Corpora for Bilingual Terminology Extraction in Cybersecurity Domain

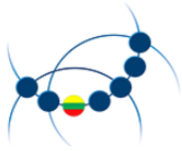
Dvikalbis automatinis terminų atpažinimas (DVITAS)

Andrius Utka, Sigita Rackevičienė, Liudmila Mockienė, Aivaras Rokas, Marius Laurinaitis and Agnė Bielinskienė

VYTAUTO DIDŽIOJO UNIVERSITETAS
MCMXXII

Research Council of Lithuania

CLARIN-LT



Aims

- The main aims of the project:
  1. **to prepare language resources for bilingual terminology extraction (BiTE),**
  2. to apply current bilingual terminology extraction methods,
  3. create an English-Lithuanian termbase of cybersecurity (CS) terminology.

Cybersecurity domain

- Cybersecurity (CS) domain is chosen, as we think:
  - that for today‘s digital world this domain is particularly relevant;
  - it is also very dynamic,
  - and Lithuanian CS terminology needs standardization.

BiTE

- **Bilingual Terminology extraction (BiTE)** is understood as a method when terms and their translations are extracted from parallel or comparable texts.
- Term extraction from parallel corpora has been already applied for several decades.
- But the importance of comparable data is increasing…

Parallel *vs* comparable data

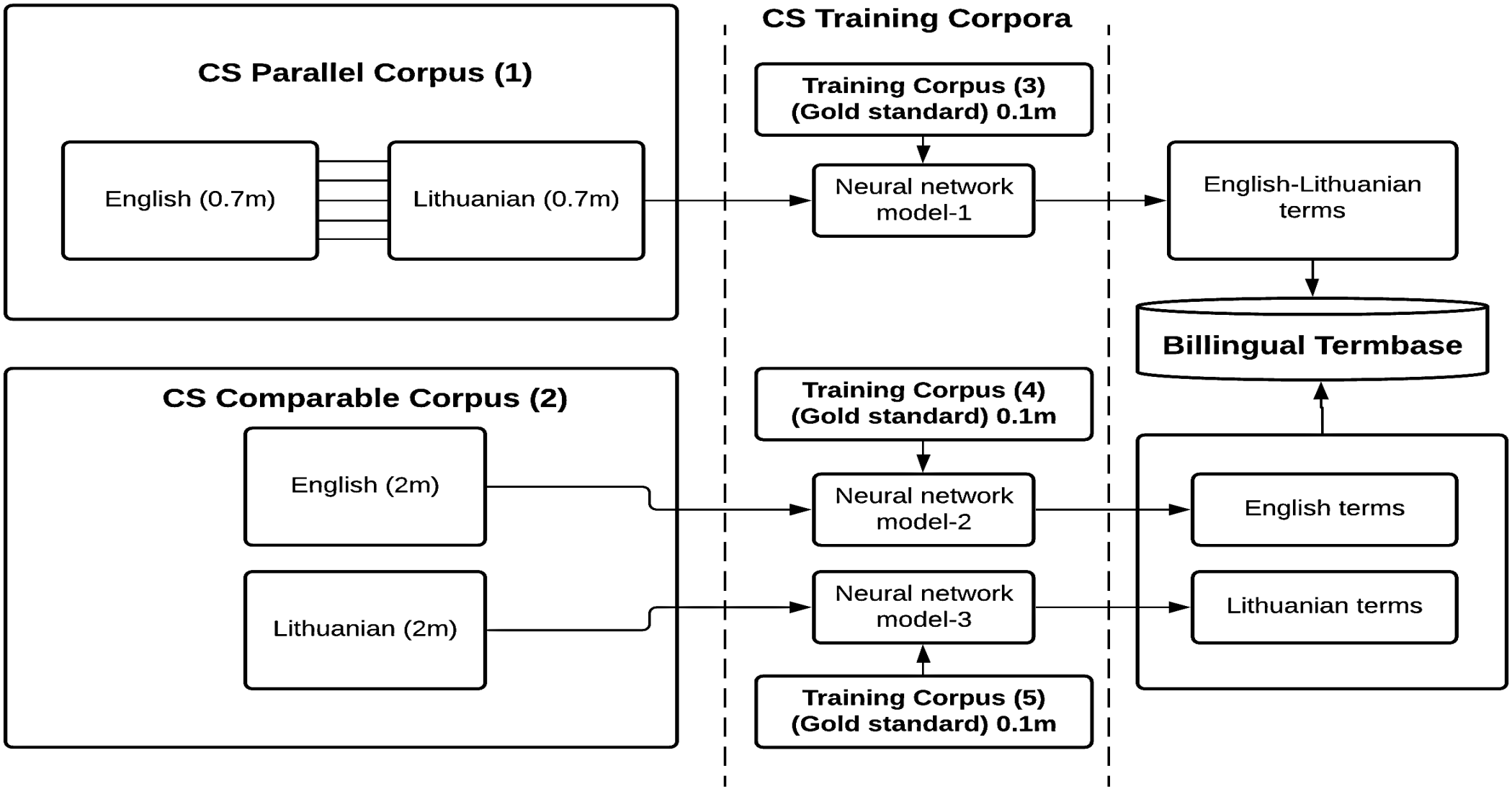
Parallel data

- Translated language is influenced by the source language;
- Shortage of parallel data;
- Lack of diversity: mostly EU documents;
- Expensive to build (needs aligning);
- **BUT**: it‘s easier to do BiTE.

Comparable data

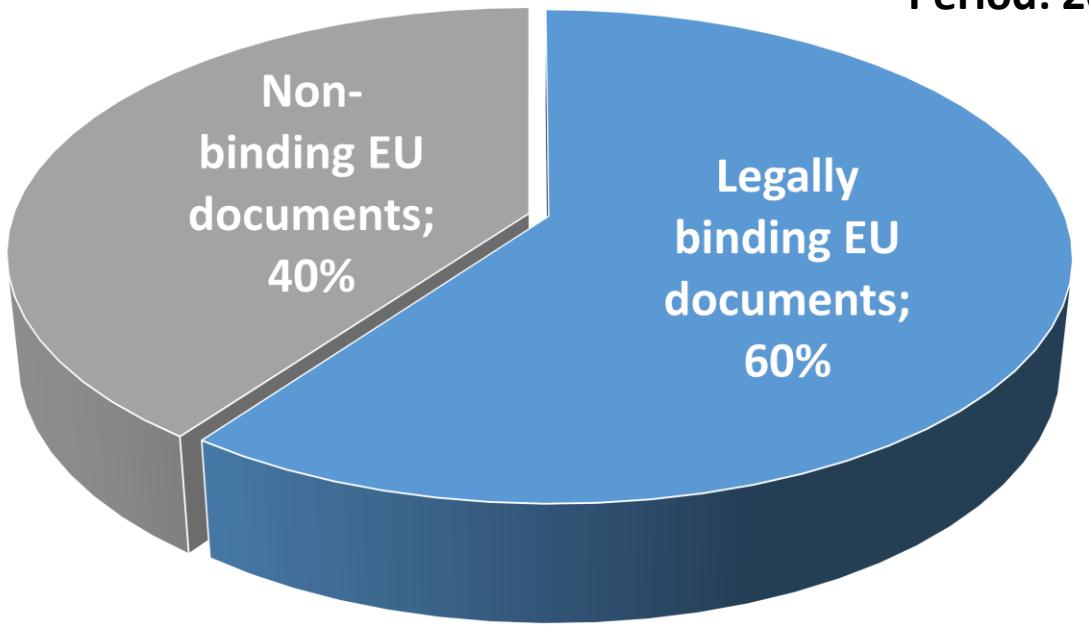
- The original language is more natural;
- More data is available;
- The data is more diverse;
- Cheaper to build;
- **BUT**: it‘s more difficult to do BiTE.

Corpora system for BiTE



Texts types in corpora

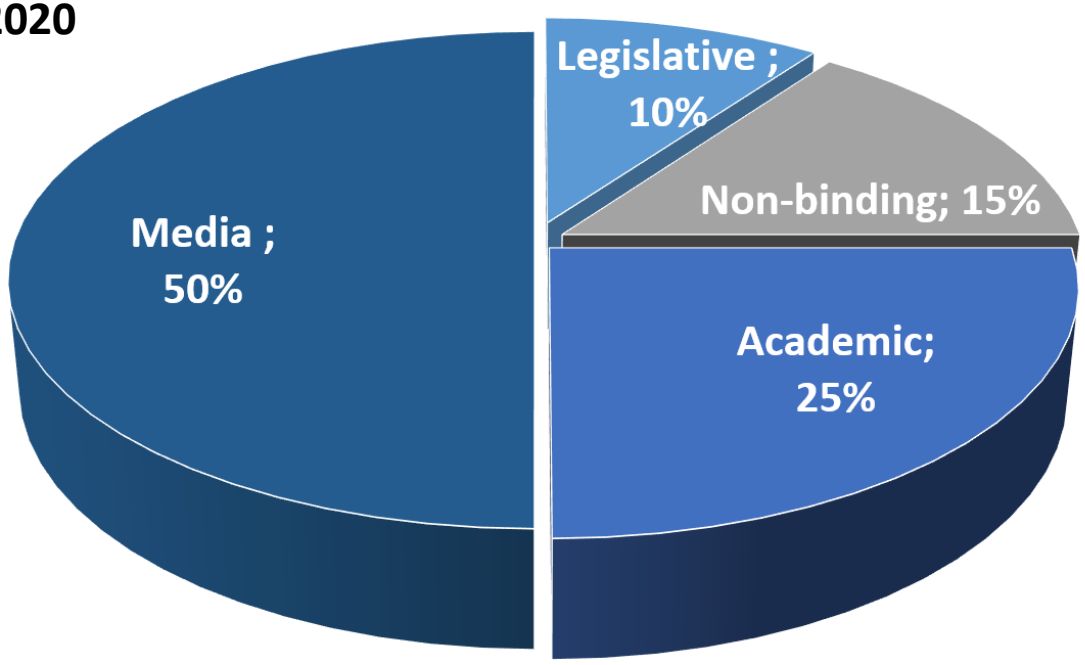
EN-LT Parallel corpus



1.4m words

EN-LT Comparable corpus

Period: 2010-2020



4m words

Thank you!

Ask questions in Q&A session and in the afternoon breakout session

15:30-16:00 Room 2 (Resources)