Corpora for Cybersecurity Term Extraction Project

**Andrius Utka (VMU), Aivaras Rokas (VMU), Agnė Bielinskienė (VMU),
Sigita Rackevičienė (MRU), Liudmila Mockienė (MRU), Marius Laurinaitis (MRU)**

The team of researchers from two universities in Lithuania (Vytautas Magnus University and Mykolas Romeris University) have started the scientific project "**Bilingual automatic terminology extraction**".

The aim of the project is to design a methodology for automatic extraction of English and Lithuanian terms in a specialised domain (cybersecurity - CS) from **parallel and comparable corpora** and create a bilingual termbase.

In the current stage of the project, the research is focused on the compilation of the corpora reflecting the usage of the CS terminology in national and international settings.

The Cybersecurity Comparable Corpus (2010-2020)

Text genres and types:

- **Legislative and executive documents** (national CS strategies, legal acts, government resoliutions, minister orders on CS issues),
- **Official reports** (reports of national CS centres),
- **Academic publications** (scientific papers, books, theses, textbooks),
- **Information publications** for the general public (brochures, posters, etc.),
- **Media articles**.

The accumulated resources will be deposited in CLARIN-LT repository.

CLARIN-LT

More info at the project site:

DVITAS

The Cybersecurity Parallel Corpus (2010-2020):

Text genres and types:

The EU legal acts and related documents extracted from EUR-Lex and other EU institutional repositories:

- regulations and directives of the European Parliament and of the Council,
- decisions of the Council;
- communications, reports and recommendations of the European Commission,
- opinions of the Committees of the EU,
- briefing papers of the Court of Auditors, etc.

Acknowledgements

The research is carried out under the project "Bilingual automatic terminology extraction" funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282). The project is also included as a use case in COST action "European network for Web-centred linguistic data science" (CA18209).

VYTAUTAS MAGNUS UNIVERSITY M C M X X I I

MYKOLAS ROMERIS UNIVERSITY

Research Council of Lithuania