faculty of social sciences, arts and humanities
ktu 1922

SAL 39/2021
Research Journal
Studies about Languages
pp. 85-92
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
DOI 10.5755/j01.sal.1.39.29156

LINGUISTICS / KALBOTYRA

Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase

Received 05/2021 | Accepted 10/2021

crossref http://dx.doi.org/10.5755/j01.sal.1.39.29156

HOW TO CITE: Rackevičienė, S., Utka, A., Mockienė, L., & Rokas, A. (2021). Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase. *Studies about Languages / Kalbų studijos, 39*, 85–92. http://doi.org/10.5755/j01.sal.1.39.29156

Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase

Anglų-lietuvių kalbų kibernetinio saugumo terminų bazės kūrimo metodikos modelis

SIGITA RACKEVIČIENĖ, Mykolas Romeris University, Lithuania

ANDRIUS UTKA, Vytautas Magnus University, Lithuania

LIUDMILA MOCKIENĖ, Mykolas Romeris University, Lithuania

AIVARAS ROKAS, Vytautas Magnus University, Lithuania

Abstract

The aim of the paper is to present a methodological framework for the development of an English-Lithuanian bilingual termbase in the cybersecurity domain, which can be applied as a model for other language pairs and other specialised domains. It is argued that the presented methodological approach can ensure creation of high-quality bilingual termbases even with limited available resources. The paper touches upon the methods and problems of dataset (corpora) compilation, terminology annotation, automatic bilingual term extraction (BiTE) and alignment, knowledge-rich context extraction, and linguistic linked open data (LLOD) technologies. The paper presents theoretical considerations as well as the arguments on the effectiveness of the described methods. The theoretical analysis and a pilot study allow arguing that: 1) a combination of parallel and comparable corpora enable to considerably expand the amount and variety of data sources that can be used for terminology extraction; this methodology is especially important for less-resourced languages which often lack parallel data; 2) 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; 3) LLOD technologies enable to integrate the terminological data into the global linguistic data ecosystem and make it reusable, searchable and discoverable across the Web.

KEYWORDS: termbase compilation, parallel and comparable corpora, terminology annotation, terminology extraction, knowledge-rich context extraction, deep-learning systems, LLOD technologies.

Introduction

The aim of the paper is to present a methodological framework for the development of an English-Lithuanian bilingual termbase in the cybersecurity domain. We argue that the methodology can be applied as a model for other language pairs and other specialised domains, as it ensures creation of high-quality bilingual termbases even with limited available resources.

Cybersecurity (CS) domain was chosen for several reasons. Firstly, this area is particularly relevant in the current information age, whereas the COVID-19 pandemic, which has accelerated digital transformation of state institutions and businesses, has further increased its significance. Global connectivity, an extensive use of cloud services and remote work pose huge challenges to the security of sensitive data on all levels: state, business and individual. Thus, cyber awareness and cyber hygiene have gained utmost importance not only for governmental institutions and companies, but also for every user of the Internet. The termbase of this domain is believed to contribute to better understanding of cyber threats and data protection measures in Lithuania. Secondly, the cybersecurity domain is particularly dynamic as new concepts are constantly developed and get new terminological designations, predominantly in English. Counterparts of these designations are constantly created in other languages. In Lithuanian, new cybersecurity concepts are often expressed by several Lithuanian term variants, English-Lithuanian hybrids or even original English terms. Therefore, the termbase based on the generalised empirical data is believed to help target users to select the most appropriate terminology for their needs: drafting of official documents and their translation, technical writing, scientific and educational writing, etc.

As the cybersecurity domain is rapidly evolving and ever changing, the methodology of termbase development should allow constant updating of its data and metadata, as well as constant monitoring of the domain and its terminological resources in other languages. The state-of-the-art technologies of machine learning and neural networks have become indispensable for effective automatisation of data and metadata extraction procedures. Therefore, a methodology based on deep learning systems was chosen for the present terminology project.

In the corresponding sections of the paper, we will discuss the following methods for the English-Lithuanian cybersecurity termbase compilation:

- Dataset collection methodology: compilation of comparable and parallel corpora; development of gold standard corpora;
- Bilingual terminology extraction (BiTE) and alignment methodology: development and application of deep learning systems using gold standard corpora as training data;
- Knowledge-rich context extraction methodology: development and application of deep learning methods;
- Development of an interlinked bilingual termbase using Linguistic Linked Open Data (LLOD).

Each of the methods will be grounded on theoretical considerations presented in scientific studies on the relevant issues, as well as a pilot study performed by the authors of the article. In addition, the authors' arguments for choosing particular methods for the on-going research on terminology of the cybersecurity domain will be presented.

Dataset Collection Methodology

The collection of datasets for BiTE encompasses two stages: 1) compilation of comparable and parallel corpora and 2) development of gold standard corpora for training of deep learning systems. Each of the stages is described in corresponding subsections.

Compilation of parallel and comparable corpora

While there is a long tradition for BiTE from parallel corpora (Kupiec, 1993), the same is not true for BiTE from comparable corpora, i.e. corpora composed of original texts in different languages which share common properties such as subject field, target audience, proportional composition of text genres/types, time period, size, etc. BiTE methodology from comparable corpora of different languages is still closely analysed with a number of significant research papers published on this topic (Vintar, 2010; Delpech et al., 2012; Gornostay et al., 2012; Aker et al., 2013; Chu et al., 2016). The research papers show common agreement that BiTE from comparable corpora provides valuable terminographic data which cannot be collected by other means. It also should be noted that comparable corpora are especially important for less-resourced languages which often lack parallel data. Thus, in order to perform more efficient data extraction, a methodology that combines parallel and comparable corpora is often used by researchers. The most important advantages of such methodology are as follows:

- Data source diversity: parallel data sources are often scarce for less-resourced languages, especially in rapidly evolving domains in which resources become obsolete very quickly. Comparable data sources, on the other hand, are highly diverse, and they include various text genres and text types. The combination of both types of data sources allows including greater variety of texts in corpora to be used for terminology extraction.
- Language diversity: as comparable corpora are composed of original texts, their language is more natural than in parallel corpora in which the target language is inevitably influenced by the source language. As McEnery and Xiao indicate, parallel corpora "alone serve as a poor basis for cross-linguistic contrasts, because translations (i.e. L2 texts) cannot avoid the effect of translationese" (McEnery & Xiao, 2007). A combination of comparable and parallel corpora allows comparing the usage of language in original and translated texts.

Thus, BiTE from comparable corpora produces terminographic data which enable constant updating of the existing bilingual termbases or compiling new ones even if parallel data are not available. Besides, BiTE from comparable corpora, used in addition to BiTE from parallel corpora, allows extracting and comparing terminology formed and used in various settings.

The process of compilation of parallel and comparable corpora differs significantly. While parallel data sources immediately provide texts, which are identical/nearly identical in their content and size, comparable data sources have to be carefully selected. Their selection contains potential pitfalls that have to be considered:

- if texts in comparable corpora cover somewhat different topics in the languages, the extracted terminology will not be easily alignable;
- if texts in comparable corpora are of different genres and time periods, the extracted terminology may differ in their pragmatic appropriateness (e.g., it may be common only to official documents in L1 and only to media texts in L2);
- if comparable corpora are different in size, it may not be possible to find equivalents for many terms.

Though data source selection is more complicated in compilation of comparable corpora, their pre-processing is much simpler as it does not require alignment of texts, which is necessary in compilation of parallel corpora.

In our project, the combination of parallel and comparable corpora is necessary. The parallel English-Lithuanian data in the cybersecurity domain are rather scarce. The only easily accessible source is EUR-Lex, the database of the European Union law and other public documents of the European Union in 24 official languages of the EU. In addition, some parallel data may be extracted from the international convention on cybersecurity translated into Lithuanian: Budapest Convention on Cybercrime, 2001, drawn by the Council of Europe.

However, these data do not contain the cybersecurity terminology used in the texts produced outside the EU institutions and international organisations. As this domain is particularly dynamic, legal, administrative, technical, scientific, educational, media and other texts, published in the international community and Lithuania, are important to capture the whole rapidly changing picture of the cyber terminology. Therefore, comparable corpora are indispensable in such research. Our English-Lithuanian comparable corpora are currently being composed of the following text genres used in the cybersecurity domain in the national and international settings:

- legislation and other legally binding documents that form a cybersecurity strategy and regulate its implementation,
- documents of cybersecurity agencies responsible for management of cybersecurity risks,
- academic literature on the cybersecurity domain,
- specialised cybersecurity media,
- mass media articles on cybersecurity issues.

Other text genres and types are being considered (e.g., technical manuals and standards). Thus, the combination of parallel and comparable corpora has widened our possibilities to include a variety of text genres and types, as well as to collect data from both national and international sources.

Development of gold standard corpora

Gold standard corpora with manually annotated terminology are widely used in development of natural language processing (NLP) systems. Their significance has increased with the usage of neural networks. Gold standard corpora allow not only validating, but also training deep learning systems and testing the results of the trained models by calculating their precision and recall. Manual terminology annotation is performed for a number of projects (Bada et al., 2010; QasemiZadeh & Handschuh, 2014; Schumann & Fischer, 2016; Hätty et al., 2017).

The definition of termhood varies in different gold standard corpora ranging from very strict to particularly liberal: in some projects, strict syntactic and semantic constraints are followed, while other projects "rely on the association an annotator has with respect to a term or to a domain (e.g. by structuring terms in a mind map) and provide theoretical background about terminology" (Hätty et al., 2017). Both approaches have their advantages: the former enables to achieve high inter-annotator agreement while the latter allows capturing a greater variety of concepts and their terminological designations.

In our project, we have planned to develop two types of English-Lithuanian gold standard corpora: comparable (100,000 words) and parallel (100,000 words). Linguistic and conceptual annotation criteria have been developed based on the pilot annotation results aiming to achieve maximum consistency in annotation work, which is critical in training deep learning systems. The linguistic criteria define the formal categories of the lexical units to be annotated, namely, nouns, noun phrases, initialisms used as nouns and noun phrases that include initialisms. The conceptual criteria determine the categories of the lexical units to be annotated according to their conceptual characteristics. These categories constitute the basis of the main tagset which comprises the following tags (c.f. Roelcke, 1999, as cited in Hätty et al., 2017):

- intra-subject terminology (terminology of the cybersecurity domain),
- inter-subject terminology (terminology of domains related to cybersecurity),
- proper names of documents, institutions, service organisations, projects, software, etc. relevant to the cybersecurity domain.

We decided to annotate both intra-subject and inter-subject terminology as it will allow us to analyse the domains that are mostly related to and dependent on cybersecurity. Guidelines for distinguishing intra-subject and inter-subject terminology have been developed based on the analysis of the existing cybersecurity ontologies and glossaries, as well as on consultations with field experts.

In addition to the main tagset, a list of attributes is provided to enable marking of additional term features: term usage variants (incomplete terms, interrupted terms), terms with specific formal structure (initialisms), terms of specific origin (English-Lithuanian hybrids, non-adapted English borrowings in the Lithuanian texts).

A special annotation software *QuickTag* has been developed for the purpose of creating training data for deep learning systems. It allows manually annotating monolingual texts and bilingual parallel texts: tagging lexical units with the tags indicating their conceptual characteristics, identifying and tagging nested terms and ascribing additional attributes to the tagged terms and proper names. The software also allows exporting tagged lexical units to a *MS Excel* spreadsheet file with rich statistical metadata for analysis purposes.

Bilingual Term Extraction (BiTE) and Alignment Methodology

After collection and pre-processing of the data, the next stage in termbase compilation is a two-step procedure involving automatic extraction of domain specific terms from comparable and parallel corpora and the alignment of extracted source language and target language terms.

Current terminology extraction methods employ machine learning and deep learning approaches. Our pilot study on automatic extraction of monolingual (Lithuanian) cybersecurity terms proved that this methodology allows achieving high results even with very limited resources (Rokas et al., 2020). In the pilot study, several setups of different neural network configurations were iteratively tested by comparing their results to the gold standard which was pre-trained on a very small manually annotated training data (66,706 words of which 1,258 cybersecurity terms were manually annotated) compiled specifically for extraction of cybersecurity terminology. The best results were achieved with Bidirectional Long Short-Term Memory model (Bi-LSTM) using multilingual Bidirectional Encoder Representations from Transformers (BERT) embeddings reaching F1 score of 78.6%. The achieved high score suggests that the semi-supervised deep learning approach is a way to go (Rokas et al., 2020).

It should, however, be noted that a number of different possibilities for neural network setups exist. These involve using different hyperparameters, optimisers, and word embeddings. Promising results for sequence labelling tasks have been reported in a study by Ulčar and Robnik-Šikonja (2020) with trilingual BERT-like models. It has been shown that the reduction of the number of languages to three (two similar less-resourced languages from the same language family and English) in multilingual models helps to achieve better results. For example, in named entity recognition (NER) task F1 score for CroSloEngual (Croatian, Slovenian, and English) model, when compared with multilingual BERT, significantly improved: from 0.790 to 0.894 for Croatian, from 0.903 to 0,949 for Slovenian, and from 0.940 to 0.949 for English (Ulčar & Robnik-Šikonja, 2020). This methodology will be tested in our project in order to select the best possible setup.

Once the source language and target language terms are extracted, they will be automatically aligned based on co-occurrence measures (such as the Dice coefficient or the weighted mutual information) of their translations. This will be performed by selecting the most probable counterpart from a set of automatically generated translations. The extracted and aligned terms will be reviewed and validated manually by a field expert.

On the basis of the comparative analysis of the most dominant terms (determined on the frequency and dispersion criteria) in the Lithuanian and English corpora, a list of no fewer than 300 most important concepts and their terminological designations in English and Lithuanian will be drafted. Synonymy cases will be registered.

Knowledge-rich Context Extraction Methodology

Once the terms and their counterparts are extracted and selected for a termbase, the formulation of their definitions has to be carried out.

In order to facilitate the formulation of definitions, knowledge-rich contexts (defining and explanatory) are often automatically extracted from available corpora. Commonly, pattern-based methods are used for the task, when pattern templates are constructed in order to identify definitions or explanatory sentences from corpora (Auger & Barrière, 2008; Walter & Pinkal, 2006; Orna-Montesinos, 2011; Bielinskienė et al., 2015). To achieve the most accurate results, the pattern identification templates have to be tested and modified grammatically and lexically. These methods are language dependent, as pattern identification procedures should be adopted for each separate language and curated accordingly for a specific corpus and a specific domain.

In recent years, we can observe the increasing tendency of using deep learning methods for extraction of knowledge-rich contexts (Petrucci et al., 2018; Ayadi et al., 2019; Navarro-Almanza, 2020). Usually, the extraction of knowledge-rich contexts using deep neural networks is a two-step procedure. In the first step, it is possible to completely eliminate hand-crafted rules and to train a neural network model without having to rely on the help of experts. Their contribution remains essential only in the second step in which the extracted definitions should be validated and refined.

This deep learning methodology, which allows simplifying the difficult and time-consuming task of generating hand-crafted rules, closely ties in with the BiTE workflow and makes collection of metadata for the cybersecurity termbase more efficient.

Development of an Interlinked Bilingual Termbase using LLOD

The last stage of the project will be development of a termbase. A modern termbase should be not only published online, but also interlinked with other language resources so that it provides possibilities to access other terminological data and collect the most possible information on a searchable concept.

The state-of-the-art methodology used for interlinking modern termbases is based on Linguistic Linked Open Data (LLOD) technologies which make them interoperable and connected to the Semantic Web. Tim Berners-Lee, the father of Linked Open Data, formulated four conditions for data to be linked data: (1) referred entities should be designated by using URIs (Uniform Resource Identifiers), (2) these URIs should be resolvable over HTTP, (3) data should be represented by means of specific W3C standards (such as Resource Description Framework (RDF)), (4) a resource should include links to other resources (Berners-Lee, 2006).

LLOD termbases have many advantages: they are linked to the global LLOD network and to other termbases, reusable, searchable and discoverable across the Web. Making any lexical database as linked data is seen as a good practice that consequently would result in the formation of a linguistic linked open data cloud (Chiarcos et al., 2013; Bosque-Gil et al., 2016; Di Buon et al., 2020; Rodriguez-Doncel et al., 2015).

Therefore, we consider the representation of English-Lithuanian terminological data in the cybersecurity domain as LLOD as the final and very important step in the workflow of the termbase creation. The integration into the global LLOD ecosystem is highly advisable to any modern online lexical data, and especially so for less-resourced languages.

Conclusions

The analysis of the related research studies, as well as the pilot study on terminology extraction performed by the authors allow arguing that the presented methodological framework would considerably enhance the quality of termbases because it allows:

- expanding the amount and variety of data sources by including both parallel and comparable corpora, which is especially important for less-resourced languages;
- facilitating term extraction by training deep learning systems with manually annotated gold standard corpora;
- regularly updating termbases by automatically extracting terminology and knowledge-rich contexts from new relevant texts;
- integrating the compiled terminological data into the global LLOD ecosystem.

The application of the presented methodology poses the following main challenges: creation of high quality gold standard corpora for training deep learning systems; development of deep neural networks for extraction of terms and knowledge-rich contexts of less-resourced and morphologically-rich languages (such as Lithuanian); alignment of terms of different languages extracted from comparable texts which deal with the same topic, but differ in their contents; application of LLOD technologies for interlinking of terminological data.

One more aspect that should be considered in the work on the compilation of a termbase is close cooperation with field experts. Their contribution is indispensable to collection of texts for corpora compilation, review and validation of annotated datasets, extracted terminology and knowledge-rich contexts, as well as formulation of final definitions of the terms selected for a termbase.

The presented methodology would allow considerably contributing to a more effective management of Lithuanian terminology, as well as terminology of other less-resourced languages which in turn will contribute to smoother communication between experts and the general public.

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).

References

1 Aker, A., Paramita, M. L., & Gaizauskas, R. (2013, August). Extracting bilingual terminologies from comparable corpora. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 402-411).

2 Auger, A., & Barrière, C. (2008). Pattern-based approaches to semantic relation extraction: A state-of-the-art. Terminology, 14(1), 1. https://doi.org/10.1075/term.14.1.02aug

3 Ayadi, A., Samet, A., de Beuvron, F. D. B., & Zanni-Merk, C. (2019). Ontology population with deep learning-based NLP: a case study on the Biomolecular Network Ontology. Procedia Computer Science, 159, 572-581. https://doi.org/10.1016/j.procs.2019.09.212

4 Bada, M., Eckert, M., Palmer, M., & Hunter, L. (2010, July). An overview of the CRAFT concept annotation guidelines. In Proceedings of the Fourth Linguistic Annotation Workshop (pp. 207-211).

5 Berners-Lee, T. (2006, July). Linked Data - Design Issues. W3. http://www.w3.org/DesignIssues/LinkedData.html.

6 Bielinskienė, A., Boizou, L., Grigonytė, G., Kovalevskaitė, J., Rimkutė, E., & Utka, A. (2015). Lietuvių kalbos terminų automatinis atpažinimas ir apibrėžimas. Kaunas: Vytauto Didžiojo universitetas.

7 Bosque-Gil, J., Gracia, J., & Gómez-Pérez, A. (2016). Linked data in lexicography. Kernerman Dictionary News, 24, 19-24.

8 Chiarcos, C., Moran, S., Mendes, P. N., Nordhoff, S., & Littauer, R. (2013). Building a Linked Open Data cloud of linguistic resources: Mo-

tivations and developments. In The People's Web Meets NLP (pp. 315-348). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-35085-6_12

9 Chu, C., Dabre, R., & Kurohashi, S. (2016, May). Parallel sentence extraction from comparable corpora with neural network features. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 2931-2935).

10 Delpech, E., Daille, B., Morin, E., & Lemaire, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora: compositional translation and ranking. arXiv preprint arXiv:1210.5751.

11 Di Buono, M. P., Cimiano, P., Elahi, M. F., & Grimm, F. (2020, May). Terme-a-llod: Simplifying the conversion and hosting of terminological resources as linked data. In Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020) (pp. 28-35).

12 Gornostay, T., Ramm, A., Heid, U., Morin, E., Harastani, R., & Planas, E. (2012, October). Terminology extraction from comparable corpora for Latvian. In Baltic HLT (pp. 66-73).

13 Hätty, A., Tannert, S., & Heid, U. (2017, September). Creating a gold standard corpus for terminological annotation from online forum data. In Proceedings of language, ontology, terminology and knowledge structures workshop (LOTKS 2017).

14 Kupiec, J. (1993, June). An algorithm for finding noun phrase correspondences in bilingual corpora. In 31st Annual Meeting of the Association for Computational Linguistics (pp. 17-22). https://doi.org/10.3115/981574.981577

15 McEnery, A. M., & Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to? In G. James, & G. Anderman (Eds.), Incorporating Corpora: Translation and the Linguist (Translating Europe), Multilingual Matters (pp. 18-31). https://doi.org/10.21832/9781853599873-005

16 Navarro-Almanza, R., Juárez-Ramírez, R., Licea, G., & Castro, J. R. (2020). Automated ontology extraction from unstructured texts using deep learning. In Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications (pp. 727-755). Cham: Springer. https://doi.org/10.1007/978-3-030-35445-9_50

17 Orna-Montesinos, C. (2011). Words and patterns: lexico-grammatical patterns and semantic relations in domain-specific discourses. Alicante Journal of English Studies / Revista Alicantina de Estudios Ingleses, 0(24), 213-233. https://doi.org/10.14198/raei.2011.24.09

18 Petrucci, G., Rospocher, M., & Ghidini, C. (2018). Expressive ontology learning as neural machine translation. Journal of Web Semantics, 52, 66-82. https://doi.org/10.1016/j.websem.2018.10.002

19 QasemiZadeh, B., & Handschuh, S. (2014, August). The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. In Proceedings of the 4th International Workshop on Computational Terminology (Computerm) (pp. 52-63).

20 Rodriguez-Doncel, V., Santos, C., Casanovas, P., Gómez-Pérez, A., & Gracia, J. (2015). A linked data terminology for copyright based on ontolex-lemon. In AI Approaches to the Complexity of Legal Systems (pp. 410-423). Cham: Springer. https://doi.org/10.1007/978-3-030-00178-0_28

21 Rokas, A., Rackevičienė, S., & Utka, A. (2020). Automatic extraction of Lithuanian cybersecurity terms using deep learning approaches. In Frontiers in Artificial Intelligence and Applications, 328 (pp. 39-46). https://doi.org/10.3233/FAIA200600

22 Schumann, A. K., & Fischer, S. (2016, May). Compasses, magnets, water microscopes: Annotation of terminology in a diachronic corpus of scientific texts. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 3578-3585).

23 Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. arXiv preprint arXiv:2006.07890. https://doi.org/10.1007/978-3-030-58323-1_11

24 Vintar, S. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 16(2), 141-158. https://doi.org/10.1075/term.16.2.01vin

25 Walter, S., & Pinkal, M. (2006, July). Automatic extraction of definitions from German court decisions. In Proceedings of the workshop on information extraction beyond the document (pp. 20-28). https://doi.org/10.3115/1641408.1641411

Santrauka

Sigita Rackevičienė, Andrius Utka, Liudmila Mockienė, Aivaras Rokas. **Anglų-lietuvių kalbų kibernetinio saugumo terminų bazės kūrimo metodikos modelis**

Straipsnio tikslas – pristatyti anglų-lietuvių kalbų kibernetinio saugumo terminų bazės kūrimo metodikos modelį, kuris galėtų būti taikomas kitų kalbų porų bei kitų specializuotų sričių terminams tvarkyti. Autorių teigimu, pateiktoji metodika gali užtikrinti aukštos kokybės dvikalbių terminų bazių kūrimą net ir turint ribotus išteklius. Straipsnyje pristatomi terminologinių duomenų ir metaduomenų rinkimo, tyrimo ir tvarkybos principai: kalbama apie tekstynų sudarymo metodus ir problemas, terminų anotavimą, automatinį dvikalbių terminų atpažinimą ir sulygiavimą, informacinių kontekstų atpažinimą ir lingvistinių atvirų susietųjų duomenų (angl. LLOD) technologijas. Straipsnyje taip pat pateikiami autorių argumentai dėl aprašytų metodų efektyvumo. Teorinė analizė ir bandomieji tyrimai leidžia teigti, kad: 1) palyginamųjų tekstynų sudarymas ir naudojimas kartu su lygiagrečiaisiais leidžia išplėsti duomenų šaltinių, skirtų terminų atpažinimui, kiekį ir įvairovę; ši metodika yra ypač svarbi kalboms, turinčioms mažiau išteklių, nes joms dažnai trūksta lygiagrečiųjų duomenų (verstinių tekstų); 2) gilaus mokymosi sistemos, apmokytos naudojant rankiniu būdu anotuotus duomenis (aukso standarto tekstynus), leidžia efektyviai automatizuoti terminologinių duomenų bei metaduomenų rinkimą ir reguliariai atnaujinti terminų bazes su minimaliomis rankų darbo sąnaudomis; 3) lingvistinių atvirų susietųjų duomenų technologijos įgalina terminologinius duomenis integruoti į globalią kalbinių duomenų ekosistemą, kurioje jie būtų susieti su kitais terminologiniais duomenimis. Ši duomenų ekosistema žymiai išplečia jų paieškos ir panaudojimo galimybes.

About the Authors

SIGITA RACKEVIČIENĖ

Prof. dr., Institute of Humanities, Faculty of Human and Social Studies, Mykolas Romeris University, Lithuania

Research interests
Multilingual terminology and terminography, corpus linguistics, computational linguistics

Address
Ateities st. 20, LT-08303 Vilnius, Lithuania

E-mail sigita.rackeviciene@mruni.eu

Orcid iD 0000-0001-5794-0296

ANDRIUS UTKA

Assoc. Prof. dr., Centre of Computational Linguistics, Vytautas Magnus University, Lithuania

Research interests
Natural language processing, terminology, computational linguistics, corpus linguistics

Address
V. Putvinskio st. 23-216, LT-44243, Kaunas, Lithuania

E-mail andrius.utka@vdu.lt

Orcid iD 0000-0001-5212-4310

LIUDMILA MOCKIENĖ

Prof. dr., Institute of Humanities, Faculty of Human and Social Studies, Mykolas Romeris University, Lithuania

Research interests
Contrastive terminology and multilingual terminography, corpus linguistics, ESP teaching

Address
Ateities st. 20, LT-08303 Vilnius, Lithuania

E-mail liudmila@mruni.eu

Orcid iD 0000-0001-7153-7276

AIVARAS ROKAS

Programmer, Centre of Computational Linguistics, Vytautas Magnus University, Lithuania

Research interests
Natural language processing, computational linguistics, deep learning systems, terminology extraction, LLOD

Address
V. Putvinskio st. 23-216, LT-44243, Kaunas, Lithuania

E-mail aivaras.rokas@vdu.lt

ORCID iD 0000-0003-3602-3872