МОВОЗНАВСТВО

UDC 811.161.2’42:004.912
DOI https://doi.org/10.18524/2307-8332.2025.2(32).350473

Mariia MALYSHEVA
PhD in Philology
Associate Professor at the Department of Applied Linguistics
Odesa I. I. Mechnikov National University
Odesa, Ukraine
mariiamalysheva@onu.edu.ua
ORCID iD: https://orcid.org/0000-0002-1910-4833

DEVELOPMENT OF UKRAINIAN-LANGUAGE CORPUS OF AGGRESSIVE TEXTS OF NETWORK DISCOURSE

The study of network discourse as a space for the deployment of aggressive communicative behavior requires the use of automated data analysis methods. This requires large arrays of texts, but Ukrainian-language corpora of texts suitable for studying aggression in network discourse are not publicly available. The investigation highlights the process of developing a Ukrainian-language corpus of aggressive texts within the project on the study of verbal aggression in network discourse. The choice of text source for the corpus is substantiated, a self-developed application for collecting comments using web-scraping is proposed, the procedure of processing the received texts is outlined, including cleaning, setting language and polarity, additional cleaning, tokenization and lematization, removal of stop words. The source of the texts is the Internet portal Censor.Net. The application was developed in the Python programming language with the Beautiful Soup module, regular expressions were used to clean texts, the langdetect library was used to set the language, polarization was determined using the polyglot tool, the simplemma library was used for tokenization and lematization, and nltk was used to identify bigrams. For the additional stage of lemmatization, a dictionary with frequent tokens (frequency> = 5) and their lemmas was created. The obtained Ukrainian-language corpus of aggressive texts of network discourse includes 16,769 comments, 188,825 tokens, of which 39,975 are unique, 168,702 unique bigrams, in the corpus the author’s nickname, comment text, language and polarity are additionally indicated. Dictionaries of tokens and bigrams indicate their frequency. The size of the corpus is not fixed, the corpus will continue to be supplemented with comments. The corpus, dictionaries and applications for their creation are available in the public access.

***Keywords**: verbal aggression, network discourse, text corpus, Ukrainian language, application for creating text corpus, computational linguistics*

Introduction

During the collection of factual material for a project on the study of verbal aggression in network discourse, the problem of unproductiveness of manual data collection was identified. This method is time consuming and requires the development of a file for manual recording and annotation of the collected material. Communication on social networks is continuous, the number of potentially aggressive comments is constantly growing, so, on the one hand, increasing the amount of factual material leads to changes in the structure of the file, and manual processing of large amounts of information increases the risk of human error. On the other hand, internal filters of social networks quickly remove content that violates the established rules, in particular, regarding the aggressive style of communication, which makes it impossible to capture a large amount of factual material. Therefore, we concluded that the study of verbal aggression in the network discourse requires the use of automated methods of data collection and analysis.

Working with text corpora provides a wide range of opportunities for data analysis. Although creating datasets is an important and time-consuming task [8], and the available datasets have a number of limitations [13], several datasets and corpora of aggressive texts in different languages are developed [9]. However, the number of investigations into the creation of Ukrainian-language corpora of aggressive texts is small. In particular, an attempt was made to create a mixed dataset of Ukrainian-language and Russian-language comments from YouTube for further automated search of offensive content [2]. Comments on 329 videos related to the Euromaidan theme were collected. The final dataset contains over 50,000 comments. 2,000 comments were randomly selected for manual annotation by native speakers on offensive and non-offensive (of which 32.7% of comments have offensive content). The resulting data set and annotated corpus are publicly available [1]. There was also an attempt to create a Ukrainian-language corpus of Twitter messages for further automated detection of toxic texts [5]. Using the twitter-scraper 1.87 million tweets with additional information about time, language, number of responses and retweets, likes, hashtags, URL links and author nicknames were received. The texts were annotated as toxic and non-toxic, a total of 55,153 tweets were annotated, however, there is no information on the percentage of aggressive texts in the corpus. The resulting corpus is available in the public access [4].

Despite the existing attempts, there is no (at least in open access) corpus of purely aggressive texts in modern Ukrainian-language linguistics, so we tried to create a Ukrainian-language corpus of aggressive texts and publish it public access. The developed corpus will be useful in solving various problems of linguistics and machine learning, e.g. natural language processing (automated search for aggressive content in texts, etc.).

Methodology

The development of an annotated Ukrainian-language corpus of aggressive texts involves the following stages:

1) selection of data sources;
2) creation of a dataset using web-scraping technology;
3) received texts processing.

Consider in more detail the work at each stage.

1. Data sources

Initially, we planned to use comments from the Internet portals Twitter, Facebook and YouTube. We sent a request on Twitter to get the API and legally download the required number of tweets for further research, however, despite the information provided about the objectives of the study, the company refused. To download comments from Facebook an API is also needed, and the tools available to collect comments from these Internet portals have a number of restrictions (e.g. allow collecting comments only under a certain hashtag, or only from a certain author or from a certain page), so we decided to develop a number of criteria for selecting a data source

Due to the objectives of the project (study of verbal aggression in network discourse), the data source must meet the following criteria: be popular (1) in the Ukrainian-language segment of the network (2) Internet portal (3), where communication takes place on socially significant potentially conflicting topics (4), with freely available comments (5). Scientists emphasize that new sources should be used when creating datasets [12], so this is the sixth criterion.

According to these criteria, we were forced to abandon the use of Facebook and YouTube, and as the final source of comments we choose the portal Censor.Net [6], because it meets all these criteria.

2. Corpus creation

Collection of comments from Internet portal is done through web scraping. Web scraping is a method of extracting data from the Internet and storing it in a file system or database for further analysis [15]. We have developed an application in the Python programming language (version 3.9.9) using the Beautiful Soup module (version 4). Beautiful Soup creates a tree with all the elements of the document and allows you to extract the necessary information from web pages. As input, our application receives the number of pages from which download comments, and the language of these comments (in our case, Ukrainian). Due to the structure of the Internet portal Censor. Net, we receive a list of links to news pages from each page of the portal, and we are looking for comments on these news pages. Each comment is considered separately. First, we use regular expressions to remove all extra spaces and special characters. We then use the langdetect library to set the comment language and use the natural language processing tool polyglot [7] to determine its polarity. The polyglot tool provides an opportunity to automatically analyze the tonality of texts based on the dictionary of polarity (three levels of polarity of words: words with positive

connotations, neutral one and negative one). Comments with negative polarity are added to the dataset. At the end we get a corpus of Ukrainian-language aggressive comments in the csv file. This corpus is suitable for further use as a source base for linguistic research, however, to solve certain problems that require the use of computer technology, additional processing of the received texts is required.

3. Additional processing

Additional processing of the received texts consists in deeper clearing of the data, lexical analysis and removal of stop words. To do this, an additional application was created in the Python programming language (version 3.9.9). First, with the help of regular expressions, we perform a deep cleansing of texts, i.e. turn into lowercase, delete non-letter characters and letters of the Latin alphabet. We understand that deep cleansing will remove some of the markers of verbal aggression (e.g. exclamation marks, intentional change of case, etc.), but these procedures are standard when creating datasets for further processing by a computer. Next, with the help of the simplemma library we perform lexical analysis and lemmatization, i.e. bringing the word form (token) obtained during lexical analysis to the lemma, its vocabulary form (e.g. for a noun it is a noun case, singular, for an adjective it is noun case, singular, masculine). Due to the fact that simplemma uses dictionaries for lemmatization, it is not possible to lemmatize occasionalisms, obscene language, neologisms, city names, surnames (e.g. rashist, orc, katsapiya, etc.), so we added an additional stage of lemmatization based on a self-created dictionary with frequent (frequency >= 5) tokens that could not be lemmatized, and their lemmas. At the final stage, we delete the stop words. To remove stop words, we took as a basis the collection of S. Kupriienko, which is available in the public access [10], and added to it additional stop words from the current dataset. Additionally, we use the nltk library to search for bigrams. At the output we get a csv file with unique tokens and a csv file with unique bigrams in descending order of their use in the corpus.

Results and discussion

The final corpus of aggressive Ukrainian-language texts at the current stage has 16,769 comments, 188,825 tokens, of which 39,975 are unique, 168,702 unique bigrams.

Texts for the corpus were collected during March-April 2022, and the reaction of the Ukrainian nation to the military aggression performed by the terrorist state Russian Federation is engraved in it. As the active phase of the war is still ongoing, and our application for collecting comments from Internet portals allows downloading comments for different periods, the corpus will be supplemented with relevant material. The dictionaries necessary for the functioning of the corpus will also be constantly updated.

Currently, the corpus provides an opportunity to explore conflict communication in situations of psychological tension on the background of war. The purpose of the corpus is to create a source base for linguistic research, in particular, to study

aggressive speech behavior, on its material you can get the necessary statistics on the use of individual tokens and collocations, compile dictionaries etc.

The developed application for creating text corpora of network discourse provides an opportunity to collect comments in different languages and for different periods of time, additionally determines the polarity of comments (positive, neutral, negative). In the output csv file we get the author's nicknames, comments, comment language and polarity (see Table 1).

Table 1. Output

Author	Text	Language	Polarity
Sergiy Bogutskiy	Ідуть у тому ж напрямку що і руZкий корабель.	uk	-1
Александр Чуднивец f1b95bc0	Лапатні дебіли знайшли та знищили мабуть біо-лабораторію...	uk	-1

After additional processing of texts (deep cleaning, lexical analysis, removal of stop words), we get two cvs files with unique tokens and unique bigrams (Table 2).

Table 2. Unique tokens, unique bigrams, their frequencies

Index	Lemma	Frequency	Bigram	Frequency
0	Україна	2273	(‘слава’, ‘Україна’)	74
1	війна	1434	(‘ядерний’, ‘зброя’)	63
2	кацап	1008	(‘український’, ‘народ’)	56
3	більший	723	(‘війна’, ‘Україна’)	49
4	питання	691	(‘окупований’, ‘територія’)	45
5	український	664	(‘територія’, ‘Україна’)	41
6	країна	609	(‘початок’, ‘війна’)	41
7	знати	594	(‘країна’, ‘нато’)	40
8	українець	590	(‘український’, ‘влада’)	38
9	ворог	571	(‘мирний’, ‘населення’)	36
10	зброя	562	(‘бойовий’, ‘дія’)	36
11	їсти	526	(‘збройний’, ‘сила’)	35
12	нато	493	(‘дати’, ‘бог’)	34
13	орк	490	(‘російський’, ‘окупант’)	34
14	росія	485	(‘вбивати’, ‘українець’)	32
15	казати	465	(‘горіти’, ‘пекло’)	31
16	місто	464	(‘військовий’, ‘злочин’)	30
17	вбивати	447	(‘військовий’, ‘злочинець’)	29
18	думати	427	(‘ракетний’, ‘програма’)	28
19	зробити	422	(‘жінка’, ‘діти’)	27

Due to the fact that simplemma uses dictionaries for lemmatization, it is not possible to lemmatize occasionalisms, obscene language, neologisms, names of cities and surnames (e.g. rashist, orc, katsapiya, etc.), so we added an additional stage of lemmatization based on a self-created dictionary with frequent (frequency >= 5) tokens that could not be lemmatized and their lemmas (Table 3). This dictionary is constantly updated with new lemmas.

Table 3. Lemma dictionary

Token	Lemma
банкової	банкова
нептунів	нептун
арахамії	арахамія
кацапію	кацапія

Conclusions and prospects for research

We developed a Ukrainian-language corpus of aggressive texts of network discourse, also we created a dictionary of unique tokens and bigrams. The developed corpus and applications need to be improved, so our next steps will be:

1) expansion of the corpus;
2) improving the dictionaries for additional lemmatization;
3) part-of-speech tagging;
4) development of detailed instructions for manual annotation of comments on aggressive and non-aggressive.

Ukrainian-language corpus of aggressive texts of network discourse, applications for collecting and processing comments and other materials are open to the general public [11].

References

1. Andrusyak, B., Rimel, M., & Kern, R. (2018). *Dataset of YouTube comments and dictionary of abusive words* [Data set]. Retrieved on May 14, 2022, from https://cutt.ly/jDSdIhc
2. Andrusyak, B., Rimel, M., & Kern, R. (2018). Detection of abusive speech for mixed sociolects of Russian and Ukrainian languages. In A. Horák, P. Rychlý, A. Rambousek (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing*, RASLAN 2018 (pp. 77–84). Tribun EU. Retrieved May 14, 2025, from https://nlp.fi.muni.cz/raslan/2018/paper04-Andrusyak.pdf
3. Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., & Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2003.07428
4. Bobrovnyk, K. (n.d.). *A corpus of Ukrainian Twitter texts + instructions for downloading and filtering texts* [Data set]. GitHub. Retrieved May 14, 2025, from https://cutt.ly/TDSdLYY
5. Bobrovnyk, K. (2019). Automated building and analysis of Ukrainian Twitter Corpus for toxic text detection. *Proceedings of the 3d International Conference 'Computational Linguistics and Intelligent Systems* (Vol. 2, pp. 55–56). Retrieved May 14, 2025, from https://ena.lpnu.ua/handle/ntb/45496
6. Censor.Net. https://censor.net/

7. Chen, Ya., & Skiena, S. (2014). Building sentiment lexicons for all major languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 383–289). Retrieved May 14, 2025, from https://cutt.ly/VF3AY60
8. Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., . . . Wu, D. M. (2017). A large labeled corpus for online harassment research. *WebSci'17: Proceedings of the 2017 ACM on Web Science Conference* (pp. 229–233). https://doi.org/10.1145/3091478.3091509
9. Hate Speech Dataset catalogue. (n.d.). Hatespeechdata. Retrieved May 14, 2025, from https://hatespeechdata.com
10. Kupriienko, S. (n.d.). *Ukrainian-Stopwords: the list of ~2000 Ukrainian stopwords (with numbers)* [Data set]. GitHub. Retrieved May 14, 2025, from https://cutt.ly/lF3IYjY
11. Malysheva, M. (n.d.). *Ukr_corpus_affression: Ukrainian-language corpus of aggressive texts of network discourse* [Data set]. GitHub. Retrieved May 14, 2025, from https://github.com/mariia-malysheva/ukr_corpus_aggression [Access date: 14.05.2022]
12. Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: garbage in, garbage out. *PLoS ONE*, *15*(12), 1–32. https://doi.org/10.1371/journal.pone.0243300
13. Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *Proceedings of the Third Workshop on Abusive Language Online* (pp. 80–93). https://doi.org/10.18653/v1/W19-3509
14. Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: a typology of abusive language detection subtasks. *Proceedings of the First Workshop on Abusive Language Online* (pp. 78–84). https://doi.org/10.18653/v1/W17-3012
15. Zhao, B. (2017). Web scraping. In L. Schintler, & C. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 1–3). Cham: Springer. https://doi.org/10.1007/978-3-319-32001-4_483-1

Марія МАЛИШЕВА

РОЗРОБЛЕННЯ УКРАЇНСЬКОМОВНОГО КОРПУСУ МЕРЕЖЕВИХ ТЕКСТІВ З АГРЕСИВАМИ

Дослідження мережевого дискурсу як простору розгортання агресивної комунікативної поведінки потребує застосування автоматизованих методів аналізу даних. Це вимагає великих масивів текстів, проте українськомовні корпуси текстів, придатні для вивчення агресії в мережевому дискурсі, у відкритому доступі відсутні. У дослідженні висвітлено процес створення українськомовного корпусу текстів з маркерами агресії у межах проєкту з вивчення вербальної агресії в мережевому дискурсі. Обґрунтовано вибір джерела текстів для корпусу, запропоновано розроблений автором застосунок для збирання коментарів за допомогою вебскрейпінгу, окреслено процедуру опрацювання отриманих текстів, що включає очищення, визначення мови та полярності, додаткове очищення, токенізацію та лематизацію, вилучення стоп-слів. Джерелом текстів є інтернет-портал Censor.Net. Застосунок розроблено мовою програмування Python із використанням модуля Beautiful Soup, для очищення текстів застосовано регулярні вирази, для визначення мови — бібліотеку langdetect, полярність визначено за допомогою інструменту polyglot, для токенізації та лематизації використано бібліотеку simplemma, а для виокремлення біграм — nltk. Для додаткового етапу лематизації створено словник частотних токенів (частота ≥ 5) та їхніх лем. Отриманий українськомовний корпус текстів мережевого дискурсу, що

містять маркери агресії, містить 16 769 коментарів, 188 825 токенів, із яких 39 975 є унікальними, а також 168 702 унікальні біграми; у корпусі додатково зазначено нікнейм автора, текст коментаря, мову та полярність. Словники токенів і біграм містять інформацію про їхню частотність. Обсяг корпусу не є фіксованим, його й надалі буде поповнювано новими коментарями. Корпус, словники та застосунки для їх створення доступні у відкритому доступі.

Ключові слова: *вербальна агресія, мережевий дискурс, корпус текстів, українська мова, застосунок для створення корпусу текстів, комп'ютерна лінгвістика.*