological entry)



Microstructure of the termbase

Kibernetiniai incidentai » Atakos

Cyber incidents » Attacks

LT **DDoS ataka** ****

paskirstytojo atsisakymo aptarnauti ataka **

paskirstytojo paslaugos trikdyimo ataka **

paslaugos trikdyimo technika, kuri atakai atlikti naudoja daugybę kompiuterių

— Pagal JAV nacionalinio standartų ir technologijos instituto glosarijų

PAVYZDŽIAI: Naujesnės, intensyvesnės <DDoS atakų> formos apima procesą, vadinamą „atminties kaupimu“, kuris naudoja neapsaugotas, atvirojo kodo objektų talpyklos sistemas, kad sustiprintų prieigos užklausas ir didesniu nei terabaitas srautu užplūstų interneto svetaines.

— KS LT PALYGINAMASIS/AKADEMINIAI

Pasak policijos atstovo, taip pat būta bandymų šturmuoti policijos svetaines rengiant vadinamąsias <paskirstytojo atsisakymo aptarnauti (DDoS) atakas>.

— KS LT PALYGINAMASIS/ŽINIASKLAIDA

Kiti sinoniminiai variantai DDoS išpuolis, paskirstyto atsisakymo aptarnauti kibernetinė ataka, paskirstyto atsisakymo aptarnauti ataka, paskirstytoji atsisakymo aptarnauti ataka, išskirstytoji atsisakymo aptarnauti sistemą ataka, paskirstyto paslaugų trikdyimo ataka, paskirstytoji paslaugų trikdyimo ataka, paskirstyta paslaugų ribojimo ataka, paskirstytoji aptarnavimo perkrovos ataka, paskirstytoji paslaugų blokavimo ataka

EN **DDoS attack** ****

distributed denial of service attack ***

a denial of service technique that uses numerous hosts to perform the attack

— Glossary of National Institute of Standards and Technology, US

EXAMPLES: One of the largest botnets for <DDoS attacks>, called Mirai, was built using insecure Internet of Things (IoT) devices, such as routers and IP cameras.

— CS EN COMPARABLE/ACADEMIC

<Distributed Denial of Service (DDoS) attacks> could cause severe damage to organizations' critical systems.

— CS EN COMPARABLE/MEDIA

Statistics of the current terminological data in the termbase:

233 concepts
514 LT terms, 414 EN terms

The number of entries is expected to increase considerably by the end of 2025, with a total of 500 entries planned.

Ongoing Work on the Termbase:

Work on Existing Entries:

- Refinement of thematic grouping of concepts
- Refinement of definitions according to ISO standards
- Enrichment of examples based on updated corpora
- Creation of structured format for source data
- Update of frequency labels of existing terms

Work on New Entries:

- Extension of concept index based on updated corpora
- Collection and structuring of necessary data for new concept entries

Additional Work:

- Comparison with **Terminų bankas** and **IATE** datasets

Survey dataset

Survey Data on Preferences of Lithuanian Cybersecurity Terminology

(Mykolas Romeris University; Vytautas Magnus University /
2024-10-04)

Author(s):

Rackevičienė, Sigita ; Utkas, Andrius



Šis įrašas turi 3 bylas (39.98 KB).

Publicly Available

The survey dataset is available in the CLARIN-LT repository in TSV (Tab-Separated Values) format, which can be opened in spreadsheet programs or read by programming languages. All responses are anonymous.

The aims of the survey

The aim of the survey was to investigate terminology preferences of different user groups, including the choice or proposal of the most suitable terms for given concepts and the rationale behind these selections.

Concepts selected for the terminology survey

Concept 1	‘cyberattack’	Concept 6	‘phishing’
Concept 2	‘spam’	Concept 7	‘botnet’
Concept 3	‘denial-of-service attack’	Concept 8	‘hacker’
Concept 4	‘man-in-the-middle attack’	Concept 9	‘honeypot method’
Concept 5	‘brute force attack’	Concept 10	‘zero-day vulnerability’

‘Zero-day vulnerability’

Sąvokos apibrėžtis: ką tik išsiaiškinta, bet dar nespėta pašalinti sistemos ar įrenginio saugumo spraga (šaltinis: IATE).

Angliškas terminas: *zero-day vulnerability*

Concept definition: vulnerability in a system or device that has been disclosed but is not yet patched (source: IATE)

English term: *zero-day vulnerability*

Koks terminas, Jūsų manymu, tinkamiausias pavadinti šią sąvoką?

Which term, in your opinion, is most suitable to designate this concept?

- ☐ „zero-day“ pažeidžiamumas
- ☐ „zero-day“ spraga
- ☐ nulinės dienos pažeidžiamumas
- ☐ nulinės dienos spraga
- ☐ ką tik nustatytas pažeidžiamumas
- ☐ Other...

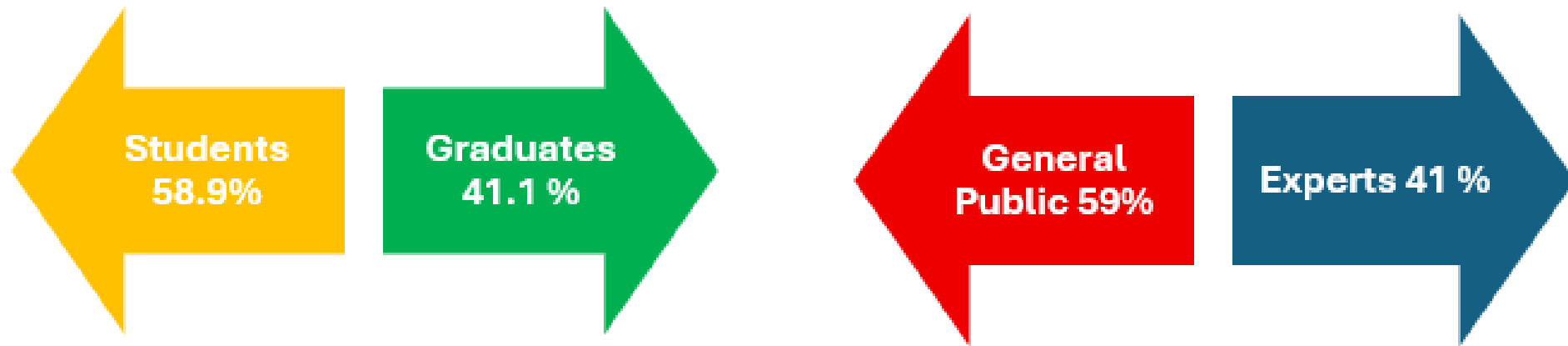
Kodėl pasirinkote šį terminą / pasiūlėte savo variantą? (galite nurodyti kelias priežastis)

Why have you chosen this term / proposed your variant? (you can select several reasons)

- ☐ Dėl jo tikslumo, aiškumo
- ☐ Dėl jo trumpumo ir vartosenos patogumo
- ☐ Dėl jo ekspresyvumo, vaizdingumo
- ☐ Dėl jo taisyklingumo (atitiktis lietuvių kalbos normoms)
- ☐ Dėl jo dažnumo
- ☐ Other...

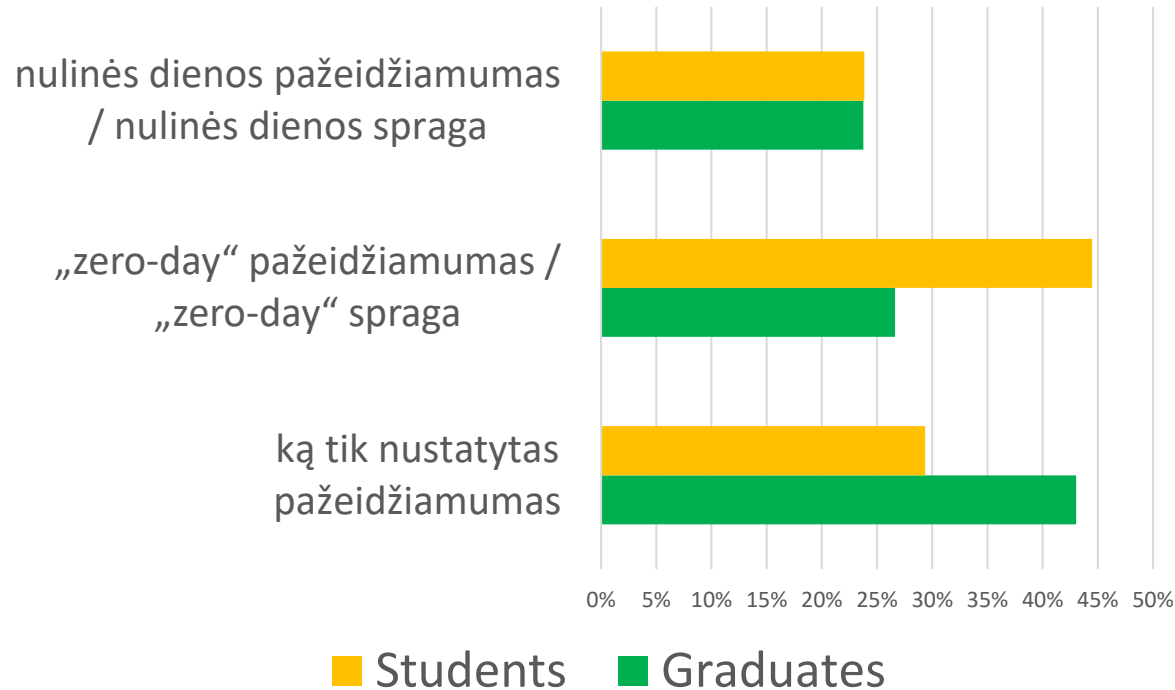
Sample of the survey

593 respondents

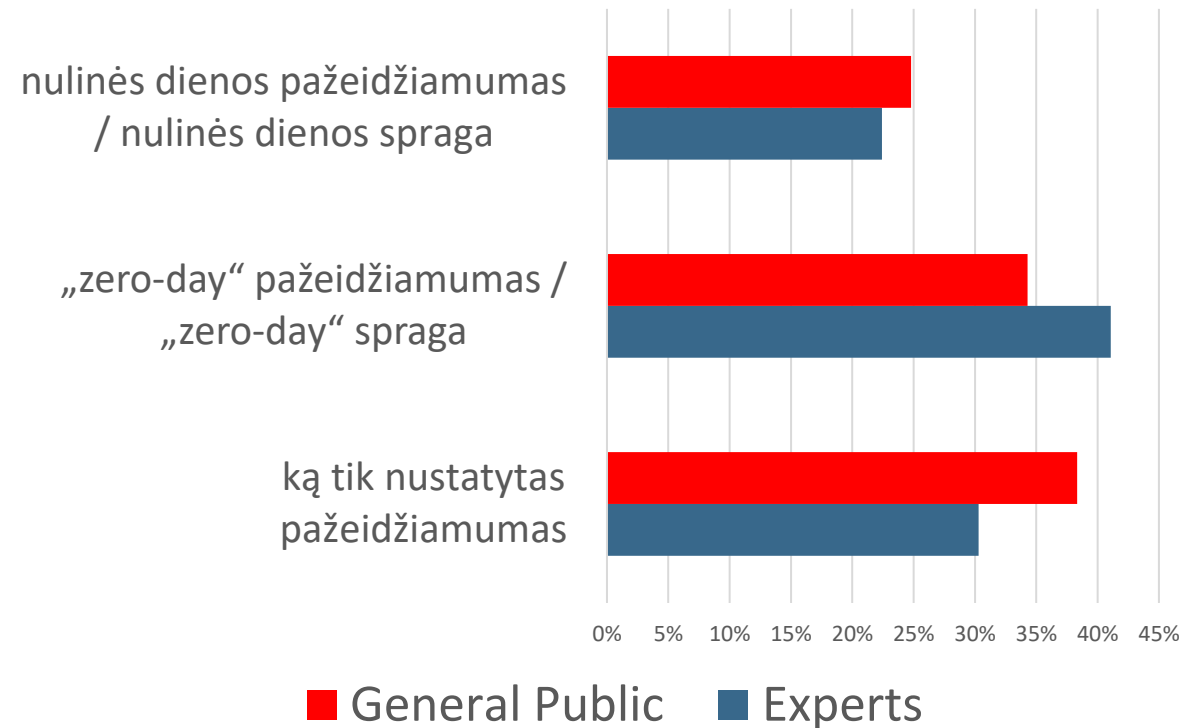


Preferences of the synonymous terms (Zero-day vulnerability)

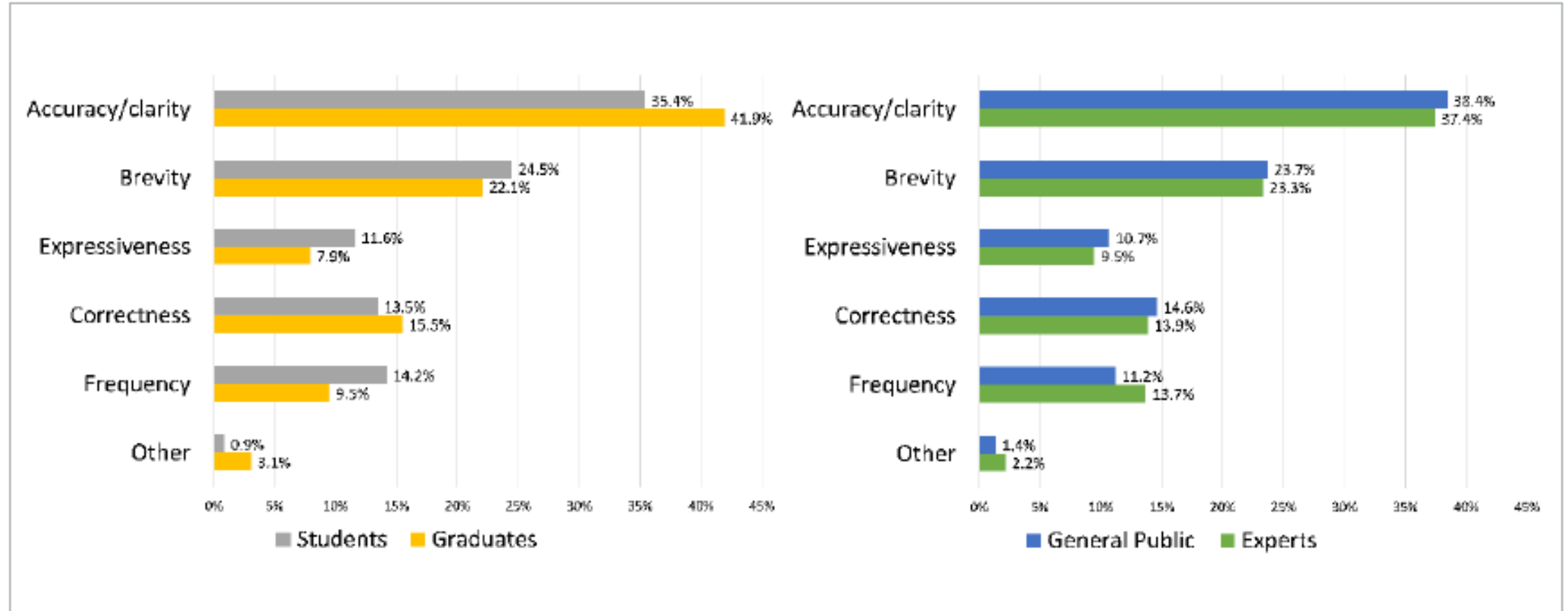
Concept 10: Students and Graduates



Concept 10: Experts and General Public

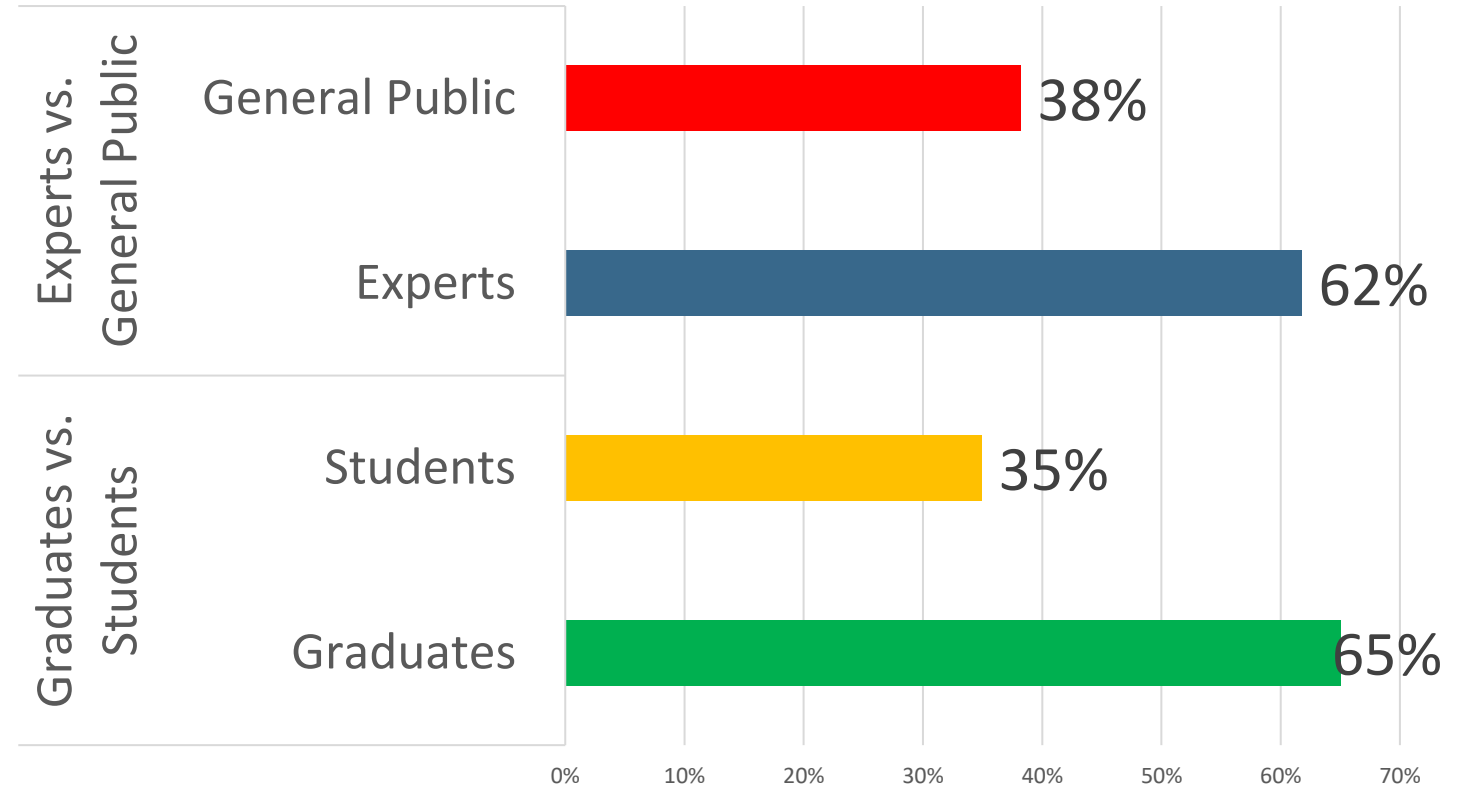


Reasons for selecting terms

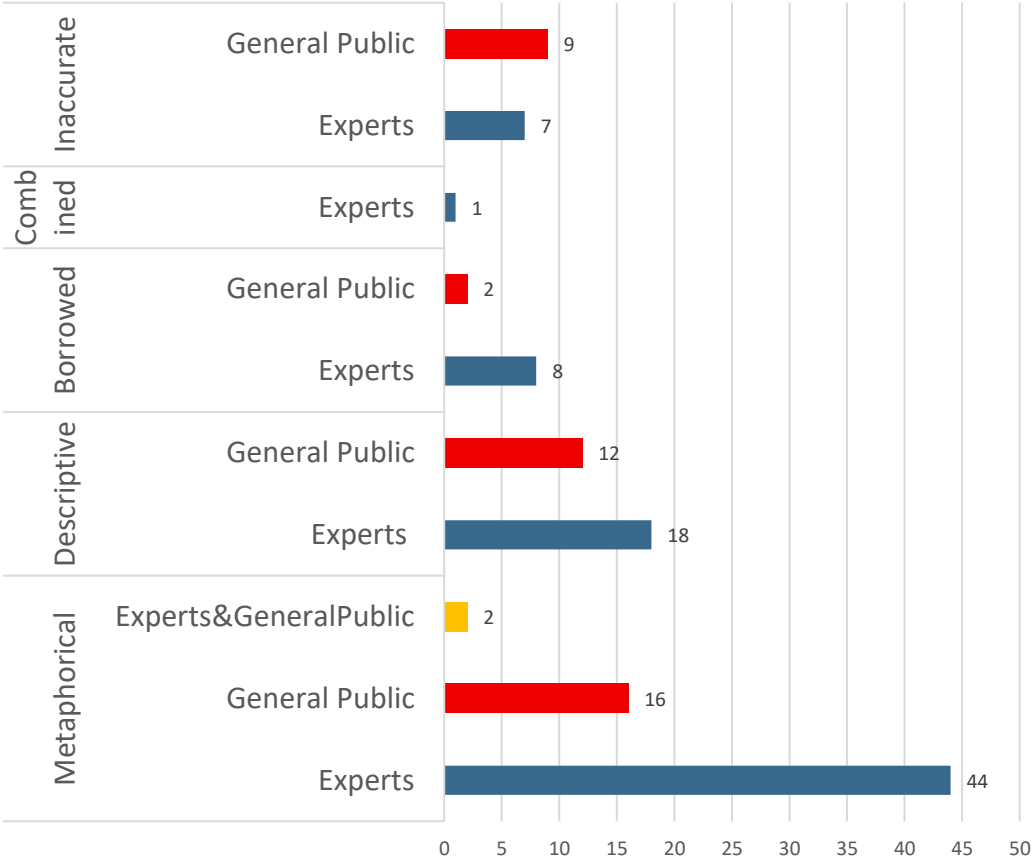
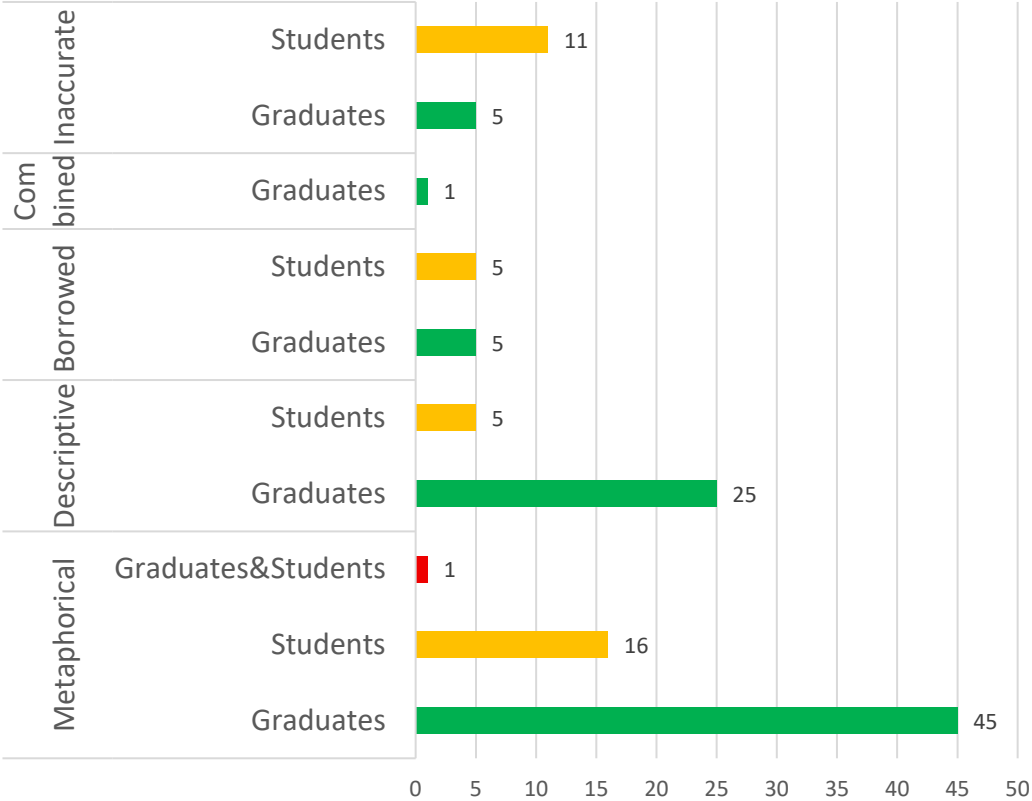


Term proposal provided by different groups of respondents

119 terminological designations were proposed by respondents.



Proposals of different formation patterns in different respondent groups



Future plans

Future plans:

- Parallel LT-EN corpus (Lithuanian national documents and their translations to EN)
- Version 2 of Comparable corpus
- Manually annotated training corpus
- Version 2 of termbase tbx
- Application of LLOD technologies for linking the LR with other resources on the web.

References of figures presented on the slides:

- 1) Utkā, A., Rackevičienė, S., Mockienė, L., Rokas, A., Laurinaitis, M., Bielinskienė, A. 2022. Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction. *Selected Papers from the CLARIN Annual Conference 2021. Linköping Electronic Conference Proceedings* 189, 126–138.
- 2) Rackevičienė, S., Utkā, A., Bielinskienė, A., Mockienė, L. 2023. Lithuanian-English Cybersecurity Termbase: Principles of Data Collection and Structuring. *Rasprave Instituta za hrvatski jezik*, 49 (2), 439-461.
- 3) Rackevičienė, S., Utkā, A. 2024. Preferences of Lithuanian Cybersecurity Synonymous Terms in Different User Groups. *Studies about Languages / Kalbų studijos* 44, 107-122.