Masking Morphosyntactic Categories to Evaluate Salience for Schizophrenia Diagnosis

Yaara Shriki
School of Computer Science
College of Management
Academic Studies
yaara.shriki@cs.colman.ac.il

Ido Ziv
Meuhedet Health Services
ido.z@meuhedet.co.il

Nachum Dershowitz
School of Computer Science
Tel Aviv University
nachum@tau.ac.il

Eiran Vadim Harel
Beer Yaakov Mental Heath Center
Beer Yaakov, Israel
eiran.harel@moh.gov.il

Kfir Bar
School of Computer Science
College of Management Academic Studies
kfirb@colman.ac.il

Abstract

Natural language processing tools have been shown to be effective for detecting symptoms of schizophrenia in transcribed speech. We analyze and assess the contribution of the various syntactic and morphological categories towards successful machine classification of texts produced by subjects with schizophrenia and by others. Specifically, we fine-tune a language model for the classification task, and mask all words that are attributed with each category of interest. The speech samples were generated in a controlled way by interviewing inpatients who were officially diagnosed with schizophrenia, and a corresponding group of healthy controls. All participants are native Hebrew speakers. Our results show that nouns are the most significant category for classification performance.

1 Introduction

Psychotic disorders such as schizophrenia are characterized by several symptoms, such as delusions, hallucinations, and thought disorders. Thought disorders are described as disturbances in the normal way of thinking, typically presented as various language impairments, such as disorganized speech, which is related to abnormal semantic associations between words (Aloia et al., 1998), and poverty of speech, a thought disorder that is associated with impairments in lexico-semantic retrieval (Nagels et al., 2016). Disorganized speech is divided into several markers, such as derailment, characterized by the usage of unrelated concepts in a conversation; tangentiality, which happens when providing oblique or irrelevant answers to a question; and incoherence, also known as "word salad", refers to speech that is incomprehensible at times due to multiple grammatical and semantic inaccuracies (Bar et al., 2019).

The diagnosis of schizophrenia is mostly based on a professional psychiatric review. However, some studies show that a computational linguistic analysis may help with diagnosis. Fraser et al. (1986), for example, demonstrated that by using a discriminant function analysis of linguistic variables it is possible to predict diagnoses with an accuracy rate of 79%.

There have been many attempts to study speech impairments that are related to thought disorders using a computational method. Some of those studies analyze the frequency of using different part-of-speech categories, such as nouns and verbs. For example, Obrębska and Obrębski (2007) reported a significantly lower frequency of adjectives in schizophrenic speech than in healthy control speech. On the other hand, they reported a higher frequency of verbs used by patients. Tang et al. (2020) measured a low frequency of adverbs in speech produced by patients with schizophrenia. Ziv et al. (2022) analyzed speech produced by Hebrew speaking patients with schizophrenia and reported low frequencies of words inflected in the third person or in the past tense. Aligned with previous work, they also reported lower frequencies of adverbs. It has been shown (Kircher et al., 2005) that patients with schizophrenia are produc-

ing grammatically simpler speech than healthy people. The results are not always consistent; Tang et al. (2021), for example, reported high frequencies of adverbs and adjectives in schizophrenic speech, in contrast to the reports made by other works. Until very recently, the large majority of those studies were conducted with English speaking patients.

One of the most popular technologies in natural language processing (NLP) is language modelling. A language model is essentially a function that assigns a probability to a given sequence of words occurring in a sentence. There are different ways to fit a language model to a certain distribution, typically using massive collections of texts. An autoregressive model conditions the probability of a word on the text that has already been seen in direction of reading. On the other hand, masked language models (MLM) are given the full sentence, while learning to assign probability to a randomly chosen hidden (masked) word. Such models are typically used as the basis for an algorithm that aims at solving a specific downstream task, such as sentiment analysis or document classification. In the first phase, the models are pre-trained for the word-probability assignment using a large unlabeled collection of texts, and later are fine-tuned on a labeled dataset for a specific downstream task. While the autoregressive models are more suitable for generation tasks, MLMs are typically the best option for fine-tuning on classification tasks.

This development of pre-trained language models provides us with the opportunity to examine the importance of certain morphosyntactic categories in speech of patients with schizophrenia, and compare it to that of a healthy control group. Specifically, we fine-tune an MLM to classify transcribed speech segments into patient or control categories, and examine its performance under extreme situations of hiding (masking) words that belong to a specific syntactic or morphological category.

While most existing techniques use some sort of counting method, in this study, we explore an alternative innovative way for assessing the salience of a specific category for detecting schizophrenic speech. We utilize the original masking technique of an MLM, by naturally masking out specific morphosyntactic categories and measure the performance of the model on a downstream classification task.

The experimental results show a decrease in prediction accuracy once nouns are masked, suggesting that nouns are more informative than other categories we tested for differentiating between patients and controls. Our participants are all native Hebrew speakers.

2 Related Work

Computational modeling has been studied in relation to cognitive disorders in order to fill the gap between theoretical models and biological evidence. Lanillos et al. (2020) reviews popular neural network models for autism spectrum disorder and schizophrenia, using different types of input. Both disorders are characterized by an altered perception of the world. According to this review, models of schizophrenia mainly concentrate on positive symptoms, such as hallucinations and delusional behavior (e.g., Hoffman and McGlashan (1997); Horn and Ruppin (1995)). However, there are also models that target other symptoms such as disturbances of attention (Cohen and Servan-Schreiber, 1992) and movement disorders (Yamashita and Tani, 2012).

The use of computational linguistic models has been applied to studying language abnormalities related to mental illness, specifically schizophrenia. Disorganized speech, including derailment, incoherence, and tangentiality, is among the common symptoms of schizophrenia being studied by researchers using computational methods (e.g., Bedi et al. (2015); Pauselli et al. (2018); Iter et al. (2018); Bar et al. (2019); Just et al. (2020)). Hitczenko et al. (2021) reviews computational methods that perform linguistic analysis of psychosis, focusing on three language abnormalities: disorganized speech, poverty of speech, and flat affect. Many studies have employed latent semantic analysis (LSA) and word embedding models (e.g., word2vec and GloVe) to measure disorganized speech. Typically, the embeddings are used to measure semantic similarity between words in the sentence, or between entire sentences or paragraphs, to assess semantic cohesion as a predictor for disorganized speech. In several studies (e.g., Elvevåg et al. (2007); Iter et al. (2018); Just et al. (2019)), psychosis patients scored significantly higher on disorganization than controls. However, Hitczenko et al. (2021) argues that the measures are not consistent across other studies.

As mentioned in the previous section, most of those works analyze transcribed speech spoken in

English, which is characterized by a relatively simple morphological system. Some recent studies have been exploring similar techniques applied to other languages, such as German (Just et al., 2020) and Hebrew (Bar et al., 2019). The latter have studied derailment, a symptom of thought disorder characterized by switching between topics and jumping from one disconnected thought to another. They measure derailment in speech through semantic similarity of adjacent words using their embeddings. It was found that patients with schizophrenia are more likely to derail than healthy controls, consistent with previous studies (Bedi et al., 2015; Iter et al., 2018). Further, they examine incoherence in schizophrenic patients, to see how they use adjectives and adverbs to describe specific nouns and verbs. Their analysis makes use of a dependency parser for Hebrew, which yields a word-dependency list for each sentence. Using dependencies, they discovered that the adjectives and adverbs used by the controls are more similar to those commonly used to describe the same nouns and verbs.

There are not many works that leverage language models to analyse text for detecting mental health symptoms, such as we do. In a recent work (Tang et al., 2021), BERT (Devlin et al., 2019), a large English language model, has been used to encode full sentences and compare the resulting embeddings of adjacent sentences for measuring tangentiality. Their results reflect increased tangentiality among patients with schizophrenia.

In our work, we use a language model as a tool for assessing the contribution of six morphosyntactic categories to the classification of transcribed speech into patients or controls.

3 Participants and Data Collection

We interviewed 49 males, aged 18–60, divided into control and patient groups, all speaking Hebrew as their first language. The patient group includes 23 inpatients from the Be'er Ya'akov–Ness Ziona Mental Health Center in Israel who were admitted following a diagnosis of schizophrenia. Diagnoses were made by a hospital psychiatrist according to the DSM-5 criteria (American Psychiatric Association DSM-5 Task Force, 2013) and a full psychiatric interview. Each participant was rewarded with approximately $8. The control group includes 26 men, mainly recruited via an advertisement that we placed on social media. Exclusion criteria for all participants were as follows:

(1) participants whose mother tongue is not Hebrew;

(2) having a history of dependence on drugs or alcohol over the past year;

(3) having a past or present neurological illness; and

(4) using fewer than 500 words in total in their transcribed interview.

Additionally, the control group had to score below the threshold for subclinical diagnosis of depression and post-traumatic stress disorder (PTSD). Most of the control participants scored below the threshold for anxiety. Most of the patients scored above the threshold for borderline or mild psychosis symptoms on a standard measure.[1] See Section 3.1 for more information about the measures we use in this study.

The demographic characteristics of the two groups are presented in Table 1.

Patients were interviewed in a quiet room at the department where they are hospitalized by one of our professional team members, and the control participants were interviewed in a similar room outside the hospital. Each interview lasted approximately one hour. The interviews were recorded and later manually transcribed by a native Hebrew speaking student from our lab. All participants were assured of anonymity, and told that they are free to end the interview at any time.

After signing a written consent, each participant was asked to describe 14 images picked from the Thematic Appreciation Test (TAT) collection; the images were presented one by one. We used the TAT images identified with the following serial numbers: 1, 2, 3BM, 4, 5, 6BM, 7GF, 8BM, 9BM, 12M, 13MF, 13B, 14, and 3GF. All images are black and white, including a mixture of men and women, children and adults. Each picture stands by itself, presented alone and has no relation to the other pictures. The participants were asked to tell a brief story about each image based on four open questions: What led up to the event shown in the picture? What is happening in the picture at this moment? What are the characters thinking and feeling? What is the outcome of the story? The

[1] Our patient group is composed of inpatients who are undergoing treatment with medications; therefore, higher scores were not expected.

	Control	Patients	Statistics
N	26	23	
Age mean (*SD*)	25.46 (6.28)	33.15 (9.72)	$t = 3.38$**
Education years mean (*SD*)	11.96 (0.15)	11.30 (1.15)	$t = 2.98$**
Place of residence (frequencies)			χ^2 (3,55) = 8.84, $p = .03$
Southern Israel	1	7	
Central Israel	22	16	
Northern Israel	2	0	
Jerusalem	1	0	
Marital status (frequencies)			χ^2 (1,49) = 0.055, $p = .81$
Single	4	3	
Married	22	20	
Income (frequencies)			χ^2 (3,49) = 3.06, $p = .38$
Low	5	4	
Lower than average	6	4	
Average	9	13	
Higher than average	6	2	
PANSS positive subscale		8.91 ± 3.91	
PANSS negative subscale		7.82 ± 3.74	
PANSS total subscale		16.73 ± 6.23	

Table 1: Demographic characteristics by group. **$p < .005$.

interviewer remained silent during the respondent's narration and offered no prompts or questions.

After describing the images, the participant was asked to answer four open questions, one by one. The four questions are listed in Table 2. As before, the interviewer remained silent during the respondent's narration and offered no prompts or questions.

Once all 18 components (14 image descriptions and 4 open questions) were answered, each participant was requested to fill in a demographic questionnaire as well as some additional questionnaires for assessing mental-health symptoms, which we describe next.

3.1 Symptom Assessment Measures

3.1.1 Control group

The control participants were assessed for symptoms of depression, PTSD, and anxiety.

Depression. Symptoms of depression were assessed using Beck's Depression Inventory–II (BDI–II) (Beck et al., 1996). The BDI–II is a 21-item inventory rated on a 4-point Likert-type scale (0 = "not at all" to 3 = "extremely"), with summary scores ranging between 0 and 63. Beck et al. (1996) suggested a preliminary cutoff value of 14 as an indicator for mild depression, as well as a threshold of 19 as an indicator for moderate depression. BDI–II has been found to demonstrate high reliability (Gallagher et al., 1982). We use a Hebrew version of BDI–II (Hasenson-Atzmon et al., 2016).

PTSD. Symptoms of PTSD were assessed using the PTSD checklist of the DSM–5 (PCL–5) (Weathers et al., 2013). The questionnaire contains twenty items that can be divided into four subscales, corresponding to the clusters B–E in DSM–5: intrusion (five items), avoidance (two items), negative alterations in cognition and mood (seven items), and alterations in arousal and reactivity (six items). The items are rated on a 5-point Likert-type scale (0 = "not at all" to 4 = "extremely"). The total score ranges between 0 and 80, provided along with a preliminary cutoff score of 38 as an indicator for PTSD. PCL–5 has been found to demonstrate high reliability (Blevins et al., 2015). We use a Hebrew translation of PCL-5 (Bensimon et al., 2013).

Anxiety. Symptoms of anxiety were assessed through the State Trait Anxiety Inventory (STAI) (Spielberger et al., 1970). The STAI questionnaire consists of two sets of twenty self-reporting measures. The STAI measure of state anxiety (S-anxiety) assesses how respondents feel "right now, at this moment" (e.g., "I feel at ease"; "I feel upset"), and the STAI measure of trait anxiety (T-anxiety) targets how respondents "generally feel" (e.g., "I am a steady person"; "I lack self-

ID	Question
1	Tell me as much as you can about your bar mitzvah.*
2	What do you like to do, mostly?
3	What are the things that annoy you the most?
4	What would you like to do in the future?

Table 2: Four open questions asked during the interview. *Bar mitzvah is a Jewish confirmation ceremony for boys who have reached the age of 13.

confidence"). For each item, respondents are asked to rate themselves on a 4-point Likert scale, ranging from 1 = "not at all" to 4 = "very much so" for S-anxiety, and from 1 = "almost never" to 4 = "almost always" for T-anxiety. Total scores range from 20 to 80, with a preliminary cutoff score of 40 recommended as indicating clinically significant symptoms for the T-Anxiety scale (Knight et al., 1983). STAI has been found to demonstrate high reliability (Barnes et al., 2002). We use a Hebrew translation of STAI (Saka and Gati, 2007).

3.1.2 Patients

Psychosis symptoms were assessed by the 6-item Positive And Negative Syndrome Scale (PANSS–6) (Østergaard et al., 2016). The original 30-item PANSS (PANSS–30) is the most widely used rating scale for schizophrenia, but it is relatively long for use in clinical settings. The items in PANSS–6 are rated on a 7-point scale (0 = "not at all" to 6 = "extremely"). The total score ranges from 0 to 36, with a score of 14 representing the threshold for mild schizophrenia, and a score between 10 and 14 defined as borderline disease or as remission. PANSS–30 has been found to demonstrate high reliability (Lin et al., 2018), while Østergaard et al. (2016) reported a high correlation between PANSS–6 and PANSS–30 (Spearman correlation coefficient = 0.86). We used the Hebrew version of PANSS–6 (Lin et al., 2018). The range of positive and negative symptoms are presented in Table 1.

4 Analysis

4.1 Preprocessing

We treat every response to any one of the 18 questions as a training/evaluation instance for our classifier. Overall we have 414 responses generated by patients, as well as 468 responses that were generated by controls. The responses are written in Hebrew, a morphologically rich Semitic language; Hebrew words are inflected for person, number, and gender, resulting in a relatively complicated word-production process. We preprocess each response using the Ben-Gurion University (BGU) morphological tagger (Adler and Elhadad, 2006), a context-sensitive morphological analyzer for Hebrew. The tagger displays morphosyntactic information for each word in the text, including part-of-speech tags, as well as information about person and number.

4.2 Classification Methodology

We use a Hebrew MLM to classify a response into the two groups, patients or controls. As mentioned before, MLMs are trained in two phases. During the first, also known as pre-training, the model is trained with a large set of text in which 15% of the input tokens are masked using a special mask token for which the model is trained to predict. In the second phase, also known as fine-tuning, the model is adapted for a downstream task using a relatively small set of annotated examples. For classification tasks, such as ours, the common practice is to add another neural dense layer connected to the output vector of the initial token. Therefore, we fine-tune a pre-trained language model using a portion of the dataset, and evaluate its performance on the remaining instances. To assess the contribution of different syntactic and morphological categories for the classification performance, we fine-tune the model several times individually, each time we mask all words of a selected category. We focus on four parts of speech including nouns, verbs, adverbs, and adjectives. Those are all considered as content words, rather than functional ones. In addition, we examine first-person and third-person words. Overall, we examine six morphosyntactic categories.

In all our experiments, we use AlephBERT (Seker et al., 2021), a pre-trained language model for Hebrew, to perform sequence classification using the Transformers library (Wolf et al., 2019). Specifically we use

`AutoModelForSequenceClassification` with the `alephbert-base` model code. The AlephBERT model was trained on data collected from three different Hebrew text sources: the OSCAR corpus (Ortiz Suárez et al., 2020), Hebrew tweets, and the Hebrew Wikipedia.

Given a category M, we begin each experiment by dividing the collection of responses into 80:20 train and test sets, respectively, by making sure the label distribution remains similar to the original dataset. We tokenize each response using the AlephBERT tokenizer, which was designed to truncate responses longer than the model's 512-token limitation. We proceed with the following three steps:

1. We iterate through all train and test responses and mask[2] all tokens that were attributed with M by the BGU Tagger. By design, the AlephBERT tokenizer may break words in the middle; therefore, to be more precise we mask all tokens that were broken from a word that was attributed with M by the BGU tagger. We then fine-tune the model on the masked train set and evaluate on the corresponding masked test set. We use accuracy as an evaluation metric.

2. As a control experiment, we mask tokens randomly by considering every token for masking using a Bernoulli trial with probability equals to the probability of occurrence of M. Same as before, we fine-tune the model on the modified train set and evaluate it on the modified test set.

3. We repeat this experiment 30 times, each time with a different random state, which affects the splitting to train and test sets, as well as on the random masking procedure, and calculate the average accuracy scores for both, M-based masking and random masking. After confirming the scores are normally distributed, we conduct a t-test in order to measure the impact of M-based masking by comparing its accuracy with the one achieved by random masking.

It should be noted that the random states that we use in the experiments are identical across different categories, to make sure that we use the same train/test splits in the 30 executions of each category.

[2]With the special token `[MASK]`.

5 Results

Figure 1 displays the probability of each morphosyntactic category to appear in the responses of patients and controls. All participants use more nouns and third-person words than verbs, adverbs, adjectives, and first-person words. The high frequency of third-person words is reasonable, since in most of the interview, the participants were asked to describe the situation as they interpret from a picture that was presented to them. Neither group uses a significant proportion of first or third person tokens. However, we can see that the inpatients use nouns and verbs slightly more often, whereas the controls use more adjectives and adverbs. The difference in adverbs has been confirmed to be statistically significant according to a Welch's unequal variances t-test (at $p < 0.0005$).

The classification results, under different masking conditions are summarized in Table 3. The table displays the difference between the mean classification accuracy of masking each morphosyntactic category (the Morph. Masking column), compared to a random masking of tokens with the same probability of occurrence (the Random Masking column). We run t-tests and provide the outcome statistics in the last two columns. The accuracy at the baseline level (i.e., no masking) is 84.4%. Standard deviations range between 5.5 and 6.5 percent for all accuracy measures. Unsurprisingly, most of the accuracy results listed in the table are below the baseline score. We expected that masking words at high rates may be detrimental for the classification performance. We do see some accuracy scores above the baseline score; however, the differences are minor and has no statistical significance.

We can clearly see the impact of masking nouns and adverbs on the classification performance. Especially when nouns are being masked, the accuracy decreases significantly compared to random masking at the same token-masking rate. The other categories do not show a significant decrease in accuracy compared to random masking at the same rate.

To confirm our results, we design another experiment in which all words in the text *except* nouns and adverbs, are masked. Like before, we compare the classification accuracy with a control model in which we use random masking at the same rate,

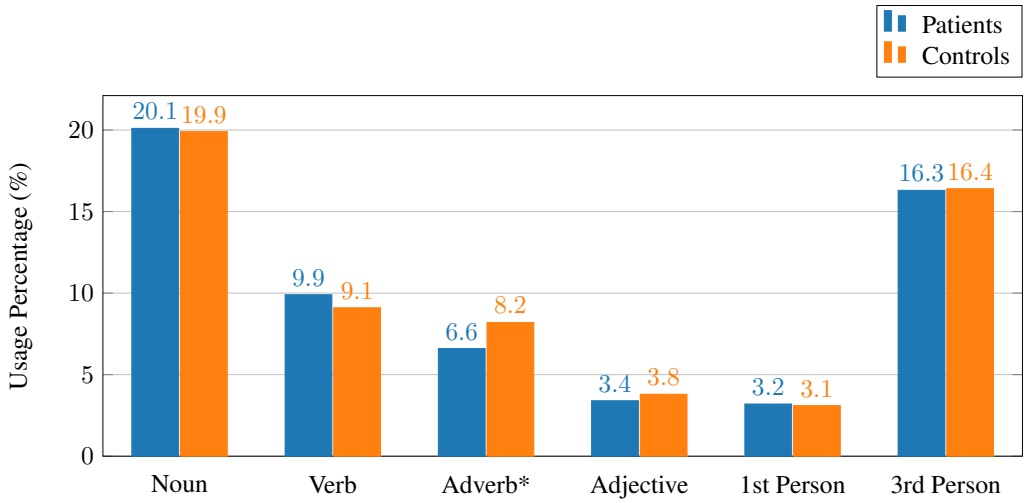

Figure 1: Usage percentage of selected syntactical and morphological categories. *$p < 0.0005$ (per Welch's unequal variance t-test).

as described before. In spite of the fact that we have masked more than 72% of the words in the text, the model has been able to achieve an accuracy of 82.8%, compared to 75% achieved by the random-masking model. This difference has been confirmed to be statistically significant by conventional standards, according to a t-test (at $p < 0.0005$). These results provide a consistent evidence that nouns and adverbs are more important than other categories for the classification task.

6 Discussion

We notice that nouns and adverbs make the biggest impact on the performance of the classifier, suggesting that those syntactic categories are the most informative to the model. Comparing with random masking of the same number of words, the accuracy drops significantly ($p < 0.0005$) when nouns are being masked. With adverbs, the difference in accuracy is less significant ($p = 0.058$). Based on the numbers assembled in Figure 1, we cannot attribute our findings to the frequency of usage of those categories. Whereas nouns are used more frequently than the other categories, adverbs are much less frequent. For adverbs, at least, we see a significant difference in the frequency of usage between the two groups; controls use them more. Adverbs are typically used in tandem with a verb; however, it turns out that the patients use slightly more verbs than the controls, although to an insignificant degree. Therefore, we believe that the significant difference in usage frequency of adverbs may be the reason for the impact that they make on classification performance.

As for nouns, we see no evidence for a usage frequency difference between the two groups. We believe that the reason for the impact made by masking nouns on the classification performance might be related to the importance of nouns in the syntactic tool set of patients with schizophrenia. Our results may suggest that the patients convey their messages more through nouns than through other linguistic categories. Nouns are considered the backbone of a language; it has been shown that English-speaking children acquire knowledge of nouns before verbs (Gentner, 1982). Nouns are considered easier to learn than verbs, probably due to their imageability (McDonough et al., 2011). Therefore, we presume that focusing more on nouns when conveying a message may be an indicator of poverty of speech. The way patients use nouns is slightly different from how controls do. This difference makes it easier for the model to predict schizophrenic symptoms. The source of the difference may be related to the type of nouns that they choose to use in a sentence, the similarity among the nouns in a sentence, or their syntactic relations with other words in the sentence. Since Hebrew is a highly inflected language, it could also be that patients inflect nouns differently than controls. We plan to further investigate the source of the difference in follow up work.

7 Ethical Considerations

This research was approved by the Helsinki Ethical Review Board (IRB) of the Be'er Ya'akov–Ness

Morph. Category	Morph. Masking	Random Masking	t	p
No masking (baseline)	84.4%	-	-	-
Noun	**82.2%**	**84.6%**	4.7809	$p < .0005$*
Verb	83.1%	84.0%	1.2646	$p = .2161$
Adverb	82.3%	83.5%	1.9739	$p = .0580$
Adjective	84.1%	84.9%	1.7963	$p = .0829$
First person	84.5%	84.9%	1.9598	$p = .0597$
Third person	83.2%	82.3%	-1.9527	$p = .0606$

Table 3: Accuracy scores under different masking conditions. *$p < 0.0005$.

Ziona Mental Health Center. Participants were guaranteed anonymity. The data was stored on a secured server, with limited access provided only to the authors of this paper.

Like with every other machine-learning model, there is a risk that the training data is unbalanced. Specifically, we do not intentionally balance the dataset for ethnicity or political affiliation. Moreover, this work is based on interviews with men only. Additionally, the language model that we use, AlephBERT, was trained on large and less controlled datasets. That may introduce some additional aspects of bias. Therefore, our study may harbor the danger of over-reliance on possibly biased machine tools.

We do not mean to suggest that an algorithm can or should be used to diagnose schizophrenia automatically. This study should not be considered as a building block for an apparatus that takes automatic decisions about topics related to mental health. Our intention is, rather, to use computational tools to identify and study the importance of various linguistic characteristics for diagnosing schizophrenia. Like other machine-learning applications, explainability is currently a problematic issue (what is it about the usage of nouns that contributes significantly to the model's success in classification?), and undue reliance on machine classification should be eschewed.

8 Conclusions

We studied the relative importance of several morphosyntactic categories for transcribed speech towards the classification task of distinguishing schizophrenia sufferers and controls. This was based on interviews of 23 male inpatients at a mental health center in Israel, officially diagnosed with schizophrenia, as well as 26 control participants; all are native Hebrew speakers. The interviews were manually transcribed and divided into individual responses that the participants provided for 18 discussion topics. Four topics were open-ended questions, and the rest were TAT images that were shown to the participants who were asked to describe the situation they see in the image.

We trained a natural-language-processing classifier by fine-tuning AlephBERT, a relatively large Hebrew language model, to distinguish between responses generated by patients and controls. To evaluate the contribution of different syntactic and morphological categories to the classification performance, we fine-tune the model each time by masking words of one specific category, and compare the classification performance with the same model trained on texts that were instead masked randomly for the same number of words. When the category-masked model performed more poorly than the randomly-masked model, we attribute it to an increased importance of the corresponding category. This new, masking method of evaluating the significance of linguistic features promises to be of use in many additional feature evaluation tasks.

Overall we examined six categories, and found (unsurprisingly) that nouns are the most important for distinguishing between patients and controls. We believe that it has to do with the idea of nouns being easier to capture in the mind due to their imageability. Given that nouns are used in comparable frequency by patients and controls, our findings reveal that the patients use nouns in a different way than do controls. We plan to investigate this further by looking more closely at the potential sources for this difference, in order to check how they may be related to poverty of speech.

Acknowledgements

This research was supported in part by grant #2168 from the Israeli Ministry of Science.

References

Meni Adler and Michael Elhadad. 2006. An unsupervised morpheme-based HMM for Hebrew morphological disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 665–672, USA. Association for Computational Linguistics.

Mark S. Aloia, Monica L. Gourovitch, David Missar, David Pickar, Daniel R. Weinberger, and Terry E. Goldberg. 1998. Cognitive substrates of thought disorder, II: Specifying a candidate cognitive mechanism. *American Journal of Psychiatry*, 155(12):1677–1684.

American Psychiatric Association DSM-5 Task Force. 2013. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, volume 5. American Psychiatric Publishing, Washington, DC.

Kfir Bar, Vered Zilberstein, Ido Ziv, Heli Baram, Nachum Dershowitz, Samuel Itzikowitz, and Eiran Vadim Harel. 2019. Semantic characteristics of schizophrenic speech. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 84–93, Minneapolis, MN. Association for Computational Linguistics.

Laura L. B. Barnes, Diane Harp, and Woo Sik Jung. 2002. Reliability generalization of scores on the Spielberger state-trait anxiety inventory. *Educational and Psychological Measurement*, 62(4):603–618.

Aaron T. Beck, Robert A. Steer, Roberta Ball, and William F. Ranieri. 1996. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment*, 67(3):588–597.

Gillinder Bedi, Facundo Carrillo, Guillermo Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália Mota, Sidarta Ribeiro, Daniel Javitt, Mauro Copelli, and Cheryl Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1:15030.

Moshe Bensimon, Stephen Zvi Levine, Gadi Zerach, Einat Stein, Vlad Svetlicky, and Zahava Solomon. 2013. Elaboration on posttraumatic stress disorder diagnostic criteria: A factor analytic study of PTSD exposure to war or terror. *Israel Journal of Psychiatry*, 50(2):84–90.

Christy A. Blevins, Frank W. Weathers, Margaret T. Davis, Tracy K. Witte, and Jessica L. Domino. 2015. The posttraumatic stress disorder checklist for *DSM-5* (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, 28(6):489–498.

Jonathan D. Cohen and David Servan-Schreiber. 1992. Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99(1):45.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN. Association for Computational Linguistics.

Brita Elvevåg, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1–3):304–316.

William I. Fraser, Kathleen M. King, Philip Thomas, and Robert E. Kendell. 1986. The diagnosis of schizophrenia by language analysis. *British Journal of Psychiatry*, 148(3):275–278.

Dolores Gallagher, Gloria Nies, and Larry W. Thompson. 1982. Reliability of the Beck Depression Inventory with older adults. *Journal of Consulting and Clinical Psychology*, 50(1):152–153.

Dedre Gentner. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. Technical Report 257, Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Kelly Hasenson-Atzmon, Sofi Marom, Tamar Sofer, Lilac Lev-Ari, Rafael Youngmann, Haggai Hermesh, Jonathan Kushnir, and Haggai Hermesh. 2016. Cultural impact on SAD: Social anxiety disorder among Ethiopian and former Soviet Union immigrants to Israel, in comparison to native-born Israelis. *Israel Journal of Psychiatry*, 53(3):48–54.

Kasia Hitczenko, Vijay A. Mittal, and Matthew Goldrick. 2021. Understanding language abnormalities and associated clinical markers in psychosis: The promise of computational methods. *Schizophrenia Bulletin*, 47(2):344–362.

Ralph E. Hoffman and Thomas H. McGlashan. 1997. Synaptic elimination, neurodevelopment, and the mechanism of hallucinated "voices" in schizophrenia. *American Journal of Psychiatry*, 154(12):1683–1689.

David Horn and Eytan Ruppin. 1995. Compensatory mechanisms in an attractor neural network model of schizophrenia. *Neural Computation*, 7(1):182–205.

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.

Sandra Just, Erik Haegert, Nora Kořánová, Anna-Lena Bröcker, Ivan Nenchev, Jakob Funcke, Christiane Montag, and Manfred Stede. 2019. Coherence models in schizophrenia. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136. Association for Computational Linguistics.

Sandra A. Just, Erik Haegert, Nora Kořánová, Anna-Lena Bröcker, Ivan Nenchev, Jakob Funcke, Andreas Heinz, Felix Bermpohl, Manfred Stede, and Christiane Montag. 2020. Modeling incoherent discourse in non-affective psychosis. *Frontiers in Psychiatry*, page 846.

Tilo T. J. Kircher, Tomasina M. Oh, Michael J. Brammer, and Philip K. McGuire. 2005. Neural correlates of syntax production in schizophrenia. *The British Journal of Psychiatry*, 186(3):209–214.

Robert G. Knight, Hendrika J. Waal-Manning, and George F. Spears. 1983. Some norms and reliability data for the state-trait anxiety inventory and the Zung self-rating depression scale. *British Journal of Clinical Psychology*, 22(4):245–249.

Pablo Lanillos, Daniel Oliva, Anja Philippsen, Yuichi Yamashita, Yukie Nagai, and Gordon Cheng. 2020. A review on neural network models of schizophrenia and autism spectrum disorder. *Neural Networks*, 122:338–363.

Ching-Hua Lin, Huey-Shyan Lin, Shih-Chi Lin, Chao-Chan Kuo, Fu-Chiang Wang, and Yu-Hui Huang. 2018. Early improvement in PANSS-30, PANSS-8, and PANSS-6 scores predicts ultimate response and remission during acute treatment of schizophrenia. *Acta Psychiatrica Scandinavica*, 137(2):98–108.

Colleen McDonough, Lulu Song, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Robert Lannon. 2011. An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental science*, 14(2):181–189.

Arne Nagels, Paul Fährmann, Mirjam Stratmann, Sayed Ghazi, Christian Schales, Michael Frauenheim, Lena Turner, Tobias Hornig, Michael Katzev, Rüdiger Müller-Isberner, Michael Grosvald, Axel Krug, and Tilo Kircher. 2016. Distinct neuropsychological correlates in positive and negative formal thought disorder syndromes: The thought and language disorder scale in endogenous psychoses. *Neuropsychobiology*, 73(3):139–147.

Monika Obrębska and Tomasz Obrębski. 2007. Lexical and grammatical analysis of schizophrenic patients' language: A preliminary report. *Psychology of Language and Communication*, 11(1):63–72.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Soren Dinesen Østergaard, Ole Michael Lemming, Ole Mors, Christoph U. Correll, and Per Bech. 2016. PANSS-6: A brief rating scale for the measurement of severity in schizophrenia. *Acta Psychiatrica Scandinavica*, 133(6):436–444.

Luca Pauselli, Brooke Halpern, Sean D. Cleary, Benson S. Ku, Michael A. Covington, and Michael T. Compton. 2018. Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. *Psychiatry Research*, 263:74–79.

Noa Saka and Itamar Gati. 2007. Emotional and personality-related aspects of persistent career decision-making difficulties. *Journal of Vocational Behavior*, 71(3):340–358.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. *AlephBERT*: A Hebrew large pre-trained language model to start-off your Hebrew NLP application with. *arXiv preprint arXiv:2104.04052*.

Charles Donald Spielberger, Richard L. Gorsuch, and Robert E. Lushene. 1970. *STAI Manual for the State-Trait Anxiety Inventory ("self-evaluation questionnaire")*. Consulting Psychologist Press, Palo Alto.

Sunny Tang, Reno Kriz, Sunghye Cho, João Sedoc, Suh Jung Park, Jenna Harowitz, Mahendra Bhati, Raquel Gur, Daniel Wolf, and Mark Liberman. 2020. Decreased speech coherence captured by novel natural language processing methods in two cohorts of individuals with schizophrenia. *Biological Psychiatry*, 87(9):S379–S380.

Sunny X. Tang, Reno Kriz, Sunghye Cho, Suh Jung Park, Jenna Harowitz, Raquel E. Gur, Mahendra T. Bhati, Daniel H. Wolf, João Sedoc, and Mark Y. Liberman. 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *npj Schizophrenia*, 7(1):1–8.

Frank W. Weathers, Brett T. Litz, Terence M. Keane, Patrick A. Palmieri, Brian P. Marx, and Paula P. Schnurr. 2013. The PTSD checklist for DSM-5 (PCL-5). Scale available from the National Center for PTSD at www.ptsd.va.gov.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuichi Yamashita and Jun Tani. 2012. Spontaneous prediction error generation in schizophrenia. *PLoS One*, 7(5):e37843.

Ido Ziv, Heli Baram, Kfir Bar, Vered Zilberstein, Samuel Itzikowitz, Eran V. Harel, and Nachum Dershowitz. 2022. Morphological characteristics of spoken language in schizophrenia patients—an exploratory study. *Scandinavian Journal of Psychology*, 63(2):91–99.