The Raymond and
Beverly Sackler Faculty
of Exact Sciences
Tel Aviv University

Tel Aviv University
Raymond and Beverly Sackler Faculty of Exact Sciences
The Blavatnik School of Computer Science

Semantic Characteristics of Schizophrenic Speech

by

Vered Zilberstein

Under the supervision of
Dr. Kfir Bar &
Prof. Nachum Dershowitz

Thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science

2020

Abstract

Natural language processing tools are used to automatically detect disturbances in transcribed speech of schizophrenia inpatients who speak Hebrew. We measure topic mutation over time and show that controls maintain more cohesive speech than inpatients. We also examine differences in how inpatients and controls use adjectives and adverbs to describe content words and show that the ones used by controls are more common than those of inpatients. We provide experimental results and show their potential for automatically detecting schizophrenia in patients by means only of their speech patterns. We then explore our findings on publicly published written text taken from social media, which does not show a significant difference in keeping cohesive discourse.

Acknowledgements

First and foremost, I would like to express my special thanks and gratitude to my advisors Dr. Kfir Bar and Prof. Nachum Dershowitz for their support and guidance and many insightful conversations during the development of the ideas in this thesis.

Their patience, motivation, and immense knowledge had made me enjoy and learn a lot during this research. I could not have imagined having better advisors and mentors for my M.Sc. study.

Special thanks to the doctors and others at the Psychology Department in Beer Yaakov-Ness Ziona Mental Health Center for their contribution in making this research possible, especially Dr. Ido Ziv for his knowledge sharing and ideas.

I would also like to acknowledge the Deutsch Institute for their financial support.

Lastly, I am extremely grateful to my partner, Roy, for his love, understanding and continuing support to complete this research.

Ethics Statement

The institutional review board of the College of Management Academic Studies of Rishon LeZion, as well as of the Beer Yaakov–Ness Ziona Mental Health Center, approved the experiments herein, and informed consent was obtained for all subjects.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Thought disorders are described as disturbances in the normal way of thinking. Bleuler [1991] original considered thought disorders to be a speech impairment in schizophrenia patients, but nowadays there is agreement that thought disorders are also relevant to other clinical disorders, including pediatric neurobehavioral disorders like attention deficit hyperactivity disorder and high functioning autism. They can even occur in normal populations, especially in people who have a high level of creativity. Bleuler focused mostly on "loosening of associations", or *derailment*, a thought disorder characterized by the usage of unrelated concepts in a conversation, or in other words, a conversation lacking coherence. The *Diagnostic and Statistical Manual of Mental Disorders (DSM 5)* [Association, 2013] outlines *disorganized speech* as one of the criteria for making a diagnosis of schizophrenia. Morice and Ingram [1982] showed that schizophrenics' speech is built upon a different syntactic structure than normal controls, and that this difference increases over time. Andreasen [1979] suggested several definitions of linguistic and cognitive behaviors frequently observed in patients, and which may be useful for thought-disorder evaluation. Among the definitions presented in that report, one finds the following, which we address in this study:

Incoherence, also known as "word salad", refers to speech that is incomprehensible at times due to multiple grammatical and semantic inaccuracies. In this dissertation, we focus mostly on the semantic inaccuracies, leaving grammatical issues for future investigation.

Derailment, also known as "loose associations", happens when a speaker shifts among topics that are only remotely related, or are completely unrelated, to the previous ones.

Tangentiality occurs when an irrelevant, or just barely relevant, answer is provided for a given question.

We focus here on derailment. But tangentiality has been addressed in some other studies. The two notions are closely related.

One of the main data sources for diagnosing mental disorders is speech, typically collected during a psychiatric interview. Identifying signals that indicate the presence of thought disorders is often challenging and subjective, especially in patients who are not undergoing a psychotic episode at the time of the interview.

In this work, we focus on schizophrenia. We investigate a number of semantic characteristics of transcribed human speech, and propose a way to use them to measure disorganized speech. Natural-language processing software is used to automatically detect those characteristics, and we suggest a way of aggregating them in a meaningful way. We use transcribed interviews, collected from Hebrew-speaking schizophrenia inpatients at a mental health hospital and from a control group. About two thirds of the patients were identified as in schizophrenia remission at the time of the interview.

Following a few previous works [Iter et al., 2018, Bedi et al., 2015], we measure Andreasen's derailment by calculating average semantic similarity between consecutive chunks of a running text to track topical mutations, and show the difference between patients and controls. For incoherence, we look at word modifiers, focusing on adjectives and adverbs, that subjects use to describe the same objects, and then learn the difference between the two groups. We then use those semantic characteristics in a classification setting and argue for their usability. As a final step, we measure derailment on public text from two social media datasets collected from English-speaking self-diagnosed depression and schizophrenia users, compare them to a matched group of controls, and find no significant difference between the groups.

This work makes the following contributions:

- We measure derailment in speech using word semantics, similar to [Bedi et al., 2015], this time on Hebrew.
- We explore a novel way of measuring one aspect of speech incoherence by measuring how similar modifiers (adjectives and adverbs) are to ones used in a reference text to describe the same words.
- Using these measures, we build a classifier for detecting schizophrenia on the basis of recorded interviews, which achieves 81.5% accuracy.
- Our approach to measuring derailment in free speech is not reproduced on social media text.

We proceed as follows: The next chapter reviews some relevant previous work. In Chapter 3, we describe how we collected the data. Our main contributions are described in Chapter 4, followed by complementary experiments on other datasets in Chapter 5, and finally, some conclusions suggested in the final chapter.

Chapter 2

Related Work

There is a large body of work that examines human-generated texts with the aim of learning about the way people who suffer from various mental-health disorders use language in different settings. For example, Al-Mosaiwi and Johnstone [2018] conducted a study in which they analyzed 63 web forums, some related to mental health disorders and others used as control. They ran their analysis with the well-known Linguistic Inquiry and Word Count [Pennebaker et al., 2015] tool to find absolutist words in free text. Overall, they discovered that anxiety, depression, and suicidal-ideation forums contained more absolutist words than control forums.

Recently, social media have become a vital source for learning about how people who suffer from mental-health disorders use language. Several studies collect relevant users from Twitter,* by considering users who intentionally write about their diagnosed mental-health disorders. For example, in [De Choudhury et al., 2013, Tsugawa et al., 2015], some language characteristics of Twitter users who claim to suffer from a clinical depression are studied. Similarly, users who suffer from post traumatic stress disorder are addressed in [Coppersmith et al., 2014]. Mitchell et al. [2015] analyze tweets posted by schizophrenics, and Coppersmith et al. [Coppersmith et al., 2016] investigate the language and emotions that are expressed by users who have previously attempted to commit suicide. Coppersmith et al. [2015] work with users who suffer from a broad range of mental-health conditions and explore language differences between groups. Most of these works found a significant difference in the usage of some linguistic characteristics by the experience group when compared to a control group. Furthermore, different levels of these linguistic characteristics are used as features for training a classifier to detect mental-health disorders prior to the report date.

Reddit† has also been identified as a convenient source for collecting data for this goal. Losada and Crestani [2016] outline a methodology for collecting posts and comments of Reddit and Twitter users who suffer from depression. Similarly, a large dataset of Reddit users with depression, manually verified (by lay annotators for an explicit claim of diagnosis), has been released for public use [Yates et al., 2017]. In that work, the authors employ a deep neural network on the raw text for detecting clinically depressed people

*https://twitter.com

†https://www.reddit.com

ahead of time, achieving 65% F1 score on an evaluation set.

A few caveats are in order when using social media for analyzing mental health conditions.

- First, self reporting of a mental health disorder is not a popular course of action. Clearly, then, the experimental group is chosen from a subgroup of the relevant population.
- Second, the controls, typically collected randomly "from the wild", are not guaranteed to be free of mental-health disorders.
- Finally, social media posts are considered to be a different form of communication than ordinary speech.

For all these reasons, in this work, we use validated experimental and control groups in an interview setting and compare the results to social media.

Measuring various aspects of incoherence in schizophrenics using computational tools has been previously addressed in [Elvevåg et al., 2007, Bedi et al., 2015, Iter et al., 2018]. Elvevåg et al. [2007] analyzed transcribed interviews of inpatients with schizophrenia to measure tangentiality. Moving along the patient's response, they calculated the semantic similarity between text chunks of different sizes and the question that was asked by the interviewer. Semantic similarity was cast by cosine similarity over the latent semantic analysis (LSA) [Deerwester et al., 1990] vectors calculated for each word, and summed across an entire chunk of words. They fitted a linear-regression line to represent the trend of the cosine similarity values, as one moves along the text. The slope of that line was used to measure how quickly the topic diverges from the original question. Overall, they were able to show a significant correlation between those values and a blind human evaluation of the same responses. Furthermore, as chunk size grows larger, the distinction between patients and controls becomes less prominent. One explanation for that could be the large number of mentions of functional and filler words, for which we typically do not have a good semantic representation. Iter et al. [2018] addressed this suggestion by cleaning the patients' responses of all those words and expressions (e.g. *uh*, *um*, *you know*) prior to calculating the semantic scores. This gave a slight improvement, although measured over a relatively small set of participants. Instead of working with chunks of text, they worked with full sentences, and replaced LSA with some modern techniques for sentence embeddings. Likewise, in our work, we use word embeddings instead of LSA.

Bedi et al. [2015] define coherence as an aggregation of the cosine similarity between pairs of consecutive sentences, each represented by the element-wise average vector of the individual words' LSA vectors. They worked with a group of 34 youths at clinical high-risk for psychosis, interviewed them quarterly for 2 1/2 years, and transcribed their answers. Five out of the 34 transitioned to psychosis. They used coherence scores, along with part-of-speech information, to automatically predict transition to psychosis with 100% accuracy.

The goal of all these works, including ours, is to automatically detect disorganized speech in a more objective and reliable way. Inspired by the last three studies described above, we analyzed transcribed responses to 18 open questions given by inpatients with schizophrenia and by controls. Instead of cleaning the text from filler words using a dictionary – as proposed by [Iter et al., 2018], we take a deeper look into the syntactic roles the words play, and calculate semantic similarity over a filtered version of the text, every time using different sets of part-of-speech categories. We report on the results of two sets of experiments:

1. We measure derailment by calculating the semantic similarity of adjacent words of various part-of-speech categories.
2. We measure semantic coherence by looking at the choices of modifiers (adjectives, adverbs) used in responses by inpatients and controls, as compared to those used in ordinary discourse.

Generally speaking, not too much is known about the role played by adjectives and adverbs in thought disorders. Modifiers are often not included in language tests, as they usually need to be presented together with the noun or verb they modify. Some previous works [Obrębska and Obrębski, 2007] have reported a significantly smaller number of adjectives used by schizophrenics. In the current study, we use computational tools to investigate the semantic relation between modifiers and objects, and its attribution to speech incoherence.

Chapter 3

Data Collection

We interviewed 51 men, aged 19–63, divided into control and patient groups, all speaking Hebrew as their mother tongue. The patient group comprised 24 inpatients at Beer Yaakov Mental Health Center in Israel who were officially diagnosed with schizophrenia. The control group includes 27 people, mainly recruited via an advertisement that we placed on social media. Most of the participants are single, with average-to-lower monthly income. Demographics for the two groups are presented in Table 3.1.

	Control	**Patients**
N	27	24
Age, Mean (SD)	30.3 (8.26)	38.3 (10.43)
Edu., HS	68%	75%
Edu., Post HS	20%	4%
Loc., South	40%	20%
Loc., Center	44%	33%
M.S., Single	80%	95%
Income, Avg/low	84%	83%

Table 3.1: Demographics by group. Edu. = Education (HS = High School); Loc. = Location in Israel; M.S. = Marital Status.

3.1 Interviews

Overall, the participants were asked 18 questions, out of which 14 were thematic-apperception-test (TAT) pictures that participants were requested to describe, followed by 4 questions that require the participant to share some personal thoughts and emotions. Both the control and patient groups completed a demographic questionnaire. To monitor the mental-health condition of the control group, they were requested to complete Beck's Depression Inventory-II (BDI-II) and the State and Trait Anxiety Inventory (STAI). The patient group also completed BDI-II, as well as a Hebrew translation [Katz et al., 2012] of the Positive and Negative Syndrome Scale–6 (PANSS-6, a shorter version of PANSS-30) questionnaire, in order to assess symptoms of psychosis [Østergaard et al., 2016]. Scores for the two questionnaires were found to be highly correlated. Out of the patient group, 66.7% were assigned a score below 14, a recommended preliminary threshold indicating schizophrenia remission.

The interviews were recorded and then manually transcribed by Hebrew-speaking students from our lab. The TAT pictures presented to participants during the interview were: 1, 2, 3BM, 4, 5, 6BM, 7GF, 8BM, 9BM, 12M, 13MF, 13B, 14, 3GF. Table 3.2 lists the questions that were presented to the participants during the interview. All the transcripts are written in Hebrew. Figure 3.1 shows average word counts by question, per group. Clearly, the patients spoke fewer words than the controls. The difference becomes less significant for the open-ended questions.

ID	Question
1	Tell me as much as you can about your bar mitzvah. אנא ספר/י ככול האפשר באריכות על מסיבת הבר/בת מצווה שלך כפי שאת/ה זוכרים אותם?
2	What do you like to do, mostly? אנא ספר/י ככול האפשר באריכות על הדברים שאת/ה אוהב/ת לעשות ואשר מסבים לך הנאה?
3	What are the things that annoy you the most? אנא ספר/י ככול האפשר באריכות על הדברים שמעצבנים אותך?
4	What would you like to do in the future? אנא ספר/י ככול האפשר באריכות הדברים אשר היית רוצה לעשות בעתיד?

Table 3.2: Four open questions asked during the interview.

3.2 Preprocessing

Hebrew being a highly-inflected language, we preprocessed the texts with the Ben-Gurion University Morphological Tagger [Adler, 2007], a context-sensitive morphological analyzer for Modern Hebrew. Given a running text, the tagger breaks the text into words and provides morphological information for every word, including the disambiguated part-of-speech tag and lemma. There were no specific instructions given to the transcribers for how to punctuate, which led to an inconsistency in the way punctuation was used in the transcriptions. We used the tags to clean up all punctuation marks by removing all tokens tagged as such.

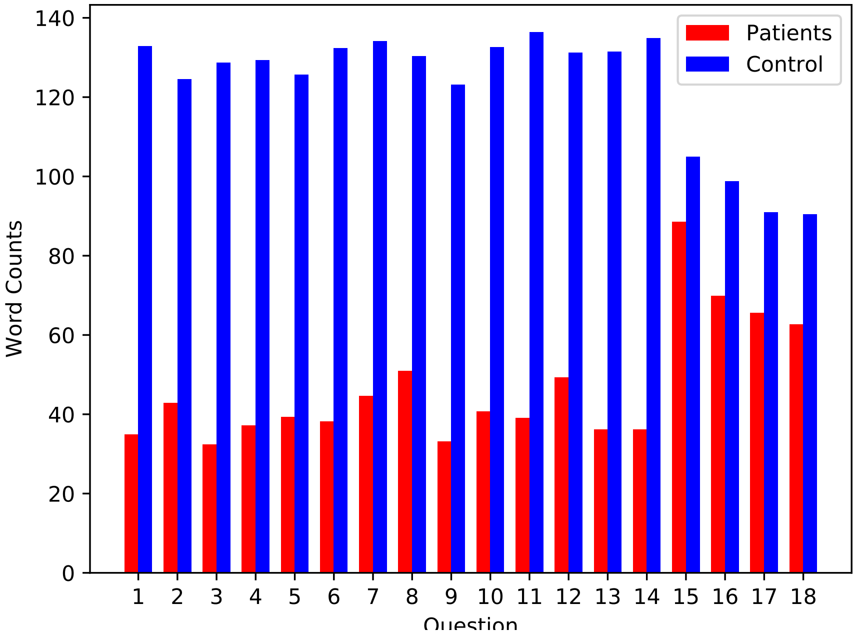

Figure 3.1: Word counts per question.

Chapter 4

Tools and Method

In this chapter we report on two sets of experiments. In the first, we measure derailment by calculating the semantic similarity between adjacent words in running text. In the second set of experiments, we investigate the modifiers that the two groups use to describe specific nouns and verbs. As a final step, we measure the contribution of the semantic characteristics that we compute in the experiments, for automatic classification of schizophrenia.

4.1 Experiment 1: Measuring Derailment

We calculate a derailment score for each response and use it to measure derailment.

Tools: To measure derailment, we calculate the semantic similarity of adjacent words in the answers provided by the participants during the interview. We use word embeddings to represent each word by means of a mathematical vector that captures its meaning. These vectors were created automatically by characterizing words by the surrounding contexts in which they are mentioned in a large corpus of documents. Specifically, we used Hebrew pretrained vectors provided by `fastText` [Grave et al., 2018], which were created from Wikipedia,[*] as well as from other content extracted from the web with Common Crawl.[†] Overall, 97% of the words in our corpus exist in `fastText`. Hebrew words are inflected for person, number and gender; prefixes and suffixes are added to indicate definiteness, conjunction, prepositions, and possessive forms. On the other hand, `fastText` was trained for surface forms. Therefore, we work on the surface-form level. To measure semantic similarity between two words, we use the common cosine-similarity function that calculates the cosine of the angle between the two corresponding vectors. The score ranges from -1 to $+1$, with $+1$ representing maximal similarity.

[*] https://www.wikipedia.org

[†] http://commoncrawl.org

Method: (1) For each sufficiently long response, R, we retrieve the `fastText` vector v_i for every word R_i, $i = 0 \dots n$, in the response. (2) For each word, we calculate the average pairwise cosine similarity between this word and the k following words. The integer k is a parameter; we experimented with different values. (3) We take the average of all the individual cosine similarity scores and form a single score for each response.

In this experiment, we consider only responses that are long enough to allow topic mutation to develop. Therefore, we use only the four questions from Table 3.2 for which the participants provided a relatively long response. Accordingly, we drop responses of fewer than 50 words. As mentioned above, we consider that the existence of some word types, like fillers and functional words, might introduce some noise, which might harm the calculation process. We would rather focus on words that convey real content. Therefore, we calculate scores separately using all words and using only *content words*, which we take to be nouns, verbs, adjectives, and adverbs. We detected a few types of text repetitions, which may bias the derailment score. One type is when a word is said twice or more for emphasis; for example, “quickly, quickly” (מהר מהר) (i.e. very quickly). To mitigate this bias, we keep only one word out of a pair of consecutive identical words. Another type is when a whole phrase is repeated; for example, “She’s in a big hurry; she’s in a big hurry” (היא ממהרת מאוד, היא ממהרת מאוד). Handling this problem is left for future work.

We calculate derailment scores for the responses provided by all participants and compare the means of the two groups.

Results: When using all words, we could not detect a significant difference between patients and controls. However, when using content words only, patients scored lower on derailment than the controls, for all window widths k, suggesting that focusing only on content words is the more robust approach for calculating derailment. This finding is consistent with previous work [Iter et al., 2018]. Overall, coherence decreases as k increases. Table 4.1 summarizes the results. To confirm the significance we are seeing in the results, is due to the diagnosis and not due to other characteristics of the participants, we aggregated the same scores for the different age groups and education levels, regardless of the diagnosis status; all these results did not appear to be significant. Figure 4.1 shows the trend of the average derailment score from Table 4.1, running with different values of k. The left plot was produced for all word types, and the right plot using only content words. We clearly observe a slight increase of the entire control curve and a slight decrease of the patients curve, when restricting to content words.

4.2 Experiment 2: Incoherence

In this experiment, we examine the way patients use adjectives and adverbs (hereafter, *modifiers*) to describe specific nouns and verbs, respectively. Our goal is to measure the difference between modifiers used by patients and the ones used by controls, when describing the same nouns and verbs. We suggest this as a tool for measuring incoherence in speech. For example, inspecting the responses for the first TAT image, we

	All Words			**Content Words**		
k	Control	Patients	t	Control	Patients	t
1	0.270 (0.014)	0.257 (0.025)	2.004*	0.265 (0.019)	0.240 (0.020)	2.968*
2	0.246 (0.017)	0.239 (0.025)	1.173	0.256 (0.018)	0.231 (0.025)	2.687*
3	0.237 (0.017)	0.233 (0.025)	0.476	0.250 (0.018)	0.225 (0.026)	2.614*
4	0.233 (0.018)	0.229 (0.025)	0.471	0.245 (0.018)	0.221 (0.026)	2.539*
5	0.230 (0.017)	0.226 (0.026)	0.528	0.241 (0.018)	0.218 (0.023)	2.598*

Table 4.1: Results for Experiment 1. Comparing average derailment scores of patients and controls. The numbers are provided as average across patients and controls, with standard deviation in parentheses, $*p < 0.05$.

learn that patients typically use the adjectives "new" (חדש) and "good" (טוב) to modify the noun "violin" (כינור), while controls use the adjectives "old" (ישן), "sad" (עצוב), and "significant" (משמעותי).

Tools: To detect all noun-adjective and verb-adverb pairs in the responses, we use a dependency parser, which analyzes the grammatical structure of a sentence and builds links between "head" words and their modifiers. Specifically, we use YAP [More and Tsarfaty, 2016], a dependency parser for Modern Hebrew, and process each sentence individually. Among other things, YAP provides a word-dependency list, shaped as a list of tuples, each includes a head word, a dependent word, and the kind of dependency. We use the relevant types (e.g. *advmod*, *amod*) for finding all noun-adjective and verb-adverb pairs. For example, Figure 4.2 shows the dependencies returned by YAP for the input sentence: "I ate a tasty candy" (אכלתי סוכריה טעימה). From this sentence we extract the noun "candy" (סוכריה), which is modified by the adjective "tasty" (טעימה).

Method: To measure the difference between the modifiers that are used by patients and controls, we compare them to the modifiers that are commonly used to describe the same nouns and verbs. For example, given an answer with only one noun "violin" (כינור) that is modified by the adjective "sad" (עצוב), we calculate a score that reflects how similar the adjective "sad" is to adjectives that are typically used to describe a violin.

We take the following steps:

[‡] https://www.doctors.co.il

[§] https://www.infomed.co.il

[¶] https://tobehealthy.co.il

[||] https://www.haaretz.co.il

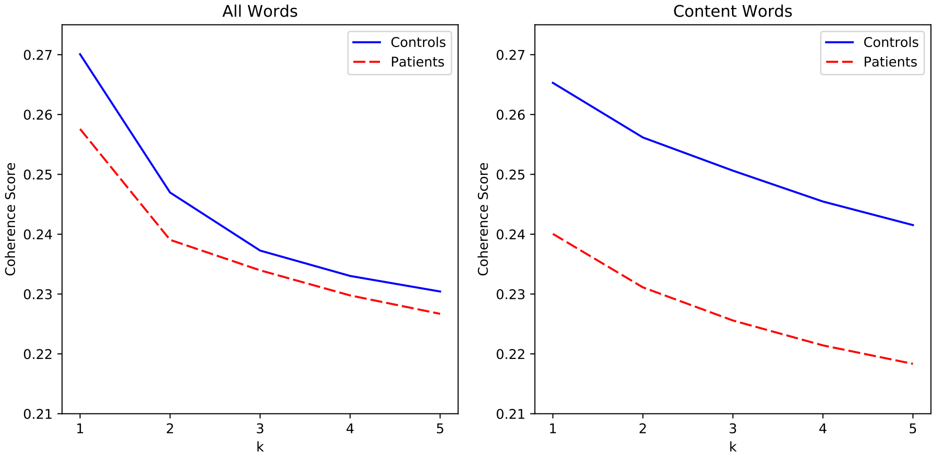

Figure 4.1: Derailment scores for different values of k. The left plot shows the results for all word types, and the right plot shows the results for content words only.

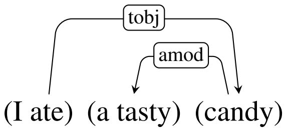

Figure 4.2: The dependencies returned by YAP for the sentence "(I ate) (a tasty) candy". The parentheses delimit the translations for each of the three Hebrew words in the sentence.

Corpus	Description	# Documents	# Words
Doctors[‡]	Articles from the Doctors medical website	239	187,938
Infomed[§]	Question-and-answer discussions from the Infomed website's medical forum, January 2006 – September 2007	749	128,090
To Be Healthy[¶]	Articles and forum discussions from the To Be Healthy (L'Hiyot Bari, 2b-bari) medical website	137	112,839
HaAretz[‖]	News and articles from the HaAretz news website, 1991	4,920	250,399

Table 4.2: The external Hebrew corpora used to collect modifiers of nouns and verbs that are typically used.

1. We convert each sentence into a list of noun-adjective and verb-adverb pairs using YAP.
2. To compare each modifier with the modifiers that are typically used to describe the same noun or verb, we use external corpora as reference. These were taken from various sources reflecting the

health domain we are working in.[**] Table 4.2 lists the sources and the corresponding number of documents and words that they contain. Each document in these sources was processed in exactly the same way to find all noun-adjective and verb-adverb pairs.

3. Given a list of noun-adjective and verb-adverb pairs of one response, we calculate the similarity score of every modifier that describes a specific noun or verb with the set of modifiers describing exactly the same noun or verb in the reference corpus. Looking at our example above, we would want to calculate a similarity score between the adjective “old” (ישן) and all the adjectives that are used to describe “violin” (כינור) in the reference corpus. Searching for instances of the same Hebrew word is challenging due to Hebrew’s rich morphology. Hebrew words are inflected for person, number, and gender; prefixes and suffixes are added to indicate definiteness, conjunction, various prepositions, and possessive forms. Therefore, we work on the lemma (base-form) level. Most vowels in Hebrew are not indicated in standard writing; therefore, Hebrew words tend to be ambiguous, and determining the correct lemma for a word is nontrivial. We use the lemmas provided by YAP.

   Another challenge is how to compare a single modifier with a group of modifiers that were taken from the reference corpus. We take the `fastText` vectors of the modifiers that were extracted from the reference corpus and aggregate them into a single vector. Then, we take cosine similarity between the modifier from the response and the aggregated vector of the modifiers from the reference corpus. As an aggregation function, we use element-wise weighted average of the individual modifiers’ `fastText` vectors, and define the weights to be the inverse-document-frequency (IDF) score to account more for modifiers that describe the noun or verb more uniquely. We calculate IDF scores using the reference corpora. For this purpose, a “qualified” word is a noun or verb that has an IDF score and that has at least one modifier linked to it in either the control or patient corpus. Most of the nouns and verbs are non-qualified; we only consider qualified words in this investigation.
4. For each response, we calculate two scores, individually. The adjective-similarity score is the IDF-weighted average of the individual adjective scores we calculate in the previous step. Similarly, the adverb-similarity score is the IDF-weighted average of the individual adverb scores we calculate in the previous step.
5. To calculate a score on the participant level, we average the scores of all the individual responses provided by the participant.

The output of this process is a pair of scores, one for adjectives and one for adverbs, calculated for each participant. The higher a score is, the more similar the modifiers are to ones that are typically used to describe the same noun or verb.

Results: Table 4.3 summarizes the results. Overall, controls have significantly higher scores for both modifier types, indicating a higher agreement on modifiers by the controls and external writers.

[**] All were downloaded from MILA Knowledge Center for Processing Hebrew: `http://mila.cs.technion.ac.il/resources_corpora.html`.

	Control	**Patients**	t
Adjectives	0.59 (0.03)	0.55 (0.03)	4.78***
Adverbs	0.69 (0.03)	0.63 (0.07)	4.30***

Table 4.3: Results for Experiment 2. The numbers are average coherence scores across patients and controls (with standard deviations); $^{***}p < 0.001$.

	Control		**Patients**	
	Total	Qual.	Total	Qual.
Nouns	934	226	242	90
Adjectives	573	371	204	127
Verbs	699	60	204	34
Adverbs	166	104	86	50

Table 4.4: Experiment 2: Counts of nouns, verbs, and their modifiers, across the two groups. Qual. = Qualified.

Classifier	**Acc.**	**Prec.**	**Recall**
Random Forest	81.5%	91.3%	71.8%
XGBoost	80.5%	86.8%	73.1%
SVM	70.4%	72.1%	47.3%

Table 4.5: Classification results for each classifier.

There are more nouns and adjectives than verbs and adverbs, as summarized in Table 4.4. On average, participants use more adjectives to describe nouns than adverbs to describe verbs. Controls use about 0.61 adjectives per noun, while patients use 0.84 adjectives on average. Similarly, patients use more adverbs to describe a verb on average than controls do. While patients use about 0.42 adverbs per verb, controls use only 0.23. However, these differences are not significant.

4.3 Classification

As a final step, we train several classifiers to distinguish between controls and patients. We represent participants with the characteristics we compute in the two experiments. Specifically, each subject is represented by the following: (1) noun and verb derailment scores; (2) coherence scores for 5 windows, using all words; and (3) coherence scores for 5 windows, using only content words. In total, we use 12 scores per subject. Each classifier was trained using a 10-fold cross-validation evaluation of prediction quality over the 51 participants. For each classifier, we report on the overall prediction accuracy, as well as precision and recall for the prediction of the patients group. The classification algorithms we tried are Random Forest [Breiman, 2001] and XGBoost [Chen and Guestrin, 2016], both based on decision trees, and, in addition, linear support vector machines (SVM) [Cortes and Vapnik, 1995]. Table 4.5 summarizes the results per classifier with respect to the different metrics.

We used the decision-tree based classifiers to calculate the most important features, that is, the ones that have the greatest impact on prediction decisions. The most important features were found to be the two derailment scores, as expected.

Chapter 5

Social Media Experiments

To further test our derailment findings, we decided to try and reproduce our results on written text, taken from two social-media corpora. Such reproduction would imply that thought derailment is not only a characteristic of free speech, but also of written text. Following conclusions from previous derailment experiment, we did not look for results using all words, and instead skipped to using only content words.

For both datasets, word embeddings were generated using Flair Embeddings [Akbik et al., 2018], by concatenating several embeddings, namely GloVe, news-forward-fast, and news-backward-fast, for a total of 2148 dimensions. In the first dataset we show that Reddit self-reported depression diagnosed users don't show derailment in text. In the first dataset we collect Twitter self-reported schizophrenia diagnosed users and also show that they show no derailment in the written texts.

We expected social media text to differ from speech due to their inherent differences. Written text, specifically social media text, is different from discourse; discourse specifies the agent of the information, making it interactive in nature in contrast to non-active social media posts. In our case, the questionnaire was conducted by interviewers asking questions. Furthermore, free speech is mostly spontaneous whereas text can be edited multiple times before being published.

5.1 Reddit Self-Reported Depression Diagnosis

We use the Reddit self-reported depression diagnosis (*RSDD*) [Yates et al., 2017] dataset as a source for users who are self-reported diagnosed with depression as well as a controls group.

Down sampling: Since RSDD contains approximately 9,000 diagnosed users and approximately 107,000 matched control users, some down sampling was required. Down sampling consisted of several steps: (1) Sampling 500 diagnosed users and 500 controls, each user with less than 500 posts and at least 10 words in each post, for approximately 17K posts. In contrast to our speech dataset, where all participants maintained the same discourse topics, social media topic vary. We needed to pick posts discussing the same topics in general, and filter out noise. (2) We use DBSCAN [Ester et al., 1996] for data clustering, and pick the

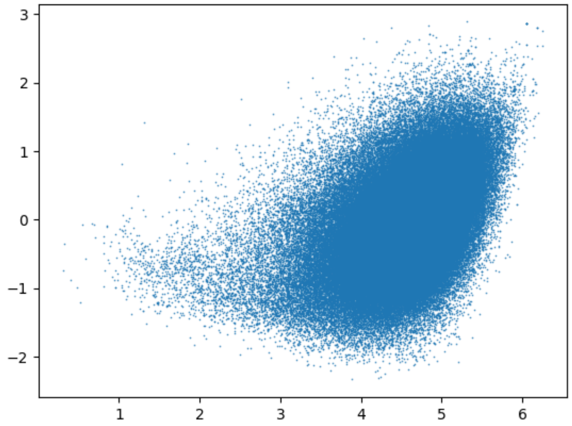

Figure 5.1: RSDD posts dimensionality as given by PCA dimensionality reduction.

	Content Words		
k	Control	Patients	t
1	0.211	0.213	-1.951
2	0.202	0.204	-1.493
3	0.199	0.200	-1.189
4	0.197	0.198	-1.074
5	0.197	0.197	-0.680

Table 5.1: Derailment results for RSDD dataset. Comparing average derailment scores of patients and controls. The numbers are provided as average across patients and controls; $*p < 0.05$.

biggest cluster. As can be seen in Figure 5.1, the data consists of one giant cluster. Running DBSCAN winnows down to 995 users with approximately 8K posts.

Derailment results: There is no significant difference between diagnosed users and controls, as can be shown in Table 5.1. This result is expected, because depression patients do not tend to show derailment symptoms.

5.2 Twitter Self-Reported Schizophrenia Diagnoses

In this section, we describe the collection of a Twitter self-reported schizophrenia diagnoses *(TSSD)* dataset. We collected data using a Twitter filter API for a total of 1431 users, separated into two groups:

- Self-reported schizophrenia spectrum disorders (schizophrenia): 179 users.
- Control group: 1252 users who are similar to the schizophrenia group.

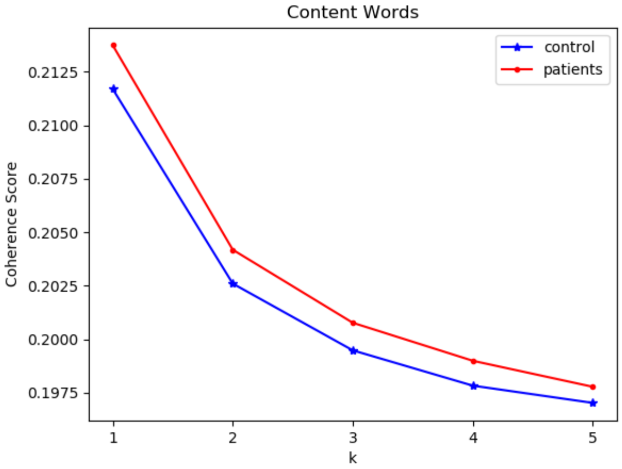

Figure 5.2: Derailment scores for the RSDD dataset for different values of k.

Diagnosed users: To create TSSD, we start by collecting a group of candidate schizophrenia users using the Twitter filter API. For filtering, all kinds of synonyms of schizophrenia were used.[*] We then apply two types of high precision patterns in a manner similar to SMHD: the first is used to match a positive diagnoses,[†] and the second to remove a negative diagnoses.[‡] These were manually reviewed and verified. A further filtering removed users with fewer than 100 public English tweets.

Control users: After having a group of diagnosed users with their respective latest 100 posts, we selected the most used 100 words, excluding stop words and mental-health related words, as a new filter for Twitter's filter API. As before, users with less than 100 public English tweets were filtered out. For each diagnosed user, a group of 7 most similar controls users were selected based on their posts cosine similarity scores. Note that the final dataset does not include duplicate users nor posts related to mental health.

Derailment results: We experimented with varying minimum post lengths. In all experiments there was no significant difference between the groups. In Figure 5.3, it is apparent that scores are not correlated. Results are given in Figure 5.2. To date, Twitter restricts tweets to a maximum of 280 characters, but the most common tweet length is 33 characters, with only 5% of tweets being longer than 190 characters. We suspect these results arise from Twitter being a medium for short, less-informative posts. It is also possible that written text is not as spontaneous as free speech, giving users time to organize their thoughts and edit multiple time before posting.

[*] E.g. schizophrenia, schizophrenic, schizo, etc. `http://ir.cs.georgetown.edu/data/smhd/conditions/schizophrenia-syns.txt`.

[†] E.g. "I am diagnosed with" `http://ir.cs.georgetown.edu/data/smhd/diagpatterns_positive.txt`.

[‡] E.g. "not formally diagnosed" `http://ir.cs.georgetown.edu/data/smhd/diagpatterns_negative.txt`.

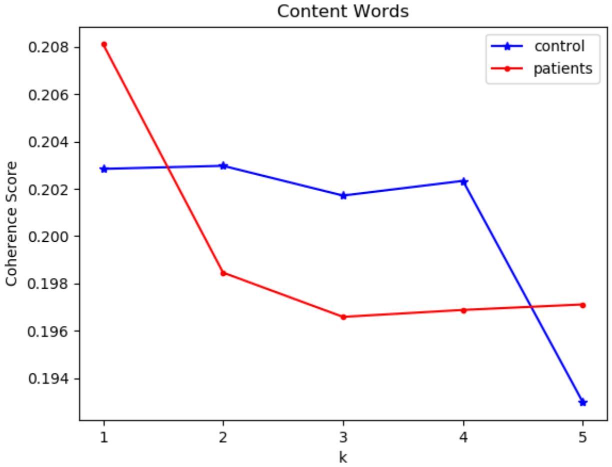

Figure 5.3: Derailment scores for the TSSD dataset for different values of k.

	Content Words		
k	Control	Patients	t
1	0.202	0.208	-2.558
2	0.202	0.198	1.873
3	0.201	0.196	1.765
4	0.202	0.196	1.589
5	0.192	0.197	-1.914

Table 5.2: Derailment results for TSSD dataset. Comparing average derailment scores of patients and controls. The numbers are provided as average across patients and controls; none are significant at $p < 0.05$.

Chapter 6

Conclusions

With the aim of detecting speech disturbances, we have analyzed transcribed Hebrew speech, produced by schizophrenia inpatients and compared it with those of controls. We believe that speech produced during a psychiatric interview is a more reliable data source for detecting disturbances than are social media posts.

Generally speaking, we find that patients talk significantly less in interviews than controls do.

In one experiment, we use word embeddings to detect derailment, that is, when a speaker shifts to a topic that is not strongly related to previously discussed ones. The results show that controls have higher scores, indicating that they keep the topic more cohesive than patients do. These results are in line with previous studies on English [Bedi et al., 2015], which showed that schizophrenics have a lower score, calculated by a similar mathematical procedure.

In a second experiment, we examine the difference in how patients and controls use adjectives and adverbs to describe nouns and verbs, respectively. Our results show that the adjectives and adverbs that are used by controls are more similar to the ones typically used to describe the same nouns and verbs. For now, we consider this difference as related to speech incoherence; however, we plan to continue investigating this direction in the near future, when more data become available.

In the last experiment, we run our derailment experiment on social media text collected separately from depression and schizophrenia self-reported users. Depression users show no derailment, in accord with derailment not being considered as a symptom of depression. Schizophrenia users also show no derailment, probably in part due to the fact that the collected social media texts are too short. Our initial hypothesis was that text is different than speech, and we believe our findings strengthens this hypothesis.

Analyzing Hebrew is more challenging than analyzing English due to Hebrew's rich morphology, as well as the absence of written vowels. In the first experiment, we work with `fastText`, which provides word embeddings on the surface-form level. In the second, we use lemmata rather than word surface forms, so we can find multiple instances of the same lexeme.

As we did not measure the IQ of participants, some of the results might, to a certain extent, be attributable to differences in intellect. Moreover, as can be seen in Table 3.1, about 20% of the control participants have some sort of post high-school education, while most of the inpatients did not continue beyond high-school. We plan to address these questions in followup work. Another limitation that we are aware of is related to the classification results, as the number of participants we use for training the classifiers might be considered relatively small.

Overall, we found the semantic characteristics that we compute in this study to be beneficial for the task of detecting thought disorders in Hebrew speech. We plan to collect speech samples from more subjects, and to continue to explore additional semantic – as well as grammatical – textual characteristics to support the automatic detection of various mental disorders.

Bibliography

Meni Adler. *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. PhD thesis, Ben-Gurion University of the Negev, Beer-Sheva, Israel, 2007.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

Mohammed Al-Mosaiwi and Tom Johnstone. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542, 2018. doi: 10.1177/2167702617747074. URL https://doi.org/10.1177/2167702617747074.

Nancy C. Andreasen. Thought, language, and communication disorders: I. clinical assessment, definition of terms, and evaluation of their reliability. *Archives of General Psychiatry*, 36(12):1315–1321, 1979.

American Psychological Association. *Diagnostic and Statistical Manual of Mental Disorders (5th ed.)*. Washington, DC, 2013.

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

Eugen Bleuler. Dementia praecox oder Gruppe der Schizophrenien. In G. Aschaffenburg, editor, *Handbuch der Psychiatrie*, volume Spezieller Teil. 4. Abteilung, 1. Hälfte. Franz Deuticke, Leipzig, 1991.

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in Twitter. In *ICWSM 2014*, 2014.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, 2015.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 106–117, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411. URL `https://doi.org/10.1023/A:1022627411411`.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *ICWSM*, 13:1–10, 2013.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391–407, 1990.

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

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Dan Iter, Jong Yoon, and Dan Jurafsky. 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. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-0615. URL `http://aclweb.org/anthology/W18-0615`.

Gregory Katz, Leon Grunhaus, Shukrallah Deeb, Emi Shufman, Rachel Bar-Hamburger, and Rimona Durst. A comparative study of Arab and Jewish patients admitted for psychiatric hospitalization in Jerusalem: the demographic, psychopathologic aspects, and the drug abuse comorbidity. *Comprehensive Psychiatry*, 53(6):850–853, 2012. ISSN 0010-440X. doi: https://doi.org/10.1016/j.comppsych.2011.11.005. URL http://www.sciencedirect.com/science/article/pii/S0010440X11002161.

David E. Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer, 2016.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the language of schizophrenia in social media. In *CLPsych@HLT-NAACL*, 2015.

Amir More and Reut Tsarfaty. Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING 2016*, December 2016.

Rodney D. Morice and John C. L. Ingram. Language analysis in schizophrenia: Diagnostic implications. *Australian & New Zealand Journal of Psychiatry*, 16(2):11–21, 1982. doi: 10.3109/00048678209161186. URL https://doi.org/10.3109/00048678209161186. PMID: 6957177.

M Obręska and T Obręski. Lexical and grammatical analysis of schizophrenic patients' language: A preliminary report. *Psychology of Language and Communication*, 11(1):63–72, 2007.

Soren Dinesen Østergaard, Ole Michael Lemming, Ole Mors, Christoph U. Correll, and Per Bech. PANSS-6: A brief rating scale for the measurement of severity in schizophrenia. *Acta Psychiatrica Scandinavica*, 133(6):436–444, 2016. doi: 10.1111/acps.12526. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/acps.12526.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. Technical report, University of Texas at Austin, Austin, TX, 2015.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing depression from Twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3187–3196. ACM, 2015.

Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. *CoRR*, abs/1709.01848, 2017. URL http://arxiv.org/abs/1709.01848.

הפקולטה למדעים מדויקים
ע״ש ריימונד ובברלי סאקלר
אוניברסיטת תל אביב

מאפיינים סמנטיים של דיבור סכיזופרני

חיבור זה
הוגש כחלק מהדרישות לקבלת תואר
מוסמך אוניברסיטה

על ידי

ורד זילברשטיין

העבודה הוכנה בהנחיית
ד"ר כפיר בר
פרופ׳ נחום דרשוביץ

אוניברסיטת תל אביב
בית הספר למדעי המחשב

אייר התש״פ

תקציר

כלים לעיבוד שפה טבעית משמשים לאיתור אוטומטי של הפרעות בדיבור מתועתק בקרב חולי סכיזופרניה מאושפזים דוברי עברית.

אנו מודדים את השינוי בנושא השיחה ומראים שלאורך זמן קבוצת הביקורת שומרת על דיבור קוהרנטי יותר מאשר קבוצת החולים.

כמו כן, אנו בוחנים הבדלים בשימוש של קבוצת החולים וקבוצת הביקורת בתארים ובתארי הפועל כדי לתאר מילות תוכן ומראים כי אלה המשמשים את קבוצת הביקורת הם שכיחים יותר מאלה של חולים.

אנו מספקים תוצאות ניסיוניות ומראים את הפוטנציאל שלהם לגילוי אוטומטי של סכיזופרניה בקרב חולים באמצעות דפוסי הדיבור שלהם בלבד.

בהמשך אנו חוקרים את הממצאים שלנו על טקסט כתוב שנלקח ממדיה חברתית, אשר אינו מציג הבדל משמעותי בשמירה על שיח קוהרנטי לאורך זמן.