ANNOTATION

of the dissertation work by Assel S. Yerbolova
«Investigation of natural language structures in Spot the bot task»
submitted for the degree of Doctor of Philosophy (PhD) in the educational program:
8D06101 (6D075100) – «Computer Science, Computer Technology and Management»

Relevance of the Research Topic is determined by the rapid development of natural language processing (NLP) technologies and artificial intelligence (AI). In the context of the increasing volume of texts generated both by humans and automated systems such as bots, the growing risks of misinformation in the digital environment, and the degradation of artificial intelligence models, there is a need to develop effective methods for identifying bot-generated texts. This is essential for ensuring information security, monitoring digital content, and maintaining the stability of AI systems. The application of modern approaches to text analysis and classification highlights the importance of this study.

The methods and approaches proposed in this work are relevant for several reasons. First, the study of intrinsic dimensions of linguistic fractal structures, combined with the analysis of semantic trajectories, makes it possible to form a holistic understanding of linguistic patterns and contributes to a deeper understanding of the dynamics and structure of language systems.

Second, the use of topological data analysis opens new opportunities for studying language systems and can improve the quality of bot identification. This is particularly important in the context of increasing threats from automated systems used for opinion manipulation and the spread of misinformation. Thus, this work emphasizes the significance and complexity of natural language by proposing effective methods and approaches for its analysis, which can significantly improve the functioning of information systems in modern digital society.

The aim of the dissertation research is to develop methods for investigating language structures and to adapt them for the identification of texts produced by humans and generated by bots.

In accordance with the aim of the study, the following objectives were defined:

1. To develop characteristics of linguistic objects and analyze their statistical properties in order to identify differences between human-written texts and bot-generated texts.
2. To adapt methods for analyzing the chaotic properties of time series to assess the complexity of semantic trajectories in natural language texts.
3. To adapt methods for estimating the intrinsic dimensionality of complex geometric objects to the study of linguistic fractal structures.
4. To adapt methods of topological data analysis for studying the semantic space of natural language, including the identification of “holes” for various n-grams (words, bigrams, trigrams), and to perform a comparative analysis of large-scale structural characteristics of natural languages across a wide range of languages.
5. To develop classification models for identifying bot-generated texts based on the extracted linguistic features and to experimentally validate their effectiveness.

Object of the research – the linguistic characteristics and structural properties of natural language that enable effective discrimination between human-written texts and texts generated by automated systems (bots).

Subject of the research – encompasses methods for the analysis of large-scale structural and geometric properties of natural language, as well as the development of approaches for identifying human-written texts and texts generated by automated systems (bots).

The research methods employed in this study include: statistical methods (hypothesis testing and distribution analysis); methods for analyzing the chaotic properties of semantic trajectories in natural languages; numerical methods for estimating intrinsic dimensionality (geometric and fractal n-gram–based approaches); topological data analysis (persistent homology); and methods for bot identification (classification models and dataset partitioning).

The scientific novelty of the obtained results is as follows:

1. An analysis of natural languages as self-organized critical systems was conducted on a large multilingual dataset. The uniqueness of the Esperanto language was established, demonstrating a Gaussian distribution in contrast to natural languages. It is shown that the statistical characteristics of power-law distributions, including power-law parameters, serve as reliable metrics for the analysis and classification of language systems.
2. The results of the analysis of semantic trajectories in natural languages are presented, revealing the chaotic nature of the underlying structure. Cluster analysis identified typological clusters correlated with entropy and complexity, highlighting the significance of chaos in language dynamics.
3. Methods for estimating the intrinsic dimensionality of natural languages were developed, revealing their multifractal nature and demonstrating the invariance of dimensionality with respect to different vector representation extraction methods.
4. A method for detecting "holes" in language based on topological data analysis was developed. The application of first-order persistent homology revealed large-scale structural features of natural language and enabled the elimination of sampling artifacts.
5. High-performance text classification algorithms were proposed for distinguishing between human-written and bot-generated texts, achieving accuracy above 96% for most languages.

The practical significance of the research lies in the development of methods for bot identification that enable effective discrimination between human-written texts and texts generated by automated systems. The developed methods for analyzing linguistic characteristics, such as entropy and the complexity of semantic trajectories, significantly improve the accuracy and efficiency of automated bot detection systems, which is particularly important in the context of the increasing automation of content generation and the spread of misinformation.

The proposed methods of topological data analysis and persistent homology are applicable to the study of structural properties of texts, enabling advanced analysis and classification across various domains.The conducted comparative analysis of natural languages contributes to the reconsideration of language classification by taking into account their large-scale structural properties.

The main provisions submitted for defense are as follows:

1. A model of natural languages as self-organized critical systems, applied to the comparative analysis of a wide range of languages.
2. A model for analyzing semantic trajectories to assess the chaotic nature of language and its application to comparative language analysis.
3. A model for analyzing linguistic fractal structures (intrinsic dimensionality and "holes") for identifying multifractality and substantiating hypotheses of linguistic relatedness.
4. A classification model for identifying human-written and bot-generated texts, the effectiveness of which has been experimentally validated and demonstrates the practical significance of the proposed methods.

Publications and Approbation of the research results

The main results of the dissertation were presented at scientific seminars at M. Kozybayev North Kazakhstan University (Department of Information and Communication Technologies) and at the National Research University Higher School of Economics. A total of seven scientific papers have been published, including five articles indexed in the Scopus database, with percentiles ranging from 39 to 80; one article in journals recommended by the Committee for Quality Assurance in Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan (MSHE RK); and one publication in the proceedings of international conferences.

- Articles published in peer-reviewed scientific journals with a non-zero impact factor, indexed in the Web of Science Core Collection and Scopus databases:

1. Gromov V.A., Borodin N.S., and Yerbolova A.S., A Language and Its Dimensions: Intrinsic Dimensions of Language Fractal Structures// *Complexity Volume 2024*, https://doi.org/10.1155/2024/8863360, *(Indexed in Scopus, Q1)*

2. Gromov V.A., Dang Q.N., Kogan A.S., Yerbolova A.S., Spot the bot: the inverse problems of NLP. *PeerJ Computer Science* 10:e2550, 2024, https://doi.org/10.7717/peerj-cs.2550, *(Indexed in Scopus, Q1)*

3. Gromov V.A., Dang Q.N., Yerbolova A.S., Language and Its Holes: the First Order Homologies of the Large-scale Geometrical Structure of a Natural Language// Complexity, 2025 (1) Article ID 9659172, https://doi.org/10.1155/cplx/9659172, *(Indexed in Scopus, Q2)*

4. Yerbolova A.S., Tomashchuk K.K., Kogan A.S., Dang Q.N., Skrynnikova I.V., and Gromov V.A., Relative Chaoticity of Natural Languages// Complexity, Volume 2026, https://doi.org/10.1155/cplx/5519690, *(Indexed in Scopus, Q2)*

5. Yerbolova A.S., Kurmashev I. G., Revealing intrinsic dimensionality patterns in semantic spaces of natural languages using graph algorithms, Eastern-European Journal of Enterprise Technologies, 1/2 (138), 2026, https://doi.org/10.15587/1729-4061.2026.351509 , *(Indexed in Scopus, Q3)*

- Articles published in scientific journals recommended by the Committee for Quality Assurance in the Field of Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan (CQAFSHE MSHE RK):

6. Yerbolova A.S., Gromov V.A., Akanova A.S. Semantic Methods for Detecting Texts Generated by Artificial Intelligence Systems // Bulletin of the Almaty University of Power Engineering and Telecommunications, No. 1(72), 2026.

- Abstracts and articles presented at international and national conferences:

7. Spot the bot: large-scale natural language structure // Conference: 7th International Conference “Futurity designing. Digital reality problems”, (February 15–17, 2024, Moscow). — Moscow: Keldysh Institute of Applied Mathematics, 2024. — pp. 281–312. — https://keldysh.ru/future/2024/6-3.pdf https://doi.org/10.20948/future-2024-6-3

Implementation of Research Results

The results of the dissertation have been implemented in research and educational activities, which confirms their theoretical and practical significance.

The practical approbation and implementation of the research results have been carried out in a number of scientific and educational institutions and are supported by official implementation acts. In particular, the results have been applied in the research activities of the National Defense University of the Republic of Kazakhstan, where they are used for the analysis of textual information, including the study of semantic structures of texts and the modeling of semantic representations in tasks of intelligent data processing.

The developed methods have been implemented in the educational process of Karaganda University of Kazpotrebsoyuz and are used in teaching, as well as in course and diploma projects in the fields related to data analysis and artificial intelligence.

Certain results of the study are applied in the research activities of the National Research University Higher School of Economics for solving problems related to semantic representation analysis and text data processing.

As a result of the dissertation research, a copyright certificate has been obtained — Certificate of Entry into the State Register of Rights to Copyright-Protected Objects No. 66441 dated January 19, 2026, titled “Analysis of Intrinsic Dimensionality of Semantic Spaces of Natural Languages.”

Personal Contribution of the Applicant. The main results of the dissertation were obtained with the direct participation of the applicant, who developed the methods and algorithms and prepared the corresponding scientific publications.

Structure of the Dissertation. This dissertation covers the full scope and structure of the research. The dissertation consists of an introduction, four chapters, a conclusion, a bibliography, and appendices.

The introduction substantiates the relevance of the chosen topic, formulates the research aim and objectives, and defines the object and subject of the study. The section presents the results submitted for defense, their scientific novelty, and practical significance. Data on the approbation of the main research results are also provided.

The first chapter presents a comprehensive review of the literature on natural language as a complex system, including its fractal structures and methods for identifying bot-generated texts. The chapter examines various aspects of the interrelation between language, semantics, and chaos, with a focus on key studies and methodological approaches.

Within the natural science paradigm, natural language is studied as a self-organized critical system. In this context, power-law regularities observed in texts are analyzed, along with psychophysiological and cognitive factors influencing their emergence. Issues related to the chaotic behavior and complexity of semantic trajectories in texts are also addressed. Existing studies, quantitative regularities, and principles governing language processes are examined, as well as methods for assessing chaotic behavior based on permutation entropy and other statistical approaches.

A significant aspect of the chapter is the investigation of linguistic fractal structures and methods for estimating their intrinsic dimensionality in natural languages. A systematic analysis of modern approaches is provided, including those based on strange attractors, graph-based representations, and persistent homology. In addition, studies on the identification of automatically generated texts are reviewed. Various approaches are examined, ranging from the analysis of simple textual features to advanced methods based on semantic spaces and the emotional characteristics of texts.

The second chapter presents a set of methods for analyzing the large-scale structure of natural language based on a representative corpus of 52 languages from 18 language families. It is demonstrated that natural languages can be considered as self-organized critical systems characterized by power-law distribution regularities. An approach to modeling text as a semantic trajectory in the embedding space is proposed, which makes it possible to establish the chaotic nature of linguistic data based on the analysis of entropy and complexity. Methods for estimating the intrinsic dimensionality of linguistic structures, as well as methods of topological analysis of semantic space, have been developed, enabling the identification of their fractal and topological properties, including stable homological structures (“holes”) of language. Based on the set of extracted features, methods for identifying bot-generated texts have been proposed, demonstrating high effectiveness when using machine learning algorithms.

The third chapter presents the results of applying the developed methods to the analysis of natural languages. Based on the study of 52 languages from 18 language families, it is established that natural languages exhibit power-law distribution regularities, which confirms their nature as self-organized critical systems, whereas Esperanto constitutes an exception, demonstrating a Gaussian distribution. The analysis of semantic trajectories confirms the chaotic nature of languages, with the vast majority exhibiting chaotic properties. Cluster analysis identifies typological groups correlated with entropy and complexity parameters and substantiates a number of linguistic hypotheses, including areal and typological regularities. The obtained estimates of intrinsic dimensionality indicate the multifractal nature of languages and their robustness with respect to the choice of methods, which allows intrinsic dimensionality to be considered an invariant characteristic of language.

Methods of topological data analysis enable the identification of stable homological structures in semantic space (“holes”), reflecting deep semantic constraints of language systems. Based on the distances of linguistic units to these structures, an approach for identifying bot-generated texts is proposed. Experimental results demonstrate statistically significant differences between human-written and bot-generated texts and confirm high classification performance.

The fourth chapter presents the results of solving the problem of identifying bot-generated texts based on the developed methods. In the course of large-scale computational experiments, it is established that classification performance significantly depends on the choice of embedding parameters and the characteristics of semantic trajectories, while the optimal parameters vary across languages. It is demonstrated that features based on entropy and complexity analysis, as well as n-gram clustering, enable effective discrimination between human-written texts and texts generated by automated systems.

It is further established that the best performance is achieved through the use of different combinations of features and classifiers for different languages, which indicates the absence of a universal model. At the same time, high classification accuracy is achieved (F1-score exceeding 0.9 for a number of languages), including cases where the test data consist of texts generated by models not used during training. It is shown that the proposed approach ensures robust identification of bot-generated texts and exhibits strong generalization capability, thereby confirming its applicability to tasks of digital content analysis and filtering.

The conclusion summarizes the main research results and presents findings on the scientific and practical significance of the work.

The appendices contain practical materials from the research.

The author expresses sincere gratitude to the scientific supervisor, Candidate of Technical Sciences, Associate Professor Ildar G. Kurmashev, as well as to the foreign scientific supervisor, Doctor of Physical and Mathematical Sciences, Professor Vasily A. Gromov, for their valuable scientific guidance, formulation of research tasks, and continuous support throughout the course of the dissertation research.