UDC 81’322

Structured, semi-structured and unstructured batch and stream data in philology

Svitlana Krasnyuk
Kyiv National University of Technologies and Design, Kyiv
https://orcid.org/0000-0002-5987-8681

Abstract. *Data in philology is an important resource for the study of language, linguistic phenomena, literature, and cultural texts. In modern philology, the use of digital technologies, such as big data processing, machine learning, and artificial intelligence, has significantly changed approaches to the analysis of linguistic and literary materials. The collection, processing and analysis of data allows researchers to conduct research at a new level, opening up opportunities for automation, more accurate analysis and interpretation of texts.*
Keywords: *philology, big data analysis & analytics, machine linguistics.*

Introduction. Data in philology play an important role in modern research, allowing effective analysis of various aspects of linguistic phenomena, linguistic structures, literary texts and other linguistic resources. The use of data in philology covers several main aspects:

1. Corpus studies. Corpora are collections of texts used to study various linguistic phenomena such as grammar, vocabulary, stylistics, and semantics. The use of cases allows:

- Analyze the frequency of use of words and constructions.
- Study language changes over time.
- Compare different dialects or styles of speech.

Corpora can be either structured (with added metadata and grammar markups) or unstructured, which requires additional processing for analysis.

2. Lexicography. Lexicographic data are dictionary resources used to create dictionaries, thesauruses, and glossaries. This data usually has a well-structured form and includes:

- Words and their meaning.
- Grammatical characteristics (part of speech, cases, forms of verbs, etc.).
- Usage examples and contexts.

3. Literary data. Literary studies can also use data in philology for:

- Text analysis: study of stylistic, rhetorical and narrative features.
- Comparative analyses: evaluation of literary works of different eras, genres or authors.
- Thematic studies: analysis of main themes or motifs in works of art.

4. Machine linguistics and NLP (Natural Language Processing). This direction is associated with the use of large volumes of data for the automation of language processes [1, 2]:

- Language recognition: analysis and interpretation of spoken or written texts.
- Syntactic and semantic analysis: automated determination of grammatical structures and word meanings in context.
- Automated translation and analysis of texts: systems capable of translating or analyzing texts in different languages.

5. Social media and Internet communication.

In modern philology, special attention is paid to the study of language processes in social media and online communications. Using data from the following sources, you can explore:

- New language trends and changes in language (words, phrases, memes).
- Tonality and emotionality of communication.
- Interaction between different language communities.

6. Dialectology and sociolinguistics

Collection and analysis of data on dialects and social variants of the language allows studying regional and social differences in languages, researching the evolution of language forms and cultural influences on the language.

7. Grammatical and morphological analysis

These studies are aimed at a detailed analysis of language structure. The data here includes different types of morphological and grammatical constructions, their variations and usage.

8. Data from digital humanities. Philological data are also used in interdisciplinary research, in particular in digital humanities (Digital Humanities), where literary and cultural texts are studied with the help of computer technologies and methods.

The Main Part. In philology, as in many other scientific fields, modern research increasingly depends on the use of large volumes of data [3]. The use of big data opens up new possibilities for the study of language, texts, communication and literature. Depending on the form of presentation and organization of information, big data can be divided into structured, semi-structured and unstructured [4, 5]. These big data are also divided into batch data and stream data [6, 7]. Let us consider each of these types of data in the context of philological research.

1. Structured data in philology. Structured data is data that has a clear, predefined organization, usually in the form of tables or databases. The author recommends using classical methods of machine learning as the main tool of descriptive and predicative analysis & analytics of structured data [8, 9, 10, 11]. And in philology, structured data can include:

- Lexicographic databases: dictionaries, thesauruses, where each word or expression has a defined structure (dictionary article with lemma, grammatical categories, usage examples, etc.).
- Text corpora: collections of texts with clear metadata such as author, date, genre, number of words, syntactic markup.
- Linguistic databases: grammatical, morphological tables for studying the structures of different languages.

Example: A structured database containing each word with its grammatical characteristics, such as part of speech, verb tense, number, etc., can be used in the study of morphological complexity of language.

2. Semi-structured data in philology. Semi-structured data has some structure, but it is not as rigid as structured data. They can be presented in a format that allows you to flexibly change the structure without losing the content. An example of semi-structured data in philology can be:

- XML or JSON files with marked up texts.

- Marked texts: texts in which part of the information (for example, grammar or semantics) is presented in a structured form, while other elements can have a free structure [12].

The author recommends using a range of Data Science methods and algorithms [13], including various architectures of artificial neural networks [14] for the analysis and analytics of such semi-structured data. In electronic text corpora, each text can be tagged to indicate grammatical structures or semantic meanings. At the same time, the text itself remains in the form of natural language, which does not have a clear structure.

3. Unstructured data in philology. Unstructured data is data that does not have a clear, predefined structure. The author recommends using deep neural networks as the main tool for predicative analytics of unstructured big data [15, 16]. In philology, such data can be:

- Literary texts: novels, poems, scientific articles, correspondence, journalism that do not have a clear structure for automatic analysis.

- Audio and video recordings: oral interviews, lectures, recordings of conversations that require decoding or additional processing for analysis.

Example: The texts of works of fiction are typical examples of unstructured data, since they do not have a clear and predictable structure, but contain various lexical and stylistic elements.

4. Batch and stream data in philology.

4.1. Batch data (Batch Data). Batch data is data that is collected and processed in batches or groups. In philology, batch data is usually used to analyze large volumes of texts or records that have already been collected.

Example: Text corpus analysis can be done as batch processing. First, texts of a certain genre or topic are collected, and then their processing is carried out using software tools: lemmatization, search for frequency of use of words or expressions, etc.

4.2. Stream data (Stream Data). Streaming data is data that comes in real time. They are constantly updated and require immediate processing.

Example: In philology, streaming data can be used to analyze live conversations on social networks, chat rooms, or other digital platforms. For example, the system can analyze texts or messages in real time to detect language trends, semantic changes or emotional tonality.

Conclusions. Philological data open up great opportunities for modern research, allowing a deeper understanding of the nature of language, its evolution, and features of communication. Due to the development of technologies and the ability to process large amounts of information, philological research is becoming more and more interdisciplinary, combining linguistics, computer science, sociology and other fields.

In philology, the use of different types of data – structured, semi-structured and unstructured – opens up new possibilities for the analysis of language and texts. The use of batch and streaming data also facilitates a flexible approach to the processing of language phenomena in various contexts: from the analysis of large corpora of literary works to the study of dynamic language processes in real time. Thanks to these methods, studies of language structures and communications become more accurate and productive.

Discussion and prospects for further research. The author puts forward the thesis that it is the hybrid methods of Data Science in the analysis and analysis of philological data that combine different approaches and tools to solve complex

problems of processing and analyzing textual resources. They allow the integration of classical statistical methods, machine learning, natural language processing (NLP), deep learning and other technologies for in-depth understanding of language phenomena, text structure [17, 18, 19]. In other words, in his subsequent research and publications, the author will try to prove that: hybrid methods of Data Science are a powerful tool for the analysis and analysis of philological data, because they allow combining the most modern technologies of machine learning and natural language processing with classical methods of linguistic analysis, opening up new opportunities for research automation, analysis of large text arrays and linguistic innovations.

References

1. Goncharenko S. Innovative architecture of large language models / S. Goncharenko, S. Krasniuk // *Лінгвістичні та методологічні аспекти викладання іноземних мов професійного спрямув*ання : 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. Krasniuk, S., & Goncharenko, S. (2024). Ethics of using large language models in machine linguistics // *Лінгвістичні та методологічні аспекти викладання іноземних мов професійного спрямування*. Національний авіаційний університет, 2024.

3. Науменко , М. (2024). Аналіз та аналітика великих даних в маркетингу та торгівлі конкурентного підприємства. *Grail of Science*, (40), 117–128. https://doi.org/10.36074/grail-of-science.07.06.2024.013.

4. 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://www.journal-grail.science/issue38.pdf.

5. Krasniuk, S., & Goncharenko, S. (2024). Big data in philology. *Collection of Scientific Papers "ΛΌГОΣ"*, (September 20, 2024; Paris, France), 159–165. https://doi.org/10.36074/logos-20.09.2024.031.

6. Maxim Krasnyuk, Dmytro Elishys (2024). Perspectives and problems of big data analysis & analytics for effective marketing of tourism industry. *Science and technology today*, 4 (32) 2024. pp. 833-857.

7. Krasnyuk M., Krasnuik I. Big data analysis and analytics for marketing and retail. *Штучний інтелект у науці та освіті:* збірник тез Міжнародної наукової конференції (*AISE*) (1-2.03.2024 р.), Київ, 2024.

8. Науменко, М. (2024). Ефективне застосування класичних алгоритмів машинного навчання при прийнятті адаптивних управлінських рішень. *Наукові перспективи*, 2024, 5 (47). https://doi.org/10.52058/2708-7530-2024-5(47)-855-875.

9. Krasnyuk M., Krasniuk S. Comparative characteristics of machine learning for predicative financial modelling. *ΛΌГОΣ*. 2020. P. 55-57.

10. Krasnyuk M., Tkalenko A., Krasniuk S. Results of analysis of machine learning practice for training effective model of bankruptcy forecasting in emerging markets. *ΛΌГОΣ*. 2021.

11. Krasnyuk M., Krasniuk S. Modern practice of machine learning in the aviation transport industry. *ΛΌГОΣ*. 2021.

12. Krasniuk, S. (2024). Advanced text mining in philology. *Collection of Scientific Papers "SCIENTIA"*, (September 27, 2024; Stockholm, Sweden), 68–72. Retrieved from https://previous.scientia.report/index.php/archive/issue/view/27.09.2024.

13. Tetiana Tsalko, Svitlana Nevmerzhytska, Svitlana Krasniuk, Svitlana Goncharenko, Liubymova Natalia (2024). Features, problems and prospects of data mining and data science application in educational management. *Bulletin of Science and Education*, №5(23), 2024. pp.637-657.

14. Krasnyuk, M., & Krasniuk, S. (2020). Application of artificial neural networks for reducing dimensions of geological-geophysical data set's for the identification of perspective oil and gas deposits. Збірник наукових праць ΛΌГОΣ, 18-19. https://doi.org/10.36074/24.04.2020.v2.05.

15. Науменко, М. (2024). Оптимальне використання алгоритмів глибокого машинного навчання в ефективному управлінні підприємством. *Успіхи і досягнення у науці*, 2024, 4 (4). https://doi.org/10.52058/3041-1254-2024-4(4)-776-794.

16. Maxim Krasnyuk, Svitlana Krasniuk, Svitlana Goncharenko, Liudmyla Roienko, Vitalina Denysenko, Liubymova Natalia (2023). Features, problems and prospects of the application of deep machine learning in linguistics. *Bulletin of Science and Education,* №11(17), 2023. pp.19-34. http://perspectives.pp.ua/index.php/vno/article/view/7746/7791.

17. M. Krasnyuk, S. Goncharenko, S. Krasniuk (2022) Intelektualni tekhnolohii v hibrydnii korporatyvnii SPPR (na prykladi Ukrainskoi naftohazovydobuvnoi kompanii) [Intelligent technologies in hybrid corporate DSS (on the example of Ukraine oil&gas production company)] Innovatsiino-investytsiinyi mekhanizm zabezpechennia konkurentospromozhnosti krainy: kolektyvna monohrafiia / za zah. red. O. L. Haltsovoi - Innovation and investment mechanism for ensuring the country's competitiveness: collective monograph / by general ed. O. L. Khultsova. – Lviv-Torun: League-Pres, 2022. – pp. 194-211. [in Ukrainian].

18. Krasnyuk, M. T. "Hibrydyzatsiia intelektualnykh metodiv analizu biznesovykh danykh (rezhym vyiavlennia anomalii) yak skladovyi instrument korporatyvnoho audytu" [Hybridization of intellectual methods of business data analysis (anomaly detection mode) as a component tool of corporate audit] // Stan i perspektyvy rozvytku oblikovo-informatsiinoi systemy v Ukraini: materialy III Mizhnar. nauk.-prakt. konf. - Status 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] - Ternopil: TNEU, 2014. - pp. 211-212. [in Ukrainian].

19. Hrashchenko I.S., Krasniuk M.T., Krasniuk S.O. (2019). Hibrydno-stsenarne zastosuvannia intelektualnykh, oriientovanykh na znannia tekhnolohii, yak vazhlyvyi antykryzovyi instrument lohistychnykh kompanii v Ukraini [Hybrid-scenario application of intellectual, knowledge-oriented technologies as an important anti-crisis tool of logistics companies in Ukraine]. Vcheni zapysky Tavriiskoho Natsionalnoho Universytetu imeni V. I. Vernadskoho. Seriia: Ekonomika i upravlinnia – Scientific notes of Tavri National University named after V. I. Vernadskyi. Series: Economics and management, 2019. Vol. 30 (69). pp.121 – 129. [in Ukrainian].