Predictive and Distinctive Linguistic Features in Schizophrenia-Bipolar Spectrum Disorders

Martina Katalin Szabó[1,2,3], **Veronika Vincze**[4], **Bernadett Dam**[5],
Csenge Guba[5], **Anita Bagi**[6], **István Szendi**[1,7]

[1]University of Szeged, Institute of Informatics, 6720 Szeged, Árpád tér 2
[2]Centre for Social Sciences, Computational Social Science
Research Center for Educational and Network Studies (CSS-RECENS)
1097 Tóth Kálmán street 4., T. building, 1st floor
[3]Tokyo University of Foreign Studies, Institute of Global Studies
3-11-1, Asahi-cho, Fuchu-shi, Tokyo 183-8534, Japan
[4]HUN-REN–SZTE Research Group on Artificial Intelligence
6720 Szeged, Tisza Lajos krt. 103, Hungary
[5]University of Szeged, Faculty of Humanities, Doctoral School of Linguistics
6722 Szeged, Egyetem utca 2, Hungary
[6]University of Szeged, Department of Hungarian Linguistics
6722 Szeged, Egyetem utca 2, Hungary
[7]Kiskunhalas Semmelweis Hospital, Teaching Hospital of the University of Szeged
Psychiatry Department
6400 Kiskunhalas, Dr. Monszpart László u. 1, Hungary

Szabo.Martina@tk.hu, vinczev@inf.u-szeged.hu, dam.bernadett@stud.u-szeged.hu,
csenge.guba@gmail.com, bagianita88@gmail.com, szendi@inf.u-szeged.hu

Abstract

In this study, we analyze spontaneous speech transcripts from Hungarian patients with schizophrenia, schizoaffective, and bipolar disorders. Our goal is to identify distinctive linguistic features in these patient groups and controls. To our knowledge, no prior study has systematically examined the linguistic features of these disorders or explored their use in distinguishing between these patient groups. We collected recordings from 77 participants during three directed spontaneous speech tasks in a clinical setting. Our research group manually transcribed the recordings. We processed the written corpus texts using Natural Language Processing methods and tools. The final corpus consists of 179,515 tokens, excluding punctuation. Using this data, we analyze different linguistic features' predictive power by computing and comparing their frequency distributions. We then attempt to automatically differentiate between patient groups and controls using our extensive set of linguistic features, employing the random forest algorithm in these experiments. Our results indicate that applying machine learning techniques based on distinctive features can effectively distinguish SZ, SAD, BD, and controls, surpassing baseline results.

Keywords: Speech Corpus, Schizophrenia, Schizoaffective disorder, Bipolar disorder

1. Introduction

Bipolar disorder (henceforth BD, previously called manic-depressive illness) is a recurrent chronic disorder characterized by episodes of mania, hypomania, and alternating or intertwining episodes of depression (Grunze, 2015). Schizophrenia (henceforth SZ) is a chronic mental health disorder characterized by symptoms of delusions, hallucinations, disorganized speech or behavior, as well as impaired cognitive ability (Patel et al., 2014). In the case of schizoaffective disorder (henceforth SAD), a person has mixed psychotic (hallucinations or delusions) and affective symptoms (mood episodes) (Malhi et al., 2008; Rose, 2014), hence it occupies an intermediate position between the two disorders in the schizophrenia-bipolar spectrum considered in a dimensional approach (Peralta and Cuesta, 2008). Cognitive impairment is a hallmark symptom of psychotic disorders including SZ and SAD, spanning verbal and non-verbal abilities (Van Rheenen et al., 2016; Little et al., 2019). Recent research findings indicate that patients with BD also have significant impairments in cognitive functioning (Van Rheenen et al., 2016). Susceptibility to psychosis spectrum disorders, including schizophrenia-bipolar spectrum, is genetically determined. These disorders usually manifest themselves during the reproductive phase of life. Among them, schizophrenia in particular leads to a significant decrease in fertility (53% for women, 77% for men) (Power et al., 2013). This reproductive disadvantage should lead to the rapid elimination of the given genes from the human

LREC-COLING 2024, pages 12938–12953
20-25 May, 2024.

genome. Conversely, cross-cultural constancy indicates that co-inheritance with some species-significant genetic variant may be the counterbalancing advantage for the entire population (Crow, 1993, 1995, 1997).

The most biologically determined characteristic of our species is language (Crow, 1996), and language separates modern humans from earlier hominids (Chomsky, 1986; Bickerton, 1995; Ganger and WOLD, 1998; Dronkers et al., 2000). During hominid evolution, the development of brain structural asymmetries is responsible for the development of human-specific components of language (Geschwind and Galaburda, 1985; Corballis, 2017), and these asymmetries are impaired in schizophrenia (Crow, 1998). According to a rather parsimonious conclusion, psychosis and language are related to genetic variation linked to the origin of the species (Berlim et al., 2003). This is why our research group has been comprehensively investigating the relationship between schizophrenia-spectrum disorders and language behavior. Since mental health influences the method of human communication, the acquisition and processing of linguistic data (spoken or written) provides an opportunity to reveal interrelation between linguistic factors and psychological aspects.

In this context, here we present a Hungarian corpus consisting of directed spontaneous speech texts produced by patients suffering from SZ, SAD or BD, as well as texts of healthy controls. Recordings transcribed later were produced in six different directed spontaneous speech tasks in a clinical environment. Our final corpus that was manually transcribed by the research group contains 458 texts and 179,515 tokens. Utilizing the corpus data we analyze and compare the speech of Hungarian SZ, SAD and BD patients. We seek to automatically identify and differentiate among them based on linguistic features of speech transcripts. Then, we analyze the predictive power of linguistic features by computing and comparing the frequency distributions of these features. We apply machine learning techniques based on a rich feature set that leads us to propose a methodology to identify and distinguish among SZ, SAD and BD on the basis of linguistic parameters of spontaneous speech.

Hence, the main contributions of the paper are the following:

- Based on a rich feature set of linguistic parameters we carry out a detailed statistical analysis that may distinguish healthy controls from SZ, SAD and BD patients.
- We use the transcripts of speech texts produced in narrative tasks.
- We perform machine learning experiments with the above-mentioned feature set for detecting SZ, SAD and BP and distinguish them from healthy controls and from each other.

2. Literature review

NLP and machine learning methods are increasingly used in the study of different types of mental health conditions. For instance, several studies have explored the possibility of utilizing acoustic features in depression detection or (Resnik et al., 2013; Akkaralaertsest and Yingthawornsuk, 2015; Taguchi et al., 2018) mild cognitive impairment and Alzheimer's disease (Haider et al., 2019; Vincze et al., 2021; Tóth et al., 2015) or Asperger syndrome (Chaput et al., 2013).

NLP and machine learning methods have been applied in the study of language usage and speech production in SZ, SAD and BD as well, but of these disorders SZ has received particular attention, and in most studies, the language use of people with SZ is compared to texts produced by healthy adults (Iter et al., 2018; Lundin et al., 2020; Mitchell et al., 2015). For a thorough review of NLP methods used in schizophrenia research, see Corcoran and Cecchi (2020). As for BD and SAD, relatively few papers have used computational methods to assess peculiarities of the language usage of patients suffering from these disorders.

Lott et al. (2002) used speech samples of 100 patients suffering from SZ, BD and major depression. They attempted to reveal linguistic abnormalities in the speech of these patients represent diagnosis-specific characteristics or constitute syndrome-like dimensions of these disorders. However, the majority of the linguistic variables did not prove to be statistically significant. Mota et al. (2017) analyse connectedness, a structural feature of speech in SZ and compare the results with BD and control groups. They aimed to verify whether speech disorganization during the first clinical contact, as measured by graph connectedness, could correctly classify negative symptoms and the SZ diagnosis 6 months in advance. Tan et al. (2021) found significant differences across five types of speech variables (utterances, single words, time/speaking rate, turns and formulation errors) between speech produced by SZ patients and healthy controls. Interestingly, the number and duration of pauses did not turn out to be significantly different variables between the two groups.

To the best of our knowledge, there is only one research work that compares linguistic features of text produced by the three patient groups in question, namely SZ, SAD and BD (Voleti et al., 2019). However, it is worth mentioning that patients with schizophrenia or schizoaffective disorder are not distinguished in this analysis. At the same time, Lundin et al. (2020) has even revealed that computational linguistic approaches are not only able to explain greater variance but even predict diagnosis better than clinician-rated scales, pointing out the importance of NLP-methods and tools in this research field.

One of the best analytical tools for text-based studies is spoken language corpus that can be analyzed using NLP methods. During the last few decades, several spoken language corpora have been created and utilized in different psycholinguistic studies in several languages as a result of research similar to the ones mentioned above (e.g. (Calvo et al., 2017; Corcoran et al., 2018; Little et al., 2019)), and among them we can find Hungarian language corpora as well (e.g. (Gosztolya et al., 2018; Bagi et al., 2019; Vincze et al., 2021; Kálmán et al., 2022)). Furthermore, there are some studies that investigate a specific linguistic feature (Szabó et al., 2023; Szabó et al., 2023). However, to the best of our knowledge, a Hungarian corpus which allows us to systematically compare the spontaneous speech of SZ, SAD, BD and controls had not been created prior to our recent research project.

As for Hungarian population, several papers deal with Hungarian patients suffering from SZ, SAD and BD (e.g. Kéri et al. (2001); Réthelyi et al. (2010); Inczédy-Farkas et al. (2010); Kocsis-Bogár et al. (2016); Döme et al. (2005); Kárpáti et al. (2018)). At the same time, so far no study has been conducted to systematically analyse linguistic features of these disorders. Furthermore, automatic discrimination among these patient groups based on linguistic features has not yet been addressed in the literature.

3. Corpus compilation

3.1. Text collection and transcription

In the present study, we employed the Hungarian database, recorded by the Prevention of Mental Illnesses Interdisciplinary Research Group (University of Szeged, Hungary) led by István Szendi. Data collection was approved by the Ethics Committee of the University of Szeged, and it was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all the participants involved in the research project. We have official written permission to use the recordings in our research. The database contains spontaneous speech recordings of people suffering from various mental disorders, as well as controls. The medical diagnosis for each person was also provided along with the speech samples. Here we focus on the linguistic characteristics of spontaneous speech produced by the four examined speaker groups. In the case of spontaneous speech, in contrast to planned speech, speakers do not have time to prepare their speech, which might truly reflect their language specificities, for instance, their difficulties in word finding (Vincze et al., 2021).

The first exercise consisted of three parts. The interviewer first asked the subjects to describe themselves (henceforth DescSelf) and then asked them to talk about their mother (henceforth DescMother) and father (henceforth DescFather). In the first part of the following task, the respondents were asked to recall the last years of their studies or the first years of their employment (henceforth YoungSelf). The interviewer then asked them to describe the same period in life of someone close to them (henceforth YoungOther). Lastly, in the third task, the subjects were asked to talk about their previous day (henceforth PrevDay). The final recorded speech corpus consists of 458 monologues from the 77 subjects all together.[1]

For demographic features (namely, gender, age[2], and education) see Table 1.

After data collection, recordings were transcribed manually. It was not necessary for us to use any specific software in this phase of the work since we did not attempt to transcribe the recordings according to the purposes of any phonological analysis. We did however mark pauses (filled and unfilled), hesitations, and false starts, as they might provide useful information about the language use of these populations. Recordings were transcribed and stored in a simple plain text format with UTF-8 character encoding.

Regarding the transcription process, transcribers worked on files from different groups interchangeably: we mixed the files to avoid potential individual biases of transcribers affecting the measurement outcomes. Then, in order to enhance the

[1] All recordings were conducted in the same city and within the same institution, ensuring consistent environmental conditions for all participants.

[2] The population under investigation exhibits considerable heterogeneity. Notably, the control group demonstrates a discernibly younger age profile compared to the other groups, whereas the BD group tends to skew older. In this study, we refrained from directly addressing this variance due to constraints posed by the limited size of our corpus; we aimed to preserve data integrity and avoid the loss of valuable information inherent in text deletion. However, in the next phase of our research, we intend to investigate the impact of using age as the sole discriminant factor in our baseline model. This will involve constructing a model where age is the only variable used to differentiate between groups. What is more, in future endeavors, we intend to enhance the robustness of our analyses. To mitigate the impact of age discrepancies more effectively, we propose expanding the corpus and implementing a stratified sampling methodology. This approach will ensure a more balanced representation of age cohorts across all groups, thus enhancing the generalizability and reliability of our findings.

	Groups				
	Control	SZ	SAD	BD	all
Number of participants	21	27	14	15	77
Number of texts	126	162	84	86	458
Age; M(SD)	36.42(10.49)	38.80(10.17)	41.43(9.73)	49.08(8.67)	40.63(10.71)
Education; M(SD)	14.76(3.05)	14.19(2.87)	14.54(2.98)	15.83(3.62)	14.73(3.09)
Sex ratio; f:m	8:13	9:18	10:4	8:7	35:42

Table 1: Basic Data of the Four Subject Groups. Education is given in years. (M = mean, SD = standard deviation)

quality of the transcription process, we employed a detailed and thorough guideline, which aided in achieving consistent and reliable results. Additionally, regular quality checks were conducted on randomly selected materials to ensure the accuracy and reliability of the transcriptions.

3.2. Corpus processing steps

Since the texts were manually transcribed there was no need for some data cleaning steps prior to the automatic analysis. Thus, as a first step, we performed on the transcribed texts an automatic linguistic analysis with magyarlanc, a linguistic preprocessing toolkit for Hungarian (Zsibrita et al., 2013) [3]. With this tool, the texts were first split into sentences, then tokenized, and finally the tokens were lemmatized and assigned a proper part-of-speech and morphological tag. Lemmatization is especially important in the case of morphologically rich languages such as Hungarian.

We extracted 17 basic statistical features, including but not limited to the number of sentences, and the number and frequency of distinct lemmas compared to the number of words.

Next, we made use of 10 speech-based features. Some examples include the number of pauses (filled and silent together), the number of hesitations compared to the number of tokens, and the number and frequency of pauses that follow an article and precede content words, as this might indicate that the given patients may have difficulties in finding the suitable content words.

During morphological and syntactic data processing, a total of 87 morphosyntactic features were extracted. These comprised, on one hand, part-of-speech features, including the count and occurrence rate of various parts-of-speech (nouns, verbs, adjectives, pronouns, numerals, adverbs, and conjunctions). On the other hand, numerous other morphological features were examined, such as the frequency of third person singular verb forms or occurrences of superlative adjectives, to name a few.

As for parts-of-speech, one of the useful features may be the number and frequency of unanalyzed words, i.e. those with an “unknown” POS tag, which could reflect whether neologisms are being created by the speaker while speaking.

Next, texts were also processed via some dictionary based analyses with which we explored some semantic and pragmatic linguistic features of the corpus, namely sentiment and emotion words, as well as the occurrence of discourse markers and intensifiers, among others. Here we analyzed 80 features altogether. The number and frequency of words belonging to several classes of linguistic uncertainty were extracted based on Vincze (2014). Positive and negative sentiment words were extracted based on two different Hungarian sentiment dictionaries: one of them was an automatic translation of an English sentiment lexicon (Liu, 2012) and the other one was a manually checked, corrected and supplemented lexicon of an automatic translation of the above-mentioned English dictionary (Szabó, 2015). The number and frequency of words belonging to the emotions were also extracted automatically with the help of a Hungarian emotion lexicon described in Szabó et al. (2016). The decision to focus on these features was justified by the fact that emotion regulation dysfunction is characteristic of psychotic disorders (Kring and Elis, 2013; Chapman et al., 2020; Green et al., 2007). For instance, based on previous research on emotions, people with schizophrenia have difficulty in sensing and predicting emotional events, integrating emotional impressions and contexts, as well as the richness and maintenance of emotional experiences (Kring and Elis, 2013). The underlying brain activities show a deficit in the functioning of the networks responsible for cognitive control, indicating insufficient integration of emotions and cognition (Kring and Elis, 2013). Moreover, an abnormally elevated mood in BD is associated with specific neurocognitive deficits consistent with neuropathology in neural networks that are critical for emotion reg-

[3] This tool has an accuracy of 96.33% in terms of POS-tagging. As for dependency parsing, magyarlanc achieved an accuracy of 91.42% (Labeled Attachment Score) and 93.22 % (Unlabeled Attachment Score), making it highly reliable (Zsibrita et al., 2013).

ulation (Green et al., 2007). Hence, we assumed that emotion, as well as sentiment analysis of this corpus may produce relevant results concerning the mental disorders in question.

As for linguistic intensification, recent research findings lead us to the conclusion that the use of intensifiers is closely related to emotion regulation (Athanasiadou, 2007; Strous et al., 2009). What is more, according to Athanasiadou (2007), intensifiers are linguistic markers of speaker subjectivity, and they have the primary function of signifying the speaker's point of view and attitude. It is worth mentioning here that there is evidence in the literature that the use of intensifiers is different in e.g. SZ (Strous et al., 2009). Then, because of the links between mental illness and emotion regulation, within the group of intensifiers, it is worth focusing on to the so-called negative emotive intensifiers (henceforth NEIs), whose prior semantic content is related to a negative emotion, but which can function as intensifiers. For the identification of linguistic intensifiers, we used a standard register (non-emotive) intensifier dictionary (Szabó et al., 2023) consisting of 125 words and a 225-item dictionary of NEIs (Szabó and Guba, 2023). (For more details about these lexicons, including their development, validation, and previous applications, refer to (Szabó et al., 2023), (Szabó and Guba, 2023) and (Szabó et al., 2022).)

As regards pragmatic features of the transcripts, we processed speech act verbs and discourse markers. The number and frequency of speech act verbs we extracted based on a manually constructed list derived from (Vincze et al., 2021). To find discourse markers in the texts we applied a word list based on Dér et al. (2007).

All these resources applied in the recent analysis were chosen for their relevance and applicability to our study's objectives and have been utilized in similar research contexts, ensuring their validity.

In Table 9, we present an example for each of the six major feature categories from the corpus, along with their English translations.

With the above described automatic analysis we extracted a rich-feature set of the corpus, consisting of 194 linguistic features altogether.

We provide a comprehensive list of features in a form of separate tables in appendix (See Section 10).

4. Statistical analysis and machine learning experiments

In order to quantify the usefulness of each feature in distinguishing SZ, SAD and BD patients and the controls, we carried out a statistical analysis of the data, namely pairwise t-tests [4] for each feature and transcript. In addition to this, we also sought to automatically discriminate the different subject groups, using the above mentioned rich feature set. In order to examine which types of linguistic features play the most important role in distinguishing among the three patient groups and the controls, the six large subsets of features were used the following way: all of the six groups were used, except for one group at a time.

We trained a random forest classifier of the WEKA package (Hall et al., 2009) (with Weka's default settings) with the above mentioned feature set. We used ten fold cross validation. Our baseline method was majority classification, which achieved an accuracy of 35.0649%.

5. Results

5.1. Results of significance tests

In order to make the results of significance tests more comprehensible, they are discussed here according to subgroups of language features.

As a first step, we compare the results measured for all patient groups with the results for the control group. Here, there is no opportunity to comprehensively review all the results; therefore, we only highlight some of the most interesting statistically significant differences.

Figure 1) presents the results using each feature group below, while Figure 3) in Appendix illustrates the results when excluding these individual groups and using all the remaining ones together (See Appendix 10).

Figure 1 shows the comparison of results across all patient groups and the control group using each feature group, excluding specific groups.

Regarding speech-based features, pause rates (filled or unfilled) differ significantly in only two tasks between all patient groups and the group of healthy controls (afterwards: HC): in the "DescMother" task the patients used a higher rate of filled pauses, while the HC group used more pauses after articles, i.e. when naming content words. However, in the "PrevDay" task, HC-s used more filled pauses. The ratio of nouns and verbs is significantly different in most the sub-corpora. The ratio of unknown words, i.e. those with an "unknown" POS tag proved to be significantly different in the case of the "PrevDay" task; patients uses more of them compared to the HC group. The frequency of pronouns significantly differs both in the "YoungOther" and the "PrevDay" subcorpora because patients use them less in both the cases. Observable differences can be noted in terms of some specific semantic feature as well. For instance, patients use more positive words in the "YoungSelf" and the "YoungOther" subcorpora, however, more negative sentiment words in the self description task ("DescSelf"). The rate of

[4] The p-value threshold for significance is 0.05.

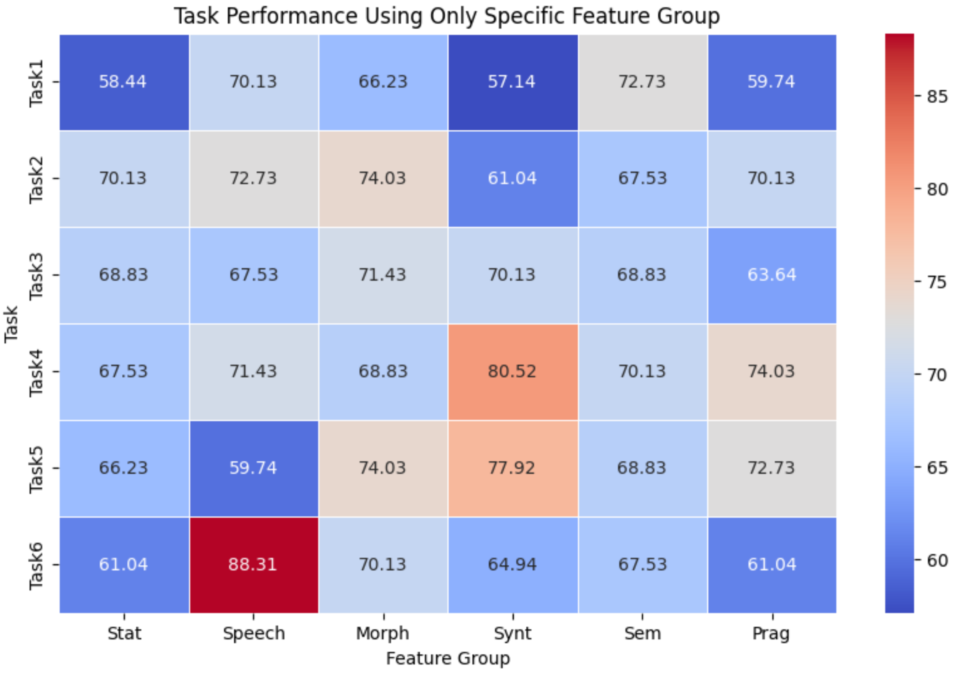

Figure 1: Comparing Results Across All Patient Groups and the Control Group Using Each Feature Group

function and content words is significantly different in all the description tasks ("DescSelf", "DescMother" and "DescFather"). Patients tend to use less function and more content words generally. As for the results of uncertainty detection, patients tend to use less doxastic-type words and the difference is significant is most subcorpora. Among the emotion words, there is no significant difference with one exception, namely words of sorrow are represented with a higher frequency in the speech of the patients in some tasks.

Now let us turn to the results broken down into individual groups.

Figure 2) presents the results using each feature group below, while Figure 4) in Appendix illustrates the results when excluding these individual groups and using all the remaining ones together (See Appendix 10). What is more, Figure 5 in the Appendix illustrates these results, but for all tasks combined (See Appendix 10).

Figure 2 shows the comparison of results across all patient groups and the control group using each feature group, excluding specific groups.

For the sake of simplicity, here we will basically compare each patient group to the HC-s.

As for statistical features, SZ patients use the least amount of sentences in each task, while the BD and the HC group use significantly more. Similarly, SZ patients produce the least amount of tokens in each task, while BD patients have the most, sometimes twice as much. On the other hand, SZ patients have the highest rate of lemmas in most tasks (except in the "PrevDay" task). SAD patients show similar patterns to SZ patients, however, when compared to the HC, no feature shows significant difference consistently between the two groups.

Regarding speech-based features, when broken down to individual patient groups, there does not seem to be a clear pattern of significant differences in terms of the rate of filled or unfilled pauses. The most apparent contrast to the HC group can be detected in BD patients, who have a lower rate of filled and unfilled pauses in several tasks.

As for morphological and syntactic features, in general, healthy controls use more function words and fewer content words than patients, which manifests in the rate of nouns and verbs. Saliently, the POS distribution used by SZ patients is notably different from healthy controls in most tasks: they use more nouns, verbs and adjectives, on the other hand, fewer adverbs, postpositions and conjunctions. It seems that their verbal thinking is centered around content, however, connectives and other function words can be detected to a lesser degree in their way of thinking, which may reflect a different mental organization of verbal information. Finally, it is also noticeable that BD patients use more complex sentences.

When it comes to sentiment analysis results, SZ and BD patients tend to employ significantly more positive words when discussing someone close to them compared to the HC group ("YoungOther"). At the same time, when discussing themselves

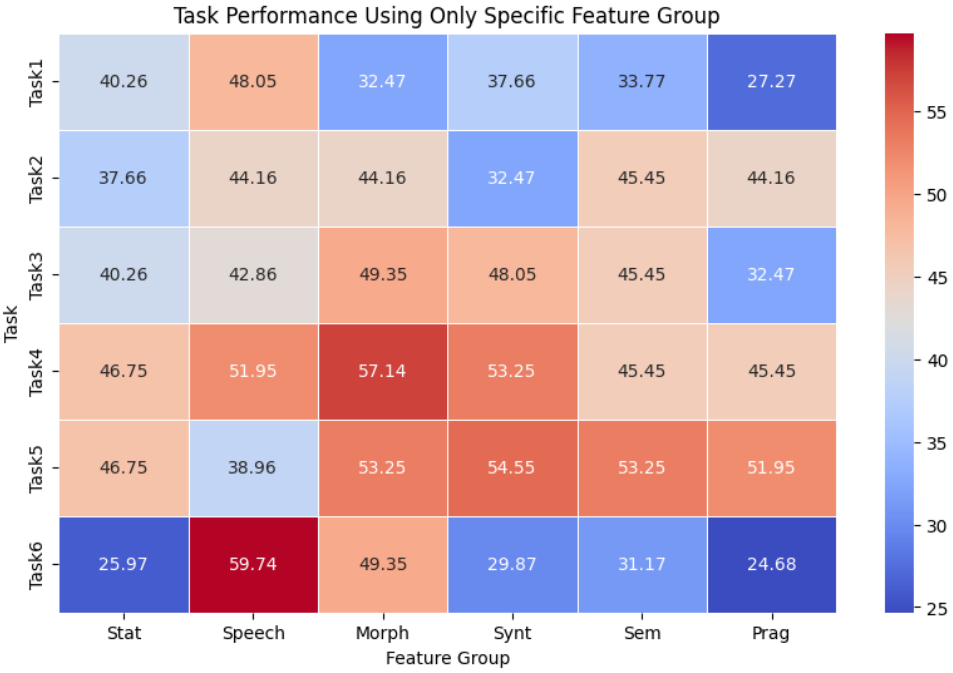

Figure 2: Comparing Each Patient Group Separately with the Control Group Using Each Feature Group

in a broader context ("DescSelf"), the same two speaker groups tend to use significantly more negative words. Regarding the two types of intensifiers, the SZ group uses fewer of them for both standard and negative emotive intensifiers, and the difference compared to the HC and BD groups appeared significant in several tasks.

When analyzing uncertainty in the transcripts, some more interesting findings could be detected. First, patients use significantly fewer words referring to doxastic uncertainty (i.e. phrases like *I think*, *I believe*) in the "DescMother" task than the HC group. Second, the BD and SAD groups employ significantly more peacock words (i.e. subjective expressions) when describing their father in the "DescFather" task than the HC group. The low number of doxastic uncertainty may refer to doxasms present in the thinking of patients with mental disorders (i.e. they sometimes lack the ability to differentiate between their thoughts and the reality). The higher number of peacock words, however, might reflect a different way of perceiving the relationship towards their father, which would need further investigation.

When examining the words related to emotions, it becomes evident that SZ speakers use significantly more words associated with joy than BD speakers in the "DescSelf" task. Additionally, they use significantly more words related to joy than both the SAD and HC groups in the "DescMother" task. In contrast, a significant difference in the use of words related to sorrow can be observed between BD speakers and the HC group in multiple tasks, including "DescSelf", "DescMother" and "DescFather". This variation in language usage may be attributed to various factors. Many BD patients report experiences of family tragedy, a challenging childhood, or strained relationships with their parents. The SAD group uses the fewest words related to fear in each task.

	Subject groups			
	BD	**SAD**	**SZ**	**Total**
DescSelf	10	15	37	62
DescMother	25	14	51	90
DescFather	10	23	11	44
YoungOther	16	10	85	111
YoungSelf	6	22	75	103
PrevDay	8	9	10	27
Total	75	93	269	437

Table 2: The Number of Statistically Significant Features for Each Task and Each Feature Group.

Examining the pragmatic features, the most striking are the differences between the use of discourse markers. The HC group typically uses the most discourse markers in each task, with one exception ("PrevDay").

In order to assess which speech tasks exhibit a higher proportion of significant differences and which ones may be less useful due to fewer significant differences, we have aggregated the results of the significance tests, see Table 2.

According to the data in Table 2, the "YoungOther"

Feature type	Metrics	without	only
Statistical	Acc	64.9351	49.3506
	F	0.625	0.467
Speech-based	Acc	50.6494	63.6364
	F	0.465	0.622
Morphological	Acc	57.1429	48.0519
	F	0.540	0.458
Syntactic	Acc	63.6364	45.4545
	F	0.591	0.418
Semantic	Acc	57.1429	54.5455
	F	0.496	0.493
Pragmatic	Acc	51.9481	42.8571
	F	0.492	0.380
All	Acc	58.4416	58.4416
	F	0.539	0.539

Table 3: Results of Ablation Analysis. Acc: accuracy, F: F-score.

task is the most effective in distinguishing categories, and the "YoungSelf" task can also be useful in our machine learning experiments. Surprisingly, in the "PrevDay" task, there are notably less significant differences compared to the other tasks. When examining the three disorders individually, it is noticeable that father's description ("DescFather") works best for SAD patients, while for BD, it is the mother's description task ("DescMother"). It also becomes evident that our feature set may be most effective in identifying SZ. What is more, if we want to reduce the data for specific speech tasks, "YoungOther", "YoungSelf", and "DescMother" may be sufficient for implementing a well-functioning machine learning.

5.2. Results of machine learning experiments

As can be seen, the overall accuracy was 58.44%, which highly outperforms the baseline. The algorithm, as expected, was the most successful at detecting SZ subjects (F-score: 0.75), while its performance was the lowest in the case of SA patients (F-score: 0.22).

In order to examine the efficiency of individual feature sets, we carried out an ablation analysis, by omitting and/or applying each and every feature group from the feature set. In this way, we found that speech based features proved to be the most effective (63.64%) when applying only this group of features, and excluding them gave us the lowest accuracy value (50.65%). We achieved better results than overall accuracy when we excluded the statistical feature group and used all the other features together (64.94%), as well as when we did the same with the syntactic features (63.64%).

In order to see how controls and patients can be distinguished, we carried out another experiment by dividing the subjects into two groups: all patient groups together vs. the control group. In this way, an accuracy of 72.73% was obtained. The algorithm was successful at detecting PT subjects (F-score: 0.84). When analysing the results in each task, it was revealed that we achieved the highest accuracy by applying the "PrevDay" speech texts (75.33%). The least useful task was "DescSelf" in this respect (68.83%). If we consider individual speech task types and feature groups, the highest efficiencies were the following: speech-based features in "PrevDay" (88.31%), syntactic features in "YoungSelf" (80.52%) and syntactic features in "YoungOther" (77.92%). These results are consistent with the findings obtained from the 4-group measurement (see above).

6. Discussion

First, let us consider the results obtained when we worked with 4 patient groups. We observe a commendable overall accuracy of 58.44%, which significantly surpasses the baseline performance. Notably, the algorithm exhibits varying degrees of success in distinguishing between different subject groups. It excels in identifying SZ subjects, as indicated by the high F-score of 0.754.

Then, the results of the ablation analysis shed light on the significance of different feature groups in our classification task. One key finding was that speech-based features emerged as the most effective, achieving an accuracy of 63.64%. This outcome underscores the pivotal role of speech-based features in enhancing our classification model's performance. Furthermore, excluding the statistical feature group yielded improved results, and the same was observed in the case of syntactic features. These outcomes emphasize the intricate interplay of different feature groups in our classification task and highlight the potential for further optimization.

An interesting finding of this research is that if we exclude NEIs (Szabó et al., 2022) from the analysis, the overall accuracy (which was 58.44% again) decreases to 54.55% (see Section1 above). So, by using the list of them, we were able to further improve the machine learning results. Szabó et al. (2023) has shown significant differences in the use of NEIs among the speaker groups studied here, although this characteristic does not hold true for all (non-negative) intensifiers, i.e. standard-register intensifiers. The results of the current study demonstrate that these findings can be applied in the automatic classification of groups, as they improve the results. In line with the above, we have also observed here that removing standard-register intensifiers from machine learning experiments leads to improved results: 61.04%. So, it is indeed the case that only a specific group of intensifiers, the NEIs, can be effectively used in the current machine learning project. These findings further highlight the complexities of automatic classification in this domain. As we have presented, we carried out another machine learning experiment by dividing the subjects into two groups, and we obtained an accuracy of 72.73%. It is interesting to note, on the other hand, that the omission of deep morphological, as well as semantic features increases the results, so it seems that they have a negative effect on identifying PT. This is a phenomenon which we plan to investigate in more details in the future.

7. Conclusions and future work

In this paper, we conducted a comprehensive analysis of spontaneous speech text produced by Hungarian patients suffering from SZ, SAD and BD. Our aim was to identify distinctive linguistic features among these patient groups and controls. For our analysis, we collected speech recordings from 77 subjects participating in three different directed spontaneous speech tasks within a clinical environment, resulting in a corpus of 458 texts. These recordings were manually transcribed by our research group and processed using Natural Language Processing methods and tools. The final corpus comprised 179,515 tokens, excluding punctuation. Leveraging this data, we examined the predictive power of various linguistic features by computing and comparing their frequency distributions. Subsequently, we employed machine learning techniques (random forest algorithm) to automatically discriminate between SZ, SAD, BD patients, and the HCs.

Our results demonstrate that by applying machine learning techniques based on the identified distinctive linguistic features, we can achieve meaningful results in the automatic discrimination among SZ, SAD, BD, and the control group, surpassing baseline results. Our study highlights the potential of linguistic features for differentiating among them.

In our experiments, the overall accuracy reached an impressive 58.44%, surpassing the baseline. The algorithm excelled in detecting SZ subjects with an F-score of 0.75. Ablation analysis revealed that speech-based features were the most effective (63.64%).Moreover, distinguishing controls from patients achieved an accuracy of 72.73%, excelling at detecting PT subjects (F-score: 0.84). Our research also highlighted that further improvements can be made with the use of additional specific dictionaries, such as the lexicon of NEIs.

While our approach may not signify a complete paradigm shift, it embodies two significant innovations. First, within the realm of the Hungarian language, our study stands out for its innovation as it delves into linguistic patterns and features in Hungarian speech that have not been systematically analyzed before in the context of the mental health disorders under consideration. Second, on an international scale, the incorporation of NEIs in our analysis represents a novel approach. This aspect of our work is pioneering as it broadens the scope of linguistic analysis within the realms of machine learning and mental health research.

In our future work, we can extend this research in several directions to gain a deeper understanding of the relationship between these mental illnesses and linguistic features. First, we may expand our linguistic feature set to conduct a more detailed examination of differences among groups. Then, a fine-tuning our machine learning methods could lead to even better results. Experimentation with different algorithms and parameter tuning may be beneficial. We shall also carry out a deeper emotion analysis (exploring their cues like linguistic intensification) and their relationship with these disorders. As another subsequent step in our research, we plan to conduct a dedicated analysis focusing solely on the most effective types of recordings and the optimal selection of features identified in our study. By narrowing down our analysis to these specific recordings and feature sets, we anticipate that the results will provide more precise insights into the linguistic patterns and variations within the studied groups. This targeted approach will help us further refine our understanding of the relationships between linguistic features and the conditions under investigation. As part of our future plans, we are actively exploring the implementation of various machine learning algorithms beyond the random forest. This includes considering algorithms such as support vector machines, neural networks, and other advanced techniques to further enhance the predictive capabilities of our models. Then, in our future research endeavors, we plan to actively experiment with different sets of hyperparameters to further optimize the performance of our model. This will include adjusting parameters such as the number of trees, maximum depth, and minimum samples split. Through this planned hyperparameter tuning process, we aim to enhance the predictive accuracy and generalization capability of our model, ensuring our findings are robust and reliable. Last, we would like to investigate data from languages other than Hungarian, reinforcing the generalizability of our findings.

We are currently in the process of masking sensitive data from the corpus in accordance with respecting ethical and privacy considerations. Once this masking process is completed, we intend to release the corpus for research purposes in order to ensure transparency and reproducibility in scientific research.

8. Acknowledgements

This work was supported by the EFOP-3.6.1-16-2016-00008 (sub-project: „Szkizofrénia – Nyelv – Innovatív gyógyítás”), HUN-REN Centre for Social Sciences, the University of Szeged, Department of Informatics and the Institute of Global Studies, Tokyo University of Foreign Studies, Japan.
Martina Katalin Szabó was funded by the OTKA Postdoctoral Research Grant, the Hungarian Research Fund of the National Research, Development and Innovation Office of Hungary (NKFIH) (grant number: PD 132312) and by the International Research Fellowship Program of Japan Society for the Promotion of Science (JSPS, Postdoctoral Fellowships for Research in Japan (Standard)).

9. Bibliographical References

Thaweewong Akkaralaertsest and Thaweesak Yingthawornsuk. 2015. Comparative analysis of vocal characteristics in speakers with depression and high-risk suicide. *International Journal of Computer Theory and Engineering*, 7(6):448.

Angeliki Athanasiadou. 2007. On the subjectivity of intensifiers. *Language sciences*, 29(4):554–565.

Anita Bagi, Gábor Gosztolya, Szilvia Szalóki, István Szendi, and Ildikó Hoffmann. 2019. Szkizofrénia azonosítása spontán beszéd temporális paraméterei alapján–egy pilot kutatás eredményei. In *XV. Magyar Számítógépes Nyelvészeti Konferencia*, pages 189–201. Szegedi Tudományegyetem, Informatikai Intézet.

Marcelo T Berlim, Betina S Mattevi, Paulo Belmonte-de Abreu, and Timothy J Crow. 2003. The etiology of schizophrenia and the origin of language: overview of a theory. *Comprehensive psychiatry*, 44(1):7–14.

Derek Bickerton. 1995. *Language and human behavior*. Seattle: University of Washington Press.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Hannah C Chapman, Katherine F Visser, Vijay A Mittal, Brandon E Gibb, Meredith E Coles, and Gregory P Strauss. 2020. Emotion regulation across the psychosis continuum. *Development and psychopathology*, 32(1):219–227.

Valérie Chaput, Frédérique Amsellem, Isabel Urdapilleta, Pauline Chaste, Marion Leboyer, Richard Delorme, and Véronique Goussé. 2013. Episodic memory and self-awareness in asperger syndrome: Analysis of memory narratives. *Research in Autism Spectrum Disorders*, 7(9):1062–1067.

Noam Chomsky. 1986. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Michael C Corballis. 2017. The evolution of language. *Annals of the New York Academy of Sciences*.

Cheryl M Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C Javitt, Carrie E Bearden, and Guillermo A Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.

Cheryl Mary Corcoran and Guillermo A. Cecchi. 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. Understanding the Nature and Treatment of Psychopathology: Letting the Data Guide the Way.

Timothy J Crow. 1995. A darwinian approach to the origins of psychosis. *British Journal of Psychiatry*, 167(1):12–25.

Timothy J Crow. 1997. Schizophrenia as failure of hemispheric dominance for language. *Trends in neurosciences*, 20(8):339–343.

TJ Crow. 1993. Sexual selection, machiavellian intelligence, and the origins of psychosis. *The Lancet*, 342(8871):594–598.

TJ Crow. 1996. Language and psychosis: common evolutionary origins. *Endeavour*, 20(3):105–109.

TJ Crow. 1998. Why cerebral asymmetry is the key to the origin of homo sapiens: how to find the gene or eliminate the theory. *Cahiers De Psychologie Cognitive-Current Psychology of Cognition*, 17(6):1237–1277.

Csilla Ilona Dér, Alexandra Markó, T Gecső, and Cs Sárdi Csilla. 2007. A magyar diskurzusjelölők szupraszegmentális jelöltsége [suprasegmental markedness of Hungarian discourse markers]. *Nyelvelmélet–nyelvhasználat [Linguistic theory and language use](Segédkönyvek a nyelvészet tanulmányozásához 74)*, pages 61–7.

P Döme, Z Rihmer, X Gonda, P Pestality, G Kovács, Z Teleki, and P Mandl. 2005. Cigarette smoking and psychiatric disorders in Hungary. *International Journal of Psychiatry in Clinical Practice*, 9(2):145–148.

Nina F Dronkers, Steven Pinker, and Antonio Damasio. 2000. Language and the aphasias. *Principles of neural science*, 4:1169–1187.

Jennifer Ganger and KARIN STROMS WOLD. 1998. Innateness, evolution, and genetics of language. *Human biology*, pages 199–213.

Norman Geschwind and Albert M Galaburda. 1985. Cerebral lateralization: Biological mechanisms, associations, and pathology: I. a hypothesis and a program for research. *Archives of neurology*, 42(5):428–459.

Gábor Gosztolya, Anita Bagi, Szilvia Szalóki, István Szendi, and Ildikó Hoffmann. 2018. Identifying schizophrenia based on temporal parameters in spontaneous speech. In *INTERSPEECH*, pages 3408–3412. International Speech Communication Association (ISCA).

Melissa J Green, Catherine M Cahill, and Gin S Malhi. 2007. The cognitive and neurophysiological basis of emotion dysregulation in bipolar disorder. *Journal of affective disorders*, 103(1-3):29–42.

Heinz Grunze. 2015. Bipolar disorder. In *Neurobiology of brain disorders*, pages 655–673. Elsevier.

Fasih Haider, Sofia De La Fuente, and Saturnino Luz. 2019. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Gabriella Inczédy-Farkas, Judit Benkovits, Nóra Balogh, Péter Álmos, Beáta Scholtz, Gábor Zahuczky, Zsolt Török, Krisztián Nagy, János Réthelyi, Zoltán Makkos, et al. 2010. Schizobank – the Hungarian national schizophrenia biobank and its role in schizophrenia research. *Orvosi Hetilap*, 151(35):1403–1408.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

János Kálmán, Davangere P Devanand, Gábor Gosztolya, Réka Balogh, Nóra Imre, László Tóth, Ildikó Hoffmann, Ildikó Kovács, Veronika Vincze, and Magdolna Pákáski. 2022. Temporal speech parameters detect mild cognitive impairment in different languages: validation and comparison of the speech-gap test® in English and Hungarian. *Current Alzheimer Research*, 19(5):373–386.

Eszter Kárpáti, Anita Bagi, István Szendi, Lujza Beatrix Tóth, Karolina Janacsek, and Ildikó Hoffmann. 2018. Rekurzió egy szkizoaffektív zavarral élő személy diskurzusaiban–esettanulmány. *Iskolakultúra: Pedagógusok Szakmai-Tudományos Folyóirata*, 28(5-6):40–54.

Szabolcs Kéri, O Kelemen, G Benedek, and Z Janka. 2001. Different trait markers for schizophrenia and bipolar disorder: a neurocognitive approach. *Psychological medicine*, 31(5):915–922.

Krisztina Kocsis-Bogár, Zsófia Nemes, and Dóra Perczel-Forintos. 2016. Factorial structure of the Hungarian version of oxford-liverpool inventory of feelings and experiences and its applicability on the schizophrenia-schizotypy continuum. *Personality and Individual Differences*, 90:130–136.

Ann M Kring and Ori Elis. 2013. Emotion deficits in people with schizophrenia. *Annual review of clinical psychology*, 9:409–433.

Bethany Little, Peter Gallagher, Vitor Zimmerer, Rosemary Varley, Maggie Douglas, Helen Spencer, Derya Çokal, Felicity Deamer, Douglas Turkington, I Nicol Ferrier, et al. 2019. Language in schizophrenia and aphasia: the relationship with non-verbal cognition and thought disorder. *Cognitive Neuropsychiatry*, 24(6):389–405.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

PR Lott, S Guggenbühl, A Schneeberger, AE Pulver, and HH Stassen. 2002. Linguistic analysis of the speech output of schizophrenic, bipolar, and depressive patients. *Psychopathology*, 35(4):220–227.

Nancy B Lundin, Jesse Hochheiser, Kyle S Minor, William P Hetrick, and Paul H Lysaker.

2020. Piecing together fragments: linguistic cohesion mediates the relationship between executive function and metacognition in schizophrenia. *Schizophrenia research*, 215:54–60.

Gin S Malhi, Melissa Green, Andrea Fagiolini, Eric D Peselow, and Veena Kumari. 2008. Schizoaffective disorder: diagnostic issues and future recommendations. *Bipolar Disorders*, 10(1p2):215–230.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Natália B Mota, Mauro Copelli, and Sidarta Ribeiro. 2017. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophrenia*, 3(1):1–10.

Krishna R Patel, Jessica Cherian, Kunj Gohil, and Dylan Atkinson. 2014. Schizophrenia: overview and treatment options. *Pharmacy and Therapeutics*, 39(9):638.

Victor Peralta and Manuel J Cuesta. 2008. Exploring the borders of the schizoaffective spectrum: a categorical and dimensional approach. *Journal of affective disorders*, 108(1-2):71–86.

Robert A Power, Simon Kyaga, Rudolf Uher, James H MacCabe, Niklas Långström, Mikael Landen, Peter McGuffin, Cathryn M Lewis, Paul Lichtenstein, and Anna C Svensson. 2013. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA psychiatry*, 70(1):22–30.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1348–1353.

János M Réthelyi, Steven C Bakker, Patrícia Polgár, Pál Czobor, Eric Strengman, Péter I Pásztor, René S Kahn, and István Bitter. 2010. Association study of nrg1, dtnbp1, rgs4, g72/g30, and pip5k2a with schizophrenia and symptom severity in a Hungarian sample. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 153(3):792–801.

D Rose. 2014. Schizophrenia/psychosis. In *Encyclopedia of the neurological sciences*, pages 99–103. Academic press.

Rael D Strous, Moshe Koppel, Jonathan Fine, Smadar Nachliel, Ginette Shaked, and Ari Z Zivotofsky. 2009. Automated characterization and identification of schizophrenia in writing. *The Journal of nervous and mental disease*, 197(8):585–588.

Martina Katalin Szabó. 2015. Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához*, 177:278–285.

Martina Katalin Szabó, Bernadett Dam, and Veronika Vincze. 2023. On the Semantic Development of Negative Emotive Intensifiers in Hungarian News. Manuscript.

Martina Katalin Szabó and Csenge Guba. 2023. Analysis of negative emotive intensifiers in Hungarian tweets. Unpublished.

Martina Katalin Szabó, Veronika Vincze, and Károly Bibok. 2022. "thank you for the terrific party!"–an analysis of Hungarian negative emotive words. *Corpus Linguistics and Linguistic Theory*, 19(0).

Martina Katalin Szabó, Veronika Vincze, Csenge Guba, Bernadett Dam, Adrienn Solymos, Anita Bagi, and István Szendi. 2023. Fokozás szkizofréniában. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 17–32.

Martina Katalin Szabó, Veronika Vincze, and Gergely Morvay. 2016. Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. *Távlatok a mai magyar alkalmazott nyelvészetben*, page 282.

Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto, Masayuki Suzuki, Toru Nagano, Ryuki Tachibana, Masafumi Nishimura, and Tetsuaki Arai. 2018. Major depressive disorder discrimination using vocal acoustic features. *Journal of affective disorders*, 225:214–220.

Eric J. Tan, Denny Meyer, Erica Neill, and Susan L. Rossell. 2021. Investigating the diagnostic utility of speech patterns in schizophrenia and their symptom associations. *Schizophrenia Research*, 238:91–98.

László Tóth, Gábor Gosztolya, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Edit Biró, Fruzsina Zsura, Magdolna Pákáski, and János Kálmán. 2015. Automatic detection of mild cognitive impairment from spontaneous speech using asr. In *Proc. Interspeech 2015*, pages 2694–2698. ISCA.

Tamsyn E Van Rheenen, Shayden Bryce, Eric J Tan, Erica Neill, Caroline Gurvich, Stephanie Louise, and Susan L Rossell. 2016. Does cognitive performance map to categorical diagnoses of schizophrenia, schizoaffective disorder and bipolar disorder? a discriminant functions analysis. *Journal of Affective Disorders*, 192:109–115.

Veronika Vincze. 2014. Uncertainty detection in Hungarian texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1844–1853.

Veronika Vincze, Martina Katalin Szabó, Ildikó Hoffmann, László Tóth, Magdolna Pákáski, János Kálmán, and Gábor Gosztolya. 2021. Linguistic parameters of spontaneous speech for identifying mild cognitive impairment and alzheimer's disease. *Computational Linguistics*, pages 1–34.

Rohit Voleti, Stephanie Woolridge, Julie M Liss, Melissa Milanovic, Christopher R Bowie, and Visar Berisha. 2019. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. *arXiv preprint arXiv:1904.10622*.

János Zsibrita, Veronika Vincze, and Richárd Farkas. 2013. magyarlanc: A tool for morphological and dependency parsing of Hungarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 763–771, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

10. Appendices

In the following tables (10, 10, 10, 10, 10), we list all the features that we processed in the corpus. Each group of features is presented in separate tables. For each feature, we separately processed its count ("Num") and its frequency relative to the number of tokens ("Rate"), which are not listed separately in the tables.
Figure 3 presents another comparison, but with specific groups excluded.

Code	Description
token	Tokens
sentence	Sentences
lemma	Lemmas
sentenceLength	Average sentence length
allUpper	All-uppercase words
firstUpper	Words with initial uppercase
declarSent	Declarative sentences
imperSent	Imperative sentences
question	Question sentences
accent	Accent feature
saturation	Saturation feature

Table 4: Statistical Features and Codes

Code	Description
hesit	Hesitations
uncertain	Uncertain instances
pause	Pauses
filledPause	Filled pauses
artPause	Artificial pauses

Table 5: Speech-based Features and Codes

Code	Description
noun	Nouns
verb	Verbs
adj	Adjectives
x	Unknown morphemes
adv	Adverbs
properNoun	Proper nouns
num	Numerals
conj	Conjunctions
punct	Punctuation marks
pron	Pronouns
relPron	Relative pronouns
demPron	Demonstrative pronouns
adpos	Adpositions
multiplePunct	Multiple punctuation marks

Table 6: Morphological Features and Codes

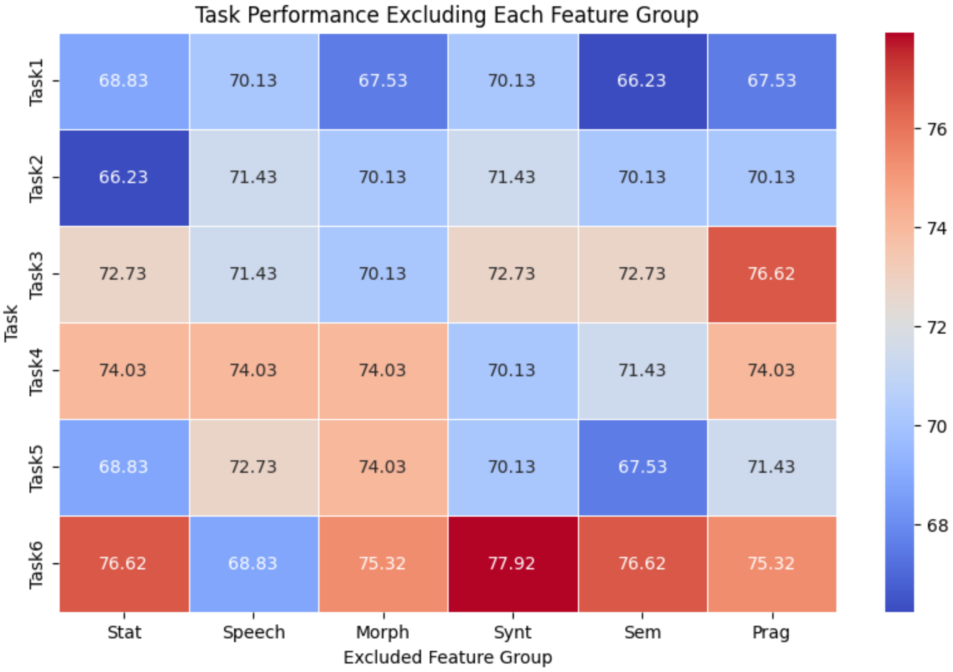

Figure 3: Comparing Results Across All Patient Groups and the Control Group Using Each Feature Group, Excluding Specific Groups

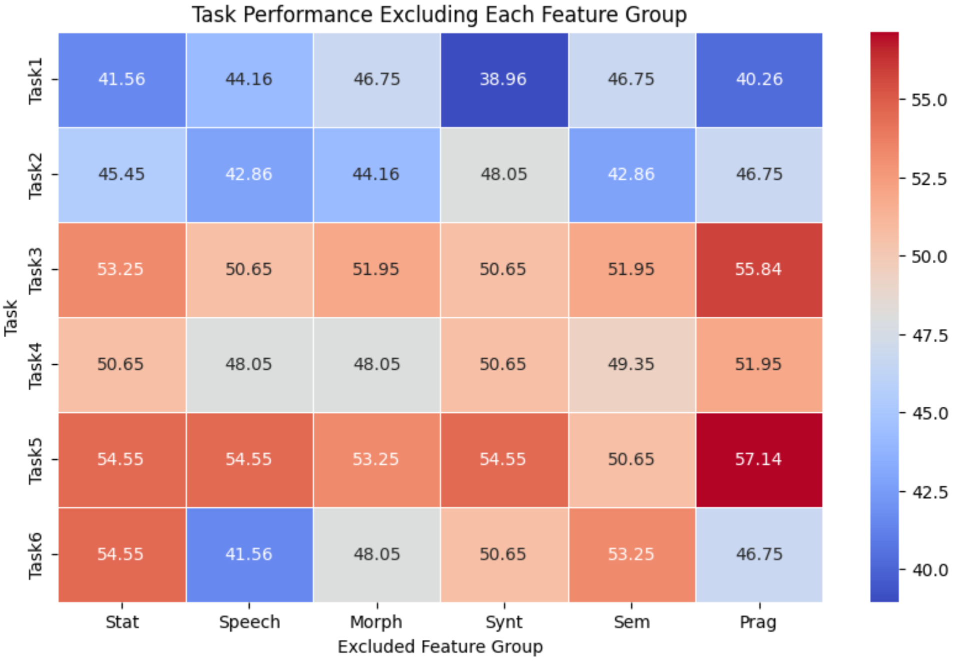

Figure 4: Comparing Each Patient Group Separately with the Control Group Using Each Feature Group, Excluding Specific Groups

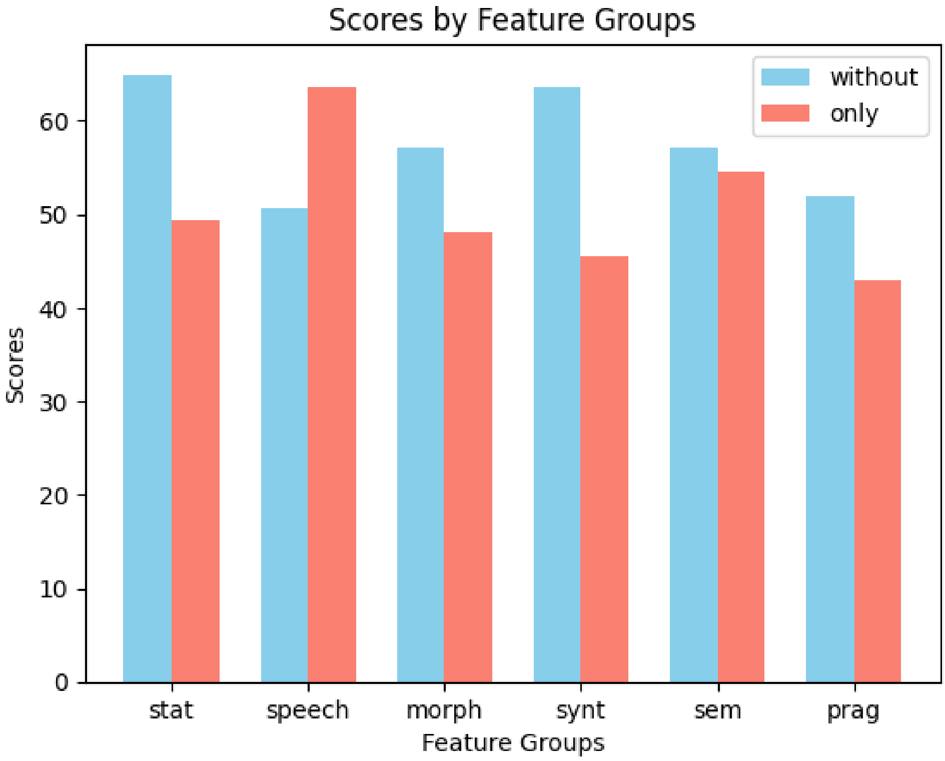

Figure 5: Comparing Each Patient Group Separately with the Control Group Using Each Feature Group for All Tasks

Code	Description
positive	Positive words
negative	Negative words
precoPositive	Precoded positive words
precoNegative	Precoded negative words
negemo	Negative emotions
negation	Negations
function	Function words
content	Content words
vulgar	Vulgar words
racist	Racist words
specStyle	Special style words
addressing	Addressing words
closing	Closing words
postscript	Postscript words
negemo2	Secondary negative emotions
intens	Intensifiers
memory	Memory-related words
epistemic	Epistemic words
invest	Investment-related words
condition	Condition-related words
weasel	Weasel words
peacock	Peacock words
hedge	Hedge words
doxastic	Doxastic words
love	Words expressing love
anxiety	Words expressing anxiety
sorrow	Words expressing sorrow
joy	Words expressing joy
disgust	Words expressing disgust
fear	Words expressing fear
surprise	Words expressing surprise
anger	Words expressing anger

Table 7: Semantic Features and Codes

Code	Description
speechact	Speech acts
quote	Quotations
dash	Dashes
public	Public references
private	Private references
suasive	Persuasive words
discMarker	Discourse markers
smiley	Smileys

Table 8: Pragmatic Features and Codes

Features	**Examples**
statistical	plain text: édesanyám második házasságából születtem lemmatized: édesanya (mother-1poss) második (second) házasság (marriage-1poss.in) születik (born-1sg.past) – word number: 4
speech-based	*Öhm* én viszonylag késői gyerek vagyok ’*Um*, I’m a relatively late child’ – number of hesitations: 1
morphological	plain text: azért lettem skizofrén, mert ilyen rossz dolgokat tettem lemmatized: azért (therefore) lesz (become-1sg.past) skizofrén (schizophrenic), mert (because) ilyen (such) rossz (bad) dolog (thing-pl.acc) tesz (do-1sg.past) – number of 1Sg verbs: 2
syntactic – number of objects: 1	(...) dolgokat tettem lemmatized: dolog (thing-pl.acc) tesz (do-1sg.past)
semantic	*undorral* mentem be dolgozni ’I went to work *with disgust*’ – number of negative emotion: 1
pragmatic	*Hát* (...) nagyjából így ennyi. ’*Well* (...) roughly that’s all.’ – number of discourse marker: 1

Table 9: A Showcase of Examples from the Six Main Feature Categories in the Corpus.