Annotation of Multiword Expressions in Lithuanian Corpus of User-Generated Comments

Giedrė Valūnaitė-Oleškevičienė, Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania, gvalunaite@mruni.eu

Key words: multiword expressions, user-generated comments, low-resourced languages, annotation guidelines.

Introduction

The research area of multiword expressions (MWEs) involves sophisticated idiosyncrasies because of their lexical, syntactic, semantic, pragmatic peculiarities. MWEs feature of the irregular nature poses complex problems in their annotation and further in NLP tasks and end-user applications such as natural language modelling and machine translation. MWE processing in low-resourced or small languages is of key importance to providing annotated datasets further used for tools allowing full MWE identification for end-user applications. One of the breakthrough projects in the MWE research area is the PARSEME shared task systems which conditioned a significant progress in MWE identification by providing datasets including low-resourced languages and tools for integrating MWE identification into end-user applications (Savary et al., 2017). Recent efforts have been directed to establishing effective methods for the automatic interpretation of MWEs as well as processing MWEs in low-resourced languages. Resource creation and sharing is of key importance for low-resourced languages and should be pursued in parallel with the development of methods enabling the use and application of small datasets.

Dataset

The data for the annotation and research is going to be obtained from the Lithuanian corpus *LITIS v.1* of user-generated comments that is freely available from the CLARIN-LT repository (Amilevičius et al., 2016). The dataset includes approximately 200 thousand comments collected between 2010-2014 from the two major Lithuanian news portals: delfi.lt and lrytas.lt. The annotation of MWEs is planned to start on the comments from "Delfi.lt" and later the other part of the dataset could be used for deep learning experiments to automatically extract MWEs.

It is important to draw attention to the fact that until December of 2021 "Delfi.lt" portal allowed the readers to publicly comment on news articles with very little restriction allowing the commentors to remain anonymous without any registration. Only the commentor IP addresses could be seen, and the content was manually moderated. Because of such liberal policies the comment sections of news portals became a platform for expressing commentor views on various societal issues with a hint of offensive language. It was estimated that about13 thousand comments were generated daily on "Delfi.lt" portal and it was identified that about 8 per cent of comments which contained vulgar or offensive information were removed (Garbačiauskaitė-Budrienė 2014).

Nowadays the portal has a much stricter policies on comments as for commenting it is required to register and reveal some private details of one's identity. Such stricter policies prevent posting more radical or conflicting comments. The corpus consists of tab separated comment files with the following information: a comment, date and time, nick name of the author, URL and a title of the article commented.

Annotation project

The part of the dataset is annotated with the INCEpTION tool (https://github.com/inception-project/inception for offensive language using Simplified Offensive Language Taxonomy (SOLT) by Lewandowska-Tomaszczyk et al. (2022) to explore the potential for identifying offensive language in Lithuanian (Valūnaitė-Oleškevičienė et al., 2023).

The current project annotation attempt is aimed at annotating the same part of the corpus for nominal MWEs using cross-lingual annotation **guidelines For MWEs** first draft produced by UniDive (COST action CA21167) (draft guidelines for nominal MWEs). The expected result is twofold:

- while working on the annotation of nominal MWEs in Lithuanian language it is expected to identify and suggest the necessary improvements to the first attempt of the cross-lingual guidelines for MWEs (draft guidelines for nominal MWEs).

- It is also expected to establish MWEs directly or idiomatically used in offensive language.

While implementing the project, the training school is of a key importance as it provides trainings in annotation of MWEs for newcomers and corpus annotation infrastructure (annotation platforms, format validators, Git etc.) It is expected to learn manually to identify MWEs in corpora, classify the multiword expressions identified in corpora according to the types defined for them in PARSEME and UniDive, and describe this phenomenon in the Lithuanian language. It is also expected to get certain insights from the annotation and to present them so as to be further taken into account for the improvement of the annotation guidelines.

Open questions

What are the peculiarities of the annotation process of nominal MWEs in Lithuanian?

What is the role of MWEs (especially idiomatic expressions) in offensive language?

How automatically identify MWEs in offensive language?

The project phase

The project is in the planning phase and ready to start at the summer school. The dataset is prepared, annotated for offensive language so it is expected to get additional layer of annotation of nominal MWEs successfully added for further research on the use of MWEs in offensive language.

References

Amilevičius, Darius and Petkevičius, Mažvydas. 2016. *LITIS v.1*, CLARIN-LT digital library in the Republic of Lithuania. Available at http://hdl.handle.net/20.500.11821/11.

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

Garbačiaukaitė-Budrienė, Monika. 2014. Kaip prisijaukinti žiniasklaidą? Žurnalistai: kovoti ar bendradarbiauti? *Valstybės tarnybos aktualijos.*

Lewandowska-Tomaszczyk, Barbara. 2022. A simplified taxonomy of offensive language (SOL) for computational applications. *Konin Language Studies*, 10:213–227.

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).

Valūnaitė-Oleškevičienė, G., Selmistraitis, L., Utka, A., & Gudelis, D. (2023). Offensive language in user-generated comments in Lithuanian. *Lodz Papers in Pragmatics*, *19*(2), 239-254.