Proceedings of the 14th International Scientific and Practical Conference «Global and Regional Aspects of Sustainable Development»

(September 16-18, 2026). Copenhagen, Denmark

No 309

PHILOLOGY AND LINGUISTICS

Leveraging and Adapting Big Data Technologies for Modern Dialectological Research: Computational Linguistics under Dynamic and Uncertain Conditions of Systemic Risks

Svitlana Krasniuk[1]

[1]Senior Lecturer, *Department of philology and translation; Kyiv National University of Technologies and Design; Ukraine*

Abstract. The article explores the specifics, features and directions of adapting Big Data technologies for modern dialectology within the framework of machine/computational linguistics. The possibilities of processing multimodal dialect data, the use of machine learning, NLP and geoinformation technologies for identifying spatio-temporal linguistic variability are considered. The need for a hybrid and risk-oriented approach in conditions of turbulence, uncertainty and systemic risks is substantiated.

Keywords: *modern linguistics; computational Linguistics; big data; dialectology; machine learning; natural language processing (NLP); dialect data; language variation; geolinguistics; digital dialectology; intelligent technologies; language change; turbulence; uncertainty; systemic risks.*

INTRODUCTION.

Modern linguistics in general (and dialectology in particular) is undergoing a significant transformation under the influence of digitalization [1], the development of artificial intelligence [2], machine learning [3], automated natural language processing [4] and Big Data technologies [5]. While traditional dialectology was mainly based on limited field materials, questionnaires, audio recordings, transcriptions and local linguogeographic studies, the digital environment forms fundamentally new and BIG DATA arrays [6], including texts, audio and video materials, digital corpora, social media [7]. This creates the prerequisites for the transition from a predominantly selective and static study of dialects to a large-scale,

PHILOLOGY AND LINGUISTICS

dynamic and computationally supported analysis of language variability.

This transformation becomes particularly relevant in the conditions of modern turbulence, uncertainty and systemic risks [8]. Migration processes, social mobility, digitalization of communication, language contacts, demographic changes, crisis phenomena and transformation of communicative practices accelerate the change of territorial and social varieties of language. As a result, traditional ideas about dialect as a stable territorially localized system require supplementation with models capable of recording the dynamics, hybridization, diffusion and gradual disappearance of individual linguistic features.

In these conditions, Big Data technologies become not just a tool for storing and processing huge volumes of linguistic semi-structured and unstructured Big Data [9] (which continue to grow), but the most important technological component of modern machine/computational linguistics. Their application allows combining heterogeneous data sources, revealing hidden patterns of variability, performing automatic classification and clustering of dialect varieties, building geolinguistic models and tracking the dynamics of changes [10]. At the same time, the direct transfer of universal Big Data technologies to dialectology is impossible without special adaptation due to the specifics of dialect material [11]. That's why, the purpose of the article is to systematically determine the specifics, features and directions of adaptation of Big Data technologies for modern dialectology within the framework of machine/computational linguistics, taking into account the dynamics of turbulence, uncertainty and systemic risks.

Main Part.

1. Dialect data as a specific object of Big Data analytics.

The use of Big Data in dialectology is primarily due to a change in the scale and structure of available language material. Modern digital communication generates significant volumes of text, language and multimodal data, potentially containing dialect features. Unlike traditional dialectological materials, digital data are characterized not only by a large volume, but also by a high rate of arrival, heterogeneity, distribution and varying degrees of structuring.

PHILOLOGY AND LINGUISTICS

The specificity of dialectological Big Data is determined by the classic characteristics of Volume, Velocity, Variety, Veracity and Value, however, each of them acquires a special meaning in this case. Large volume is associated with the need to process millions of language units; high speed - with the constant emergence of new communicative data; diversity - with a combination of written and oral speech, transcriptions, audio, video, cartographic and sociolinguistic information; reliability - with the need to assess the origin and quality of dialect data; value - with the possibility of identifying patterns of territorial, social and temporal language variability.

At the same time, dialect data are fundamentally heterogeneous. They can contain phonetic, lexical, morphological, syntactic, semantic and pragmatic features, and a significant part of dialect characteristics is manifested in speech. Therefore, Big Data in dialectology is not reduced to the processing of large text corpora. They should be considered as multimodal and multilayered language data, in which language material is associated with geographical, social, historical and communicative parameters.

2. Technological specificity of Big Data in modern computer dialectology.

Adaptation of Big Data to dialectology involves the formation of a multi-level technological infrastructure. At the first level, data is collected from various sources: digital corpora, archives, field research, social networks, user content platforms, audio archives and other digital resources. At the second level, material cleaning, normalization, markup and integration are performed. The third level is associated with the application of methods of intellectual analysis, machine learning and statistical modeling. At the fourth level, the results are visualized using digital maps, graphs, cluster models and interactive geolinguistic interfaces.

The transition from simple digitization of traditional dialectological materials to the creation of an intelligent infrastructure of computer dialectology is becoming fundamentally important. Such an infrastructure should provide not only data storage, but also automatic detection of linguistic features, establishment of correlations between features, detection of spatial patterns and modeling of their dynamics.

PHILOLOGY AND LINGUISTICS

The integration of Big Data with technologies of natural language processing (NLP), automatic speech recognition, machine translation, computer phonetics, intelligent search and generative language models is of particular importance. For dialectology, this opens up the possibility of automatically converting large arrays of linguistic material into structured data suitable for further quantitative analysis.

3. Features of adapting machine learning to dialect data.

One of the key features of computational dialectology is the need to adapt machine learning algorithms to the high variability of speech material. Universal NLP models are mainly trained on standardized speech data, therefore they may not be effective enough in recognizing regional variants, non-standard spelling, phonetic features and mixed forms of speech.

To solve this problem, the use of specialized dialect corpora and transfer learning methods is promising, which allow adapting pre-trained speech models to specific dialect varieties. Self-supervised and semi-supervised learning methods are of additional importance, since in dialectology the volume of unlabeled data usually significantly exceeds the volume of qualitatively labeled materials.

Machine learning can be used for automatic identification of dialect features, classification of texts and speech fragments, determination of the regional affiliation of the speaker, clustering of speech varieties and prediction of language changes. In this case, the classification model should be perceived as the final description of the dialect system. The resulting algorithmic classes are data models that require linguistic interpretation and comparison with dialectological concepts.

The combination of supervised, unsupervised and hybrid machine learning is especially promising. If supervised learning is effective in the presence of labeled corpora, then clustering and dimensionality reduction methods allow you to identify previously unobvious groups of linguistic features. The hybrid approach provides a combination of automatic detection of patterns with expert interpretation by a linguist.

4. Geolinguistic and spatial potential of Big Data.

One of the most promising areas is the combination of Big Data with geoinformation technologies. Traditional dialectology has long used maps of the distribution of

PHILOLOGY AND LINGUISTICS

linguistic features, but digital technologies allow you to move from static maps to dynamic spatial models.

The integration of language data with geographic coordinates, demographic characteristics, migration flows and social parameters allows us to investigate not only where a certain linguistic feature is widespread, but also how, why and at what speed it spreads or disappears.

In this context, dialect can be modeled as a dynamic spatio-temporal system. Big Data allows us to analyze language areas as networks of interacting territories and communicative groups, to identify zones of language contact, transitional areas and centers of innovation. The study of dialect continua becomes especially important, where a rigid division of the language into separate territorial varieties turns out to be insufficient.

5. Big Data and the study of the dynamics of dialect changes.

In conditions of social and technological turbulence, dialectological analysis must take into account the time factor. Traditional studies often record the state of the language system in a certain historical or field period. Big Data allows you to form time series and observe changes in the frequency, prevalence and interrelationship of dialect features.

This is especially important for studying the processes of dialect leveling, convergence, divergence, language contact, hybridization and the disappearance of local features. Comparing data from different periods can allow you to determine the directions of language evolution and form predictive models.

The use of methods of temporal analysis, dynamic clustering and predictive machine learning is promising. At the same time, the forecast should not be understood as a deterministic prediction of the future dialect. Rather, it is about forming probabilistic scenarios for the development of linguistic variability, taking into account various factors of uncertainty.

6. Big Data in the context of turbulence, uncertainty and systemic risks.

Modern dialectology operates under conditions where the sources of linguistic data themselves are exposed to external crisis factors. Population migration changes the geography of linguistic groups; military, economic and social crises disrupt the continuity of field observations; digital

PHILOLOGY AND LINGUISTICS

communication simultaneously expands the volume of available data and changes the natural structure of communication.

Therefore, the adaptation of Big Data should include a risk-aware approach that involves assessing the uncertainty, incompleteness and possible representational bias of the data. A large volume of data does not automatically mean their high scientific representativeness. For example, digital platforms may predominantly reflect the language of certain age, social or territorial groups.

Systemic risk arises from technological dependence on automatic processing algorithms. Language recognition errors, incorrect normalization of non-standard forms, or excessive focus on normative language can lead to the loss of precisely those features that are of greatest value for dialectological analysis.

Therefore, reliable computer dialectology should be based on the principle of "automation + linguistic expertise", and not on the complete replacement of the researcher by the algorithm.

7. Hybrid intellectual model of modern Big Data-dialectology.

The most promising direction is the formation of a hybrid intellectual architecture that combines Big Data, machine learning, NLP, geoinformation technologies, statistical modeling and expert linguistic knowledge.

In such a model, Big Data provides the scale and diversity of the material; NLP is responsible for automatic language processing; machine learning reveals hidden patterns; geoinformation technologies model spatial distribution; statistical and probabilistic methods assess the stability of the identified dependencies; an expert linguist provides interpretation and verification of the results.

It is fundamentally important that such a system should be adaptive, that is, able to take into account the emergence of new language data, changes in communicative practices and the transformation of dialect systems themselves. In the case of turbulence, this is a transition from a static model of dialectological description to a dynamic system of continuous knowledge updating.

In the future, it is possible to create intelligent dialectological platforms capable of automatically collecting, structuring, classifying, and mapping new data, identifying new linguistic features, and forming hypotheses about the directions of language evolution. The final

PHILOLOGY AND LINGUISTICS

scientific interpretation should remain the result of the interaction of computational methods and professional linguistic expertise.

Conclusions.

The analysis shows that Big Data technologies form fundamentally new opportunities for modern dialectology, translating it from a predominantly local and selective study of language varieties to a large-scale, dynamic and intellectually supported analysis of language variability. The specificity of dialectological Big Data is determined not only by its volume, but primarily by multimodality, spatio-temporal structure, heterogeneity, incompleteness and high dependence on the socio-cultural context.

The adaptation of Big Data to dialectology requires the integration of natural language processing technologies, machine learning, automatic speech recognition, geographic information systems, statistical and spatio-temporal modeling. Hybrid methods that combine automatic pattern detection with expert linguistic interpretation are of particular importance.

In conditions of turbulence, uncertainty and systemic risks, the quality, representativeness, origin and stability of data are of particular importance. Therefore, a promising computer dialectology should take into account not only the technological efficiency of Big Data, but also the risks of algorithmic errors, digital elimination, loss of non-standard language forms and insufficient representativeness of individual digital sources.

Thus, Big Data should be considered not just as a tool for processing large arrays of dialect material, but the basis for the formation of a new intellectual infrastructure of modern dialectology. Its further development is associated with the creation of adaptive and hybrid information and analytical systems capable of simultaneously taking into account linguistic, geographical, social and temporal variability. This opens up the prospects for the beginning of dynamic computer dialectology, in which machine methods and expert knowledge form a single research system.

References:

[1] Smits, T., & Wevers, M. (2023). A multimodal turn in Digital Humanities: Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. Digital Scholarship in the Humanities, 38(3), 1267-1280. https://doi.org/10.1093/llc/fqad008

PHILOLOGY AND LINGUISTICS

[2] Belém, C. G., Kelly, M., Steyvers, M., Singh, S., & Smyth, P. (2024). Perceptions of linguistic uncertainty by language models and humans. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 8467-8502). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.483

[3] Goncharenko S. (2025). Big semi-structured data & deep ANN in computational linguistics. Science and Global Challenges in the Modern World : Proceedings of the 2nd International Scientific Conference (Leicester, United Kingdom, 5 October 2025). - Lulu Press, Inc., 2025. - pp. 133-136.

[4] Zubiaga, A. (2024). Natural language processing in the era of large language models. Frontiers in Artificial Intelligence, 6, Article 1343587. https://doi.org/10.3389/frai.2023.1343587

[5] Kaplan, F. (2015). A map for Big Data research in digital humanities. Frontiers in Digital Humanities, 2, Article 1. https://doi.org/10.3389/fdigh.2015.00001

[6] Blasi, D. E., & Roberts, S. G. (2018). Studying language evolution in the age of big data. Journal of Language Evolution, 3(2), 94-129. https://doi.org/10.1093/jole/lzy004

[7] Ganascia, J.-G. (2015). The logic of the Big Data turn in digital literary studies. Frontiers in Digital Humanities, 2, Article 7. https://doi.org/10.3389/fdigh.2015.00007

[8] Krasnyuk M.T. (2006). Problemy zastosuvannia system upravlinnia korporatyvnymy znanniamy ta yikh taksonomiia [Problems of applying corporate knowledge management systems and their taxonomy]. Modeliuvannia ta informatsiini systemy v ekonomitsi - Modeling and information systems in the economy, vol. 73, p. 256 [in Ukrainian].

[9] Goncharenko S. (2025). BIG Data in modern literary studies. Science: Development and Factors its Influence : Proceedings of the 6th International Scientific and Practical Conference (October 6-8, 2025; Amsterdam, Netherlands). - Amsterdam: Scientific Collection «InterConf», 2025. - P. 82-85.

[10] Krasnyuk M, Elishis D (2024). Perspectives and problems of big data analysis & analytics for effective marketing of tourism industry. Наука і Техніка Сьогодні, 4(32). https://doi.org/10.52058/2786-6025- 2024-4(32)-833-857

[11] Baayen, R. H. (2024). The wompom. Corpus Linguistics and Linguistic Theory, 20(3), 615-648. https://doi.org/10.1515/cllt-2024-0053