

Dvikalbis automatinis terminų atpažinimas (DVITAS)

Andrius Utkas (VDU) ir Sigita Rackevičienė (MRU)



VYTAUTO
DIDŽIOJO
UNIVERSITETAS
MCMXXII



Kompiuterinės
lingvistikos
centras



MYKOLO ROMERIO
UNIVERSITETAS



Research
Council of
Lithuania

CLARIN-LT



Nexus
Linguarum

<https://klc.vdu.lt/dvitas/lt>

DVITO komanda



Tikslai

- Pagrindiniai projekto tikslai:
 1. **Paruošti kalbinius išteklius DVikalbiam automatiniam Terminų Atpažinimui (DVITAS),**
 2. Pritaikyti šiuolaikinius dvikalbio terminologijos atpažinimo metodus (giliojo mokymosi sistemas),
 3. Sukurti anglų-lietuvių kibernetinio saugumo (KS) terminų bazę.



Kibernetinio saugumo sritis

- Tyrimui pasirinkta kibernetinio saugumo (KS) terminologijos sritis, nes mes galvojame:
 - _ kad ši sritis yra ypatingai svarbi šiandienos skaitmeniniame pasaulyje;
 - _ ši sritis yra greitai besikeičianti;
 - _ lietuviška KS terminologija turėtų būti standartizuota.



DVITAS (BiTE)

- **DVIkalbis automatinis Terminų Atpažinimas** (angl. Bilingual Terminology extraction (BiTE)) yra suprantamas kaip metodas, kurį naudojant terminija automatiškai atpažįstama lygiagrečiuosiuose ir palyginamuosiuose tekstuose.
- Automatinis terminų atpažinimas lygiagrečiuosiuose tekstynuose jau yra naudojamas keletą dešimtmečių.
- Tačiau palyginamųjų tekstynų svarba labai auga...

Lygiagretieji vs palyginamieji duomenys

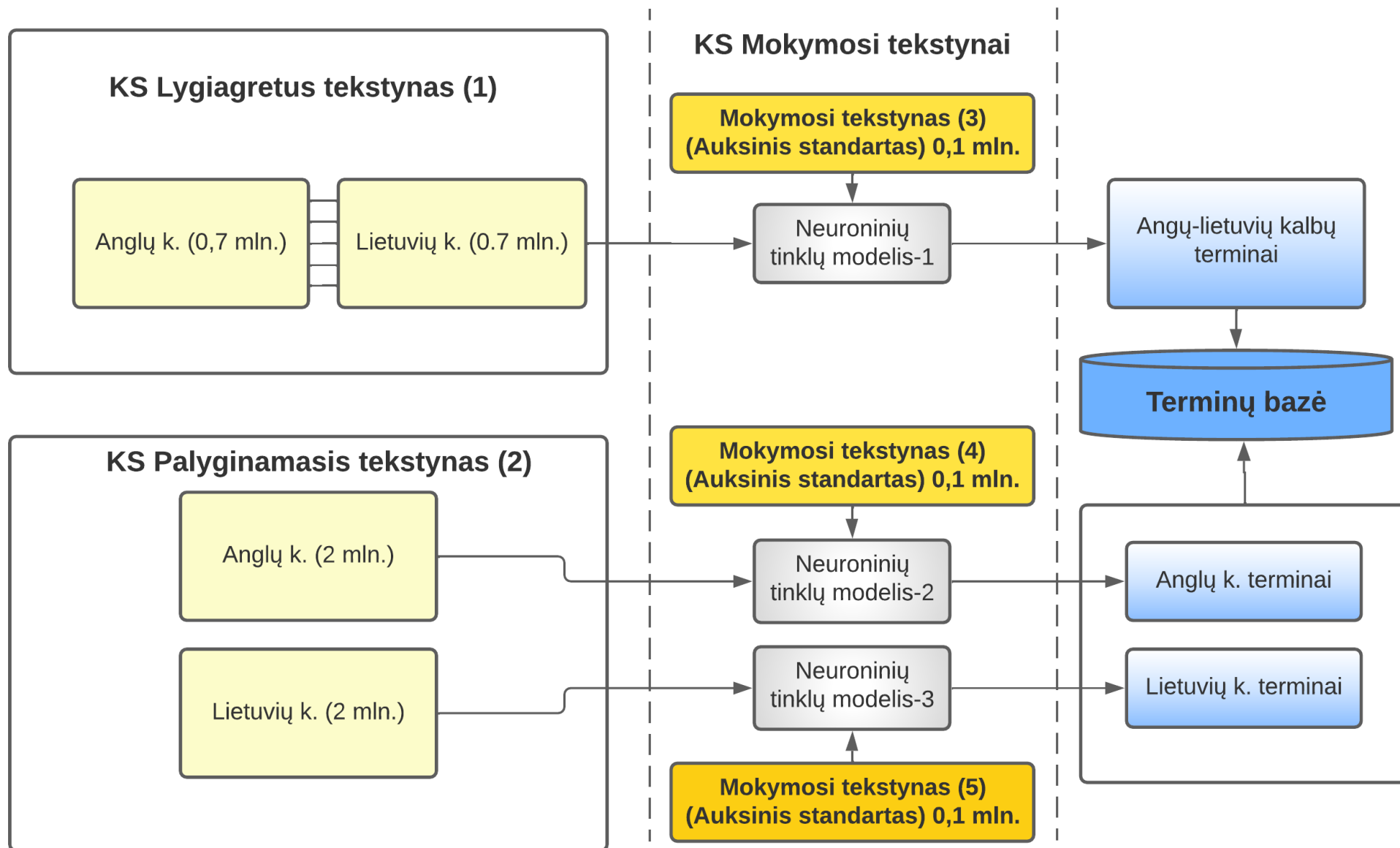
Lygiagretieji duomenys

- Vertimo kalba yra veikiamą originalios kalbos;
- Lygiagrečiųjų duomenų trūkumas;
- Įvairovės trūkumas: daugiausia ES dokumentai;
- Sunkesnis ir brangesnis kūrimas (reikalingas lygiavimas);
- **BET**: lengviau yra atlikti DVITA.

Palyginamieji duomenys

- Originali kalba yra natūralesnė;
- Tyrimams prieinama daugiau duomenų;
- Duomenys yra įvairesni;
- Lengvesnis ir pigesnis kūrimas;
- **BET**: juose sunkiau yra atlikti DVITA.

DVITAS tekstynų sistema

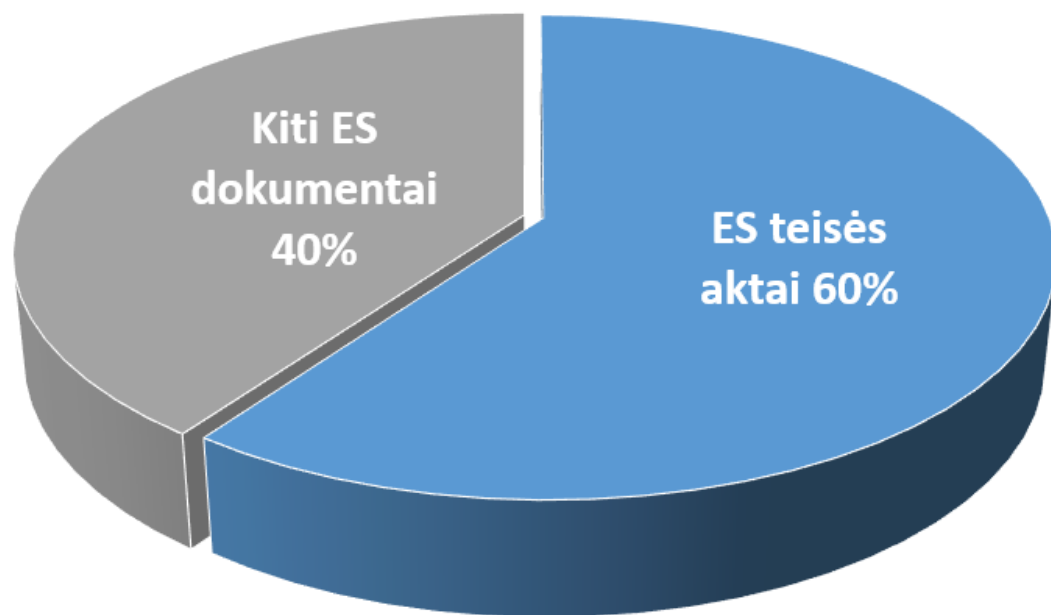


Tekstų tipai

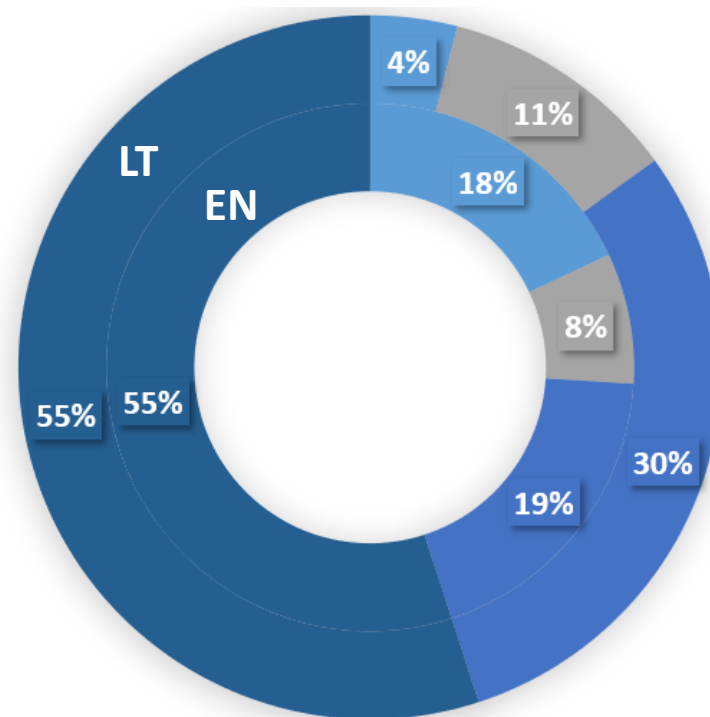
Anglų-lietuvių lygiagretusis tekstynas

Anglų lietuvių palyginamasis tekstynas

Laikotarpis: 2010-2021



1,4 mln. žodžių



■ Teisiniai ■ Administraciniai-informaciniai ■ Akademiniai ■ Žiniasklaidos

4 mln. žodžių

Tekstynai CLARIN-LT saugykloje

Anglų-lietuvių palyginamasis ir lygiagretusis tekstynai šiuo metu yra laisvai prieinami CLARIN-LT saugykloje:

- Palyginamasis tekstynas: neanotuos ir morfologiškai anotuos versijos prieinamos:

<https://clarin.vdu.lt/xmlui/handle/20.500.11821/47>

- Anglų-lietuvių lygiagretusis: nelygiuotos ir lygiuotos TMX formato versijos prieinamos:

<https://clarin.vdu.lt/xmlui/handle/20.500.11821/46>

What's New

Corpus

CLARIN-LT

English-Lithuanian Comparable Cybersecurity Corpus - DVITAS

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

Description:
The English-Lithuanian comparable corpus (DVITAS COMPARABLE) is morphologically annotated. It includes English and Lithuanian original texts on cybersecurity from the time period of 2010-2021. The corpus was compiled for ...

Šis įrašas turi 4 bylas (36.66 MB).

Publicly Available

Corpus

CLARIN-LT

English-Lithuanian Parallel Cybersecurity Corpus - DVITAS

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

Description:
English-Lithuanian parallel corpus DVITAS includes original English texts on cybersecurity and their Lithuanian translations aligned on the sentence level. The corpus was compiled for the bilingual terminology extraction ...

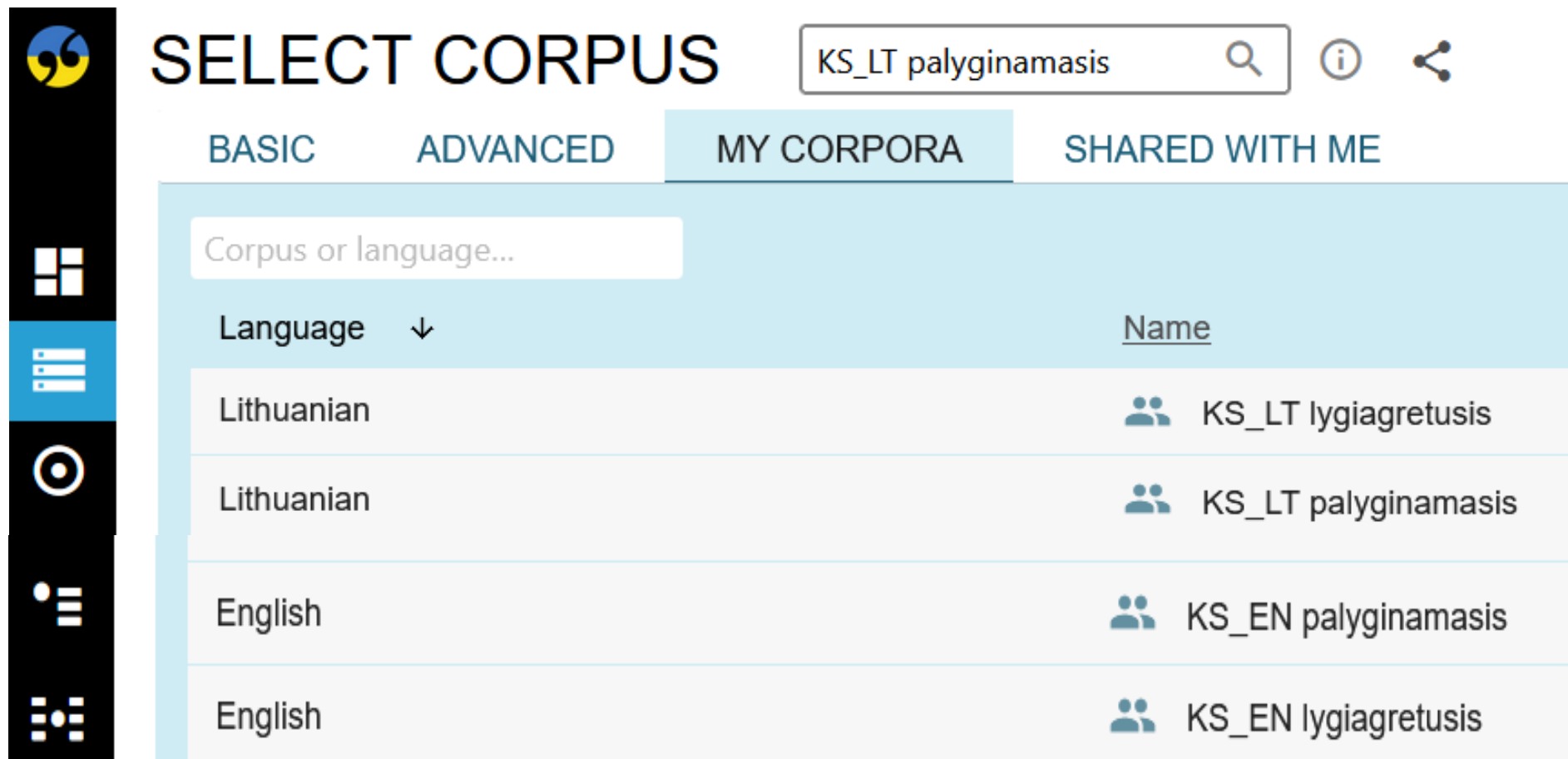
Šis įrašas turi 3 bylas (6.56 MB).

Publicly Available

ToolService

CLARIN-LT

Tekstynai *Sketch Engine* platformoje



The image shows the Sketch Engine web interface. On the left is a vertical sidebar with icons for home, search, corpora, settings, and help. The main area is titled 'SELECT CORPUS' and has a search bar containing 'KS_LT palyginamasis'. Below the search bar are four tabs: 'BASIC', 'ADVANCED', 'MY CORPORA' (which is selected), and 'SHARED WITH ME'. Under the 'MY CORPORA' tab, there is a table with the following data:

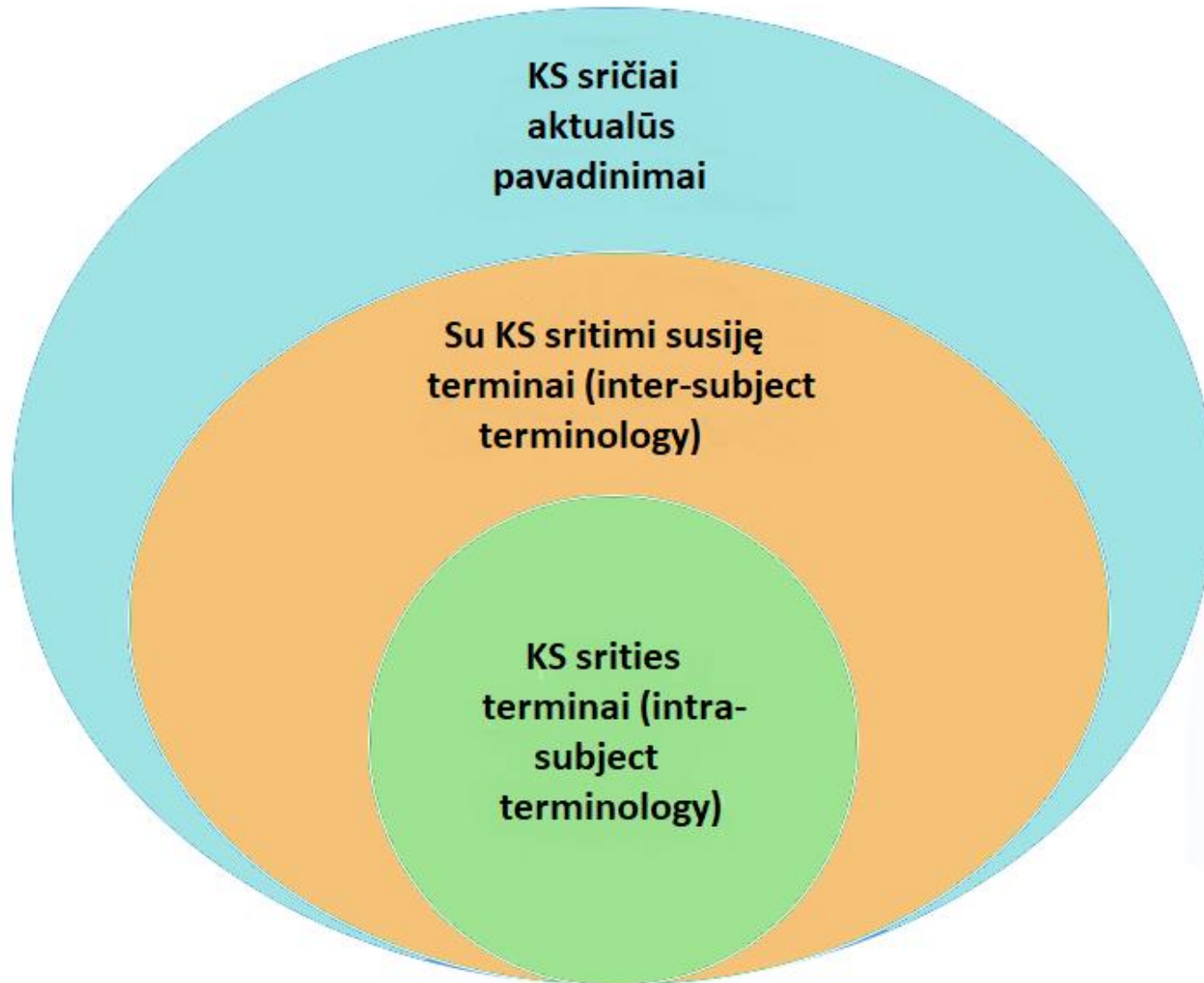
Language	Name
Lithuanian	KS_LT lygiagretusis
Lithuanian	KS_LT palyginamasis
English	KS_EN palyginamasis
English	KS_EN lygiagretusis

Aukštinio standarto tekstynų kūrimas

Aukšinio standarto tekstynai

Lygiagretusis aukšinio standarto tekstynas	Palyginamasis aukšinio standarto tekstynas
Laikotarpis: 2011-2021, dydis: po 0,1 mln. žodžių	
• ES reglamentai; • ES direktyvos; • ES Komisijos komunikatai; • ES Komisijos rekomendacijos.	• LT: LR KS strategija ir įstatymas; Vyriausybės nutarimai; • EN: JAV KS strategija; • LT: NKSC ataskaitos ir rekomendacijos; • EN: ENISA ataskaitos; • LT&EN: KS vadovėliai, moksliniai darbai; • LT&EN: Žiniasklaidos straipsniai (naujienų portalai ir specializuoti portalai).

Sąvokiniai anotavimo kriterijai ir žymų rinkinys



Žymų rinkinys (angl. *tagset*):

ks,
kse,
kss,
ksse,
ksdp.

Kibernetinis saugumas = KS terminas

Kibernetinis saugumas EN = KS angliškas terminas

KS s. = Su KS susijęs terminas

KS s. EN = Su KS susijęs angliškas terminas

Pavadinimas = Dokumento pavadinimas

Anotavimo programinė įranga *QuickTag*

Eilutė: 13 iš 658



EN

(3) Cybersecurity incidents can trigger a broader crisis , impacting sectors of activity beyond network and information systems and communication networks ; any appropriate response must rely upon both cyber and non-cyber mitigation activities .

LT

(3) kibernetinio saugumo incidentai gali sukelti platesnę krizę ir paveikti sektorius , kurių veikla neapsiriboja tinklais ir informacinėmis sistemomis ar ryšių tinklais ; bet koks tinkamas reagavimas turi būti grindžiamas ir kibernetinėmis, ir nekibernetinėmis poveikio mažinimo priemonėmis ;

Anotuoti tekstai neuroninių tinklų mokymuisi

(3) <START:kse|undefined|undefined>Cybersecurity incidents<END> can trigger a broader crisis, impacting sectors of activity beyond <START:ksse|undefined|undefined>network and information systems<END> and <START:ksse|undefined|undefined>communication networks<END> ; any appropriate response must rely upon both <START:kse|undefined|underfined>cyber and non-cyber mitigation activities<END> .</seg></tuv>

(3) <START:ks|undefined|undefined>kibernetinio saugumo incidentai<END> gali sukelti platesnę krizę ir paveikti sektorius, kurių veikla neapsiriboja <START:kss|undefined|undefined>tinklais ir informacinėmis sistemomis<END> ar <START:kss|undefined|undefined>ryšių tinklais<END> ; bet koks tinkamas reagavimas turi būti grindžiamas ir <START:ks|undefined|undefined>kibernetinėmis, ir nekibernetinėmis poveikio mažinimo priemonėmis<END> ;</seg></tuv> </tu>

Papildomos anotacijos

Nepilni terminai	<i>ataka</i> [žyma: nepilnas terminas, komentaras: pilnas terminas <i>kibernetinė ataka</i>]
Akronimai	<i>CERT</i> [žyma: akronimas, komentaras: <i>Cyber Emergency Response Team</i>]
EN terminai, EN-LT hibridai lietuviškuose sakiniuose	<i>„Man in the middle“ ataka</i> [žyma: hibridinis terminas]
Pavadinimų semantinės grupės	<i>ENISA</i> [žyma: institucijos pavadinimas]
Lizdiniai terminai	sudėtiniai terminai/pavadinimai, į kuriuos įeina KS terminai, išskaidomi ir anotuojami atskirai

Anotuotų terminų eksportavimas tyrimams

P	Q	R	S	T
Nr.	Kibernetinio saugumo terminai (visi LT)	Ypatybės	Sak. eil. nr.	Sakiniai
1	atsako		6	dėl koordinuoto atsako į didelio masto kibernetinio saugumo
2	didelio masto kibernetinio saugumo incidentus ir krizes		6	dėl koordinuoto atsako į didelio masto kibernetinio saugum
3	Kibernetinio saugumo incidentas		11	Kibernetinio saugumo incidentas , paveikiantis organizacijas
4	kibernetinio saugumo incidentas		12	(2) kibernetinio saugumo incidentas gali būti laikomas Sąjun
5	Sąjungos lygmens krize		12	(2) kibernetinio saugumo incidentas gali būti laikomas Sąjung
6	incidento	Pagrindinis: [6. kibernetinio saugumo incidentas]	12	(2) kibernetinio saugumo incidentas gali būti laikomas Sąjungo
7	kibernetinio saugumo incidentai		13	(3) kibernetinio saugumo incidentai gali sukelti platesnę kriz
8	kibernetinėmis, ir nekibernetinėmis poveikio mažinimo priemonėmis		13	(3) kibernetinio saugumo incidentai gali sukelti platesnę krizę
9	reagavimo į kompiuterių saugumo incidentus tarnybų		23	(9) Komisija su valstybėmis narėmis konsultavosi dviejuose ko
10	CSIRT		23	(9) Komisija su valstybėmis narėmis konsultavosi dviejuose ko

Anotavimo problemos

- **KS ar KSS terminas?**

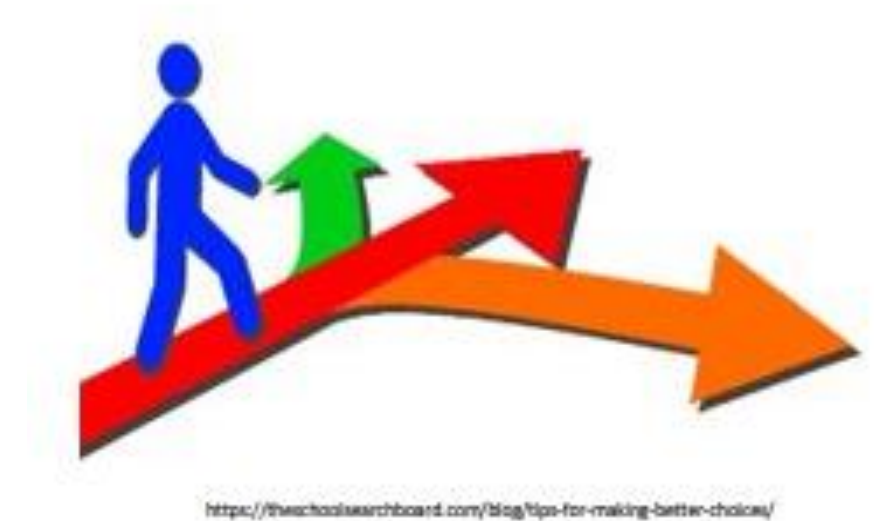
*ypatingos svarbos informacinė infrastruktūra
valstybės informacinis išteklius*

- **Terminas ar pavadinimas?**

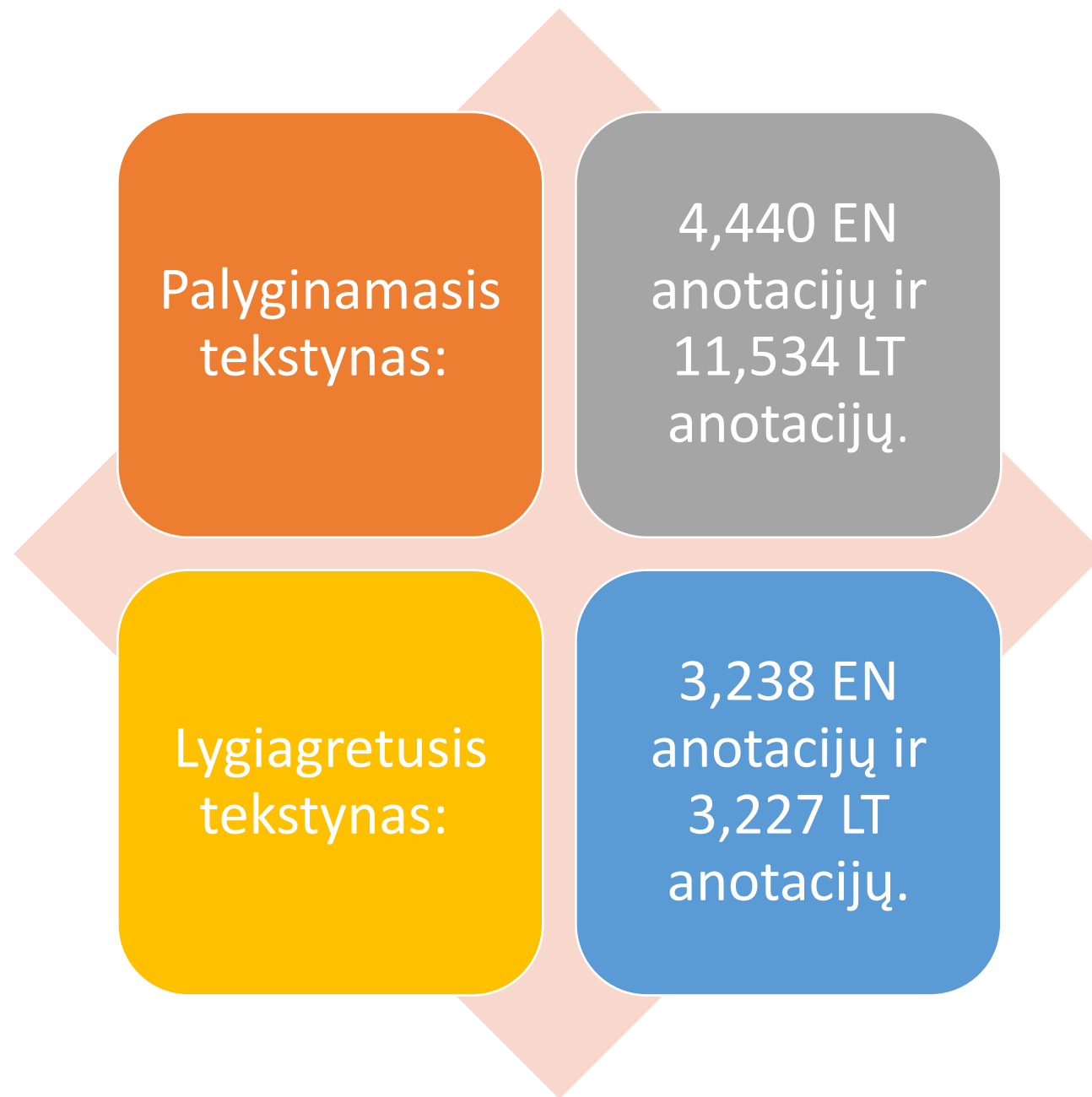
TDSS, AAEH; CSIRT, CERT

- **EN terminas ar LT terminas?**

"phishing", phishing'as, phishingas

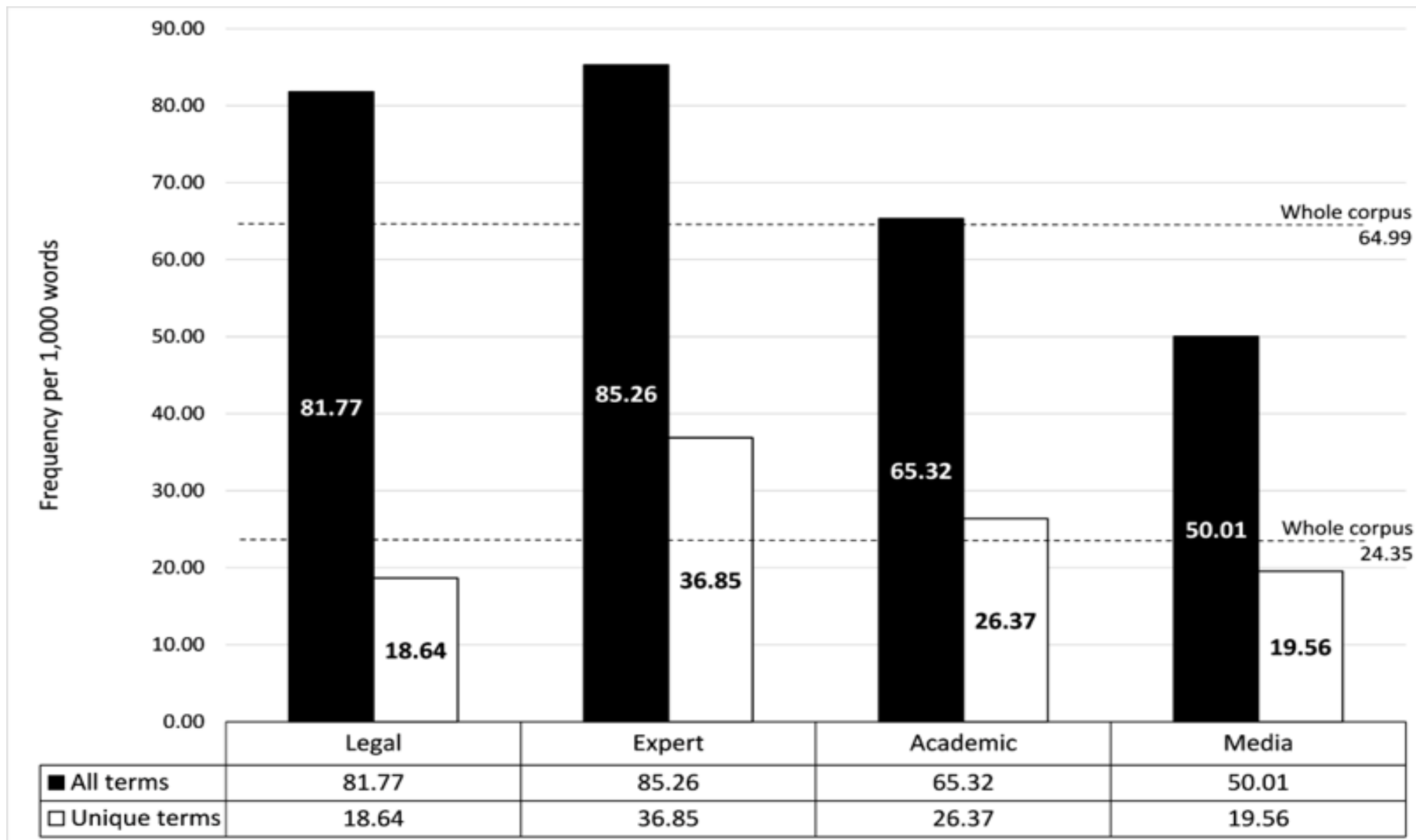


Anotavimo rezultatai



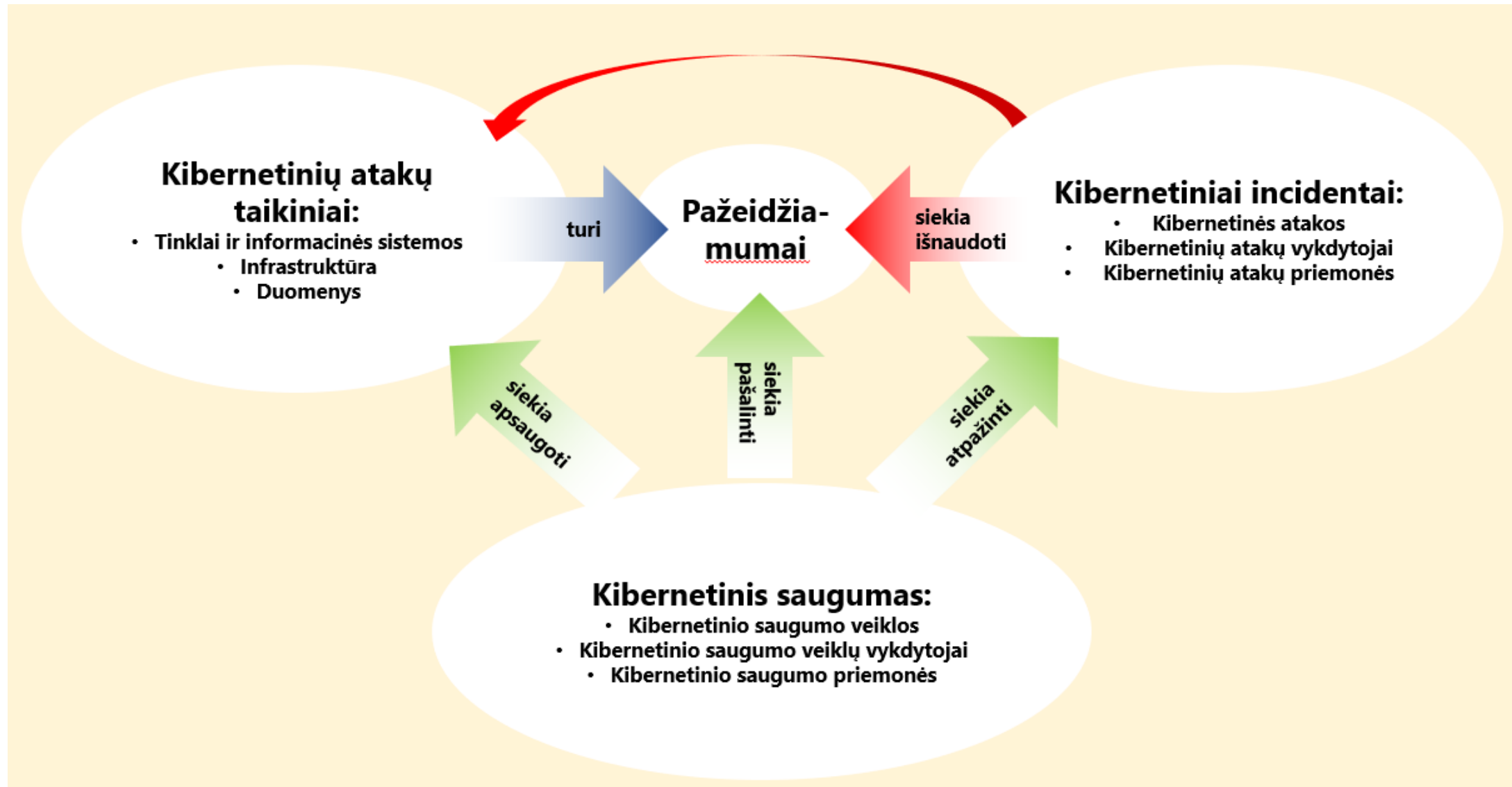
KS palyginamojo tekstyno LT dalies anotacijų tyrimo rezultatai

KS terminų tankis ir įvairovė skirtingo žanro tekstuose



Terminų bazės kūrimas

Terminų klasifikacijos kūrimas



Terminų tvarkybos programinės įrangos platforma

Terminologue

DVITO terminų bazė

Editing Administration Configuration

English sigita.rackeviciene@mruni.eu

—

smart search

LT

8

Advanced Persistent Threat

APT

DEF XMPL NOT

✗ ✓ ✓ 2021-12-15

Kibernetiniai incidentai » Kibernetinės atakos

LT slaptažodžių parinkimo ataka

„brute force“ ataka

EN brute force attack

brute force password attack

DEF XMPL NOT COLL

✓ ✓ ✓ 2021-12-14

New ID 5 Edit Clone Delete

Kibernetiniai incidentai » Kibernetinės atakos

LT slaptažodžių parinkimo ataka

— [www.nksc.lt/doc/biuletiniai/NKSC_Slaptazodziu_saugumo_biuletenis.pdf](#)

„brute force“ ataka

— [www.nksc.lt/rekomendacijos/brute_force_atakos.html](#)

bandymai atspėti vartotojo prisijungimo duomenis prie informacinės sistemos, įvedinėjant atsitiktines simbolių sekas ir dažnai naudojamas kombinacijas

— [www.nksc.lt/rekomendacijos/brute_force_atakos.html](#)

Natūralu, kad kuo slaptažodis ilgesnis ir kuo jame daugiau skirtingų ženklų, raidžių ir skaitmenų, tuo ilgiau užtrunka jį atspėti. Toks slaptažodis yra atsparesnis slaptažodžių parinkimo (angl. brute force) atakoms.

— [www.nksc.lt/doc/biuletiniai/NKSC_Slaptazodziu_saugumo_biuletenis.pdf](#)

EN brute force attack

— [lastbit.com/octopass.asp](#)

brute force password attack

— [dataplane.org/vnc-tac.html](#)

a method of accessing an obstructed device by attempting multiple combinations of numeric/alphanumeric passwords

— [nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-101r1.pdf](#)

A brute force attack is a strategy that can in theory be used by an attacker who is unable to take advantage of any weakness in a system.

— [www.pentest-standard.org/index.php/PTES_Technical_Guidelines](#)

100

Ačiū už dėmesį!

Projekto svetainė:

<https://klc.vdu.lt/dvitas/lt>

