Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2025”

April 23–25, 2025

Computer tools in mental disorders diagnostics by oral speech

Khomenko Anna
HSE University
Nizhny Novgorod, Russia
akhomenko@hse.ru

Komratova Anastasia
HSE University
Nizhny Novgorod, Russia
akomratova@hse.ru

Isakov Danila
HSE University
Nizhny Novgorod, Russia
disakov@hse.ru

Balba Daria
HSE University
Nizhny Novgorod, Russia
dasha.balba@gmail.com

Shishkovskaya Tatiana
HSE University
Moscow, Russia;
Mental Health Research Center
Moscow, Russia
tszyszkowska@gmail.com

Khudyakova Maria
HSE University
Nizhny Novgorod, Russia
mariya.kh@gmail.com

Abstract

The integration of automated speech analysis in diagnosing mental health disorders is becoming increasingly significant in both clinical and computational linguistics. This study aims to construct linguistic profiles for individuals with neurocognitive and affective mental disorders. Using speech transcriptions and relevant to the study computational techniques like lexical clustering and stylostatistical analysis, this research looks for characteristics capable of distinguishing speech patterns indicative of various mental health conditions. A text corpus of oral speech from 136 people diagnosed with schizophrenia, schizotypal disorder, schizoaffective disorder, borderline personality disorder, other personality disorders, primary depressive episode, recurrent depressive disorder, bipolar affective disorder and 210 participants with no diagnosed diseases in the control group was used in the research. As a result of the study, it was proved that people with mental disorders display specific features in oral speech, that can be used in creation of an automatic mental disorders diagnostic model.

Keywords: mental disorders; early automatic diagnostics; oral speech; clustering; stylistics
DOI: 10.28995/2075-7182-2025-23-XX-XX

Применение компьютерных инструментов в диагностике психических расстройств по устной речи

Хоменко Анна
НИУ ВШЭ
Нижний Новгород, Россия
akhomenko@hse.ru

Комратова Анастасия
НИУ ВШЭ
Нижний Новгород, Россия
akomratova@hse.ru

Исаков Данила
НИУ ВШЭ
Нижний Новгород, Россия
disakov@hse.ru

Бальба Дарья
НИУ ВШЭ
Нижний Новгород, Россия
dasha.balba@gmail.com

Шишковская Татьяна
НИУ ВШЭ
Москва, Россия;
Научный центр
психического здоровья
Москва, Россия
tszyszkowska@gmail.com

Худякова Мария
НИУ ВШЭ
Нижний Новгород, Россия
mariya.kh@gmail.com

Аннотация

Диагностика ментальных расстройств по продуцируемой речи является важнейшей задачей среди технологических разработок моделей нейрокогнитивных и психолого-психиатрических процессов. Целью данного исследования является построение лингвистических профилей для лиц с нейрокогнитивными и аффективными психическими расстройствами. Используя транскрипты устной речи и релевантные для исследования компьютерные технологии, такие как лексическая кластеризация и стилостатистический анализ, разрабатывается пул характеристик, способных различать речевые паттерны, указывающие на различные состояния психического здоровья. В работе использовался текстовый корпус устной речи 136 человек с диагнозом шизофрения, шизотипическое расстройство, шизоаффективное расстройство, пограничное расстройство личности, другие расстройства личности, первичный депрессивный эпизод, рецидивирующее депрессивное расстройство, биполярное аффективное расстройство и 210 участников контрольной группы без диагностированных заболеваний. В результате исследования было доказано, что устная речь людей с психическими расстройствами имеет особые диагностические признаки, которые могут быть использованы при создании автоматической модели определения психических расстройств.

Ключевые слова: ментальные расстройства; ранняя автоматическая диагностика; устная речь; кластеризация; стилостатистика

1 Introduction

Early diagnosis of cognitive impairments is a critical task in psychiatry, neurology, and pedagogy, as timely intervention can prevent complications and support recovery (Khudyakova et al., 2023). Diagnosing mental disorders through speech analysis is an emerging field within technological advancements focused on neurocognitive processes. Clinical linguistics, leveraging the analysis of oral speech, plays a pivotal role at both microlinguistic (phonetic, semantic, lexical) and macrolinguistic levels. Traditionally, this process involves manual profiling — expert interpretative processing of experimental material—which can be subjective and time-consuming. However, advances in computational linguistics now allow for automated tools to analyze and evaluate speech patterns objectively, addressing the growing need for efficient and standardized diagnostic methods (Malgaroli et al. 2023; Le et al. 2021; Ribeiro et al., 2024).

Speech profiling involves constructing a linguistic portrait of individuals, comparing those with diagnosed cognitive impairments to the control group. These profiles capture the distinctive linguistic characteristics associated with mental disorders, offering insights into cognitive and emotional states (Mota et al., 2012). Many existing models focus primarily on acoustic characteristics of audio signals, such as pitch, tempo, and pause duration, making them relatively universal and non-language-specific (Koops et al., 2023; Ramanarayanan et al., 2022). However, such approaches often overlook the broader set of linguistic data, including syntactic complexity and discourse coherence, which are critical for comprehensive diagnostics.

This research explores speech profiling for generalized language personalities of individuals with mental disorders, focusing on Russian-language oral discourse. Unlike prior studies predominantly conducted on English-language material, this study addresses a significant gap in research by constructing speech profiles for individuals with borderline personality disorder, recurrent depressive disorder, bipolar affective disorder, and other conditions within the Russian-speaking population. By integrating microlinguistic and macrolinguistic features, the research aims to develop automated models capable of identifying speech patterns indicative of neurocognitive and affective mental disorders.

The potential of speech profiling has been demonstrated in studies examining various linguistic features to diagnose schizophrenia, bipolar disorder (Mota et al., 2023), and depression (Khomenko et al., 2023).

The inclusion of linguistic markers, such as syntactic complexity and semantic coherence, enriches the diagnostic process. For Russian-speaking discourse, this approach remains largely unexplored, despite its promise for creating functional diagnostic models tailored to specific populations.

By building on methods from forensic authorship linguistics and speech profiling, this research aims to advance clinical linguistics through the creation of digital tools. The constructed algorithm paves the way for early diagnosis and personalized interventions and supports the rapid and objective assessment of speech data. This study focuses on Russian-language discourse, broadening the scope of current research by highlighting the potential for developing language-specific diagnostic solutions and addressing the problem of creating a set of diagnostic parameters for analyzing the discourse of Russian-language speaking people with mental disorders.

2 Related work

One of the approaches to analyzing speech in mental disorders is interpretive processing through clustering, as it is established that respondents tend to produce semantically or phonetically related groupings of words in bursts over time, described as “clustering”. In this case the researchers pay attention to semantic cohesion of an individual’s speech and semantic similarity between the produced words, as patients with psychotic disorder may display distinguished from healthy patients semantic networks (Aloia et al., 1996; Paulsen et al., 1996; Pintos et al., 2022). Traditionally the aforementioned processing is done manually, however recent studies tend to automate this procedure in order to get more objective and repeatable results (Corcoran et al., 2020; Lundin et al., 2022).

Another approach to speech analysis in cognitive disorders is based on speech profiling. In this case researchers tend to create a generalized linguistic personality model – a set of linguistic traits of individuals within the studied group (here – clinical one), as it is widely recognized, that language personality may be affected by psychiatric disturbances (e.g. Sizova & Semenova, 2023). To build clinical and control group profiles researcher often resort to quantitative methods and stylostatistics, as it may offer a faster, more efficient and objective way to retrieve data (Mota et al., 2012).

Several studies have identified distinctive speech patterns associated with schizophrenia. Acoustic markers such as pitch variability, speech rate, and pause duration have been linked to symptoms like flat affect and alogia. For example, analyses of these features have achieved classification accuracies exceeding 90% when distinguishing schizophrenia patients from healthy controls (Espinola et al., 2020; Pan et al., 2023). Similarly, linguistic markers, including reduced syntactic complexity and decreased discourse coherence, have been instrumental in identifying psychotic symptoms and differentiating between patient and control groups (Baklund et al., 2022).

Advancements in computational methods, particularly machine learning, have facilitated the use of vocal biomarkers for differential diagnosis. Features such as Mel-frequency cepstrum coefficients (MFCCs), fundamental frequency, and formant patterns have been leveraged to classify individuals into groups such as schizophrenia, bipolar disorder, and depression. Classification accuracies of up to 91% have been reported, underscoring the effectiveness of these approaches (Espinola et al., 2020). Additionally, variability in vocal features such as pitch and speech tempo has proven useful in distinguishing between different psychiatric conditions (Pan et al., 2023).

Research into schizophrenia within Russian-language materials has explored various linguistic and cognitive aspects of the disorder. A notable study by Panicheva and Litvinova (Panicheva, & Litvinova, 2019) examined semantic coherence in written texts by Russian-speaking individuals diagnosed with schizophrenia. Their findings revealed that, contrary to studies in English, patients' texts exhibited higher minimum semantic coherence values compared to those of healthy controls. This suggests that written narratives of Russian-speaking patients may display less topic variability, potentially reflecting a tendency toward perseveration or focused thought patterns.

Additionally, the study highlighted the significance of the relative position of semantic coherence minima within texts, indicating that patients' writings often showed distinct patterns in the flow of ideas. These insights underscore the importance of considering linguistic and cultural contexts when analyzing cognitive impairments in schizophrenia.

Further research has involved the adaptation and testing of diagnostic tools for Russian-speaking populations. For instance, the Russian-language version of the Thought, Language, and Communication (TLC) Scale has been evaluated for its effectiveness in detecting early cognitive disorders in individuals at risk of schizophrenia. The results support the TLC scale's utility in clinical settings, facilitating early intervention strategies (Omelchenko et al., 2022).

3 Methods

3.1 Participants

For the study speech samples from the Discourse Diversity Database (3D; Khudyakova et al., 2022) were taken representing a diverse collection designed to explore linguistic markers indicative of mental health conditions. It comprises speech samples by 136 people with mental disorders, such as schizophrenia (32), schizoaffective disorder (17), schizotypal disorder (22), bipolar disorder (36), borderline personality disorder (8), recurrent depressive disorder (12), primary depressive episode (4), other personality disorder (5), and 210 individuals from a control group with no diagnosed mental health conditions. The diagnosis of disorders in the clinical group was carried out in accordance with ICD-10 by qualified psychiatrists and the information about the control group was self-reported[1].

3.2 Elicitation stimuli

The dataset contains three discourse types: picture based narratives, personal stories, and picture-based instructions (procedural discourse), each containing three variants of stimuli: comics by Herluf Bidstrup («Superman», «Discovery of the World», «Wonderful Day») – for picture-elicited narratives task; questions about the best or the most memorable gift, trip or holiday – for personal stories; IKEA's self-assembly furniture manuals – for the picture-based instructions (Khudyakova et al., 2022: 38).

3.3 Transcription

Audio recordings were transcribed for the control and clinical group using the machine learning model for speech recognition and transcription Whisper[2]. Transcriptions were checked manually according to pre-established rules used in compiling verbatim contents in forensic examination of oral speech (Kurachenkova et al., 2007). Thus, speech fragments, the content of which could not be determined are marked as *нрзб*. (illegible). Words with contracted word pronunciation are indicated in the form of orthographic transliteration, for example: *че, гит, гру, ниче.* Unfinished words are accompanied by the sign =: *сба= сбавлять* (slo= to slow down). Unfinished remarks are accompanied by three dots: *Я не знаю*… (I don't know…). Filled hesitation pauses are indicated by stretching vowels and consonants like *аа, ээ, ем, мм* (aa, ee, em, mm): *Я, мм, пошла, э, в склад* (I, mm, went, e, on the warehouse). Manual correction of transcripts was carried out by students in linguistics. Verification was made by specialists in language with a qualification of at least a candidate of sciences (PhD).

4 Data analysis

The primary research method used in this study is speech profiling (Argamon et al., 2009) within the paradigms of linguistic personality theory, idiolect (Litvinova et al., 2016; Litvinova, 2019), and idiostyle (Karaulov, 2010; Wright, 2017), aimed at addressing diagnostic tasks in forensic authorship linguistics (Coulthard, 2004). These methods are implemented through general principles of text processing and machine learning, including topic modeling, lexical vectorization, and neural network

[1] The demographic data of the participants is located by the link https://github.com/KhomenkoAnna/3D_research.

[2] https://github.com/openai/whisper

training. Oral speech processing is conducted within the framework of experimental linguistics in practical use (Kaganov, 2005). The diagnostic potential for identifying clinical disorders is supported by previous research using non-Russian language materials (Corcoran et al., 2020) and the transitive principle, which allows for the identification of individuals with auto- and hetero-aggressive behavioral tendencies through speech analysis (Khomenko et al., 2023).

In this paper, the construction of speech profiles for clinical and control groups is based on a combination of automatic lexical clustering and stylostatistical (stylometric) analysis. Clustering allows to establish the semantic structure of discourse: lexical clusters describe semantic groups present in a particular piece of speech. Calculating stylometric indices give an opportunity to track the so-called uncontrolled characteristics of speech associated with deep thinking processes.

4.1 Preprocessing

With the help of the NLTK library[3], the texts were tokenized in two formats: with the preservation of stop words (also taken from the NLTK library) and without for clustering and with the preservation of stop words for stylostatistical analysis, in both cases all punctuation marks were removed. Using the NLP library SpaCy (ru_core_news_sm model[4]), transcriptions were also lemmatized to normalize words to their base forms, reducing linguistic variability while preserving semantic meaning.

4.2 Clustering

To the resulting four groups of words, the clustering algorithm was applied according to the rule proposed in the work of Lundin et al. (Lundin et al., 2022). According to this rule, for a sequence of words A, B, C, D, the switch follows B when S (A, B) > S (B, C) and S (B, C) < S (C, D), where S (A, B) is the cosine similarity between the vectors of the words A and B. To estimate the cosine similarity, the pre-trained static embedding model geowac_lemmas_none_fasttextskipgram_300_5_202024[5] of the RusVect_rēs service was used, which was used to obtain embeddings of the fastText type (Bojanowski et al., 2017). For the first word in the sequence, the proximity was taken as 0.

Based on the clustering results, the following metrics were calculated:

1. Number of switches for each person in each category.
2. The average distance between clusters for each person in each category (calculated as the distance between cluster centroids; the average value of vectors located in the same cluster is taken as the centroid).
3. The average value of the T-score in the cluster for each person in each category (a metric that shows how non-random the strength of the association between collocates is; all sequences of two words inside one cluster are taken as collocates).
4. The average value of the silhouette-score (a metric that shows how close an object is to its cluster compared to other clusters: the higher the value of this metric, the closer the object is to its own cluster).
5. The average cluster size for each person.

For each metric in each set, the mean, median, and standard deviation were calculated.

4.3 Stylostatistics

Statistical information was extracted from the texts, including n-grams — sequences of two (bigrams) and three (trigrams) words that were obtained using n-gram models. These models were based on calculating the frequency of word combinations in the text and allowed us to identify frequently occurring sequences characteristic of each group.

Keywords were highlighted using various metrics such as Log-Likelihood, T-score, Mutual Information (MI), Dice and Frequency Ratio. These metrics helped to identify the most significant and frequently occurring words in the texts of each group, taking into account their frequency of occurrence both within the group and in relation to the total frequency in the corpus. Additionally, collocations -

---

[3] https://github.com/nltk/nltk
[4] https://spacy.io/models/ru#ru_core_news_sm
[5] https://github.com/avidale/compress-fasttext

stable phrases characteristic of these texts - were identified for the top 20 extracted keywords according to the T-score metric. The statistical T-score method used to identify collocations made it possible to identify the most frequent and significant combinations of words that play a key role in the speech structure of both groups.

As part of the study, various stylostatistical parameters were also calculated in order to analyze the texts more deeply in terms of their structure and content. Among these parameters (Golovin 1971) are the coefficients of objectivity, *Pr* (the ratio of nouns plus pronouns amount to the sum of adjectives and verbs), quality, *Qu* (the ratio of adjectives plus adverbs amount to the sum of verbs and nouns), activity, *Ac* (the ratio of verbs plus all verb forms amount to the word number in the text), dynamism, *Din* (the ratio of verbs plus all verb forms amount to the sum of nouns, adjectives and pronouns) and coherence of the text, *Con* (the ratio of prepositions plus conjunctions amount to the number of sentences). In addition to these parameters, readability indices such as the Flesch-Kincaid index and the Gunning Fog index were calculated, which give an idea of how easy or difficult it is to perceive texts from both groups. These indexes are based on characteristics such as sentence length and the complexity of the words used.

In addition, the type-token ratio (TTR), the average word length (in letters) and the average sentence length (in words) were calculated, which made it possible to assess the richness of vocabulary and the structural complexity of the text.

5 Results

Statistical analysis was performed using the Pandas[6], NumPy[7], SciPy[8] and Matplotlib[9] libraries. First of all, the statistical significance of the feature parameters for clustering and stylostatistics was checked. To assess the significance of differences between the groups, two types of statistical tests were applied:

- Independent samples t-test: used when data followed a normal distribution
- Non-parametric Mann-Whitney U-test: applied for data that did not follow a normal distribution

A criterion was considered statistically significant if its p-value corrected for multiple comparisons was less than 0.05, indicating meaningful differences between the groups with and without mental disorders.

Significant differences between the two groups were found for the following metrics: the number of switches, distance between clusters, and t-score (see Table 1). Cluster size metric (M-control=4,95, M-disorder=4,9) and silhouette-score (M-control=0,21, M-disorder=0,21) have not shown statistical significance[10].

For the twogroup comparison, the following criteria were statistically significant: coefficients of objectivity, quality, activity and dynamism, Flesch-Kincaid Index, Gunning Fog Index, type-token ratio (TTR).

Next, threshold values for each metric were determined to predict group membership using two statistical parameters: Youden’s J index and the sum/difference between the mean and standard deviation of the metric in the control group.

Youden’s J index was computed based on the AUC-ROC between the target variable and the metric, representing the optimal threshold where the true positive rate exceeds the false positive rate. The mean and standard deviation of the control group were used such that if the control group's mean was higher than that of the disorder group, the standard deviation was subtracted from the mean to define the lower threshold. Otherwise, the standard deviation was added to the mean.

Threshold values were calculated by both methods, and their efficacy was compared. Each metric was labeled 0 (control group) or 1 (disorder group) based on the threshold, and the labels were evaluated using recall, a key metric for predicting the positive class. The optimal threshold for each metric was selected based on the recall results. Statistical analysis was conducted using the Scikit-Learn[11] library.

---

[6] https://pandas.pydata.org/
[7] https://numpy.org/
[8] https://scipy.org/
[9] https://matplotlib.org/
[10] Full data are available by the link https://github.com/KhomenkoAnna/3D_research.
[11] https://scikit-learn.org/stable/

	Control Group	Disorder Group	Threshold Value	Metric	Recall
Number of switches	49.12	28.16	16.68	Mean - Std	0.3
Cluster distance	0.51	0.52	0.53	Younden's J	0.4
T-score	897.98	420.47	275.8	Mean - Std	0.32

Table 1: Average values of statistically significant metrics, threshold values, and recall for clustering

For stylostatistics metrics, the threshold value was determined only based on Youden's J index with correction for multiple comparisons (see Table 2).

	Control Group	Disorder Group	Threshold Value	Recall
Flesch-Kincaid readability index	6.34	4.35	1.49	0.92
Gunning fog index	5.41	3.94	inf	0
TTR	0.68	0.74	0.72	0.58
Pr	1.68	1.75	1.78	0.46
Qu	0.1	0.11	0.08	0.77
Din	0.06	0.05	0.17	0.01

Table 2: Average values of statistically significant metrics, threshold values, and recall for stylostatistics

As indicated in the table, for one criterion, it was not possible to calculate the threshold value. This is because there is a high degree of overlap between the classes for Gunning fog index, making it impossible to identify a clear threshold for class separation.

6 Discussion

Speech profiles have been created for generalized language personalities of all types of mental and behavioral disorders: schizophrenia, schizotypal disorder, schizoaffective disorder, borderline personality disorder, other personality disorders, primary depressive episode, recurrent depressive disorder, bipolar affective disorder – and for each disorder separately. A speech profile was also created for the language personality of the control group – a group of individuals who had not been diagnosed with mental disorders.

Previous research has established that individuals with these conditions exhibit certain linguistic abnormalities, such as reduced semantic coherence, syntactic simplifications, narrative fragmentation, and excessive self-referential language (Tang et al., 2021; Vlasova & Akhmetova, 2015). Schizophrenic speech is fragmented, incoherent, and marked by neologisms and perseveration, reflecting disorganized thinking (Akhmetzyanova & Tregubenko 2021; Vlasova & Akhmetova, 2015). Personality disorders feature frequent present-tense usage, self-referential language, and an observer perspective (Smerchinskaya et al., 2024). Bipolar speech is verbose and grammatically complex, while depressive speech is elliptical, impersonal, and rich in idiomatic expressions (Smirnova et al., 2013).

The results of this study support these observations, as statistical analysis has revealed that different patterns of clustering and stylostatistical indicators are observed in various disorders. Therefore, when summarizing the findings of the stylostatistical and clustering analysis, the following descriptions of the speech patterns of the analyzed conditions can be given. Speech in depression is complex and detailed, featuring long sentences, large number of lexical clusters and a high lexical variety. Patients with depression also tend to provide more standardized associations. Schizophrenic spectrum and personality disorder patients tend to speak in shorter, simpler sentences, though they may not always be logical. Depressive patients are more likely to employ abstract constructions, while in schizotypal and schizoaffective conditions, speech is more specific and focused on details. Personality disorders are characterized by a high level of evaluative and active language, while depression and borderline disorders exhibit a static quality. Text coherence is greater in schizoaffective disorder, although semantic gaps may occur. Patients with schizophrenic spectrum disorders tend to switch between word clusters more often, showing loose associations effect. The mean distance between associations is larger for respondents with bipolar and personality disorders. Although the specific types of speech varied depending on the diagnosis, all the disorders showed significant differences from the speech patterns of the control group.

Thus, for the final automated diagnostic model vocabulary clustering metrics do not show the required level of statistical significance for heterogeneous patient groups. The best metrics is mean cluster distance. It has the highest recall. Nevertheless, for individual patient groups, clustering shows positive results in class division. Thus, for speech of individuals with bipolar disorder mean cluster size shows a recall higher than 0.5, for schizophrenia and schizoaffective disorder – mean distance between clusters; for schizotypal disorder – mean cluster size and silhouette score. For the generalized group with different types of disorders, stylostatistical metrics show their effectiveness in separating the clinical and control groups. Thus, Flesch-Kincaid readability index, type-token ratio (TTR), coefficient of quality (Qu) (the metrics have the recall value equal or greater than 0.5) could be considered as possible for inclusion in diagnostic model of psychiatric disorders and used in creating an automatic model for classifying speech as a control group or a group with thought and mood disorders. Naturally, for working psychiatric diagnostics, more customized models should be used, configured to determine a specific disorder. Nevertheless, for a fundamental study of the Russian-language discourse of people with mental and mood disorders, this work can be useful.

7 Conclusion

This study contributes to the growing body of research aimed at automating the diagnosis of mental health conditions through speech analysis. The findings demonstrate that speech profiling can show differences between individuals with and without mental health disorders, particularly within the Russian-speaking population, which has been underexplored in previous studies. By integrating both microlinguistic and macrolinguistic features the study offers a comprehensive approach to diagnosing mental disorders.

Future research should focus on expanding the dataset to include more varied linguistic features and refining the algorithms with the aim of further model customization for diagnosing specific disorders. Additionally, incorporating speech data from multiple languages and cultural contexts will enhance the model's generalizability, ultimately making it a valuable tool for mental health professionals worldwide.

8 Limitations

This article will focus mainly on generalized results for two groups: the control group and the group with mental disorders. Such a division is undoubtedly a limitation of the study. The concept of indivisible psychosis (E.A. Zeller, H. Neumann и W. Griesenger) is not accepted as leading due to its proved inconsistency. Nevertheless, the unification of heterogeneous disorders seems possible within the framework of a pilot study. Firstly, such unification is carried out in psychiatrics works (e.g. Sychugov et al., 2019), and secondly, the study aims to identify pathological speech as a class (Pashkovsky et al., 2013) and to understand whether it is possible to distinguish the norm group from the abnormal. Thus, there are three main causes of mental disorders – exogenous, endogenous and stress-related. "In all these cases, the brain is unable to function normally, and pathological mental production occurs. With the exception of obvious deficit processes associated with progressive morphological degeneration of brain cells and the formation of dementia, it can be said that pathological production is aimed at compensating for disadaptation. Thus, specific disorders (e.g., dementia) are possible only with organic damage to the cortex. All other organic mental disorders <…> are nonspecific” (Ten et al., 2008:152). It is clear that due to the heterogeneity of the sample, the final result of the study will be influenced by the largest classes (schizophrenia, schizoaffective disorder, schizotypal disorder, bipolar disorder), but within these classes it is also possible to predict the presence or absence of statistically significant parameters capable of detecting the speech of people with a certain diagnosis.

The conclusion about the presence or absence of a disease is always made by a psychiatrist. At this stage of their development automatic tools can only give a signal to a specialist that it is necessary to continue working with a particular patient. As a result of the study, it was found that different types of disorders show different typification of specific characteristics, however, in each case this typification differs from the typification of a control group.

Acknowledgements

This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

References

[1] Akhmetzyanova, E. M., & Tregubenko, I. A. (2021). Psycholinguistics Analysis of schizophrenia spectrum disorders patients’ writing and oral speech (on the example of the autobiographical texts). *The Scientific Notes of the Pavlov University*, 28(2), 38–44. https://doi.org/10.24884/1607-4181-2021-28-2-38-44

[2] Aloia, M. S., Gourovitch, M. L., Weinberger, D. R. & Goldberg, T. E. (1996). An investigation of semantic space in patients with schizophrenia. Journal of the International Neuropsychological Society, 2(4), 267-273. https://doi.org/10.1017/S1355617700001272.

[3] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119–123. https://doi.org/10.1145/1461928.1461959

[4] Baklund, L., Røssberg, J. I., & Møller, P. (2022). Linguistic markers and basic self-disturbances among adolescents at risk of psychosis. A qualitative study. *EClinicalMedicine, 55*, 101733. https://doi.org/10.1016/j.eclinm.2022.101733

[5] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://doi.org/10.1162/tacl_a_00051

[6] Corcoran, C. M., & Cecchi, G. A. (2020). Using language processing and speech analysis for the identification of psychosis and other disorders. *Biological Psychiatry Cognitive Neuroscience and Neuroimaging, 5*(8), 770–779. https://doi.org/10.1016/j.bpsc.2020.06.004

[7] Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics, 25*(4), 431–447. https://doi.org/10.1093/applin/25.4.431

[8] Espinola, C. W., Gomes, J. C., Pereira, J. M. S., & Santos, W. P. D. (2020). Vocal acoustic analysis and machine learning for the identification of schizophrenia. *Research on Biomedical Engineering, 37*(1), 33–46. https://doi.org/10.1007/s42600-020-00097-1

[9] Golovin, B. N. (1971). *Yazyk i statistika* [Language and statistics]. Prosveshchenie.

[10] Kaganov A. Sh. (2005) *Kriminalisticheskaya ekspertiza zvukozapisej* [Forensic examination of recordings]. Moscow: Yurlitinform.

[11] Karaulov Yu. N. (2010). *Russkij yazyk i yazykovaya lichnost'* [Russian language and linguistic personality] (7th ed.). LKI Publishing House.

[12] Khomenko A. Yu., Kasimova L. N., Sychugov E. M., Komratova A.D., Galkin N. V., & Titova A. A. (2023). Diagnostika auto- i geteroagressii metodami psiholingvistiki [Psycholinguistics methods in diagnostics of auto- and heteroagression]. In P. A. Katyshev & I. V. Ogorelkov (Eds.), *Sovremennyj mediatekst i sudebnaya ekspertiza: mezhdisciplinarnye svyazi i ekspertnaya ocenka*, 12, 337-361.

[13] Khudyakova, M., Antonova, N., Nelubina, M., Surova, A., Vorobyova, A., Minnigulova, A., Gronskaya, N., Yashin, K., Medyanik, I., Shishkovskaya, T., Ryazanskaya, G., Zuev, A., Dragoy, O., & Skorobogatova, A. S. (2022). Discourse Diversity Database (3D) for Clinical Linguistics Research: Design, Development, and analysis. *Bakhtiniana Revista De Estudos Do Discurso, 18*(1), 32–57. https://doi.org/10.1590/2176-4573e55885

[14] Khudyakova, M. V. (2023). *Methods for analysis and assessment of spoken discourse by people with speech-language disorders* (Doctoral dissertation, HSE University).

[15] Koops, S., Brederoo, S. G., de Boer, J. N., Nadema, F. G., Voppel, A. E., & Sommer, I. E. (2023). Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets-CNS & Neurological Disorders)*, *22*(2), 152-160.

[16] Kurachenkova, N. B., Baicharov, N. V., & Ermakova, M. A. (2007). *Identifikaciya lic po ustnoj rechi na russkom yazyke. Metodika" Dialekt": posobie dlya ekspertov* [Identification by oral speech in Russian. Methodology «Dialect». Manual for experts]. IK CST FSB Rossii.

[17] Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., ... & Lemey, C. (2021). Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, *23*(5), e15708.

[18] Litvinova, T., Seredin, P., Litvinova, O., & Zagorovskaya, O. (2016). Profiling a set of personality traits of text author: what our words reveal about us. *Research in Language, 14*(4), 409–422. https://doi.org/10.1515/rela-2016-0019

[19] Litvinova T. A. (2019). Idiolekt kak ob"ekt korpusnoj idiolektologii: k stanovleniyu novogo lingvisticheskogo napravleniya [Idiolect as an object of corpus and idiolectology: toward the formation of a new linguistic trend]. *Uchenye zapiski Novgorodskogo gosudarstvennogo universiteta, 7*(25), 1–5.

[20] Lundin, N. B., Jones, M. N., Myers, E. J., Breier, A., & Minor, K. S. (2022). Semantic and phonetic similarity of verbal fluency responses in early-stage psychosis. *Psychiatry Research, 309*, 114404. https://doi.org/10.1016/j.psychres.2022.114404

[21] Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, *13*(1), 309.

[22] Mota, N. B., Vasconcelos, N. a. P., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., & Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS ONE, 7*(4), e34928. https://doi.org/10.1371/journal.pone.0034928

[23] Mota, N. B., Weissheimer, J., Finger, I., Ribeiro, M., Malcorra, B., & Hübner, L. (2023). Speech as a graph: Developmental perspectives on the organization of spoken language. *Biological Psychiatry Cognitive Neuroscience and Neuroimaging, 8*(10), 985–993. https://doi.org/10.1016/j.bpsc.2023.04.004

[24] Omelchenko, M. A., Arutyunova, E. V., Butashin, A. D., & Ivanova, E. M. (2022). Testing of the Russian-language version of the Thought, Language and Communication Scale. *S S Korsakov Journal of Neurology and Psychiatry, 122*(10), 116. https://doi.org/10.17116/jnevro2022122101116

[25] Pan, W., Deng, F., Wang, X., Hang, B., Zhou, W., & Zhu, T. (2023). Exploring the ability of vocal biomarkers in distinguishing depression from bipolar disorder, schizophrenia, and healthy controls. *Frontiers in Psychiatry, 14*. https://doi.org/10.3389/fpsyt.2023.1079448

[26] Panicheva, P., & Litvinova, T. (2019). Semantic Coherence in Schizophrenia in Russian Written Texts. *IEEE*, 241–249. https://doi.org/10.23919/fruct48121.2019.8981535

[27] Pashkovsky, V. E., Piotrovskaya, V. R., & Piotrovsky, R. G. (2013). *Psihiatricheskaya lingvistika* [Psychiatric linguistics]. Moscow: LENAND.

[28] Paulsen, J. S., Romero, R., Chan, A., Davis, A. V., Heaton, R. K., & Jeste, D. V. (1996). Impairment of the semantic network in schizophrenia. Psychiatry research, 63(2-3), 109-121. https://doi.org/10.1016/0165-1781(96)02901-0

[29] Pintos, A. S., Hui, C. L. M., De Deyne, S., Cheung, C., Ko, W. T., Nam, S. Y., ... & Chen, E. Y. H. (2022). A longitudinal study of semantic networks in schizophrenia and other psychotic disorders using the word association task. Schizophrenia Bulletin Open, 3(1), sgac054. https://doi.org/10.1093/schizbullopen/sgac054

[30] Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F., & Green, J. R. (2022). Speech as a biomarker: Opportunities, interpretability, and challenges. *Perspectives of the ASHA Special Interest Groups*, *7*(1), 276-283.

[31] Ribeiro, M., Malcorra, B., Mota, N. B., Wilkens, R., Villavicencio, A., Hubner, L. C., & Rennó-Costa, C. (2024). A Methodology for Explainable Large Language Models with Integrated Gradients and Linguistic Analysis in Text Classification. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2410.00250

[32] Sizova N.S., Semenova N.D. (2022). Issledovaniye ekspressivnoy rechi u lits s shizofreniyey i rasstroystvami shizofrenicheskogo spektra [Study of expressive speech in persons with schizophrenia and schizophrenia spectrum disorders]. *Russian Journal of Psychiatry*. No. 6, 53-60.

[33] Smerchinskaya E. M., Tregubenko I. A., Isaeva E. R. (2024). Osobennosti pis'mennoj rechi pacientov s razlichnymi psihicheskimi rasstrojstvami [Written Narrative in Patients with Different Mental Disorders]. *SibScript, 26*(5), 739–747.

[34] Smirnova, D., Sloeva, E., Kuvshinova, N., Krasnov, A., Romanov, D., & Nosachev, G. (2013). 1419 – Language changes as an important psychopathological phenomenon of mild depression. *European Psychiatry, 28*, 1. https://doi.org/10.1016/s0924-9338(13)76454-x

[35] Sychugov E.M., Petrova D.N., Morozova A.S. (2019). Psiholingvisticheskij analiz patologicheskih tekstov kak metod ob"ektivizacii diagnostiki rasstrojstv rechemyslitel'noj deyatel'nosti v psihiatricheskoj praktike [Psycholinguistic analysis of pathological texts as a method of objectification of diagnostics of mental and speech activity disorders in psychiatric practice]. *Fundamental aspects of mental health, 3*, 52 – 58.

[36] Tang, S. X., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R. E., Bhati, M. T., Wolf, D. H., Sedoc, J., & Liberman, M. Y. (2021). Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. Schizophrenia, 7(1). https://doi.org/10.1038/s41537-021-00154-3

[37] Ten V.I. (2008). Koncepciya «edinogo psihoza» Nejmana-SHrizingera: adaptacionnaya paradigma v psihiatrii [The concept of «indivisible psychosis» by Neumann-Schrizinger: an adaptation paradigm in psychiatry]. *Bulletin of Kyrgyz-Russian Slavic University, 8*(4), 151–154.
[38] Vlasova V. P., & Akhmetova E. R. (2015). Osobennosti rechi pri shizofrenii [Features of speech in schizophrenia]. *In The Eusebian Readings. Series: Actual problems of special and inclusive education*, 12-17.
[39] Wright, D. (2017). Using word n-grams to identify authors and idiolects. *International Journal of Corpus Linguistics, 22*(2), 212–241. https://doi.org/10.1075/ijcl.22.2.03wri