Digital Glossary of War in Ukraine (LLM-assisted Extraction to Lexonomy Interface Pipeline) *

Rusudan Makhachashvili[1,*,†], Anastasiia Marchenko[1,†] and Nataliia Bober[1,†]

[1] *Borys Grinchenko Kyiv Metropolitan University, Bulvarno-Kudriavska-st. 18/2 Kyiv, Ukraine*

Abstract

This study presents the interim results of the *Digital Glossary of War in Ukraine* project, focusing on the design and implementation of an end-to-end lexicographic pipeline that integrates LLM-assisted term extraction, digital processing tools, and deployment through the Lexonomy platform. The study demonstrates how large language models can support neoterminology disambiguation and extraction from heterogeneous digital corpora related to contemporary warfare discourse. The proposed workflow combines AI-enhanced automated extraction, expert-driven curation, and platform-based publication, ensuring both scalability and lexicographic reliability.

Keywords

digital terminography, LLM-assisted term extraction, digital glossary

1. Introduction

Global armed conflicts continue to shape multiple spheres of the human endeavour, including politics, economy, culture, and, of course, language, which is an integral part of the latter. The most dynamic structure within the system, the lexicon, is the most heavily and notably influenced by the societal changes that reflect directly in speech. Considering the recent wars, the vocabulary of the English language has been enriched with loan words, jargon, and slang, most actively coined across news outlets and on social media – in particular, on such platforms as X (formerly Twitter), Facebook, Instagram, YouTube, TikTok, and whatnot. All of these can be defined under the same umbrella term of Cyberspace – a complex, multidimensional sphere of synthesis of reality, human experience and activity mediated by the digital and information technologies, a component of the technosphere of human existence [1, 2, 3, 4].

Along with the military parlance that has been used for centuries, the English lexicon is now being actively enriched with neologisms, often of a foreign origin. One of the most striking examples is the terms that have been adopted as a result of the Russian war in Ukraine, which has been ongoing since 2014 and escalated into a full-scale invasion of Ukraine on February 24, 2022. The Glossary of War is an overarching digital lexicographic project implemented primarily via the Lexonomy interface, designed to document and semantically organize emergent terminology, neology, and cultural references arising from the full-scale Russian invasion of Ukraine since 2022 onwards. The project foregrounds semiotic layering, pragmatic functions, and modality through annotated metadata, marking its relevance in sociopolitical linguistics and digital humanities research.

Media outlets, both printed and web-based, are beaming with new military slang, transliterated into English directly from Ukrainian and, consequently, adopted not only by the Ukrainians themselves but by the non-indifferent people abroad. The examples below (Figure 1 and Figure 2) demonstrate a use case of the term “Bavovna” by Espresso.tv, a Ukrainian media outlet focused on

* *5th International Conference on “Multilingual digital terminology today. Design, representation formats and management systems” (MDTT) 2026, June 25-26, 2026, Zadar, Croatia.*

[1,*] Corresponding author.

r.makhachashvili@kubg.edu.ua (R. Makhachashvili); avmarchenko.frgf24m@kubg.edu.ua (A. Marchenko); n.bober@kubg.edu.ua (N.Bober)

0000-0002-4806-6434 (R. Makhachashvili); 0000-0002-9639-0562 (N.Bober)

news from Russia and Ukraine's war in the region, and the following usage of the term by an X account Wium Lacock, based in Newcastle, South Africa.

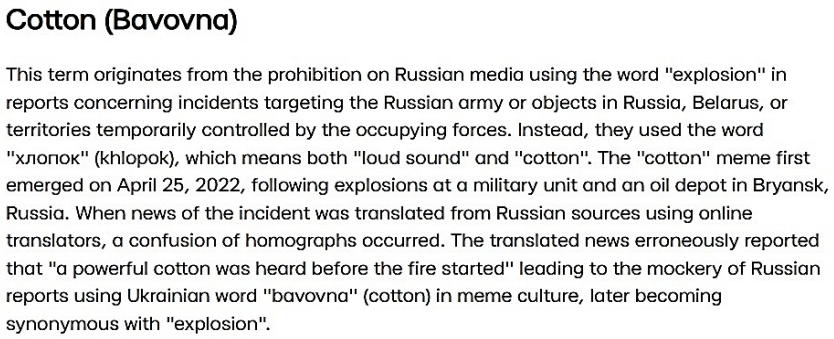

Figure 1: "Bavovna" in a news article.

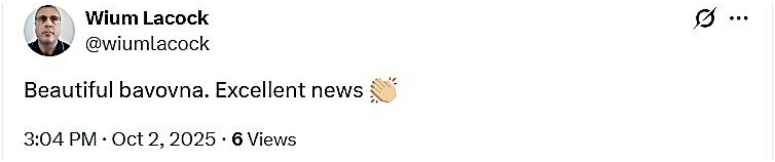

Figure 2: "Bavovna" used in an X post.

Military-specific vocabulary that has emerged in the English language as a direct consequence of the new military technology is not limited to slang alone. For example, after an experimental launch of the newest Russian ICBM (Intercontinental Ballistic Missile), infamously known as "Oreshnik", at the city of Dnipro, Ukraine, on November 21, 2024, the abbreviation "ICBM" has been adopted for active use both on news platforms and across social media (Figure 3).

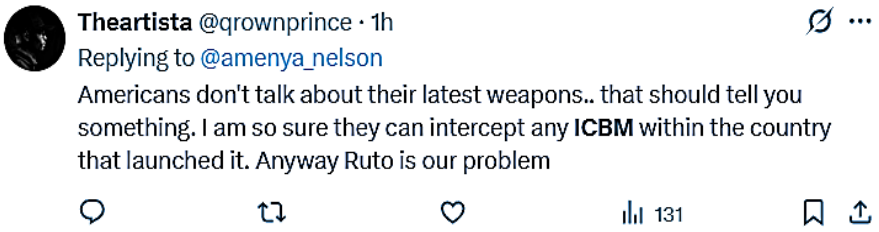

Figure 3: "ICBM" in social media.

Some of these terms are purely informal and humorous, while others are technical and heavily context-dependent. Either way, both types constitute inseparable parts of the contemporary military discourse in cyberspace, which is an integral environment, demanding new cognition and perception ways via complex philosophic, cultural, social, linguistic approaches, providing unlimited opportunities for human intellect, language development, and research [5, 6, 7, 8].

The war-related layer of the English lexicon, in turn, will continue to expand, develop, and progress for as long as armed conflicts worldwide keep unfolding and for generations afterwards. For linguists and, in particular, lexicographers, this causes a prominent and long-standing issue, which is the absence of a clear, structured, flexible, and unified glossary of military terminology compiled with both long-adopted terms and the newest lexical units, which continue to appear every day.

Therefore, the overall objective of the study is to evaluate an end-to-end pipeline for compiling a digital glossary of war in Ukraine, integrating LLM-assisted term extraction, digital processing tools, and deployment through the Lexonomy digital lexicographic platform.

2. Study design and methodology

Properly reflecting the aim of this research requires a structured, carefully curated methodology, as well as a set of lexicographic tools that would be flexible and expandable, and provide open access to anyone wishing to look up an entry or contribute to the glossary's compilation. Therefore, the main tool used at the core of the practical project of the Multimodal Glossary of War is Lexonomy (https://lexonomy.eu/), a cloud-based dictionary-writing and online-dictionary-publishing system developed by the ELEXIS Horizon 2020 project (https://project.elex.is/tools-and-services/). Lexonomy is suitable for editing and online publishing of domain-specific glossaries or terminology resources. Developed in the European Union, it is one of the best contemporary tools for compiling general and domain-specific dictionaries of any size. In this research, Lexonomy is the core platform for building the final version of The Multimodal Glossary of War.

At the same time, the compilation process encompasses a larger arsenal of tools and web-based platforms, such as a pre-collected and compiled corpus of war-related news pieces and blog articles from sources such as The Economist, Reuters, The Kyiv Independent, Espresso.tv, The New York Times, Ukrinform, RBC News, Defence Express, Diplomatic Courier, and others, as well as AI tools like OpenAI (ChatGPT), Gemini, and Perplexity AI for extracting terms from the corpus and providing their definitions.

The use of Large Language Models (LLMs) is commonly linked to the contemporary military discourse. Social media users and other interested parties often refer to tools like ChatGPT, Gemini, or Grok (X's native language model) for information and clarification on the topic of the ongoing armed conflicts, including the Russo-Ukrainian war and the war in the Middle East. However, the scope and precision of the responses given require a considerable amount of correction via manual intervention, both accuracy- and clarity-wise. The reason this research employs AI in the course of the glossary compilation can be broken down into two major aspects: first, to test how well the LLMs already respond to the requests related to war and whether they use the correct terminology in every given case, and second, to further utilize the ready-made glossary in the process of AI machine teaching. This teaching process occurs through manually feeding the vague and newly coined lexicon to the LLMs in order to improve the quality of their responses. The image below provides an illustrative example of an AI model, Perplexity AI, being unable to discern the meaning of a war-related slang term of Ukrainian origin (Figure 4).

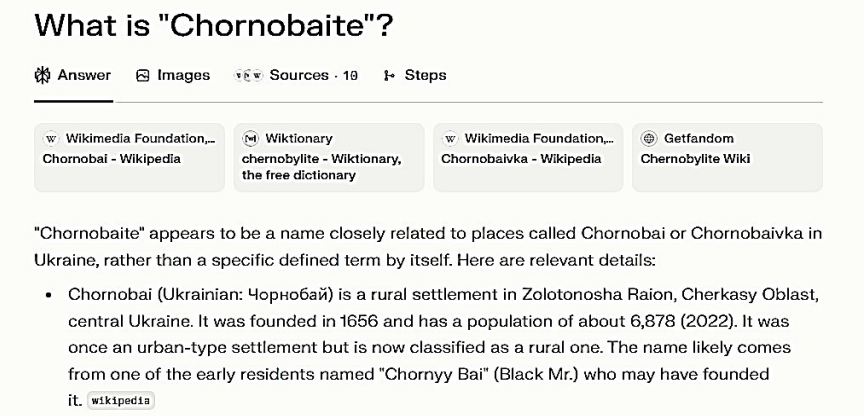

Figure 4: "Chornobaite" interpreted by Perplexity AI.

Evidently, there is still a long road ahead before Large Language Models learn to correctly discern the meaning of all war-related lexicon, and the Multimodal Glossary of War, as an ongoing electronic

lexicographic project [3, 1], aims to contribute to the AI training process. In this way, the final product of the research can become an indispensable consulting tool for humans and machines alike. In addition, the open nature of Lexonomy allows anyone to contribute to the expansion of the glossary with new entries. As the final result, we are hoping to create a dynamic, open-access online glossary of the military lexicon, which can be consulted for the purposes of teaching, learning, writing, and self-education.

Once a list of terms is extracted from the corpus, with a number of manually added words found on social media, the AI tools are employed to provide clear and concise definitions of each. ChatGPT, Gemini, and Perplexity AI each participate in the creation of every entry: the definitions are then compared both between themselves and, when possible, with the entries from a conventional dictionary. Word definitions, along with usage examples, translation, and source, are then added to the Glossary using a pre-configured markup (Figure 5):

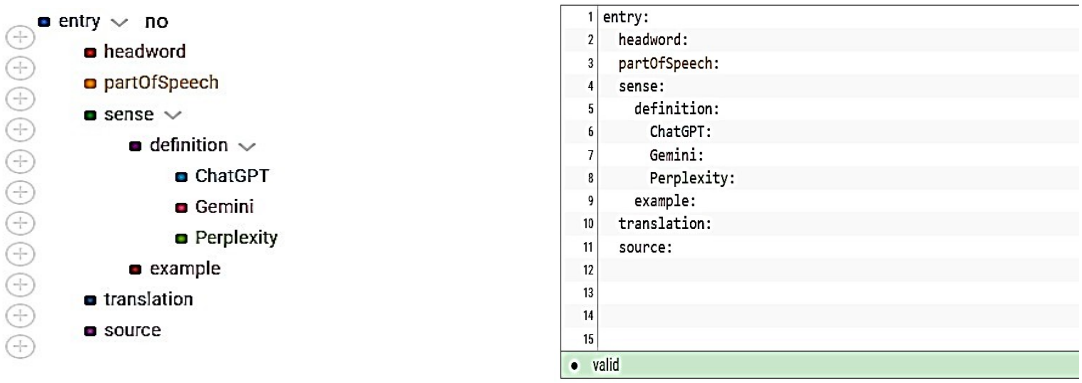

Figure 5: Lexonomy Markup for the Multimodal Glossary of War.

This consistent markup makes it possible for anyone interested to supply the glossary with new entries at their discretion. A universal entry template is ideal for a project in the sector of digital lexicography, as it makes the glossary simple and intuitive in use, even for non-language specialists. As such, the purpose of making the Multimodal Glossary of War universally accessible and editable to anyone in the cyberspace is achieved in full.

3. Findings

At the beginning of the research, a survey on the Military-Themed Lexicon Use Online was conducted among sixteen participants aged from 18 to 40. The chosen respondents were professionals in the fields of linguistics and translation, IT and cybersecurity, fine arts, and social studies. The survey set the stage for further exploration of the topic and helped develop the roadmap for building a thematic military-centred glossary for bilingual speakers of English and Ukrainian.

The responses have demonstrated that more than half of the participants (56.3%) prefer to receive relevant information on the ongoing armed conflicts from multiple sources at once (Figure 6), including social media, online news outlets, and even public news channels on Telegram, especially relevant for the coverage of the War in Ukraine (see chart below). Such a breakdown allowed us to allocate the research time and resources to retrieving related vocabulary equally from all the aforementioned sources. Public media outlets provided formal, domain-specific terminology that has emerged in the English language as a result of the new inventions and modifications in the global military sector (terms like ICBM, Shahed, etc). Social media and Telegram public channels, on the other hand, contributed to the search for war-related slang and neologisms, predominantly used by the younger generation (words like bavovna, Chornobaite, and so on).

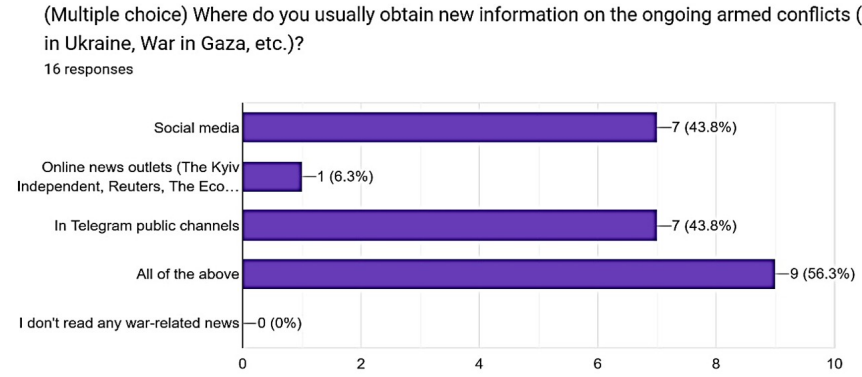

Figure 6: Chart on the Main Sources of Information about the Ongoing Armed Conflicts in the World.

The majority of respondents (56.3%) have assessed the importance of understanding war-related lexicon by ordinary people, not affiliated with armed conflicts directly (as in, non-journalists, writers, and war researchers) as high (4 or 5 on the Likert scale) (Figure 7).

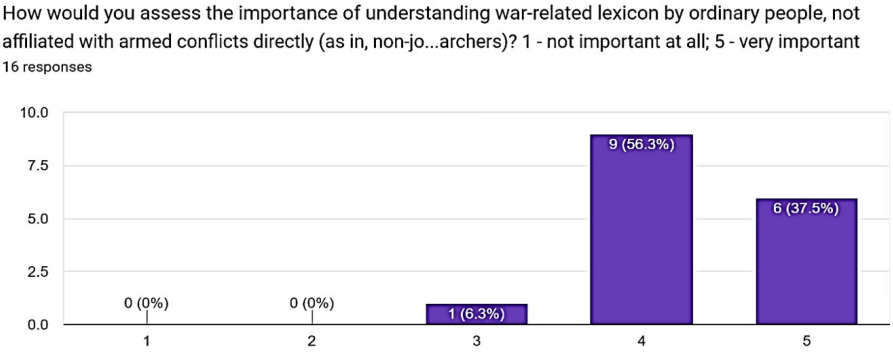

Figure 7: The Importance of Understanding War-Related Lexicon by Those Not Directly Affiliated with Armed Conflicts.

Additionally, half of the respondents (50%) claimed that knowing and using war-related slang in the military discourse is important for people both in Ukraine and abroad. This statistic further highlighted the need for contemporary, digitalized, and universally accessible lexicographic tools, such as the Multimodal Glossary of War.

Another important finding in the survey process was that only 31.3% of the respondents found textbooks effective in teaching students war-related lexicon in English. Instead, the majority (87.5%) preferred military blogs and war-themed documentaries as the main source of the contemporary military lexicon. Web-based blogs and videoblogs, including social media profiles and channels of real-life soldiers and war reporters, are among the sources used in the compiling process for the Multimodal Glossary of War.

A high percentage of the respondents (68.8%) also identified web-based glossaries that collect and summarize all relevant terms as an effective method of vocabulary teaching. In fact, the Multimodal Glossary of War extracts and connects relevant lexical elements from all the relevant sources at once, which eliminates the need to use and juggle multiple tools at once. In addition, textbooks quickly get outdated, as the military parlance transforms on a daily basis, and new slang terms appear faster than linguists can register them and timely update all the corresponding educational and lexicographic sources.

Henceforth, the web-based nature of this project is superior in usability, flexibility, and scale to any other teaching and learning method. Instead of surfing multiple sources at once in the search for the newly emerged war terms, users will be able to consult the Multimodal Glossary of War at any time, getting all the relevant information about a specific entry under one umbrella: definition in three interpretations, an example of use in the context, and a Ukrainian translation.

Finally, 75% of the survey respondents claimed that they would use an online tool, such as the Multimodal Glossary of War, for educational and other purposes (Figure 8). The project, then, is devised in accordance with the reflected demand, prioritizing relevant, frequently used terms that can be used for text interpretation, supporting war-related conversations, expressing public opinions, etc.

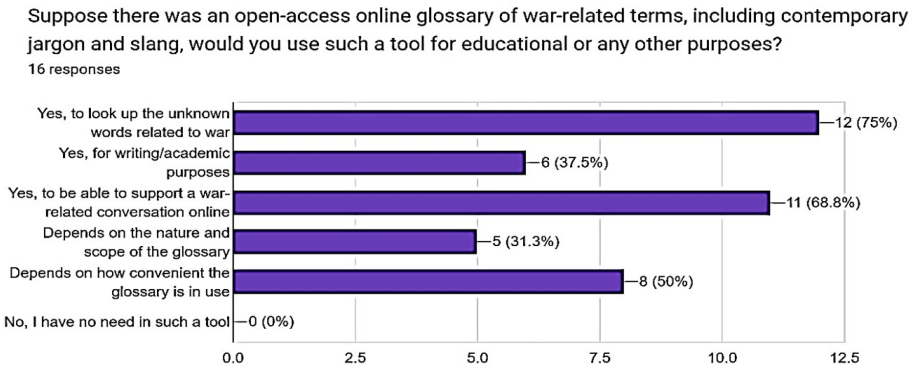

Figure 8: The Respondents' Demand for a Multimodal Glossary of War.

At the moment, the final version of the project is in a stage of active development and amendment: terms are being extracted and processed by AI tools, and definitions are being properly organized and added as separate entries, along with examples of use, translations, and sources.

The process of candidate terms selection for the Multimodal Glossary of War follows several key principles that ensure the collected material is relevant, reliable, and representative of how war is discussed in different contexts today:

1. Relevance to the Field of War Studies. Each term must have a clear connection to the topic of war. This includes words that describe military actions, strategies, technologies, and their effects, as well as terms that reflect the political, social, and emotional sides of conflict. This principle helps to show war as a complex phenomenon that extends beyond the battlefield.
2. Presence Across Different Modes of Communication. Since the glossary is multimodal, priority is given to terms that appear in various forms of media – written texts, images, and digital platforms. For example, a term may be found in a news headline, a social media post, or a political cartoon. This shows how war language functions across different types of communication.
3. Balance Between Historical and Modern Terms. The glossary includes both older, well-established terms (such as armistice or trench warfare) and newer expressions that reflect recent conflicts (like cyberwarfare or information front). This combination allows for observing how the language of war evolves over time and adapts to technological and ideological changes.
4. Frequency and Visibility in Public Discourse. Preference is given to terms that appear often in public communication – news, speeches, online discussions, and media reports. The more visible a term is, the more it influences how people understand and talk about war.
5. Emotional and Ideological Value. Many war-related terms carry emotional or ideological meanings. For instance, some words may serve propaganda purposes or express sympathy

or hostility. Including such terms helps to explore how language can be used to influence opinions and shape collective attitudes toward conflict.

6. Cross-Cultural and International Use. Some terms are used across different countries and languages, especially in the global media space. Including these internationally recognized expressions helps to show how war discourse circulates globally and how certain concepts become shared across cultures.
7. Variation in Meaning and Use. The same term can be used in different ways – literally, metaphorically, or even ironically. The glossary considers this variation to show how language about war changes depending on context, intention, and audience.
8. Authenticity and Reliability of Sources. All terms are taken from real, verifiable sources, including online news, official documents, and public social media posts. Using authentic materials ensures that the glossary reflects how people actually use war-related vocabulary in real communication.

Together, these principles ensure that the Multimodal Glossary of War captures the complexity of contemporary war discourse. By combining linguistic, cultural, and multimodal perspectives, the glossary not only documents relevant vocabulary but also highlights the ways in which language and imagery shape public understanding of conflict. This systematic approach allows the collected terms to serve as both linguistic data and cultural evidence, forming a solid foundation for further analysis of how war is represented and interpreted across different media contexts.

Any necessary corrections to the structure of the glossary and separate entries are being added during the periodic reviews of the project's technical and semantic integrity. The periodic reviews are conducted on a weekly basis or, alternatively, after each full batch of new entries added to the glossary. The open nature of Lexonomy allows for quick and effortless error correction at any point of the project's development.

Furthermore, the source code template was applied to train the AI model (based on few-shot and RAG approach) to automatically identify, extract, and tag linguistic innovations according to specifications in random and customized corpora. The same protocol was applied to automate the extraction of the ready-for-dictionary entries in NVH format. The suggested protocol significantly streamlines the digital neographic workflow and provides grounds for multi-faceted enhancement of AI-powered neology and neoterminology encoding.

4. Conclusions

As this project reflects the dynamic nature of the English war-related vocabulary, the glossary is intended to remain open and adaptable, continuously evolving along with changes in global discourse and technology.

Future research may focus on several directions. One possible area is the comparative study of war-related terminology across different languages to identify cross-cultural patterns and semantic shifts. Another promising direction involves exploring the interaction between linguistic and visual elements in multimodal representations of war, which could deepen the understanding of how conflict is framed and perceived in modern media. Additionally, expanding the glossary with corpus-based frequency data or discourse annotations could enhance its analytical value and make it a useful tool for interdisciplinary studies in linguistics, communication, and media analysis.

Overall, the Multimodal Glossary of War serves as both a linguistic resource and a methodological experiment, illustrating how digital tools can support the study of contemporary language in motion. Its open-ended design ensures that the project will continue to grow and remain relevant as new conflicts, technologies, and forms of expression emerge in the ever-changing landscape of war discourse.

Acknowledgements

Empirical findings and theoretical procedures have been conducted under the auspices of the projects COST Action CA21167 UniDive: Universality, Diversity and Idiosyncrasy in Language Technology, COST Action CA22126 ENEOLI: European Network on Lexical Innovation, COST Action CA23105 PLURILINGMEDIA: Language Plurality in Europe's Changing Media Sphere.

The authors extend a special acknowledgement to the Armed Forces of Ukraine for providing safety to complete this work.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI-GPT-5 and Grammarly in order to conduct grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. Makhachashvili, I. Semenist, V. Klochkov, AI-enhanced multilingual lexicography for digital communication, in: N. Callaos, N. Lace, B. Sánchez, M. Savoie (Eds.), Proceedings of the 16th International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2025), IIIS, USA 2025, pp. 247–253. URL: https://www.iiis.org/DOI2025/DR701VS/.

[2] N. Lazebna, English Language as Mediator of Human-Machine Communication, Royal Book Publishing, Mysore, India, 2021.

[3] R. Makhachashvili, Cyber-speak Dictionary (ELEXIS), Slovenian Language Resource Repository CLARIN.SI, 2020. URL: http://hdl.handle.net/11356/1610.

[4] R. Makhachashvili, Models and digital diagnostic tools for the innovative polylingual logosphere of computer being dynamics, in: S. del Gaudio (Ed.), Italian-Ukrainian Contrastive Studies: Linguistics, Literature, Translation, Peter Lang, Berlin, 2020, pp.99-124.

[5] R. Makhachashvili, I. Semenist, Linguistic Philosophy of Cyberspace, in N. C. Callaos, N. Lace, B. Sanchez, M. Savoie (Eds.), Proceedings of the 25th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2021), IIIS, USA, 2021, pp. 191–207.

[6] R. Makhachashvili, S. Kovpik, I. Semenist, A. Bakhtina, Hieroglyphic Semiotics of Emoji Signs in Digital Communication, in: V. Hamaniuk, S. Semerikov, S. Amelina, R. Makhachashvili (Eds.) Proceedings of the International Conference on New Trends in Languages, Literature and Social Communications (ICNTLLSC 2021), Volume 557 of Atlantis Highlights in Social Sciences, Education and Humanities, Atlantis Press, 2021, pp. 182–192. doi:10.2991/assehr.k.210525.023.

[7] R. Makhachashvili, N. Bober, AI-Enhanced Transdisciplinary Data Encoding for LLMs Training, in: N. Callaos, N. Lace, B. Sánchez, M. Savoie (Eds.), Proceedings of the 29th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2025), IIIS, USA, 2025, pp. 327–333. URL: https://www.iiis.org/DOI2025/SA747NT/#FullText.

[8] A. Ruch, E. Kirkland (Eds.), Posthumanity: Merger and Embodiment, Brill, Berlin, 2020. doi:10.1163/9781848880184.