UDC 811.161'32]:004.032.26:004.055:004.423.46
DOI https://doi.org/10.24919/2308-4863/74-1-32

Oleksii DOVHAN,
orcid.org/0000-0002-6728-818X
PhD in Philology,
Doctoral student at the Department of Slavic, Romance and Oriental Languages
Mykhailo Dragomanov Ukrainian State University
(Kyiv, Ukraine) oleksiivdovhan@gmail.com

WORKING WITH LINGUISTIC MARKERS AND ITS ROLE IN TEXT ANALYSIS BY ARTIFICIAL NEURAL NETWORK CONCERNING CLEANROOM METHOD

The article argues that the development of the cleanroom method concerning modern linguistic research produces its further genesis, providing new opportunities, algorithms, vision, etc. of the usual processes (analysis, processing, representation, etc. of textual data). It is emphasized that training neural network models according to the above methodology contributes to the understanding of the peculiarities of the existence, genesis, and other features of the language poly system, which results in the deepening of the modern innovative methodology of working with data.

Accordingly, the use of the above makes it possible to localize, track the dynamics, analyze, etc., language poly systems with different degrees of distribution, which produces the identification of gaps in traditional methods of modeling the latter. An important advantage of updating the cleanroom method is the ability to identify and avoid possible fluctuations of sense arising from the influence of cultural, social, etc. perceptions on the modeling process.

Thus, the actualization of this methodology opens up several new perspectives for the study of linguistic phenomena, the dynamics of language functioning, the development of language technologies, and so on. This is because the cleanroom method contributes to ensuring the objectivity, accuracy, validity, representativeness, etc. of modern linguistic research. In turn, this becomes an important step in the organic development of modern science (in particular, linguistics), producing a change in its approaches, methods, etc. without losing its traditional achievements.

Thus, the prospect of applying the cleanroom method to work with artificial neural networks is to improve existing models that are productive for use in a) language technologies (speech recognition, machine translation, etc.); b) distance education (the latter is extremely relevant due to the russian-Ukrainian war, which has forced some pupils and students to study remotely at home or even in shelters); c) language analysis (in fact, the entire innovation block that brings modern humanities research, which is common for Ukrainians, closer to digital humanities, which is common for Europeans), etc.

Key words: *sense, neural network modeling, linguistic markers, artificial neural networks, cleanroom method.*

Олексій ДОВГАНЬ,
orcid.org/0000-0002-6728-818X
кандидат філологічних наук,
докторант кафедри слов'янських, романських і східних мов
Українського державного університету імені Михайла Драгоманова
(Київ, Україна) oleksiivdovhan@gmail.com

РОБОТА З МОВНИМИ МАРКЕРАМИ ТА ЇЇ РОЛЬ В АНАЛІЗІ ТЕКСТУ ШТУЧНОЮ НЕЙРОННОЮ МЕРЕЖЕЮ ЩОДО МЕТОДУ «ЧИСТОЇ КІМНАТИ»

У статті стверджується, що розвиток методу «чистої кімнати» щодо сучасних лінгвістичних досліджень продукує їх подальшу генезу, забезпечуючи нові можливості, алгоритми, бачення тощо узвичаєних процесів (аналізу, обробки, репрезентації тощо текстових даних). Підкреслено, що навчання нейромережевих моделей за вищезазначеним методом сприяє розумінню особливостей побутування, генези та іншого мовної полісистеми, наслідком чого стає поглиблення сучасної інноваційної методології роботи з даними.

Відповідно, використання вищезазначеного дозволяє локалізувати, відстежити динаміку, проаналізувати тощо мовні полісистеми з різним ступенем поширення, що продукує виявлення лакун традиційних методів моделювання останніх. Важливою перевагою актуалізації методу «чистої кімнати» є можливість виявлення та уникнення можливих флуктуацій смислу, які виникають через вплив культурних, суспільних тощо уявлянь на процес моделювання.

Отже, актуалізація цього методу відкриває низку нових перспектив для дослідження мовних явищ, динаміки функціонування мов, розвитку мовних технологій та іншого. Це пов'язано з тим, що метод «чистої кімнати» сприяє забезпеченню об'єктивності, точності, валідності, репрезентативності тощо сучасних мовних досліджень. Своєю чергою, це стає важливим кроком органічного розвитку сучасної науки (зокрема, лінгвістики), продукуючи зміну її підходів, методів та іншого без втрати традиційних надбань.

Таким чином, перспективою застосування методу «чистої кімнати» щодо роботи з штучними нейронними мережами є удосконалення наявних моделей, які продуктивні для застосування у: а) мовних технологіях (розпізнавання мови, машинний переклад тощо); б) дистанційній освіті (остання є вкрай актуальною через російсько-українську війну, через яку частина учнів та студентів змушені навчатися у віддаленому форматі вдома, або навіть в укриттях); в) мовному аналізі (власне, весь інноваційний блок, який наближає сучасні гуманітарні дослідження, узвичаєні для українців, до цифрових гуманітарних наук, звичних для європейців) тощо.

Ключові слова: *смисл, нейромережеве моделювання, мовні маркери, штучні нейронні мережі, метод «чистої кімнати».*

Statement of the problem in general and its connection with important scientific or practical tasks. The modifications of modern linguistic science under the influence of digitalization processes are manifested, first of all, in the transformations of approaches to the study of certain phenomena of the language poly system. In turn, the above-mentioned changes entail the need to use new tools, corresponding changes (adaptation) of the methodology, and the original genesis of the basic concept of sources (e.g., artificial neural networks (ANNs)) of such research, understanding of its parameterization, etc.

At the same time, it is natural that the actualization of innovative tools leads to a restructuring of the entire algorithm of the above work. Thus, if we choose the aforementioned artificial neural networks (ANNs) as the core tool for the latter, which are mostly used here to operate with the probabilities of occurrence of certain words, phrases, sentences, etc. in a given context through analysis, then the entire range of their actualization should be taken into account. We are talking about parameterization features: *type* (deep neural networks, recurrent neural networks, etc.), *training* (deep and machine learning) and *retraining* (if necessary), *analysis of results*, *correction* (if necessary), and *obtaining results*. The latter should be pre-validated and represented in a certain way, and, of course, reliable and reproducible.

The most common application of artificial neural networks (ANNs) is predicting the occurrence of the next element (the above probabilities) in a sentence, which an average person can do using available data sets: background knowledge, language national worldview, etc. In the case of updating this toolkit, we are talking about working at the level of certain words without regard to previous configurations. Thus, we are talking about updating the frequency of use (e.g., the unigram model), or taking into account the above combinations (e.g., the n-gram language model).

Characteristically, in the second case, large language models are mostly used: *character language model* (predicts occurrences by character) and *statistical language model* (calculates the probability of occurrence of certain word combinations based on large corpora). The latter is related to their representativeness and productivity in the context of textual analysis, in particular, working with senses in different environments (including political Internet discourse) (Zhao, 2024).

The above is the parameterization features of updating innovative tools: a statistical language model can be created using several machine learning algorithms. Instead, *the n-gram language model* calculates the probability of occurrence of a certain word relative to N previous words, *the Markov chain* updates the probability of occurrence of certain words relative to those analyzed earlier, and *the recurrent neural network (RNN)* localizes the chance of occurrence of a lexical unit by taking into account the entire series of previous ones, etc. Such a wide range of applications allows us to assert the productivity of large language models in natural language processing. First of all, we are talking about generation (e.g., writing a text of a certain style, topic, etc.), processing (e.g., reading and paraphrasing, answering questions about the content, etc.), and representation of textual data (e.g., the type of information presented: texts, graphs, tables, drawings, etc.).

Naturally, the core of the above-mentioned text data processing (neural network modeling of linguistic units) is the consideration of *linguistic markers*. In this context, they are words, phrases, or distinctive structural features of sentences that are representative of the context. For example, the frequency of use of a certain construction (word, phrase, etc.) can be a sign of text data generation using an artificial neural network (ANN). Also, the use of certain words, phrases, sentences, etc. may be a sign of the actualization of a particular communication strategy or tactic. In particular, it may be the actual repetition of a question without a direct answer to it, indicating the importance of the problem and emphasizing the need to solve it (manipulative techniques common in political discourse and political Internet discourse).

Thus, the above shows the special role of these markers in the process of analyzing textual data with an artificial neural network (ANN). The point is that linguistic markers are pivotal in comprehending the specifics of sense construction in such data. First of all, their role is determined by actualizing the identity of semantics, syntax, etc. (in turn, working with context, etc.). For example, the actualization of word sequences, and the specifics of the relationship between them with the extrapolation of the frequency of use of certain constructions is productive in the case of recurrent neural networks. Whereas deep neural networks calculate the peculiarities of semantic relations between words, localizing contextual specifics, etc., allowing one to comprehend the actualized senses and their parameterization.

Therefore, the use of linguistic markers for textual analysis in the context of neural network modeling will significantly increase the efficiency, accuracy, representativeness, predictability, reproducibility, etc. of its results. First of all, the above correlates with the specifics of textual data analysis (syntax, semantics, etc.), which, in turn, allows us to position language markers as pivotal in the neural network modeling of linguistic units. This is related to the development of more efficient, accurate, valid, etc. natural language processing systems.

Analysis of the latest research and publications that initiated the solution of this problem and on which the author relies. The problem of the originality of work with linguistic markers and its role in the analysis of artificial neural network (ANN) text about the cleanroom method (CM) is complex. The aforementioned property of the latter leads to the poly-interpretability of its research on the dynamics, specifics of deployment, possible empirical realizations, etc. in several works (mostly interdisciplinary, related to the digital humanities). These works are characterized by a special discursiveness, which, in turn, constitutes the specificity of the analyzed issues.

First of all, the aforementioned property is determined by the production of an original process of actualization in terms of linguistic analysis, machine, and deep learning, data science, as well as reverse engineering, etc. In turn, this suggests the need to use modern system optimization methodology in the context of linguistic research. That is, we are talking about the deployment of such research within the framework of formal methods, incremental implementation in terms of statistical quality control, and statistical testing.

The study of the specifics of the chatbot functioning designed to simulate human conversation and answer user queries is presented in the study by T. Goh et al. (Goh, 2024). In the analyzed study, the authors point out the inefficiency of the vast majority of such bots, arguing that they can only answer predefined questions. The researchers emphasize that the purpose of the paper is to develop and create a chatbot for the Faculty of Information Science & Technology (FIST) of Multimedia University that facilitates the study life of FIST students. The researchers position the originality of the chatbot they have developed in that they use the methods of the aforementioned natural language processing to obtain input data that can be used to train a neural network model (a multilayer perceptron model). We are talking about tokenization, lemmatization, and the "bag of words" model. First of all, this allows the chatbot to understand the intentions of users by analyzing their questions, which, in turn, makes it possible to process a wide range of student requests regarding the above-mentioned faculty.

The study of the possibilities of automating sign language (a visual-manual modality designed to exchange data using gestures, postures, and movements) through the implementation of systems that can convert it into text is presented in the study by V. Durdi et al. (Durdi, 2024). The authors note that automation of sign language helps to overcome the communication barrier between people with disabilities and others. The researchers emphasize that their attention was focused on a hand gesture recognition system that has an accessible, understandable method of interacting with a computer in real-time. The system was implemented in the Jupyter Notebook environment using convolutional neural networks (CNN), which showed good results for each of the studied datasets.

The detection of hate speech in social networks in the example of X is presented in the study by I. Abasan, E. Setiawan (Abasan & Setiawan, 2024), in which the authors present the developed system for localizing the above data. The basis for this system is hybrid deep learning (HDL) and solo deep learning (SDL) with the mentioned convolutional neural networks (CNN) and bidirectional gated recurrent unit (Bi-GRU) algorithm. The researchers have built four artificial neural network models (CNN, Bi-GRU, CNN+Bi-GRU, and Bi-GRU+CNN), and the term frequency-inverse document frequency (TF-IDF) has been updated to extract the analyzed linguistic features, while FastText is used to expand the features (minimize irrelevant lexis).

The specifics of natural language processing (in particular, sentiment analysis) in terms of updating deep and machine learning algorithms in the context of Monkeypox analysis of Arabic-language text data are presented in the study by H. Gharaibeh et al. (Gharaibeh, Mamlook, Emhamed, Samara, Nasayreh, Smadi, ... & Abualigah, 2024). The authors note that their goal was to develop an accurate, efficient neural network model capable of classifying the above data into sentiment categories. The scientists emphasize that they used machine learning algorithms (Support Vector Machines, Naive

Bayes, and Random Forest), as well as deep learning methods (recurrent neural networks and transformer models). In addition, the researchers believe that the actualization of hyperparameter optimization methods used to fine-tune the above models is productive. According to the authors, hyperparametric optimization helped to select the optimal configuration of neural network models, which, in turn, increased the accuracy of their results and reduced the need for retraining.

The originality and complexity of the analysis of vision-based deep neural networks are discussed in the study by R. Mangal et al. (Mangal, Narodytska, Gopinath, Hu, Roy, Jha, & Păsăreanu, 2024), in which the authors propose to use new emerging multimodal, vision-language, foundation models (VLMs) as a basis for studying the prospects of technical vision models. The researchers describe the logical language of Conspec specifications, which is designed to facilitate their writing in updated terms and was tested using a map built by the researchers. The specificity of the latter was to actualize the connections between the internal representations of a given vision model and the aforementioned multimodal, vision-language, foundation model. This approach, according to the authors, was productive in verifying the properties of natural language in the context of this type of neural network model.

The problems of the above deep learning are discussed in the study by F. Aguirre et al. (Aguirre, Sebastian, Le Gallo Song, Wang, Yang, ... & Lanza, 2024), which presents the latest approaches to the creation of hardware-based memristive artificial neural networks (HBMANNs). The authors describe in detail the parameterization, principles of operation, design options, etc., noting the peculiarities of their actualization. The researchers focus on the process of data validation and the tools required to accurately assess the performance metrics of the above models. The researchers emphasize that they aimed to present a comprehensive overview of materials and methods related to memristive artificial neural networks (MANNs).

The problem of the existence of large language models (LLMs) about recommender systems (RS) is presented in the study by Z. Zheng et al. (Zheng, Chao, Qiu, Zhu, & Xiong, 2024), where the cornerstone of the use of the former is presented – the need for rich textual information to achieve the proper result (in particular, displaying the historical sequence of user behavior). The authors see the solution to the above problem in the framework they have developed for using large language models for text-rich sequential recommendation (LLM-TRSR). The researchers propose to segment the above data (historical user behavior) first, and then use an LLM-based summarizer to summarize these user behavior blocks. This approach has allowed us to successfully apply convolutional neural networks (CNNs) and recurrent neural networks (RNNs) models in modeling user behavior, introducing two unique summarization techniques: respectively hierarchical summarization and recurrent summarization.

The study of L. Messeri and M. Crockett (Messeri, & Crockett, 2024) is devoted to the use of innovative tools in modern scientific research, in which the authors reflect on the genesis of scientific activity through the prism of digitalization changes. The scientists note that a potential threat to the modern scientific community is the formation of monocultures with an established set of techniques and methods, regardless of possible alternative approaches. According to the researchers, the above-mentioned trend produces a loss of innovation in scientific activity and a tendency to make mistakes. The latter is associated with the use of artificial intelligence tools, which, in their opinion, will produce an increase in the volume of existing research without a deep understanding of the cause-and-effect relations between them.

The correlations of deep neural network (DNN) operation with human cognitive and conceptual processing in terms of the impact on the latter language poly systems are discussed in the study by P. Nguyen, M. Henningsen-Schomers, F. Pulvermüller (Nguyen, Henningsen-Schomers, & Pulvermüller, 2024). In the analyzed study, the authors use a brain-constrained deep neural network model of category formation and symbol learning, with the help of which scientists study the specifics of its internal mechanisms at the neural circuit level. The researchers claim that in one series of simulations, the neural network model was given similar patterns of neural activity indexing instances of objects and actions that belong to the same categories. The authors emphasize that biologically realistic Hebbian learning produced the formation of specific neurons (the latter were distributed in different parts of the network) and cellular assembly patterns of "common" neurons (responding to all category instances), which are network correlates of conceptual categories. It is noteworthy that in two separate sets of simulations, the artificial neural network (ANN) studied the same patterns along with symbols for individual instances ("proper names" (PN)) or symbols that correlate with classes of instances that share common features ("category terms" (CT)).

M. Carreiras et al. (Carreiras, Quiñones, Chen, Vázquez-Araujo, Small, & Frost, 2024) continue to investigate the connection of an individual's primary experience in the context of the specifics of chemosensory processing and/or modulating how sensory data correlate with the semantic network (SM). Using a combination of diffusion-weighted images and fMRI (activation and connectivity), the authors analyzed the brain's response to wine tasting between a professional (sommelier) and ordinary connoisseurs. Scientists focused on sensory neural circuits responsible for taste and/or those responsible for language and memory. Using mediation analysis, the researchers demonstrated a correlation between fractional anisotropy and behavior, noting the dependence of the former on the latter's features. The authors found functional differences between sommeliers and ordinary connoisseurs that actualize the sensory chain of taste and areas involved in semantic operations. The former is representative of the ability to differentiate sensory processing, while the latter is representative of the professional's ability to actualize relevant sensory data and transform them into complex verbal descriptions. The enhanced synchronization between the above seemingly independent circuits indicates that the sommelier combines these descriptions with previous semantic data to optimize his/her ability to distinguish subtle differences in the characteristics of the drink (wine).

The specifics of the development of computational neural network models of visual and linguistic processes (image description) are covered in the study by E. Takmaz (Takmaz, 2024), in which the author bases his work on several cognitive sciences and psycholinguistics. The scientist aims to present the correlation between non-linguistic modalities and the language poly system concerning artificial neural networks (multimodal natural language processing). The researcher organizes his work in two parts: a) modeling human gaze in language use (production and comprehension), where he focuses on the genesis of neural network models for generating image descriptions using eye tracking data, estimates variations in human signals during the above activity, and predicts human reading behavior based on the above eye tracking; b) modeling communication strategies in referential tasks in visually grounded dialogue, where it builds neural network models for quantifying, generating, recognizing, and adapting utterances in referential tasks localized in visual and conversational contexts (visuo-linguistic processes). In turn, the results obtained are relevant for the development of research on cross-modal processes (in particular, the actualization of artificial neural networks (ANNs)) and vice versa.

The originality of sociocultural paradigms and their correlation with the genesis of artificial neural networks (ANNs) is presented in the study by L. Zhao (Zhao, 2024), in which the author analyzes innovations and improvements in painting techniques of generative adversarial network (GAN). The scientist focuses on the one-dimensional midpoint substitution method and the dichotomous method to generate the rock outline in creating a painting, combining the above with the parameterization of the generative adversarial network (GAN). The latter is used by the researcher to localize the style migration model of modern painting creation techniques and morphological language.

Thus, the analysis of the historiography on the subject of the article has shown the originality of the work with linguistic markers and its role in the analysis of artificial neural networks (ANNs) text and revealed several gaps in existing research. In particular, despite the pivotal role of this process, we note the lack of a comprehensive study of its use in terms of analysis, processing, and representation in the context of the digital humanities. In addition, there is a noticeable tendency to refuse to use innovative tools due to ethical norms, prejudices, etc., which can be considered a kind of sinister valley effect. The above naturally leads to the focus of our research on filling these gaps, in particular, on studying the peculiarities of actualizing linguistic markers in the context of text analysis and the cleanroom method (CM) to avoid apophaticizing the outlined issues.

Identification of previously unresolved parts of the general problem to which this article is devoted. The exponential development of information technologies (in particular, artificial neural networks (ANNs)) has led to their active involvement in conducting, planning, and validating, etc. scientific research. In the context of the above, modern linguistic science is marked by a certain fluidity of its boundaries: the latter are mostly associated and localized in the context of the digital humanities. In turn, this leads to the integration of such works, and their focus on grants, research projects, etc. Despite the above, the selection, phasing, and relevance of certain tools remain debatable, as the selection of approaches, methods, tools, etc. is open under such conditions.

We see the way out of this situation in the actualization of the cleanroom method (CM) (Clean Room Design, 2024), which we will discuss in detail below, which, in turn, will "remove" some of the problems associated with the use of the above-mentioned artificial neural networks (ANNs). First and foremost, we are talking about the accuracy, objectivity, representativeness, validity, etc. of their results, which, in turn, will improve natural language processing (in particular, sentiment analysis), machine translation, generation, analysis, and visualization of text data, its recognition, etc.

Formulation of the article's objectives (statement of the task). *The purpose* of the article is to consider the originality of the work on linguistic markers and its role in the analysis of artificial neural networks (ANNs). *The subject* is the specifics of the above-mentioned work on the cleanroom method (CM) as an innovative tool of linguistic science. In turn, the aforementioned purpose and subject of the study allowed us to formulate its *tasks:*

1. To highlight the theoretical achievements of neural network modeling of linguistic units (text data analysis) and extrapolate their specificity to work with linguistic markers.
2. To analyze the algorithms, methods, approaches, etc. of text analysis using artificial neural networks (ANNs) and to highlight its features.
3. To present the work with linguistic markers as a determinant of neural network modeling of linguistic units concerning the cleanroom method (CM) and outline the prospects of its use for modern linguistic research (including within the digital humanities).

Presentation of the main research material with full justification of the scientific results obtained. Today, digitalization processes are becoming increasingly widespread: e-government, the concept of the state in a smartphone, etc. are being actively developed on a Ukrainian basis. The latter, in the first place, is becoming increasingly important for modern linguistic research, which is relevant in several areas (deep and machine learning, machine translation, etc.). At the same time, a specific feature of social services commissioning for research in general (including linguistic research) is its practical orientation (recall the aforementioned digital humanities) (Durdi, 2024).

Notably, modern information technologies are characterized by a special specificity: they are integrated by nature, as they are focused on user requests rather than on a specific subject area (which was typical for many research studies). Thus, they are not limited, which naturally leads to a change in modern scientific research: the latter is increasingly integrated, telling a holistic story of the analyzed issue, not limited to a specialty. In this regard, artificial neural networks (ANNs) are illustrative, as their work seems to be concentrated around a certain functionality: for example, it is convenient for linguists and other specialists to divide them into certain updated types, which, in turn, are productive in working with data types. For example: the aforementioned deep neural networks, recurrent neural networks, etc., which are commonly used to work with text data, analyze it, and so on.

However, the very functioning of such networks demonstrates the need for openness about the tools used by modern researchers. For example, on April 18 this year, Meta Platforms released early versions of the Llama 3 large language model and image generator. It seems clear that such a model is of interest to linguists and can be used in certain studies, but this article is about something else. We are talking about the specifics of the aforementioned image generator, which, like the Google search engine, produces query results in the process of formulating and correcting them. The latter demonstrates a powerful system of language data processing, which, in our usual approach, would not have been actualized, since its parameterization does not seem to include our subject field (Gharaibeh, Mamlook, Emhamed, Samara, Nasayreh, Smadi, ... & Abualigah, 2024).

In turn, there is a certain paradox here: no science can cover all the options for interpreting the use of innovative tools, which, moreover, are constantly changing. The solution to this situation, in our opinion, is a seeming departure from the subject field, which is carried out to return it. Thus, the software development methodology comes to the aid of modern linguists who aim to effectively, representatively, qualitatively, etc. actualize artificial neural networks (ANNs) (in particular, neural network modeling of linguistic units) in their research (Aguirre, Sebastian, Le Gallo, Song, Wang, Yang, ... & Lanza, 2024).

We are talking about the use of a specific cleanroom method (CM) (Overview of Clean Room Software Engineering, 2024), which is designed to eliminate several possible errors, shortcomings, etc. in the process of working with artificial neural networks (ANNs). The essence of the aforementioned method is strict inspection, the purpose of which is to build the correct process of interaction between the developer (researcher) and his product (in this case, an artificial neural network (ANN)). Of course, the use of the latter has a certain algorithm, specifics, etc., which are not the purpose of this research paper (but are reflected in many previous ones).

The cleanroom method (CM) is an element of software reverse engineering, which is a special approach to development aimed at creating a specific software product. The main difference between this approach and the conventional ones is that Quality Assurance (QA) is updated at each stage of development, while in the standard approach it is used only at the end. The peculiarity of the proposed approach is also the isolation of the developer from the background knowledge of the internal workings of the system or the updating of certain tools (Clean Room Design, 2024). Such isolation is intended to provide a “clean” result, not contaminated by the influence of the external environment (in this case, the noosphere), which is marked by the dynamics of human-mind interaction.

In the context of linguistics, the above method means avoiding the use and consideration of existing neural network models, algorithms, frameworks, datasets, etc., as well as knowledge about the language poly system. Researchers fill in the gaps by developing neural network models based solely on the input data for analysis, processing, and representation. This avoids the influence of previous work experience on the actual preparation, course, and results of the study, which could potentially distort them (Zheng, Chao, Qiu, Zhu, & Xiong, 2024).

We consider the cleanroom method (CM) to be the most productive for training neural network models designed to perform certain specific tasks (text classification, machine translation, speech recognition, etc.). In this case, only the data from the analyzed language poly system in the current state, i.e. without taking into account diachrony, is updated in the process of training these models. Thus, the stages of using the cleanroom method (CM) in linguistics may look like this (see Fig. 1).

Let’s take a closer look at the milestones outlined in Figure 1: yes, they are pivotal in the development of the above cleanroom method (CM). In particular, they are:

1. *A formal specification* allows you to capture the current state of the product using a state model that reflects its responses to the stimuli used.

2. *Step-by-step development* consists of localizing the stages of the entire process of working with the neural network model, each of which is verified by the cleanroom method (CM) independently for each such element.

3. *Structural programming* consists of a clearly defined number of control structures and abstractions of the analyzed text data (e.g., in Python).

4. *Static verification* – checking the neural network model by the static method of strict inspection (it is important that testing is not performed for modules or separate elements (e.g., individual network layers)).

5. *Static testing* is carried out using statistical methods that allow to assess the reliability of the neural network model. As shown in Figure 1, statistical tests are based on an operational profile, which should be developed in parallel with the creation of a neural network model specification.

Thus, the actualization of the cleanroom method (CM) for training neural network models designed to perform certain specific tasks (text classification, machine translation, speech recognition, etc.) is planned in such a way as to ensure strict inspection of the work and results. Instead, the specification of a neural network model is represented by a state model, which gradually acquires the desired parameterization through a series of successive models. It is noteworthy that the aforementioned inspection of the neural network model should be combined with a clear vision of the causal relations since retraining will take time and lead to the localization of the solution to a particular natural language processing task (Takmaz, 2024).

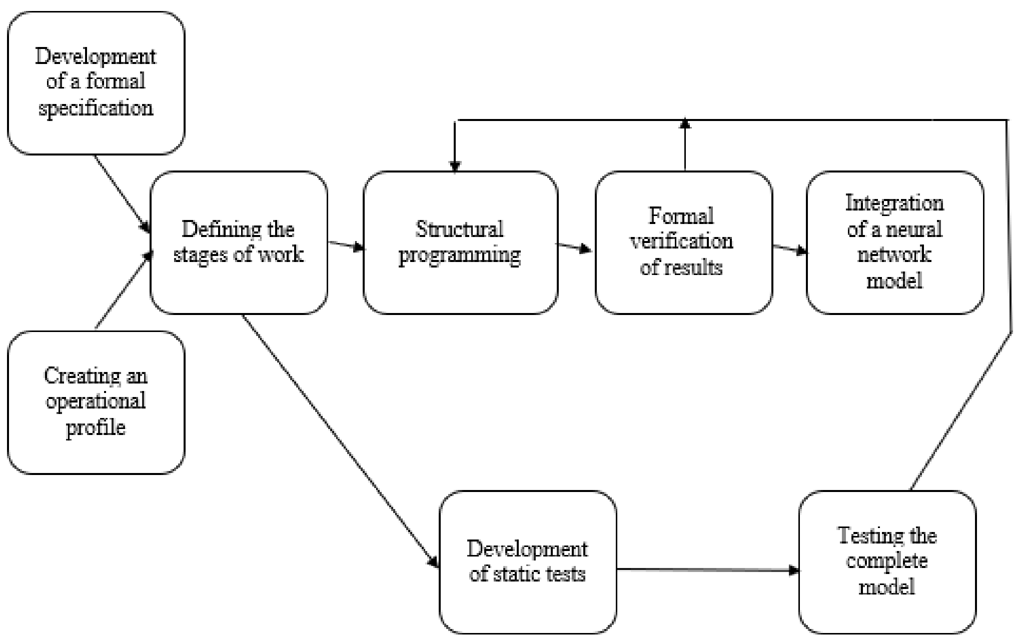

Fig. 1. The process of updating the cleanroom method in linguistics

At the same time, updating the aforementioned methodology will avoid many mistakes, significantly reduce time, etc., and optimize linguistic research. For example, thanks to the use of the cleanroom method (CM), most defects, learning errors, etc. are detected during the development process. First of all, this will save the researcher time for testing and correcting errors (e.g., the aforementioned relearning). Thus, the use of this method in modern linguistic research on neural network modeling of linguistic units, text analysis, etc. correlates with the quality, accuracy, and validity of the results of this process. In turn, this approach will ensure the objectivity, representativeness, speed, etc. of neural network models, which is crucial for their actualization in linguistic research.

Conclusions from this study and prospects for further research in this area. Thus, the development of the cleanroom method (CM) produces further genesis of linguistic research, providing new opportunities, algorithms, visions, etc. regarding the conventional processes (analysis, processing, representation, etc. of textual data). Also, training neural network models using the cleanroom method (CM) contributes to the understanding of the peculiarities of the existence, genesis, and other features of the language poly system, which results in the deepening of the modern innovative methodology of working with data.

So, the use of the above-mentioned approach allows us to localize, track the dynamics, analyze, etc. language polysystems with different degrees of distribution, which produces the identification of gaps in traditional methods of modeling the latter. An important advantage of updating the cleanroom method (CM) is the ability to identify and avoid possible fluctuations of sense that arise due to the influence of cultural, social, etc. perceptions on the modeling process. For example, this may include the spread of stereotypes, prejudices, etc. that the neural network model learned during its training and that were not taken into account by the researcher. For the sake of objective coverage of the problem, we note that any methodology (including the cleanroom method (CM)) is not universal, requiring flexibility of approach. That is why, naturally, the best results for modern linguists will come from the integration of the tools they use, where the above method is combined with others to obtain comprehensive and objective results of neural network modeling.

Nevertheless, the actualization of this method opens up many new perspectives for the study of linguistic phenomena, the dynamics of language functioning, the development of language technologies, and so on. This is because the cleanroom method (CM) contributes to ensuring the objectivity, accuracy, validity, representativeness, etc. of modern linguistic research. In turn, this becomes an important step in the organic development of modern science (in particular, linguistics), producing changes in its approaches, methods, etc. without losing its traditional achievements.

Fundamentally unresolved aspects of the functioning of the approach in modern linguistics are:

1. The impossibility of developing an optimal way to isolate researchers from the existing achievements in the field in general (in particular, in the analyzed problem).

2. The problem of datasets, which each researcher solves individually, putting certain aspects of the problem under study as the cornerstone (accordingly, we can talk about representativeness only in the context of the spectrum of the problem's reflection in the data).

3. The fundamental openness of the issue of the influence of input data on the work of the neural network model, which is related to the previous aspect): e.g., stereotypes in datasets, etc. are not always the result of researchers' miscalculations, as they can be part of the problem under analysis (in particular, in the case of studying expressive texts, etc.).

4. The appropriateness of updated metrics for assessing the quality of models can also be positioned spectrally, since some standard metrics (e.g., accuracy or training losses) are not always representative of the actual performance of a neural network model, etc.

The perspective of applying the cleanroom method (CM) to work with artificial neural networks (ANNs) is to improve existing models that are productive for use in a) *language technologies* (speech recognition, machine translation, etc.); b) *distance education* (the latter is extremely relevant due to the russian-Ukrainian war, which has forced some pupils and students to study remotely at home or even in shelters); c) *language analysis* (in fact, the entire innovation block that brings modern humanities research, which is common for Ukrainians, closer to digital humanities, which is common for Europeans), etc.

BIBLIOGRAPHY

1. A Campus-based Chatbot System using Natural Language Processing and Neural Network / T. J. Goh et al. *Journal of Informatics and Web Engineering*. 2024. Volume 3, № 1. P. 96–116. https://doi.org/10.33093/jiwe.2024.3.1.7 *Journal of Informatics and Web Engineering* : website. URL: https://goo.su/gVj0l (date of application: 30.04.24).

2. A Novel Method for Recognizing Hand Gestured Sign Language Using the Stochastic Gradient Descent Algorithm and Convolutional Neural Network Techniques / V. B. Durdi et al. *International Journal of Intelligent Systems and Applications*

in Engineering. 2024. Volume 12, № 7. P. 529–538. *International Journal of Intelligent Systems and Applications in Engineering* : website. URL: https://goo.su/U1rw8Sy (date of application: 30.04.24).
3. Abasan I. G. B. J., Setiawan E. B. Empowering hate speech detection: leveraging linguistic richness and deep learning. *Bulletin of Electrical Engineering and Informatics*. 2024. Volume 13, Issue 2. P. 1371–1382. https://doi.org/10.11591/eei.v13i2.6938 *Bulletin of Electrical Engineering and Informatics* : website. URL: https://goo.su/yj65bmp (date of application: 30.04.24).
4. Arabic sentiment analysis of Monkeypox using deep neural network and optimized hyperparameters of machine learning algorithms / H. Gharaibeh et al. *Social Network Analysis and Mining*. 2024. Volume 14, Article 30. P. 1–18. https://doi.org/10.1007/s13278-023-01188-4 *Springer Link* : website. URL: https://goo.su/0BUPN (date of application: 30.04.24).
5. Clean Room Design. *DevX* : technology glossary. website. URL: https://goo.su/P3gvV (date of application: 30.04.24).
6. Formal Concept-based Analysis of Neural Networks via Vision-Language Models / R. Mangal et al. *arXiv* : website. URL: https://goo.su/Bo6t6O1 (date of application: 30.04.24).
7. Hardware implementation of memristor-based artificial neural networks / F. Aguirre et al. *Nature Communications*. 2024. Volume 15. Issue 1. https://doi.org/10.1038/s41467-024-45670-9 *Nature Communications* : website. URL: https://goo.su/8tEZgl9 (date of application: 30.04.24).
8. Harnessing Large Language Models for Text-Rich Sequential Recommendation / Z. Zheng et al. *arXiv* : website. URL: https://goo.su/F1bqd (date of application: 30.04.24).
9. Messeri L., Crockett M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature.* 49–58. https://doi.org/10.1038/s41586-024-07146-0 *Nature* : website. URL: https://goo.su/5p8LhvD (date of application: 30.04.24).
10. Nguyen P. T., Henningsen-Schomers M. R., Pulvermüller F. Causal influence of linguistic learning on perceptual and conceptual processing : A brain-constrained deep neural network study of proper names and category terms. *Journal of Neuroscience*. 2024. Volume 44, Issue 9. https://doi.org/10.1523/JNEUROSCI.1048-23.2023 *The Journal of Neuroscience* : website. URL: https://goo.su/DwRlyA8 (date of application: 30.04.24).
11.Overview of Clean Room Software Engineering. *GeekForGeeks* : website. URL: https://goo.su/glOsLQ6 (date of application: 30.04.24).
12. Sniffing out meaning: Chemosensory and semantic neural network changes in sommeliers / M. Carreiras et al. *Human Brain Mapping.* Volume 45, Issue 2. https://doi.org/10.1002/hbm.26564 *Wiley* : online library. URL: https://goo.su/qAoifF (date of application: 30.04.24).
13. Takmaz E. Visual and Linguistic Processes in Deep Neural Networks: A Cognitive Perspective : doctoral thesis, Universiteit van Amsterdam. Amsterdam, 2024. 227 p. *EPrints* : repository software. URL: https://goo.su/ytR9u (date of application: 30.04.24).
14. Zhao L. Analysis of the Technique and Morphological Language of Modern Painting Creation Based on Generative Adversarial Networks. *Applied Mathematics and Nonlinear Sciences.* 2024. Volume 9, Issue 1. https://doi.org/10.2478/amns-2024-0279 *Sciendo* : website. URL: https://goo.su/pK6p (date of application: 30.04.24).

REFERENCES

1. Goh, T. J., Chong, L. Y., Chong, S. C., & Goh, P. Y. (2024). A Campus-based Chatbot System using Natural Language Processing and Neural Network. *Journal of Informatics and Web Engineering*, *3(1),* 96–116.
2. Durdi, V. B., Kiran, A., Rao, A., Bhat, S. S., Santhosh, B., & Kumar, M. (2024). A Novel Method for Recognizing Hand Gestured Sign Language Using the Stochastic Gradient Descent Algorithm and Convolutional Neural Network Techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *12(7s)*, 529–538.
3. Abasan, I. G. B. J., & Setiawan, E. B. (2024). Empowering hate speech detection: leveraging linguistic richness and deep learning. *Bulletin of Electrical Engineering and Informatics*, *13(2)*, 1371–1382.
4. Gharaibeh, H., Mamlook, A., Emhamed, R., Samara, G., Nasayreh, A., Smadi, S., ... & Abualigah, L. (2024). Arabic sentiment analysis of Monkeypox using deep neural network and optimized hyperparameters of machine learning algorithms. *Social Network Analysis and Mining*, *14(1)*, 1–18.
5. Clean Room Design. *DevX*: technology glossary. website. URL: https://goo.su/P3gvV (date of application: 30.04.24).
6. Mangal, R., Narodytska, N., Gopinath, D., Hu, B. C., Roy, A., Jha, S., & Păsăreanu, C. (2024). Formal Concept-based Analysis of Neural Networks via Vision-Language Models. *arXiv*: website. URL: https://goo.su/Bo6t6O1 (date of application: 30.04.24).
7. Aguirre, F., Sebastian, A., Le Gallo, M., Song, W., Wang, T., Yang, J. J., ... & Lanza, M. (2024). Hardware implementation of memristor-based artificial neural networks. *Nature Communications*, *15(1)*. URL: https://goo.su/8tEZgl9 (date of application: 30.04.24).
8. Zheng, Z., Chao, W., Qiu, Z., Zhu, H., & Xiong, H. (2024). Harnessing Large Language Models for Text-Rich Sequential Recommendation. *arXiv*: website. URL: https://goo.su/F1bqd (date of application: 30.04.24).
9. Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature,* 49–58.
10. Nguyen, P. T., Henningsen-Schomers, M. R., & Pulvermüller, F. (2024). Causal influence of linguistic learning on perceptual and conceptual processing: A brain-constrained deep neural network study of proper names and category terms. *Journal of Neuroscience*, *44(9)*. URL: https://goo.su/DwRlyA8 (date of application: 30.04.24).
11. Overview of Clean Room Software Engineering. *GeekForGeeks* : website. URL: https://goo.su/glOsLQ6 (date of application: 30.04.24)
12. Carreiras, M., Quiñones, I., Chen, H. A., Vázquez-Araujo, L., Small, D., & Frost, R. (2024). Sniffing out meaning: Chemosensory and semantic neural network changes in sommeliers. *Human Brain Mapping, 45 (2)*. URL: https://goo.su/qAoifF (date of application: 30.04.24).
13. Takmaz, E. (2024). Visual and Linguistic Processes in Deep Neural Networks: A Cognitive Perspective (Doctoral dissertation, Universiteit van Amsterdam). 227 p.
14. Zhao, L. (2024). Analysis of the Technique and Morphological Language of Modern Painting Creation Based on Generative Adversarial Networks. *Applied Mathematics and Nonlinear Sciences*, *9(1)*. URL: https://goo.su/pK6p (date of application: 30.04.24).