ВІСНИК науки та освіти
ISSN 2786-6165 (ONLINE) №4(46) 2026
Bulletin of Science and Education

UDC 811.111'373.23:004.9

https://doi.org/10.52058/2786-6165-2026-4(46)-163-184

Krasniuk Svitlana Senior Lecturer, Department of Philology and Translation of Kyiv National University of Technologies and Design, Kyiv, https://orcid.org/0000-0002-5987-8681

Zaitseva Nataliia Candidate of Philology, Associate Professor, Department of English Philology of National University of Life and Environmental Sciences of Ukraine, https://orcid.org/0009-0008-4030-7185

Shcherbyna Svitlana Candidate of Philology, Associate Professor, Department of Finance and Marketing of Private Higher Educational Establishment «Institute of Ecology, Economics and Law», Department of Foreign Languages for Sciences of Yuriy Fedkovych Chernivtsi National University, https://orcid.org/0000-0001-7317-7921

BIG LINGUISTIC DATA AS A PARADIGM SHIFT IN MODERN PHILOLOGY

Abstract. The article conceptualizes Big Linguistic Data (BLD) as a fundamental factor driving a systemic paradigm shift in modern philology. It is argued that at the beginning of the 21st century, philological science entered a phase of large-scale transformation due to the changing ontological nature of linguistic material—from static, closed sign systems to dynamic, non-linear, and continuously updated data flows. The research explicates the methodological transition from traditional qualitative analysis and corpus-based validation to a data-driven paradigm. A core contribution of the study is the adaptation of the classical Big Data parameters—volume, velocity, variety, and veracity—to the linguistic dimension by introducing a crucial fifth element: semantic and pragmatic complexity.

The study highlights the integration of artificial intelligence, machine learning, and transformer architectures (such as GPT and BERT) into a broader Data Science ecosystem. This shift redefines the object of philology, moving from the analysis of some text as a static final product to the critical analysis of the models that generate and interpret this text. The authors examine the rise of "platform philology," where scientific success increasingly depends on access to complex digital infrastructures and super-powerful computing resources. This transformation is accompanied by acute structural inequalities, such as the digital

endangerment of low-resource languages (including Ukrainian) and the persistence of Anglocentrism in global linguistic technologies.

Furthermore, the article scrutinizes serious epistemological and ethical challenges, including the "black box" dilemma of algorithmic opacity, privacy risks, and the reinforcement of social biases through original training corpora. The authors emphasize the necessity of developing a hybrid methodology that synthesizes the colossal computing power of AI with the hermeneutic depth and ethical completeness of classical philology. Ultimately, the research posits that philology is experiencing a rebirth as a high-tech discipline within the Digital Humanities, capable of operating with global digital arrays while preserving the human dimension and linguistic diversity.

Keywords: Big Linguistic Data (BLD), Modern Philology, Paradigm Shift, Data-driven Linguistics, Digital Humanities, Artificial Intelligence, Platform Philology, Low-resource Languages, Algorithmic Bias, Digital Sovereignty, Linguistic Ecology

Краснюк Світлана Олександрівна старший викладач кафедри філології та перекладу Київського національного університету технологій та дизайну, м. Київ, https://orcid.org/0000-0002-5987-8681

Зайцева Наталія Петрівна кандидат філологічних наук, доцент, кафедра англійської філології Національного університету біоресурсів і природокористування України, https://orcid.org/0009-0008-4030-7185

Щербина Світлана Миколаївна кандидат філологічних наук, доцент, кафедра фінансів і маркетингу ПВНЗ “Інститут екології, економіки і права”, кафедра іноземних мов для природничих факультетів Чернівецького національного університету імені Юрія Федьковича, https://orcid.org/0000-0001-7317-7921

ВЕЛИКІ ЛІНГВІСТИЧНІ ДАНІ ЯК ЧИННИК ПАРАДИГМАЛЬНОЇ ТРАНСФОРМАЦІЇ СУЧАСНОЇ ФІЛОЛОГІЇ

Анотація. У статті концептуалізуються Великі Лінгвістичні Дані (ВЛД) як фундаментальний фактор, що зумовлює системну зміну парадигми в сучасній філології. Стверджується, що на початку 21 століття філологічна наука вступила у фазу масштабної трансформації через зміну онтологічної природи лінгвістичного матеріалу — від статичних, закритих знакових систем до динамічних, нелінійних та постійно оновлюваних потоків даних. Дослідження пояснює методологічний перехід від традиційного якісного

аналізу та корпусної валідації до парадигми, керованої даними. Основним внеском дослідження є адаптація класичних параметрів Великих даних — обсягу, швидкості, різноманітності та правдивості — до лінгвістичного виміру шляхом введення вирішального п'ятого елемента: семантичної та прагматичної складності.

У дослідженні підкреслюється інтеграція штучного інтелекту, машинного навчання та трансформаторних архітектур (таких як GPT та BERT) у ширшу екосистему науки про дані. Цей зсув переосмислює об'єкт філології, переходячи від аналізу тексту як статичного кінцевого продукту до критичного аналізу моделей, які генерують та інтерпретують ці тексти. Автор досліджує піднесення «платформної філології», де науковий успіх дедалі більше залежить від доступу до складних цифрових інфраструктур та надпотужних обчислювальних ресурсів. Ця трансформація супроводжується гострою структурною нерівністю, такою як цифрова загроза для мов з низькими ресурсами (включаючи українську) та збереження англоцентризму в глобальних лінгвістичних технологіях.

Крім того, у статті ретельно розглядаються серйозні епістемологічні та етичні виклики, включаючи дилему «чорної скриньки» алгоритмічної непрозорості, ризики для конфіденційності та посилення соціальних упереджень через оригінальні навчальні корпуси. Автор наголошує на необхідності розробки гібридної методології, яка синтезує колосальну обчислювальну потужність штучного інтелекту з герменевтичною глибиною та етичною повнотою класичної філології. Зрештою, дослідження постулює, що філологія переживає відродження як високотехнологічна дисципліна в рамках цифрової гуманітарної науки, здатна працювати з глобальними цифровими масивами, зберігаючи при цьому людський вимір та лінгвістичне різноманіття.

Ключові слова: Великі лінгвістичні дані (ВЛД), Сучасна філологія, Зміна парадигми, Лінгвістика, керована даними, Цифрова гуманітарна наука, Штучний інтелект, Платформна філологія, Мови з низьким рівнем ресурсів, Алгоритмічне упередження, Цифровий суверенітет, Лінгвістична екологія

INTRODUCTION

At the beginning of the 21st century, philological science entered a phase of the most large-scale and multi-vector transformation in the entire history of its development, which is due not only to rapid technological progress, but also to a fundamental change in the very ontological nature of linguistic material. Within the framework of the classical philological tradition, language has been considered for centuries mainly as a relatively stable and closed system of signs,

recorded in canonical written texts and subject to interpretation through the prism of the historical and cultural context. The researcher in this model invariably acted as a hermeneutic, whose efforts were aimed at extracting deep meanings from a limited set of sources. However, in the conditions of global digitalization, language finally loses the features of a static object of description, acquiring the characteristics of a dynamic, nonlinear and continuously updated data flow.

This information flow, generated by billions of users in real time, covers all spheres of human activity - from microdiscourses in social networks to complex multimodal practices in which verbal text is inextricably merged with visual images, audio and video sequences. Social networks, instant messengers, digital archives and global information systems form a qualitatively new linguistic reality in which the traditional boundaries between oral and written language, normative and deviant use, as well as individual and collective discourse become extremely blurred. In the modern situation, language appears as a high-speed process that requires fundamentally different analytical tools and innovative research strategies capable of covering the scale of Big Linguistic Data (BLD). A key role in the development of this new type of linguistic material is played by artificial intelligence and machine learning technologies, which are a fundamental catalyst for the restructuring of the methodological apparatus of linguistics. Natural language processing methods, neural network architectures and transformative models provide the opportunity to detect hidden regularities, complex semantic connections and discursive patterns that remain inaccessible to traditional philological analysis [1]. As a result of these processes, a fundamentally different research logic is being formed: a transition from the analysis of local, pre-selected text samples to the modeling of global language processes based on massive and heterogeneous data. This shift marks the emergence of a data-centric paradigm, within which the primary source of scientific knowledge is the data itself, and not a priori theoretical assumptions. Such a transformation inevitably leads to a change in the epistemological status of philology and gives rise to a number of fundamental theoretical problems. First of all, there is a clear tension between qualitative and quantitative knowledge: if the classical tradition strives for interpretative accuracy and contextual depth, then the data-driven approach is focused on statistical reliability, scalability and representativeness of the results. At the same time, there is a radical transformation of the status of the researcher himself, who ceases to be the only subject of interpretation and becomes the operator of the most complex analytical systems. This requires a deep rethinking of the philosophical foundations of philology as a science that has found itself in a situation of paradigmatic transition. The key task of modern linguistics in the conditions is not to abandon the centuries-old tradition, but to find ways to form an integrative approach that can effectively combine the wisdom and ethics of

classical humanitarian analysis with the colossal capabilities of computational methods and algorithms of artificial intelligence [2]. Thus, sustainable development of the discipline in the digital age is possible only through a synthesis of interpretive depth and technological scale, which will allow for the creation of a new structure of knowledge about language as a living, constantly evolving global system.

FORMULATION OF THE PROBLEM

Modern philology faces a fundamental challenge - the inconsistency of classical tools to the scale of global digital communication. Traditional hermeneutics, oriented towards microanalysis of canonical texts, turns out to be insufficiently effective in the face of nonlinear flows of Big Linguistic Data. The main contradiction lies in the need to integrate algorithmic AI methods into the humanitarian sphere without losing the depth of interpretation. The emerging gap between the statistical power of macroanalysis and qualitative philological reflection requires a revision of the scientific paradigm itself, as well as finding ways to overcome digital inequality and preserve linguistic diversity in the conditions of algorithmization of discourse.

ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

The theoretical foundation of Big Linguistic Data research is formed at the intersection of computational linguistics, digital humanities and critical media theory. The global shift from traditional “close reading” to macroanalysis was initiated by the works of Moretti F. [3], whose concept of distant reading allowed us to consider literature as a large-scale system subject to statistical modeling. In the development of this line, Manovich L. [4] proposed methods of “cultural analytics” to identify patterns in global visual and textual content. The empirical basis of the work is based on the classic works on corpus linguistics by Sinclair J. [5] and Biber D. et al. [6], which set the standards for the analysis of real language use. However, in the era of Big Data, classic corpora are transformed into dynamic systems, which is described in the works of Schellffleysh T. [7]. The methodological revolution caused by the introduction of the transformer architecture is recorded in the fundamental article by Vaswani A. et al. [8]. The impact of these technologies on philological practice is analyzed in the works of Andrenko, K. V. [9] and Derevianko Iu. [10], where the emergent properties of large language models (LLM) are considered.

Issues of digital hermeneutics and computer interpretation of texts are deeply developed in the works of Rockwell G. & Sinclair S. in [11]. The features of the functioning of language in digital networks and the phenomenon of “network linguistics” are investigated in the works of Crystal D. [12].

A critical reflection on the development of AI is presented in the works of Bender E. M. et al. [13], who introduced the metaphor of "stochastic parrots". The problems of algorithmic bias and "digital colonialism" are raised in the works of Crawford K. [14], Noble S. U. [15], as well as in the concept of "surveillance capitalism" by Zuboff Sh. [16].

The issue of low-resource languages and ways to overcome the digital divide are analyzed in collective research under the auspices of ACL (Joshi P. et al.) [17] and the works of Magidson M. [18].

The issues of preserving cultural heritage in the digital age are considered in the works of Van de Velde, E. [19].

The connection of language, power and technology in conflict situations is reflected in the studies of Pomerantsev P. [20] and works on digital diplomacy by Bjola C. & Zaiotti R. [21]. Discursive changes in war conditions are analyzed through the prism of the dynamics of network communities in the publications of Castells M. [22].

The analytical review is completed by sources devoted to the transformation of the academic environment and writing under the influence of generative AI (Lopez J. [23], Wheeler S. [24]), issues of digital sovereignty (Couldry N. & Mejias U. A. [25]), and the epistemology of digital humanities (Drucker J. [26], Hayles N.K. [27]).

PURPOSE OF THE ARTICLE

The aim of the work is to conceptually substantiate big linguistic data as a factor of paradigm shift in philology. The research is aimed at explicating methodological changes caused by the introduction of data-driven approaches, and at analyzing the ethical and linguocultural challenges of the digital age. The author seeks to prove the need to synthesize computing power with classical philological reflection to form a new episteme of knowledge about language, capable of ensuring the sustainable development of cultures in conditions of total algorithmization.

RESULTS

1. The concept of Big Linguistic Data (BLD) is not just a quantitative extension of the classical paradigm of corpus linguistics, but a qualitative leap in understanding and processing language resources in the digital age. This concept appears as a conceptual extension of the Big Data methodology to the field of philological and linguistic research, where the content of the definition is not reduced solely to increasing the volume of processed text arrays. In modern scientific discourse, Big Linguistic Data is interpreted as a complex, multi-layered digital ecosystem that captures language in all its syncretic diversity. This system

combines not only written texts of various genres, functional styles and registers, but also speech arrays, multimodal forms of communication, including video, audio and visual-verbal combinations, as well as colossal volumes of accompanying metadata and behavioral characteristics of communication subjects.

Within this approach, language ceases to be perceived as an abstract system of signs or a static code; it is fixed as a living social practice, inextricably embedded in and conditioned by digital environments of interaction.

Adapting the classical parameters of Big Data [28] - volume, velocity, variety and veracity - to the linguistic dimension requires the inclusion of a fundamentally important fifth element - semantic and pragmatic complexity (Table 1).

Linguistic data in the context of BLD are not neutral "raw materials" in the traditional sense of the term: they are deeply filled with cultural codes, ideological guidelines, communicative intentions and cognitive structures. Such specificity dictates the need to develop new methods of interpretation that can effectively take into account the contextual and discursive ambiguity of data. The most important characteristic of Big Linguistic Data is their dynamism and continuity. Unlike traditional static linguistic corpora that function as archives, BLD is a continuous flow that is constantly updated and redistributed in the global digital space.

This necessitates a methodological transition from the analysis of discrete “slices” of language to the study of language evolution and communicative processes as a real clock.

The scientific novelty and theoretical significance of the BLD concept lies in the fundamental rethinking of the epistemological status of linguistic data. They are considered not only as an empirical basis for verifying existing linguistic hypotheses, but as an independent object and an active source of knowledge formation. In this context, a new epistemological model is being formed in which scientific knowledge is derived directly from data through algorithmic detection of hidden patterns, and not only from a priori theoretical constructs. Thus, Big Linguistic Data become a basic element of the data-centric paradigm of language science, radically transforming both the tools and the teleology of philological research, opening new horizons for interdisciplinary research at the junction of linguistics, cognitive science and computer science. The transition to working with BLD means recognizing language as a dynamic data stream, where the complexity of structure and context requires a symbiosis of traditional hermeneutic analysis and powerful computational methods. This transformation affirms the status of philology as a high-tech discipline capable of operating with meanings on the scale of global digital arrays.

Table 1.

Conceptual Parameters of Big Linguistic Data vs. Classical Big Data

Parameter	**Classical Big Data Interpretation**	**Linguistic Specificity of BLD**
Volume	Terabytes/petabytes of raw data	Vast arrays of text, speech, video, audio, and metadata
Velocity	High-speed data processing	Continuous flow of linguistic evolution in real-time
Variety	Multiple data formats	Multimodality: a syncretism of genres, styles, and visual-verbal codes
Veracity	Data accuracy and "cleanliness"	Representativeness of living social practices and communication
Complexity	—	Key element: Deep saturation with cultural codes, intentions, and contextual ambiguity

Source: authors' results of systematization and analysis

2. The comparison of traditional, corpus-based and data-driven linguistics reveals a deep evolution of the logic of scientific knowledge. Each strategy in its own way defines the relationship between the scientist and the text, transforming the mechanisms of knowledge verification.

Traditional linguistics is based on qualitative analysis, hermeneutics and research intuition. Its advantage is in the depth of interpretation and subtle accounting of the pragmatic context. However, the method is limited by the cognitive resource of a person: a scientist analyzes dozens of texts in detail, but is not able to detect statistical patterns in arrays of millions of documents.

Corpus linguistics has become a stage of empirical validation of humanitarian knowledge. It introduced quantitative methods and structured corpora, which made it possible to find frequency patterns and collocations. However, here the data remain only a testing ground for testing pre-formulated hypotheses, and the researcher retains full control over the process of increasing knowledge.

Data-driven linguistics radicalizes the empirical approach, relying on machine learning algorithms and AI. They are able to automatically extract hidden patterns from unstructured data chaos without prior theoretical frameworks. There is a transition to “algorithmic abduction”, when explanatory models are formed taking into account the results of the machine.

This shift gives rise to serious epistemological challenges. The researcher loses the status of the sole interpreter, becoming the operator of complex systems. There is a risk of a “black box”: the results are accurate, but their logic is difficult

to explain in traditional categories. The solution is to create hybrid methodologies that combine the computational power of algorithms with the hermeneutic depth of classical philology. Such a synthesis allows us to overcome the limitations of individual approaches, providing a new level of understanding of language as a global digital ecosystem (Table 2).

Table 2.

Evolution of Linguistic Paradigms

Criterion	**Traditional Linguistics**	**Corpus Linguistics**	**Data-driven Linguistics**
Methodology	Hermeneutics, intuition, qualitative analysis	Quantitative methods, frequency patterns, collocations	Machine learning algorithms, AI, neural networks
Object of Analysis	Individual texts (dozens/hundreds)	Structured corpora (millions of words)	Unstructured "chaos" of global data arrays
Researcher's Role	Sole interpreter and meaning-maker	Controller of hypothesis validation	Operator of complex algorithmic systems
Type of Knowledge	Qualitative depth and pragmatic context	Empirical validation and statistical patterns	Algorithmic abduction (hidden pattern discovery)

Source: authors' research results

3. The modern methodology of linguistic research is undergoing fundamental transformations under the powerful influence of integration with computer science and the rapid development of artificial intelligence technologies in the context of Big Data. Machine learning methods, including deep neural networks, transformative architectures (such as GPT or BERT), probabilistic models and tools for complex statistical analysis, are now being actively implemented in the current research practice of philology and linguistics. As a result of these processes, classical language science is gradually incorporated into a broader and more dynamic Data Science ecosystem, where natural language begins to be considered as a specific type of data subject to high-precision computational processing, digital modeling and quantitative analysis.

This methodological transformation is accompanied not only by the expansion of the applied instrumental apparatus, but also by a radical change in the very epistemological foundations of humanitarian knowledge. If in the traditional paradigm the researcher directly interacted with the text as the primary

object of analysis (the "living" word), then in modern conditions he increasingly works with its digital representations - multidimensional vector spaces, embeddings and probability distributions obtained as a result of the work of algorithmic systems. Such a shift creates an urgent need for a new methodological reflection: the question arises of how to interpret the conclusions of neural network models in the humanitarian field, especially in the conditions of their colossal structural complexity and cognitive opacity.

Of particular importance in this context is the dilemma of interpretability and explainability of algorithms. Modern high-performance models, despite their unprecedented accuracy in processing linguistic structures, most often function according to the principle of a "black box". This creates a certain barrier to their full use in academic humanities research, which is historically oriented towards evidential, logically argued explanation of cause-and-effect relationships, rather than simple fixation of statistical correlations. Thus, the problem of validity of conclusions is actualized: how scientific is the knowledge obtained on the basis of correlation patterns devoid of transparent linguistic logic?

The theoretical novelty of this approach lies in a significant shift in the focus of scientific research: from the analysis of the text as a static final product to the critical analysis of the models themselves that generate, transform and interpret this text. Thus, the object of modern philology becomes not only language in its classical sense, but also its algorithmic simulations and computational metamorphoses. This significantly expands the boundaries of the discipline, turning it into an interdisciplinary field, where the technical accuracy of machine linguistics must be balanced by the semantic richness and depth of interpretive analysis. In the conditions, the development of a hybrid methodology capable of integrating the colossal computing power of AI with the traditions of hermeneutic reading becomes a strategically important task, which will allow maintaining the necessary balance between the scale of the processed data and the qualitative depth of scientific knowledge (Table 3). The integration of these approaches marks the transition to a new stage in the development of philological thought, where digital tools become not just an auxiliary tool, but also an organic part of the process of cognition of linguistic reality.

Table 3.

Methodological Transformation of Philology in the AI Era

Component	**Traditional Approach**	**Modern (Data Science) Approach**
Primary Object	"Living" word, static text as a product.	Digital representations (vectors, embeddings).

Component	Traditional Approach	Modern (Data Science) Approach
Toolbox	Philological/hermeneutic analysis.	Neural networks (GPT, BERT), transformers.
Logic of Explanation	Causal relationships (Cause-Effect).	Statistical correlations (the "Black Box" problem).
Research Goal	Interpretation of a finished text.	Analysis of models that generate and interpret text.

Source: authors' results of systematization and analysis

4. The development of Big Data technologies has a systemic and comprehensive impact on the formation of modern philological trends, radically transforming the scale, pace and the very nature of scientific research in the field of humanitarian knowledge. The transition to macroanalysis of colossal arrays of linguistic data allows researchers to go beyond local contexts and focus on identifying global patterns, transnational discursive practices and the dynamics of language change in real time. As a result, a fundamentally new research optics is being formed, oriented towards the study of language as a global process, operating in the digital environment “here and now”, which automates routine analytical operations and opens access to previously inaccessible statistical generalizations.

One of the key consequences of this technological expansion is the emergence of the concept of the so-called “platform philology”. Within this paradigm, the success and depth of scientific research increasingly depend on direct access to complex digital infrastructures, specialized cloud services, super-powerful computing resources, and unique proprietary data corpora. Large technology corporations and leading global academic centers, which have the necessary tools for data accumulation and processing, are becoming the dominant centers of gravity of scientific activity, effectively dictating new standards, methodological frameworks, and priority areas of world-class philological research. However, such a profound transformation is inevitably accompanied by the growth of acute structural inequality in the international scientific environment. Researchers representing different countries and institutional contexts have fundamentally different levels of access to critical data and analysis technologies. This leads to the formation of a rigid scientific hierarchy in which global centers of knowledge production, primarily concentrated in the USA and China, concentrate in their hands the main intellectual and technical resources. In such conditions, the technological dependence of peripheral scientific communities on global nodes of digital infrastructure is increasing, which seriously

questions the possibility of equal participation of different cultures and national philological schools in the formation of the current world order. This situation dictates the need to develop international strategies for a fairer and more democratic distribution of digital resources to ensure the inclusive development of modern language science.

5. Modern global trends in the field of Big Linguistic Data (BLD) reflect not only rapid technological development, but also fundamental transformations in the structure of production, management and verification of language resources at the global level. The key vector in this area is the desire to universalize and automate natural language processing processes, which finds its embodiment in the development of super-powerful multilingual models. These systems are capable of simultaneously operating with hundreds of languages in a single semantic space, expanding the horizons of intercultural communication and high-quality automatic translation. Nevertheless, such technological expansion inevitably reveals deep imbalances in the processing of different language groups, which requires further methodological correction.

An important direction in the development of the global infrastructure of language data is the active promotion of Open Data initiatives and the creation of open linguistic corpora. These processes are aimed at democratizing access to resources and are accompanied by strict standardization of data formats, annotation protocols and processing methods, which ensures the necessary compatibility of the results of interdisciplinary research. However, in parallel with the trend towards openness, there is also an alarming counter-dynamic: the concentration of power over colossal arrays of linguistic information in the hands of a limited circle of transnational corporations and technological platforms. Becoming key actors in the collection and analysis of BLD, these entities actually monopolize the right to shape the research agenda, setting priorities for the development of technologies that often conflict with the interests of individual national cultures and academic communities.

In this context, the concept of digital and linguistic sovereignty acquires critical relevance. In the conditions of total digitalization and unification of the global information space, there is an urgent need to ensure effective control over national linguistic data. Preservation of linguistic diversity and protection of unique cultural identity become possible only if independent mechanisms for managing linguistic resources are formed. Thus, modern trends in the field of Big Linguistic Data dictate the need to find a balance between technological universalization and preservation of the sovereign right of nations to represent their linguistic heritage in the digital world, which transforms BLD from a purely technical category into an object of strategic and cultural importance.

5.1. *Despite the global information explosion, modern linguistics faces a chronic shortage of high-quality and representative corpora for most of the world's languages.* This problem is especially acute for Ukrainian, Belarusian, Balkan and other regional languages, which remain underrepresented in global digital resources. The shortage is not only quantitative, but also structural and semantic in nature: existing data sets are often fragmentary, poorly annotated and devoid of complete metadata.

Insufficient representativeness of training samples leads to systematic errors in the work of machine learning algorithms, which do not see the specific grammatical and discursive patterns of these languages. This distorts the results of statistical analysis and limits the possibilities of detecting frequency patterns. To overcome this gap, it is necessary to develop unified standards for data collection, markup and verification, combining deep philological expertise with engineering solutions. Without the creation of full-fledged national corpora, the development of data-driven linguistics of less represented languages will be virtually paralyzed, which requires the immediate implementation of interdisciplinary approaches to the formation of language ecosystems.

5.2. The use of Big Linguistic Data inevitably poses a difficult ethical choice for the researcher, raising issues of privacy, legitimacy and cultural responsibility. Huge amounts of linguistic data are most often collected and analyzed without the explicit consent of the authors, which creates risks of de-anonymization of communication participants through metadata and violation of their security. From a cultural point of view, algorithmic text processing carries the threat of homogenization: the desire of algorithms to average can lead to the erasure of unique discursive practices, which is especially dangerous for historically marginalized communities. Strict ethical standards should be introduced into scientific and educational practice, including transparency of collection methods, mandatory anonymization of data and obtaining informed consent. Only such an approach allows combining the analytical power of big data with respect for the cultural uniqueness and social security of language communities.

5.3. *The globalization of digital space radically transforms the circulation of linguistic data, acting as a process with an ambivalent nature.* On the one hand, it significantly facilitates access to multilingual resources, accelerates intercultural exchange and the dissemination of scientific results. On the other hand, the imposition of uniform formats and standards, oriented towards dominant language groups, leads to the leveling of local features. There is a risk that dialects, stylistic nuances and unique cultural meanings will be filtered out by

algorithms as statistical noise, turning living diversity into unified patterns. This creates a threat of impoverishment of the global context and loss of cultural identity. In these conditions, philology has the role of defender of "linguistic ecology", developing strategies for the digital preservation of diversity and ensuring adequate representation of the specifics of each culture in global data sets.

5.4. Automation of text analysis and the active use of machine learning algorithms inevitably lead to a reduction of the semantic level, creating a threat of loss of depth of interpretation. Modern algorithms effectively detect frequency patterns and structural patterns in large data sets, but they often remain “blind” to irony, metaphor, subtext and complex cultural allusions. In the pursuit of the scale of processing, the interpretive depth that has historically been the hallmark of classical philology is lost.

Quantitative conclusions, not supported by a qualitative analysis of hidden semantic layers, risk being superficial or false, not reflecting the true semantic nuances of language. This poses the task of developing hybrid methodologies for the scientific community, in which computing power serves only as a tool for deep philological penetration into the essence of the text. Only the integration of computational models with high-quality analytics will allow us to preserve the semantic richness and adequacy of scientific search in the era of Big Data dominance.

5.5. *Natural language processing algorithms are neutral tools: they inevitably inherit, reinforce and often reinforce biases present in the original BLD training corpora.* If the data sets contain social, gender, racial or cultural stereotypes, machine learning models will systematically reproduce them, which leads to distortion of the analysis results and automatic marginalization of certain groups in the translation and classification systems of texts.

The problem of algorithmic bias is not only a technical but also a deep socio-humanitarian challenge that can exacerbate existing social inequality. Its solution requires an interdisciplinary approach that involves the active participation of philologists in the ethical audit of language models. Specialists should carry out sociocultural control of data at the stage of their preparation, identifying hidden imbalances and developing verification standards. Only a combination of computational methods with high-quality humanitarian expertise will minimize the risks of discrimination in the digital environment.

5.6. *Anglocentrism remains a fundamental problem in modern language technologies, as English dominates the digital space and forms the basis of most*

open corpora and model training standards. This causes a significant shift in research priorities and technological resources towards English-language linguistics, creating systemic distortions in the development of NLP technologies. Methods that work effectively for English often turn out to be inadequate for languages with different morphology or syntax, which complicates the study of low-resource and regional languages.

This state of affairs exacerbates digital inequality and endangers global cultural and linguistic diversity, contributing to the marginalization of non-English-speaking discourses. Overcoming this dominance requires focused efforts to create representative multilingual resources, adapt algorithms to the specific needs of national languages, and introduce international standards for the fair distribution of language data. Only through the development of inclusive digital infrastructures is it possible to ensure the full scientific and technological presence of the entire diversity of the world's languages.

5.7. Low-resource languages are in a state of pronounced "digital danger" under the dominance of Big Linguistic Data. Their extremely low representation in global data sets makes them practically invisible to modern AI systems and NLP tools. A vicious circle arises: the shortage of representative corpora leads to systemic marginalization and low-quality automatic processing, which, in turn, reduces the motivation to use these languages in the digital environment.

The problem is complicated not only by the small volume of texts, but also by the lack of high-quality markup, metadata and corpus standards. To overcome this gap, a comprehensive strategy is needed, including the creation of national digital archives, standardization of annotations and the development of models capable of effectively learning on small samples. Without active interdisciplinary intervention by philologists and the state, low-resource languages risk completely disappearing from the global information space, which threatens the world's linguistic and cultural diversity.

5.8. Modern military and hybrid conflicts have a profound impact on language dynamics, transforming language into a tool for strategic communication and the formation of national identity. In the era of Big Data, war is instantly reflected in linguistic arrays: new terms emerge, the semantic meanings of words are transformed, and specific discourses of information warfare and propaganda are formed. For a linguist, this opens up a unique opportunity to study the accelerated evolution of language and network practices in extreme conditions.

However, working with such data requires special methodological caution and a high level of ethical control. Automatic text processing algorithms are often

unable to adequately interpret acute emotional, ideological, and political contexts, which threatens to produce superficial or distorted conclusions. Studying the "language of war" through the prism of large data corpora requires a deep understanding of the sociocultural background, which allows integrating the computational power of analysis with high-quality philological expertise to correctly assess the changes taking place.

5.9. The modern era of dominance of Big Data and artificial intelligence marks not the decline of classical philology, but its fundamental transformation and rebirth. Deep text analysis, consideration of sociocultural context and hermeneutic tradition remain the unshakable foundation of science, but today they must be integrated with powerful tools of computational linguistics and machine learning.

At the junction of traditional methods and quantitative data-driven approaches, Digital Humanities is born, where algorithmic analysis complements research intuition. Such a hybrid approach allows to radically expand the scale and efficiency of research, while preserving the critical and interpretative function of philology. The main conceptual challenge of modernity is to find a balance between the speed of processing colossal data sets and the semantic richness of human understanding. The philology of the future is a high-tech discipline capable of operating with trillions of words without losing the deep meanings and ethical completeness of humanitarian knowledge.

Table 4.

Systemic Challenges and Threats in the BLD Sphere

Challenge	Essence of the Problem	Potential Consequences
Platform Inequality	Access to resources is concentrated in tech giants.	Rigid scientific hierarchy; dependence of peripheral research centers.
Digital Endangerment	Underrepresentation of low-resource languages.	Marginalization of languages; systemic errors in NLP models.
Ethical Risks	Data collection without consent; deanonymization.	Privacy violations; homogenization of unique discourse practices.
Anglocentrism	English dominance in training datasets.	Systemic bias in the development of linguistic technologies.
Algorithmic Bias	Stereotypes present in training corpora.	Reproduction of racial, gender, and cultural biases by AI.

Source: authors' research results

CONCLUSIONS

The conducted research irrefutably proves that BLDs form a new paradigm in modern philology, capable of radically transforming traditional approaches to language learning. This shift from classical and corpus linguistics to data-driven models is accompanied by a profound change in the scientific logic itself: data itself becomes primary and self-sufficient, while the researcher increasingly acts as an operator of complex algorithms that reveal hidden statistical and semantic patterns. In this new reality, data becomes the main source of knowledge, requiring from the scientist not only philological erudition, but also fundamentally new competencies, as well as an interdisciplinary approach at the junction of linguistics, mathematics and information technologies.

However, the analysis showed that technological progress does not cancel the need to maintain interpretive depth. This justifies the urgent need to develop hybrid methodologies that harmoniously combine computing power and traditional humanitarian approaches. Only such a combination allows us to avoid the reduction of meanings and preserve the quality of text analysis.

The study identified a set of critical challenges facing modern science. Key among them are the acute shortage of high-quality and representative corpora for "low-resource" languages, as well as the total dominance of English-language data in the global digital space. This state of affairs generates structural inequality in access to resources and creates ethical risks related to privacy and the correct representation of users of different cultures. The globalization of linguistic data, on the one hand, contributes to their rapid spread, but on the other hand, imposes standardized and simplified forms of communication, which creates a real threat to the preservation of world linguistic and cultural diversity.

BLD acquires particular importance in the context of modern global crises, wars and hybrid conflicts. Language in these conditions is finally transformed into a field of struggle for meanings and a powerful tool of information warfare. This not only forms specific new discourses, but also unprecedentedly accelerates the dynamics of language changes in real time, requiring special efficiency and social responsibility from philological science.

In conclusion, it can be argued that modern philology is experiencing not the decline or end of classical science, but a phase of its deep transformation. A new epistemic structure of language knowledge is being formed, combining the scale of data-driven methods with the centuries-old wisdom of the classical tradition. The future of the discipline and its sustainable development directly depend on our ability to integrate quantitative and qualitative methods, develop strict ethical standards for data processing and, most importantly, protect the digital sovereignty of national cultures. Big Linguistic Data opens up unique horizons for global research, but at the same time poses a fundamental task for

science: to preserve the human dimension and depth of interpretation in the world of dominant algorithms.

PROSPECTS FOR FUTURE RESEARCH

The prospects for further research in the field of innovative machine linguistics open up qualitatively new horizons for understanding how language functions and transforms in conditions of global instability. The use of hybrid intelligent technologies is becoming not just a trend, but a necessity for the analysis of BLD, when standard statistical methods are powerless in the face of high uncertainty and rapid change in discourses [29]. Key research will focus on the development of neurosymbolic systems that combine neural networks with logical inference to achieve interpretability of results, and the creation of predictive analytics tools capable of predicting social crises from micro-shifts in the semantic field. An important direction will be cognitive-affective modeling of multimodal discourse to understand the emotional state of society and the operational adaptation of technologies for low-resource languages in conflict zones. In addition, the development of adversarial linguistics will allow for the effective detection of manipulative patterns and synthetic content, ensuring the protection of cognitive sovereignty. Ultimately, the synthesis of computing power and deep philological expertise will transform the information flow into a strategic resource for decision-making in times of global upheaval.

References:

1. S. Goncharenko, S. Krasniuk (2024). Innovative architecture of large language models. *Linhvistychni ta metodolohichni aspekty vykladannia inozemnykh mov profesiinoho spriamuvannia - Linguistic and methodological aspects of teaching foreign languages for professional purposes*: materials of the V International Scientific and Practical Conference (Kyiv, 28-29 March 2024), Kyiv, NAU, 2024, pp. 25-26.

2. S. Krasniuk, S. Goncharenko (2024). Ethics of using large language models in machine linguistics. *Linhvistychni ta metodolohichni aspekty vykladannia inozemnykh mov profesiinoho spriamuvannia - Linguistic and methodological aspects of teaching foreign languages for professional purposes*: materials of the V International Scientific and Practical Conference (Kyiv, 28-29 March 2024), Kyiv, NAU, 2024, pp. 43-44.

3. Moretti, F. (2013). *Distant reading*. Verso. https://www.versobooks.com/products/1633-distant-reading

4. Manovich, L. (2020). *Cultural analytics*. MIT Press. https://culturalanalytics.info/

5. Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press. https://archive.org/details/corpusconcordanc0000sinc

6. Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804496

7. Schellfleysh, T. (2022). From static corpora to linguistic streams. *Journal of Digital Philology*, *11*(2), 45-62. https://muse.jhu.edu/journal/524

8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NIPS 2017), 30*, 5998-6008. https://arxiv.org/abs/1706.03762

9. Andrenko, K. V. (2024). Veliki movni modeli v linhvistytsi [Large language models in linguistics]. *Visnyk MSLU. Seriia 1: Filolohiia*, (128), 15-24. https://mslu.by/science/periodicheskie-izdaniya-mglu/vestnik-mglu

10. Derevianko, Iu. (2023). Transformatsiia filolohichnoho metodu v epokhu ShI [Transformation of the philological method in the era of AI]. *Digital Philology*, (4), 112-121. https://cyberleninka.ru/

11. Rockwell, G., & Sinclair, S. (2016). *Hermeneutica: Computer-assisted interpretation in the humanities*. MIT Press. http://hermeneuti.ca/

12. Crystal, D. (2011). *Internet linguistics*. Routledge. https://www.davidcrystal.com/books-and-articles/internet-linguistics

13. Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., Ide, J. S., & Williams, N. M. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610-623. https://doi.org/10.1145/3442188.3445922

14. Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press. https://www.atlasofai.org/

15. Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press. https://nyupress.org/9781479837243/algorithms-of-oppression/

16. Zuboff, Sh. (2019). *The age of surveillance capitalism*. PublicAffairs. https://www.shoshanazuboff.com/book/

17. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 6282-6293. https://aclanthology.org/2020.acl-main.560/

18. Magidson, M. (2023). NLP for low-resource languages. *Journal of Natural Language Processing*, *29*(3), 201-218. https://www.jstage.jst.go.jp/browse/jnlp/-char/en

19. Van de Velde, E. (2021). Digital archiving and linguistic heritage. *Heritage Science*, *9*(1). https://heritagesciencejournal.springeropen.com/articles/10.1186/s40494-021-00518-w

20. Pomerantsev, P. (2019). *This is not propaganda: Adventures in the war against reality*. Faber & Faber. https://www.faber.co.uk/product/9780571338634-this-is-not-propaganda/

21. Bjola, C., & Zaiotti, R. (Eds.). (2020). *Digital diplomacy: Theory and practice*. Routledge. https://www.routledge.com/Digital-Diplomacy-Theory-and-Practice/Bjola-Zaiotti/p/book/9780367134372

22. Castells, M. (2015). *Networks of outrage and hope: Social movements in the internet age* (2nd ed.). Polity Press. https://www.politybooks.com/bookdetail?book_id=9780745695754

23. Lopez, J. (2023). Generative AI in academic writing: A linguistic perspective. *Linguistics and Education*, *75*, 101183. https://doi.org/10.1016/j.linged.2023.101183

24. Wheeler, S. (2019). *Digital learning in higher education*. SAGE. https://uk.sagepub.com/en-gb/eur/digital-learning-in-higher-education/book259461

25. Couldry, N., & Mejias, U. A. (2019). *The costs of connection*. Stanford University Press. https://www.sup.org/books/title/?id=28515

26. Drucker, J. (2021). *The digital humanities coursebook: An introduction to digital methods for humanities research*. Routledge. https://www.routledge.com/The-Digital-Humanities-Coursebook-An-Introduction-to-Digital-Methods-for/Drucker/p/book/9780367565503

27. Hayles, N. K. (2012). *How we think: Digital media and contemporary technogenesis*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/H/bo12891316.html

1. Maxim Krasnyuk, Svitlana Nevmerzhytska, Tetiana Tsalko. (2024). Processing, analysis & analytics of big data for the innovative management. *Grail of Science*, #38, April 2024. pp. 75-83. https://archive.journal-grail.science/index.php/2710-3056/article/view/2230

28. Krasnyuk, M. (2014). Hibrydyzatsiia intelektualnykh metodiv analizu biznesovykh danykh (rezhym vyiavlennia anomalii) yak skfladovyi instrument korporatyvnoho audytu [Hybridization of intelligent methods of business data analysis (anomaly detection mode) as a standard tool of corporate audit]. *Stan i perspektyvy rozvytku oblikovo-informatsiinoi systemy v Ukraini - The state and prospects of the development of the accounting and information system in Ukraine*: materials of the III International science and practice conf. [m. Ternopil, October 10-11. 2014]. TNEU, 2014. pp. 211-212 [in Ukrainian].

Література:

1. S. Goncharenko, S. Krasniuk. Innovative architecture of large language models // Лінгвістичні та методологічні аспекти викладання іноземних мов професійного спрямування : матеріали V Міжнародної науково-практичної конференції, м. Київ, 28-29 березня 2024 року / за заг. ред. О. М. Акмалдінової. - Київ : НАУ, 2024. - С. 25-26.

2. S. Krasniuk, S. Goncharenko. Ethics of using large language models in machine linguistics // Лінгвістичні та методологічні аспекти викладання іноземних мов професійного спрямування : матеріали V Міжнародної науково-практичної конференції, м. Київ, 28-29 березня 2024 року / за заг. ред. О. М. Акмалдінової. - Київ : НАУ, 2024. - С. 43.

3. Moretti F. Distant Reading. London : Verso, 2013. 254 p. URL: https://www.versobooks.com/products/1633-distant-reading

4. Manovich L. Cultural Analytics. Cambridge, MA : MIT Press, 2020. 304 p. URL: https://culturalanalytics.info/

5. Sinclair J. Corpus, Concordance, Collocation. Oxford : Oxford University Press, 1991. 179 p. URL: https://archive.org/details/corpusconcordanc0000sinc

6. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge : Cambridge University Press, 1998. 312 p. URL: https://doi.org/10.1017/CBO9780511804496

7. Schellfleysh T. From Static Corpora to Linguistic Streams. *Journal of Digital Philology*. 2022. Vol. 11, No. 2. P. 45-62. URL: https://muse.jhu.edu/journal/524

8. Vaswani A. et al. Attention Is All You Need. *Advances in Neural Information Processing Systems (NIPS 2017)*. 2017. Vol. 30. P. 5998-6008. URL: https://arxiv.org/abs/1706.03762

9. Андренко К. В. Великі мовні моделі в лінгвістиці. *Вісник МДЛУ*. Серія 1 : Філологія. 2024. № 1 (128). С. 15-24. URL: https://mslu.by/science/periodicheskie-izdaniya-mglu/vestnik-mglu

10. Дерев'янко Ю. Трансформація філологічного методу в епоху ШІ. *Цифрова філологія*. 2023. Вип. 4. С. 112-121. URL: https://cyberleninka.ru/

11. Rockwell G., Sinclair S. Hermeneutica: Computer-Assisted Interpretation in the Humanities. Cambridge, MA : MIT Press, 2016. 304 p. URL: http://hermeneuti.ca/

12. Crystal D. Internet Linguistics. London : Routledge, 2011. 192 p. URL: https://www.davidcrystal.com/books-and-articles/internet-linguistics

13. Bender E. M., Gebru T. et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of FAccT '21*. 2021. P. 610-623. URL: https://doi.org/10.1145/3442188.3445922

14. Crawford K. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, 2021. 336 p. URL: https://www.atlasofai.org/

15. Noble S. U. Algorithms of Oppression: How Search Engines Reinforce Racism. New York : NYU Press, 2018. 256 p. URL: https://nyupress.org/9781479837243/algorithms-of-oppression/

16. Zuboff Sh. The Age of Surveillance Capitalism. New York : PublicAffairs, 2019. 704 p. URL: https://www.shoshanazuboff.com/book/

17. Joshi P. et al. The State and Fate of Linguistic Diversity in the NLP World. *Proceedings of ACL*. 2020. P. 6282-6293. URL: https://aclanthology.org/2020.acl-main.560/

18. Magidson M. NLP for Low-Resource Languages. *Journal of Natural Language Processing*. 2023. Vol. 29, No. 3. P. 201-218. URL: https://www.jstage.jst.go.jp/browse/jnlp/-char/en

19. Van de Velde E. Digital Archiving and Linguistic Heritage. *Heritage Science*. 2021. Vol. 9, No. 1. URL: https://heritagesciencejournal.springeropen.com/articles/10.1186/s40494-021-00518-w

20. Pomerantsev P. This Is Not Propaganda. London : Faber & Faber, 2019. 256 p. URL: https://www.faber.co.uk/product/9780571338634-this-is-not-propaganda/

21. Bjola C., Zaiotti R. Digital Diplomacy: Theory and Practice. London : Routledge, 2020. URL: https://www.routledge.com/Digital-Diplomacy-Theory-and-Practice/Bjola-Zaiotti/p/book/9780367134372

22. Castells M. Networks of Outrage and Hope. Cambridge : Polity Press, 2015. URL: https://www.politybooks.com/bookdetail?book_id=9780745695754

23. Lopez J. Generative AI in Academic Writing. *Linguistics and Education*. 2023. Vol. 75. URL: https://doi.org/10.1016/j.linged.2023.101183

24. Wheeler S. Digital Learning in Higher Education. London : SAGE, 2019. URL: https://uk.sagepub.com/en-gb/eur/digital-learning-in-higher-education/book259461

25. Couldry N., Mejias U. A. The Costs of Connection. Stanford University Press, 2019. URL: https://www.sup.org/books/title/?id=28515

26. Drucker J. The Digital Humanities Coursebook. London : Routledge, 2021. URL: https://www.routledge.com/The-Digital-Humanities-Coursebook-An-Introduction-to-Digital-Methods-for/Drucker/p/book/9780367565503

27. Hayles N. K. How We Think. Chicago : University of Chicago Press, 2012. URL: https://press.uchicago.edu/ucp/books/book/chicago/H/bo12891316.html

28. Maxim Krasnyuk, Svitlana Nevmerzhytska, Tetiana Tsalko. Processing, analysis & analytics of big data for the innovative management. Grail of Sciencею - №38. - April 2024. - P. 75-83. - URL: https://archive.journal-grail.science/index.php/2710-3056/article/view/2230

29. Краснюк М.Т. Гібридизація інтелектуальних методів аналізу бізнесових даних (режим виявлення аномалій) як складовий інструмент корпоративного аудиту. Стан і перспективи розвитку обліково-інформаційної системи в Україні: матеріали III

Міжнар. наук.-практ. конф. (м. Тернопіль, 10-11 жовт. 2014 р.). Тернопіль: ТНЕУ, 2014. С. 211-212.

Дата першого надходження статті до видання: 13.04.2026
Дата прийняття статті до друку після рецензування: 27.04.2026