Annotation of Multiword Expressions in the Lithuanian Corpus of the Cyber-Security Domain

Liudmila Mockienė, Institute of Humanities, Faculty of Human and Social Studies, Mykolas Romeris University, Vilnius, Lithuania, liudmila@mruni.eu

Key words: multiword expressions, cyber-security, low-resourced languages.

Introduction

As multiword expressions (MWEs) can refer to a variety of multiword units that have a common feature of idiosyncrasy, or idiomaticity, which can be of several types, i.e. lexical, morphological, syntactic, semantic, pragmatic, and statistical (Baldwin and Kim, 2010), it is crucial to apply the same methodology and classification of MWEs for the linguistic data to be interoperable across different languages and linkable as LLOD. Universality of language descriptions that are applicable and consistent across languages advances NLP, aids in contrastive analyses, and provides insightful information about linguistic issues, including idiosyncrasy (Savary et al., 2023). Thus, development of linguistic resources and linking them is of utmost importance for under-resourced or low-resourced languages. One of the language annotation frameworks PARSEME developed during a dedicated COST action that focused on parsing and MWEs (Savary et al., 2017; Savary et al., 2023) sought to annotate MWEs in numerous languages, including low-resourced ones, such as Lithuanian.

Dataset and annotation project

The data for the annotation and research is going to be obtained from the English-Lithuanian Parallel Cybersecurity Corpus DVITAS which is freely available from the CLARIN-LT repository https://clarin.vdu.lt/xmlui/handle/20.500.11821/46 (Utka et al, 2022b).

The English-Lithuanian parallel corpus DVITAS consists of original English texts on cybersecurity and their translations into the Lithuanian language aligned on the sentence level. The corpus was compiled for the bilingual terminology extraction project. The corpus comprises the EU legal acts and other documents from the time period of 2006-2021. The documents were extracted from the EUR-Lex database and other EU institutional repositories. There are 80 aligned files in TMX format in English and Lithuanian, as well as 160 raw files (80 in English, and 80 in Lithuanian) in the dataset. The total size of the corpus is 1.4m words (EN-0.77m; LT-0.63m). The corpus contains 35,415 aligned segments.

The data in the corpus was annotated both manually and automatically because as Lithuanian is still a low resourced language, the researchers had to apply supervised and semi-supervised deep learning methods for automatic extraction of Lithuanian terminology (Rokas et al., 2020). First, the gold standard corpus was created to train neural networks, validate and test the data based on manually annotated linguistic data. Next, various deep learning models were tested to pre-process the data and extract terms automatically. The manually annotated linguistic data in the gold standard corpus included one-word and multi-word terms related to the cyber-security domain.

The project results revealed that 'deep learning systems trained by using gold standard corpora (manually annotated data) allow effective automatization of extraction of terminological data and metadata, which enables to regularly update termbases with minimised manual input' (Rackevičienė, 2021, p. 38).

Thus, annotation of MWEs could be carried out using the same gold standard corpus which was used for training neural networks to extract terminology automatically and could be used to extract MWEs (that not only relate to the cyber-security domain specifically). Moreover, the extracted terminological units in the project are of nominal character only, but other types of MWEs (verbal and modifier) are also of great importance for the analysed domain. Such analysis would enrich the insights on the usage of lexical units in the cyber-security domain regarding MWEs across genres (Rackevičienė, 2022).

The annotation project within the UniDive summer school would focus on the same part of the corpus (the gold standard corpus) for nominal, verbal, and modifier MWEs relying on the cross-lingual annotation guidelines for MWEs as developed by UniDive (COST action CA21167) (draft guidelines for nominal MWEs) and could prospectively contribute to the guidelines for MWEs.

The annotated part of the corpus could be further used for automatic extraction of MWEs not only from the parallel corpus of the cyber-security domain, but also from the comparative corpus, which was created during the same project (Utka, 2022b, 2022c) and also could be applied in other domains. As a result, annotation of nominal, verbal and modifier MWEs in the domain of cyber-security would contribute to the development of automatic extraction of MWEs in the Lithuanian language.

UniDive summer training school is crucial to the project's onset and further development as it will provide the indispensable training on the universal typology of MWEs as developed by PARSEME to be applied cross-linguistically and the annotation infrastructure that is available. It will also provide the necessary practical skills of recognising, categorising and annotating MWEs in the Lithuanian language to be further on shared with other members of the project.

Open questions

What are the difficulties of the annotation process of nominal, verbal, and modifier MWEs in different languages?

What is the tendency of using nominal, verbal, and modifier MWEs in the cyber-security domain? Is there anything domain-specific?

How could the manual annotation of MWEs in the cyber-security domain contribute to developing automatic extraction of MWEs not only in the same domain, but also in other domains?

The project phase

The project of annotating MWEs in the cyber-security domain is in the phase of planning. It has been discussed with members of the team who have carried out the project DVITAS and are eager

to continue annotation work based on the created corpora. The project of annotating MWEs will be started during the summer school. The acquired knowledge and skills will be used to further develop the project with the same team members with the aim to expand the application of the developed methodology to other domains and genres. The dataset is ready for annotation of MWEs.

References

Baldwin T. and Su Nam Kim, 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, Handbook of Natural Language Processing, 2 edition, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA.

Draft guidelines for nominal MWEs (https://docs.google.com/document/d/1bvjSwHpj8I2zJXmftCpx19u3BNWdKtdeg21f4YVHhWw/edit#heading=h.gt53hu7d9q5p)

Rackevičienė, S.; Utka, A.; Mockienė, L.; Rokas, A., 2021. Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase // Kalbų studijos. Kaunas: Technologija. ISSN 1648-2824. eISSN 2029-7203. no. 39, p. 85-92. DOI: 10.5755/j01.sal.1.39.29156. Available at: https://kalbos.ktu.lt/index.php/KStud/article/download/29156/15155

Rackevičienė, S.; Utka, A.; Bielinskienė, A.; Rokas, A., 2022. Distribution of Terms Across Genres in the Annotated Lithuanian Cybersecurity Corpus // Respectus philologicus. Kaunas: Vilniaus universiteto Kauno fakultetas. ISSN1392-8295. eISSN2335-2388., vol. 41(46), p. 26-42. DOI: 10.15388/RESPECTUS.2022.41.46.105. [Humanities International Complete; Scopus; CEEOL – Central and Eastern European Online Library] [CiteScore: 0,10, SNIP: 0,000, SJR: 0,139, Q3 (2020, Scopus Sources)] [M.kr.: H 004]. Available at: https://www.zurnalai.vu.lt/respectus-philologicus/article/view/24950/26155

Rokas, A.; Rackevičienė, S.; Utka, A., 2020. Automatic Extraction of Lithuanian Cybersecurity Terms Using Deep Learning Approaches // Human Language Technologies – The Baltic Perspective Proceedings of the Ninth International Conference Baltic HLT 2020 / edited by Andrius Utka, Jurgita Vaičenonienė, Jolanta Kovalevskaitė, Danguolė Kalinauskaitė. Amsterdam; Berlin; Washington: IOS Press. ISBN 9781643681160. eISBN 9781643681177. p. 39-46. (Frontiers in artificial intelligence and applications, ISSN 0922-6389, eISSN 1879-8314; vol. 328). DOI: 10.3233/FAIA200600. Available at: http://ebooks.iospress.nl/volumearticle/55521

Savary, A., Ramisch, C., Cordeiro, S. R., Sangati, F., Vincze, V., Qasemi Zadeh, B., ... & Doucet, A., 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL* (pp. 31-47).

Savary, A., Stymne, S., Mititelu, V. B., Schneider, N., Ramisch, C., & Nivre, J. 2023. PARSEME meets universal dependencies: getting on the same page in representing multiword expressions. Northern European Journal of Language Technology, 9(1).

Utka, A.; Rackevičienė, S.; Mockienė, L.; Rokas, A.; Laurinaitis, M.; Bielinskienė, A., 2022a. Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction // Selected Papers from the CLARIN Annual Conference 2021. Edited by Monica Monachini and Maria Eskevich. Linköping Electronic Conference Proceedings 189, ISBN: 978-91-7929-444-1, ISSN: 1650-3686 (print), 1650-3740 (online), p. 126–138. DOI: https://doi.org/10.3384/ecp18912

Utka, Andrius; Rackevičienė, Sigita; Rokas, Aivaras; Bielinskienė, Agnė; Mockienė, Liudmila; Laurinaitis, Marius, 2022b, English-Lithuanian Parallel Cybersecurity Corpus - DVITAS, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/46.

Utka, Andrius; Rackevičienė, Sigita; Rokas, Aivaras; Bielinskienė, Agnė; Mockienė, Liudmila and Laurinaitis, Marius, 2022c, English-Lithuanian Comparable Cybersecurity Corpus - DVITAS, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/47.