UDC 811.111’22 *https://www.doi.org/10.21272/Ftrk.2023.15(1)-1*

# NEWSPAPER SUBCORPUS (SUBCORPUS OF THE MODERN EUROPEAN MEDIA) IN THE STRUCTURE OF THE MULTILINGUAL CORPUS

**Anokhina Tetiana,**
Dr. in Philology
ORCID ID 0000-0002-8859-5568
Kyiv National Linguistic University
73, Velyka Vasylkivska st., 03150, Kyiv, Ukraine
tetiana.anokhina@knlu.edu.ua

***Abstract.*** *The research of the European Media comprises plenty of precious documents compiled into our educational corpora. This study represents how the corpus has been compiled. Also, we find it necessary to show the tools for this compilation which were described and analyzed. The first step of our compilation was collecting raw data in the library. The second step was selecting the format of the selection to compile by the chosen tool (the Sketch Engine). Then we made the following selection of files containing European Media content allowing it to go into the EU collection.*

*The selected files were from the library of the popular media of Europe. It has been selected to have comprised highly cited articles from British sources: the BBC, the Sun, the Daily Mail, the Guardian, the Times, and the Economist. All the mentioned newspapers are known for their investigative journalism and critical analysis of current affairs. Our selected tools for our media subcorpus were the web-based tool for creating corpora Sketch Engine. Also, we used the offline corpus manager AntConc, the open-source software program for corpora analysis.*

***Keywords:*** *EU collection, the Sketch Engine, AntConc, respected media, web-based tools, corpus analysis.*

# ГАЗЕТНИЙ ПІДКОРПУС (ПІДКОРПУС СУЧАСНИХ ЄВРОПЕЙСЬКИХ ЗМІ) У СТРУКТУРІ БАГАТОМОВНОГО КОРПУСУ

**Анохіна Тетяна,**
доктор філологічних наук,
ORCID ID 0000-0002-8859-5568
Київський національний лінгвістичний університет
вул. Велика Васильківська, 73, Київ, 03150, Україна,
tetiana.anokhina@knlu.edu.ua

***Анотація.*** *Корпус європейських ЗМІ, якому присвячено наше дослідження, містять велику кількість цінних документів, зібраних у цей навчальний корпус. Наша розвідка демонструє, яким чином було укладено цей корпус. В цьому дослідженні ми вважали за необхідне описати інструменти для створення корпусу сучасних європейських ЗМІ в межах більшого мультилінгвального корпусу, які були описані та проаналізовані. Першим кроком нашої компіляції був збір метаданих у бібліотеку європейських ЗМІ. Другим кроком був вибір формату виділення методів компіляції корпусу: для компіляції вибраним інструментом (Sketch Engine). Щоб поповнити бібліотеку цього ми створювати Інтернет запити в межах Sketch engine, створюючи вибірку з файлів ЗМІ ЄС.*

*Вибрані файли були бібліотекою популярних ЗМІ Європи. Він був обраний таким чином, щоб він складався з популярних статей з британських джерел: BBC, Sun, the Daily Mail, The Guardian, The Times, The Economist. Усі згадані джерела є відомі журналістськими виданнями та місять критичний аналіз поточних подій. Обрані нами інструменти для укладання нашого газетного підкорпусу ЄС медіа – це веб-інструмент для створення корпусу Sketch Engine, а також офлайн-менеджер AntConc, програму з відкритим кодом для аналізу корпусів в режимі офлайн.*

***Ключові слова**: колекція ЄС, Sketch Engine, AntConc, авторитетні ЗМІ, веб-інструменти, аналіз корпусу.*

**Introduction**

The problem of creating a database of the European Union was an actual research problem that has arisen in terms of our grant topic. The Media corpus we have been selecting is part of the larger corpus containing the units of the modern European media corpus. The problem of corpus creation is relatively new and it is developing in the Ukrainian linguistic circles (Bober, Cherkhava, Hryshchuk, Zhukovska, Kapranov, Korolyova, Liashko, Meleshkevych, Mosiyuk, Vasko).

Some issues are still staying unsolved. We are aiming at developing the educational corpora in which we see the potential and EU elements study facilitation.

The purpose of the study is to create an analysis of the media corpus which gives information on political, economic, and social issues in the EU environment. It serves an understanding of the EU values and problems in terms of our strivings to enter the European zone we need much to be done in the general scopes of the EU studies to fulfill the standards of the EU. A newspaper subcorpus refers to a subset of a larger corpus that consists specifically of newspaper articles. This subcorpus is usually created by extracting newspaper articles from a larger collection of text, such as a general corpus or web corpus. The articles in the subcorpus are typically selected based on criteria such as publication date, source, and topic.

**Methods of research**

The methods applied were broad context and corpus based search automatic and semiautomatic search. Newspaper corpora commonly used in the corpus linguistics research, were used to compile our educational subcorpus of the EU media of our multilingual corpus of the EU studies. We have researched the usage of the popular newspaper subcorpora (BNC) to investigate patterns of language use, such as discourse features of the media text, following the different types of newspapers in Europe.

We have used the modern corpus methodology and tools to apply to the newspaper subcorpus of the EU. The concordance analysis enables us to make the EU media selection from the British National Corpora (BNC) and Sketch Engine tool into the larger corpus “the Europeans multilingual corpus”. The idea of compiling popular media articles is making us familiar with a European heritage. We have relied on the structural approach to collect and compile data, also the applied methods used were data storage and corpus compilation which are mathematically oriented.

## The EU newspaper subcorpus

The newspaper subcorpus we have compiled can be used in teaching European studies. By analyzing the language patterns in newspaper articles, we can learn to identify key topics on European social matters, track climate changes in Europe, economics, culture, and style or general public opinion of Europeans.

The corpus contains the individual documents added manually to the corpus and also it contains other texts from the internet added automatically. The corpus texts are added in multiple languages. These documents may are grouped based on their newspaper-oriented

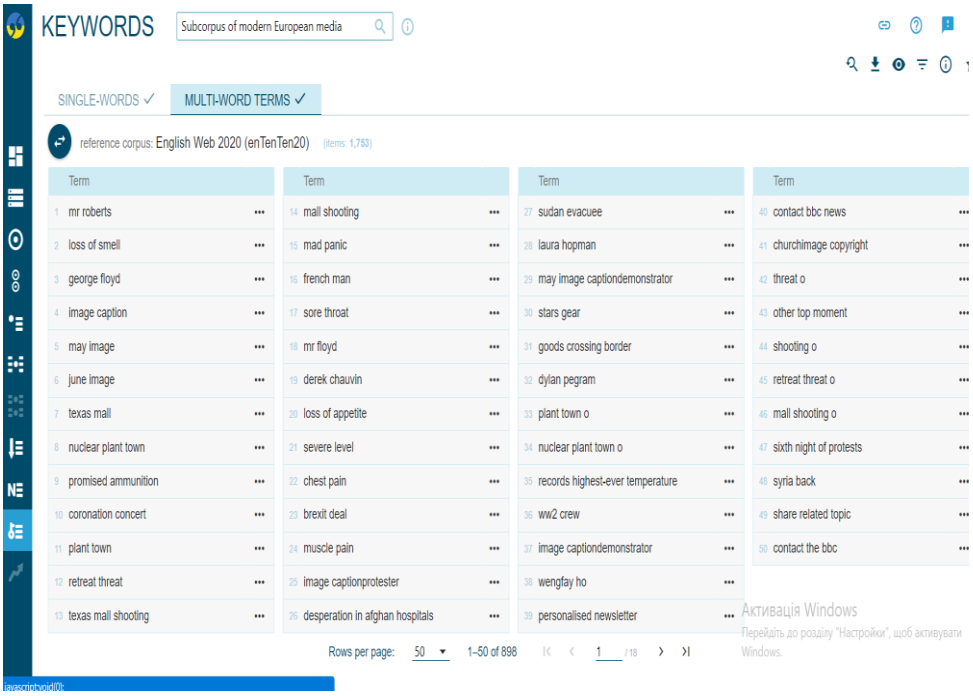
discourse of the EU. It is possible to go on with the educational corpus to add some additional features such as translation alignment to use the multilingual corpus both as an educational and translation corpus. Using various sources the texts of the EU are compiled into the data set of the multilingual corpus with the perspective plan to be aligned at the sentence level or word level.

The multilingual corpora of the newspaper corpus of the EU include metadata, which provides additional information about the documents, such as publication date, author, and source. Sketch Engine enables the newspaper corpus compiled by any sort of document as it runs different encoding formats but it doesn't work with scans. Typically we include text in the Unicode and they are the library of raw texts.

While processing by Skecth Engine our corpus is acquiring additional layers, added by the Skecth Engine system, such as part-of-speech tags, named entities, or sentiment labels. Thus this sort of tagging enables CQL search. The units of our newspaper corpus make up the subcorpus of the modern European media which contains full texts searched for within the Sketch Engine system environment and in an extra way sorting the selection by the *Sketch Engine* into mono units and multi terms (Figure 1-2). Subcorpus of modern European media: mono and multi units

The selection is performed automatically, with such additional possibilities as a download in CSV format, and good academic examples to teach and study EU materials.

**Figure 1 Subcorpus of multi terms of the modern European media**




A subcorpus of modern European media consists of a subset of text data from a larger corpus of European media sources, such as newspapers, magazines, or online news websites. The subcorpus is created by selecting articles based on specific criteria "news spread by media sources in Europe" in the pragmatic proposition. So we aimed at the European Union selection covering a specific period starting from 1st of November 1993 year up to now. The EU corpus captures the recent trends and events in European media.

The subcorpus includes articles in English and in future it will add multiple languages spoken in Europe, such as English, French, German. The subcorpus could focus on media sources from a particular region of Europe, such as Western Europe, Eastern Europe, or the Nordic countries. The subcorpus could be centered on a particular theme or topic, such as politics, economics, sports, or entertainment. Once the subcorpus has been selected, it can be further processed and analyzed using various computational tools and methods, such as text mining, natural language processing, or machine learning.

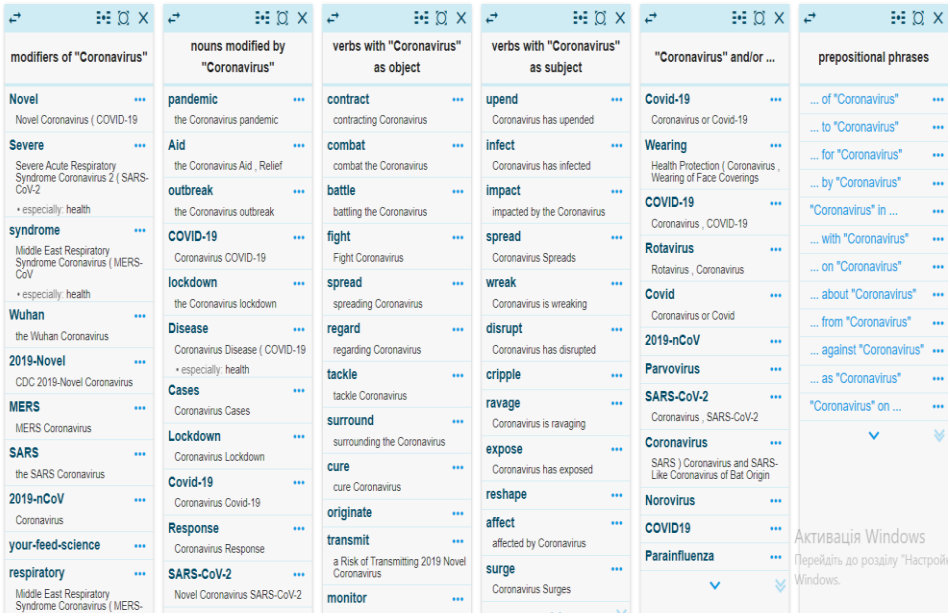
The resulting insights can help researchers better understand the language use, discourse patterns, and cultural trends of modern European media.

The most popular European media outlets can vary depending on the country and language, but several sources have a wide readership or viewership across Europe. The most popular British media outlets are added to the EU media corpus.

BBC: The British Broadcasting Corporation (BBC) is a public service broadcaster that provides news, entertainment, and educational content across various media platforms, including television, radio, and online. The BBC has a wide audience in the UK and is regarded as a trusted source of news and information. The highly cited articles from the BBC contain information on health, e.g. coronavirus, making the thematic section HEALTH of the EU corpus (Figure 2).

The "Coronavirus: What are the symptoms?" article was published in January 2020 and provided an overview of the symptoms of COVID-19, including fever, cough, and difficulty breathing. It was widely cited in the early stages of the pandemic as people sought information about the virus.

**Figure 2 HEALTH in the modern European media**



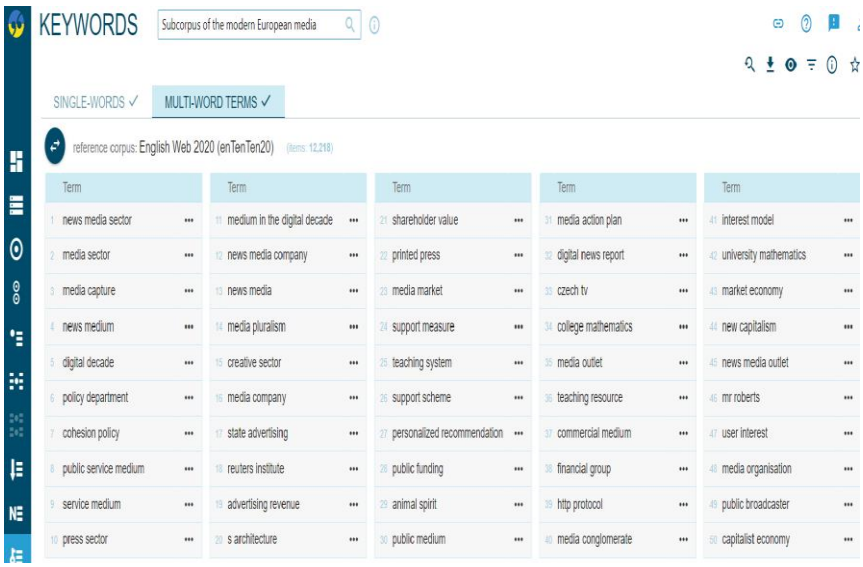

Another section devoted to EU matters is the BRITISH NEWS, containing materials in the EU scope, e.g. "Brexit: All you need to know about the UK leaving the EU" article, published in January 2020, provided an overview of the UK's departure from the European Union.

It was widely cited throughout the year as the Brexit process unfolded.

**Figure 3 BREXIT in the modern European media**

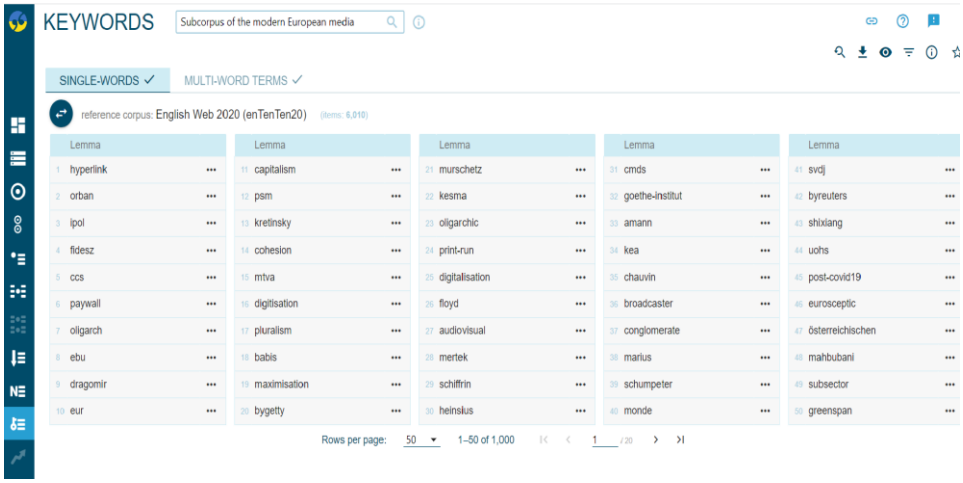

Among other popular tabloids that we have included in our subcorpus is The Sun, a British tabloid newspaper known for its sensationalist headlines and coverage of celebrity gossip, crime, and politics. It is the highest-circulating daily newspaper in the UK and has a large online following. Also, we have included the Daily Mail, which is a British tabloid newspaper known for its conservative editorial stance. The *Guardian* is a British daily newspaper known for its liberal and progressive editorial stance. Its online edition has a large following in the UK and beyond, particularly among younger and more politically engaged readers. *The Times* was also included for it is a popular British daily newspaper known for its quality journalism and coverage of politics, business, and culture. It has a large online following and is regarded as one of the most influential newspapers in the UK. The economics-oriented issues were added to the subcorpus (Figure 4).

**Figure 4 Corpus contains economic issues**



The special layer of the subcorpus are articles to have examined the global economy covered by the Economist. The articles added to the corpus (e.g. “The Future of Capitalism: Rent Collectors” published in September 2020) are widely cited and sold out. The next layer of the corpus is devoted to environmental issues, e.g. “Why the world is running out of sand” (This article, published in May 2017), reported on the growing demand for sand and the environmental impact of sand mining. It was widely cited as an example of the Economist's ability to tackle complex issues in an accessible and engaging way. The environment-oriented issues were added to the subcorpus (Figure 5).

**Figure 5 Environmental issues in corpora**




Among other Economist papers are worth mentioning “The Rise of the rich world’s new aristocracy”. This article, published in September 2020, examined the growing concentration of wealth and power among a small group of super-rich individuals. It was widely cited as an important commentary on income inequality and the changing nature of class in the 21st century. The article “A world without work”, published in May 2017, examined the impact of automation on the global workforce. It was widely cited as an important contribution to the debate on the future of work.

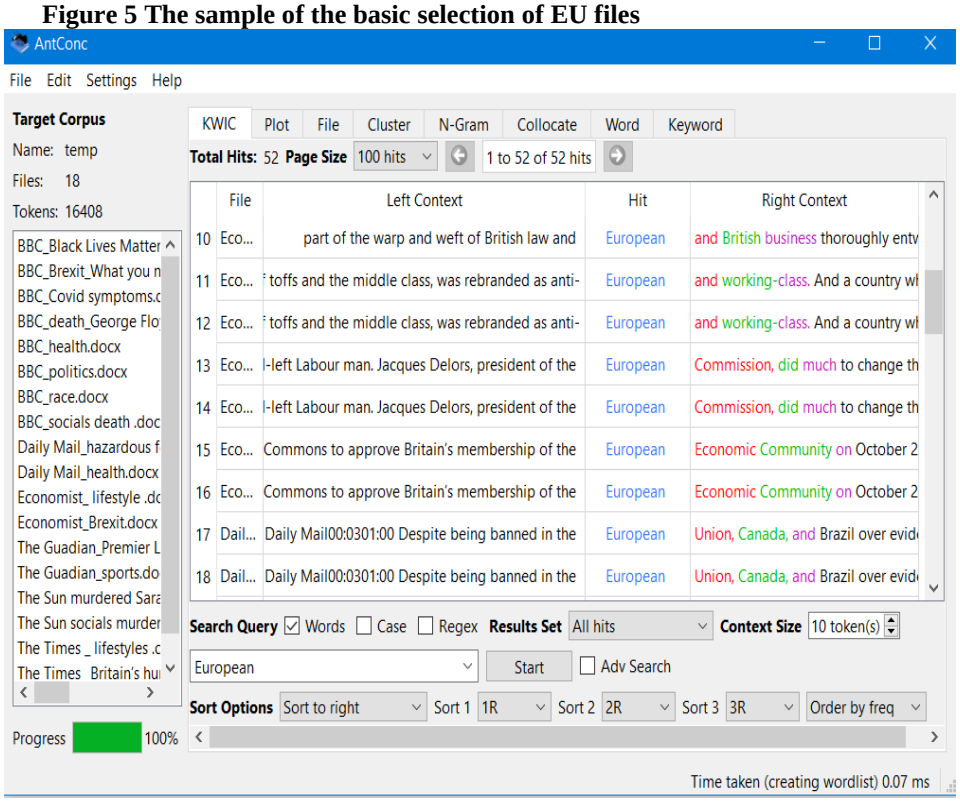
The Economist article “Why is America’s economy so resilient?”, published in October 2019, examined the reasons behind the US economy’s ability to weather economic shocks and downturns. It was widely cited as an important analysis of the strengths and weaknesses of the American economy. When we deal with economic issues, the information on changes and tendencies added to the corpus data sets (Figure 4).

***Figure 4 Economics issues in corpora***


**Tools for the subcorpus compilation**

In this study we used two popular tools *AntConc* and *Sketch Engine* to compile and analyze our corpus. AntConc was used to verify our text selection based on the broad context search. Then the texts were sorted, and downloaded to the subcorpus area of the *Sketch Engine. AntConc* as a free and open-source software program was used for corpus analysis. We used to create a corpus collection of the importing text files with EU scope by the *Sketch Engine* tool.

**Figure 5 The sample of the basic selection of EU files**



Sketch Engine is a web-based tool for creating and analyzing corpora. It provided access to a wide range of pre-built corpora, as well as tools for compiling your corpus by importing text files. We have been accessing single words and multiword terms after downloading and program processing (Figure 6).

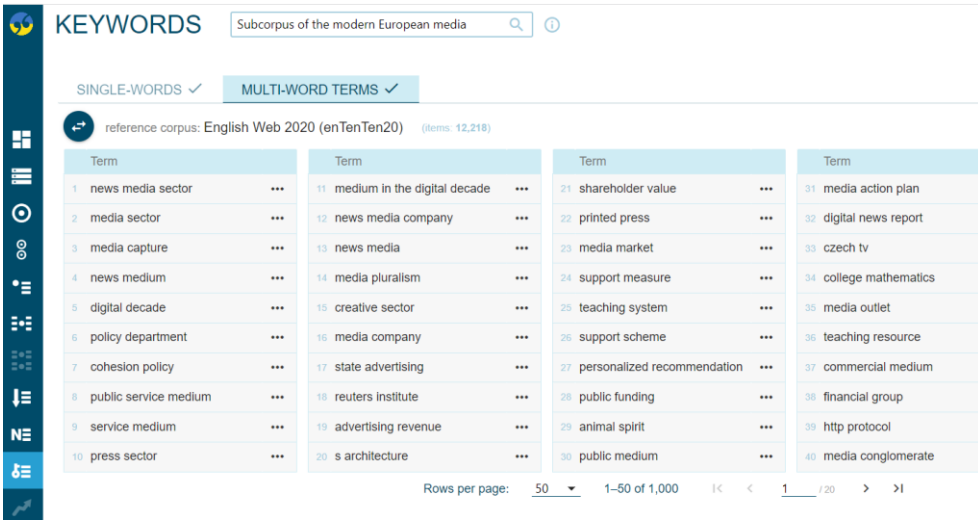
*Sketch Engine* data storage is prepaid thus it takes place on the server. This tool has many built-in tools, for example, keyword search, corpus markup, part-language analysis, accumulation and addition of texts.

As we can follow the compilation of the corpus in the *AntConc* manager collecting various files from various sources manually and then we have something to compare when the texts from web are added tour corpus automatically.

Deliberately we have selected the library of media files verified in the AntConc (containing Europe in the semantic core) and then adding some verified soucer to the library by the Sketch Engine tool by one click we have our corpus ready to use.

Our selection comprised BBC, The Daily Mail, the Economist, the Guardian, the Sun files all of them containing the semantic orientation of the European affairs, e.g.: the risky food additives banned in Europe (the Daily Mail), health issues, murder and crime (the Sun), sport affairs (the Guardian), etc.

**Figure 6 After processing single words and multi terms of the modern European media corpus**




**Conclusions**

This article discusses the compilation and analysis of a newspaper subcorpus within the larger multilingual corpus of the European Union media. The study describes the methods used to compile the corpus, including the collection of raw data and the selection of files from popular European newspapers known for their investigative journalism and critical analysis. The tools employed for corpus compilation and analysis include the web-based tool *Sketch Engine* and the offline corpus manager *AntConc*.

The newspaper subcorpus is intended for use in teaching and studying European studies, providing insights into political, economic, and social issues in the EU. The corpus contains documents manually added and automatically gathered from the internet in multiple languages. It can be further enhanced with additional features such as translation alignment. The subcorpus is processed using *Sketch Engine*, which adds layers such as part-of-speech tags and named entities, enabling advanced searches.

The selected newspapers for the subcorpus include respected sources like *BBC, The Guardian, The Times, and The Economist*, covering various topics such as *health* issues, *social* issues (e. g. gender, race), *economics* (e.g. *Brexit)*, and the *environment*. The articles from these sources are widely cited and provide valuable insights into European media discourse. The tools used, *AntConc* and *Sketch Engine*, facilitate the compilation, analysis, and search capabilities of the corpus.

In conclusion, the creation of the newspaper subcorpus within the larger EU media corpus using the described methods and tools offers a valuable resource for studying and understanding the language use, discourse patterns, and cultural trends of modern European media.

**Acknowledgements**

Co-funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

**REFERENCES**

1.Vasko, R., Korolyova, A., Hryshchuk, Y., & Kapranov, Y. (2021, September). Transfer of Mathematical Formulas and Computer Algorithms into Macrocomparative Studies. In 2021 11th International Conference on Advanced Computer Information Technologies (ACIT) IEEE, 2021. p. 642-647.

2.Liashko, O., Bober, N., Kapranov, Y., Cherkhava, O., & Meleshkevych, L. (2022). Interpretation of Keywords as Indicators of Intertextuality in English New Testament Texts (Antconc Corpus Manager Toolkit). *WISDOM*, *22*(2), 193-207.

3.Zhukovska V. English detached adjectival constructions with an explicit subject: A quantitative corpus-based analysis. Journal of Linguistics (Jazykovedný časopis), ROČNÍK 72 (2), 2021. P. 465–477.

4.Zhukovska V. Quantitative Corpus-Driven Approach to Disambiguation of Synonymous Grammatical Constructions. Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). Volume I: Main Conference, Lviv, Ukraine, April 23-24, 2020. CEUR Workshop Proceedings 2604, CEUR-WS.org 2020. P. 507–522.

5.Zhukovska V.V., Mosiyuk O. O. Statistical software R in corpus-driven research and machine learning. Information Technologies and Learning Tools. 2021. Vol. 86, № 6. P. 1–18.

6.Daily Mail. – The mode of access: https://www.dailymail.co.uk/health/article-11777037/The-risky-food-additives-banned-Europe-legal-US.html– Accessed: 25.05.2023.

7.The Sun. – The mode of access: https://www.thesun.co.uk/news/22433895/onlyfans-met-cop-guilty-misconduct-wayne-couzens/– Accessed: 25.05.2023.

8.The Guardian. – The mode of access: https://www.theguardian.com/football/european-super-league– Accessed: 25.05.2023.

9.The Economist. – The mode of access: https://www.economist.com/europe/2008/08/14/unacceptable-prejudice– Accessed: 25.05.2023.

10.The BBC. – The mode of access: https://www.bbc.com/news/world-europe-65681806– Accessed: 25.05.2023.

11.SketchEngine. – The mode of access: https://www.sketchengine.eu/– Accessed: 25.05.2023.

12.AnConc. – The mode of access: https://www.laurenceanthony.net/software/antconc/ – Accessed: 25.05.2023.

*Received: 27 May, 2023*