Prediction of schizophrenia with Ngram Tracing and Word Graph Analysis - an exploratory approach

Project Report by Lisa Becker (775242)

Project at the research group CLiPS at the University of Antwerp, supervised by Prof. Dr. Walter Daelemans. Course Assessment for an *Individual Module* for the Masters Program *Cognitive Systems* at the University of Potsdam, supervised by Prof. Dr. Manfred Stede.
Submitted February 6th, 2020.

1 Abstract

Ngram Tracing and Word Graph Analysis are two novel methods in the area of authorship classification (Grieve et al., 2018; Mota et al., 2012). We compared those approaches to a baseline of bag of words with Naïve Bayes and Support Vector Machines in order to evaluate the most suitable method for the task of schizophrenia prediction. Both methods tend to overfit on the small amount of data and not lead to more valuable results than the baseline.

2 Schizophrenia in Natural Language Processing

People with schizophrenia can show a variety of symptoms but a central one seems to be the impairment of language and communication in terms of problems distinguishing between internal speech and verbalised thought. This leads to a disorganized language output, also known as positive thought disorder (Kuperberg, 2010). This usually shows as disorganized and difficult to follow discourse and can entail derailment (spontaneous speech, obliquely- or non-related ideas), neo-logisms or non-words or schizophasia (unintelligible speech). On the other side, negative symptoms contain alogia (poverty of speech) which often appears together with other non-linguistic negative symptoms.

These symptoms indicate that computational linguistic methods might be able to predict schizophrenia which is potentially interesting for future diagnostic methods. According to our knowledge, computational methods have not been used yet to help with the diagnosis of psychological differences based on a person's language output.

Lorien Verachtert investigated in her Master's thesis three approaches to predict schizophrenia (Verachtert, 2017). The first one employs statistical measurements such as the Type Token Ratio (Manschreck et al., 1991), quantifying the number of associations in text (Han et al., 2003) and Latent Semantic Analysis (Elvevåg et al., 2007) which showed that patients with schizophrenia usually produce language with a lower TTr-score, more associations and a lower LSA-score when compared with controls. The second approach employing lexical and syntactic structure found that patients with schizophrenia produce lower syntactic complexity (Sanders et al., 1995). The third approach investigates mostly cohesion measures which show a certain referential impairment in people with schizophrenia.

We investigated the same data as Verachtert but worked with two different methods in order to see whether they perform better than Verachtert's approach. The first one had not been used on schizophrenic data before (Ngram Tracing), the other one was specifically designed for schizophrenia prediction and performed with high classification accuracy (Word Graph Analysis).

3 Data and Tools

47 schizophrenic patients and 55 controls (short: "C") took part in the study. The patients could be both in the acute phase (sudden onset of the disorder and quickly deteriorating, short: "PA") and in remission (temporary decrease in symptoms of the disorder, short: "PR"). All participants were Dutch speaking adults from a variety of educational levels and between 18 and 50 years old. The

average age was 28.5 years in the control group and 32.1 years in the patient group. The majority of the participants were men. Furthermore, participants were not matched on age and gender.

The participants performed either one language task or both of them, depending on how long they stayed in the study. They were asked to participate in two narrative tasks: the first one was to describe "how to make coffee", the second was describing a complex drawing that showed a person making pizza in a restaurant setting.

All tasks were manually written down in Dutch by the participants and the data was collected by CAPRI[1].

Task	Class	C	PA	PR
1	Instances	50	30	4
	Mean length in words	97.04	65.63	34.25
	STD	±72.45	±42.11	±13.01
2	Instances	46	5	28
	Mean length in words	72.60	51.20	41.79
	STD	±46.26	±35.05	±25.40
both	Instances	96	35	32
	Mean length in words	88.35	63.57	40.84
	STD	±61.97	±41.48	±24.33
Task	Class	C+PR	PR+PA	
1	Instances	50	30	
	Mean length in words	92.39	61.94	
	STD	±71.71	±41.07	
2	Instances	76	32	
	Mean length in words	64.86	43.21	
	STD	±46.02	±27.29	
both	Instances	128	67	
	Mean length in words	76.48	52.72	
	STD	±61.15	±40.04	

Table 1: Amount of instances per class and class grouping, their mean length in words including the standard deviation.

Classifier and evaluation packages were obtained from Scikit-Learn[2]. FROG[3] was used for lemmatization since it has been proven to be the best performing tool for lemmatizing Dutch currently available (Bosch et al., 2007).

[1]The three datasets are based on the data collected by Dr. Livia De Picker from the Collaborative Antwerp Psychiatric Research Institute (CAPRI), a scientific research centre of (neuro-)psychiatry and mental health.

[2]Scikit-Learn: https://scikit-learn.org/stable/, last accessed January 25th, 2020.

[3]FROG: https://languagemachines.github.io/frog/, last accessed January 25th, 2020.

4 Models

The methods Ngram Tracing and Word Graph Analysis seemed to be promising approaches for authorship detection and schizophrenia prediction, which will be in detail explained later. In order to gather further evidence regarding their performance on our data, we compared those two approaches to a supervied machine learning baseline with a bag of words as features classified with a Multinomial Naïve Bayes (short: "NB") and a Linear Support Vector Machine classifier (short: "SVM").

We tested the methods on different combinations of participant groups. First, we compared controls versus acute patients versus patients in remission ("C vs PR vs PA").

Since patients in remission are under therapy which includes medicine to suppress typical schizophrenic symptoms and thus are anticipated to act more similar to controls than to acute patients, the second combination grouped controls together with patients in remission and compared them with the acute patients ("C+PR vs PA").

As a third constellation, we grouped together the two patient groups and compared them with the control group ("C vs PR+PA") in order to evaluate the influence of schizophrenia on language production regardless of treatment.

Similarly, we tested both language tasks individually but also added both together to a third corpus to improve classifier performance by the increase of test size and a more balanced dataset.

5 Supervised Machine Learning

We chose a Multinomial Naïve Bayes classifier and a Linear Support Vector Machine with a bag of words based on word counts and lemma counts as features. This served as a baseline to compare to the two novel methods.

5.1 Evaluation and Results

Both models were evaluated with a repeated random test-train split which splits the data into a 67% / 33% train/test split and repeats the process ten times. This evaluation method proved to be the most stable with the lowest standard deviation compared to other evaluation methods. The Support Vector Machine achieved a better performance across all trials and combinations (Figure 2 and 4, except for task 2 with controls and patients

in remission tested against acute patients (Figure 2).

Task, Clf	CvsPRvsPA	C+PRvsPA
1, NB	.61 (±.11)/.59 (±.08)	.65 (±.09)/.64 (±.07)
SVM	.67 (±.06)/.71 (±.06)	.73 (±.06)/**.74 (±.04)**
2, NB	.55 (±.09)/.56 (±.09)	**.94 (±.04)/.94 (±.04)**
SVM	.63 (±.05)/.60 (±.07)	.93 (±.04)/.92 (±.04)
both, NB	.57 (±.06)/.57 (±.07)	.76 (±.06)/.75 (±.05)
SVM	.65 (±.03)/.67 (±.04)	**.82 (±.03)**/.81 (±.04)
Task, Clf	CvsPR+PA	
1, NB	.63 (±.11)/.63 (±.09)	
SVM	.70 (±.06)/.73 (±.06)	
2, NB	.58 (±.11)/.59 (±.10)	
SVM	.67 (±.07)/.66 (±.07)	
both, NB	.60 (±.06)/.59 (±.07)	
SVM	.70 (±.03)/.72 (±.03)	

Table 2: Accuracy of Naïve Bayes and Support Vector Machines. Left accuracy shows the performance based on raw word counts, right accuracy is based on lemmata counts.

As seen in the classification report in Table 3, the Naïve Bayes performs better regarding the f1-score and overall less misclassification compared to the Support Vector Machine:

Clf	Class	Precision	Recall	F1-Score	Support
NB	C+PR	.93	1.	.97	14
	PA	1.	.50	.67	2
SVM	C+PR	.81	1.	.90	13
	PA	.00	.00	.00	3

Table 3: Classification report for Naïve Bayes and Support Vector Machine performing on C+PR vs PA in language task 2 with raw words.

Lastly, we combined the bag of words of raw word counts and of lemmata and used them as a combined feature. The accuracy increased over all combinations of patients and language tasks, as seen in Table 4:

Task, Clf	CvsPRvsPA	C+PRvsPA	CvsPR+PA
1, NB	.77 (±.05)	.81 (±.05)	.80 (±.05)
1, SVM	.87 (±.06)	.89 (±.07)	**.90 (±.05)**
2, NB	.67 (±.05)	.95 (±.03)	.72 (±.04)
2, SVM	.83 (±.03)	**.96 (±.01)**	.85 (±.04)
both, NB	.71 (±.04)	.83 (±.03)	.74 (±.05)
both, SVM	.84 (±.04)	**.91 (±.02)**	.89 (±.04)

Table 4: Accuracy (in %) of Naïve Bayes and Support Vector Machines with both raw word and lemmata counts as features.

When raw word counts and lemmata are combined, the classification report shows a lower f1-score for Naïve Bayes, but an increased performance for the Support Vector Machine (Table 5):

Clf	Class	Precision	Recall	F1-Score	Support
NB	C+PR	.87	1.	.93	27
	PA	1.	.20	.33	5
SVM	C+PR	.93	1.	.96	27
	PA	1.	.60	.75	3

Table 5: Classification report for Naïve Bayes and Support Vector Machine performing on C+PR vs PA in language task 2 with raw word counts and lemmata combined.

6 Ngram Tracing

Grieve and colleagues introduced a new authorship attribution method for short texts in 2018 which they called *Ngram Tracing*. Instead of the frequency of ngrams, this method is solely based on the overlap of ngrams in the test and train document, attributing the document to the author whose overlap coefficient is the highest. In this method, a random subset of ngrams is chosen from every possible author (A) such that each test subset is equal in size. Every subset of ngrams is compared with all the ngrams in the document of question (Q) by calculating the intersection of A and Q. The coefficient calculated by dividing this intersection by the amount of ngrams in the document of question:

$$\frac{|Q \cap A|}{|Q|}$$

The overlap coefficient is calculated between a test item (one participant) and each participant group. The classification is based on the highest overlap between a test set and the groups in question. By iterating this process with all participants it is possible to calculate accuracy, f1-score and extract a confusion matrix. Grieve and colleagues observed the best results between 7-8-character-ngrams (f1-scores 0.96) and 2-word-ngrams (f1-score 0.97) when testing single documents written by either Abraham Lincoln or John Hay against (1,085 texts with median length of 125, ranging from 5 to 17,003 words, totaling 400,747 words. Hay: 557 texts with median 159, ranging from 9 to 8,954 words, totaling 261,126). They applied their method on all 1- to 3-word-ngrams and all 3- to 16-character-ngrams for applying the Ngram Tracing on the Bixby letter, which authorship remains unclear. All seventeen analyses attributed the Bixby letter to Lincoln.

6.1 Implementation and Results

Based on the discription of the algorithm in their paper (Grieve et al., 2018), we reconstructed the

method and tested and evaluated it on our data. The extracted ngrams do not span sentence boundaries and all characters are lower cased. We tested both raw words and their lemmata with this approach.

Our implementation of the Ngram Tracing yields the highest performance around 7-8 character ngrams in most cases, similar to the findings of Grieve and colleagues (Figures 1, 2 and 3). The second language task (Figure 2) yields slightly better classification results than the first one (Figure 1). The combination of both tasks averages their accuracy (Figure 3). Raw words performed on average better than lemmata. All numbers are reported in detail in Tables 9, 10 and 11.

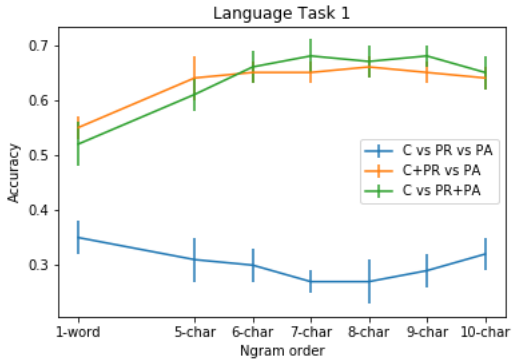

Figure 1: Accuracy and standard deviation per ngram order in language task 1. One graph per participant grouping.

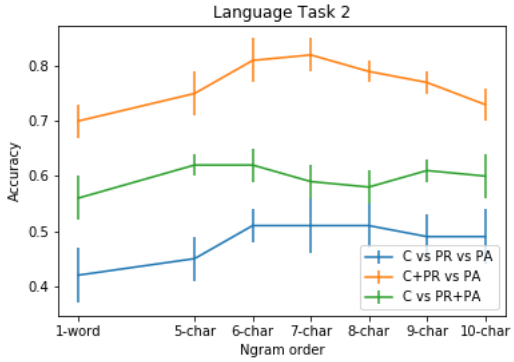

Figure 2: Accuracy and standard deviation per ngram order in language task 2. One graph per participant grouping.

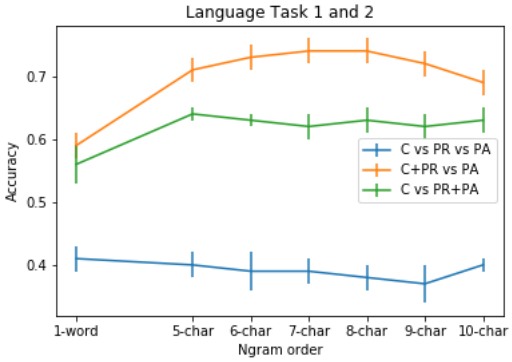

Figure 3: Accuracy and standard deviation per ngram order in both language tasks combined. One graph per participant grouping.

Figure 4 and 5 illustrate the importance of choosing random samples of each group of the same size. This is to avoid overfitting by the document always being attributed to the class with the highest amount of ngrams.

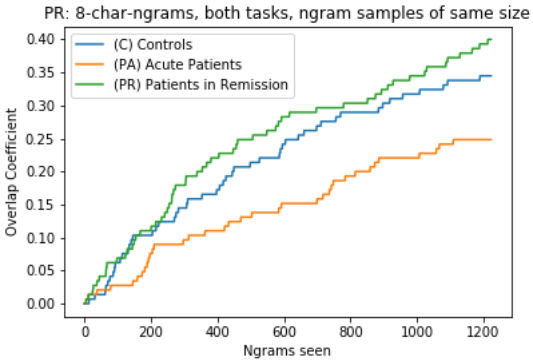

Figure 4: Raise of overlap coefficient by incremental increase of ngrams with a random sample of ngrams per group of the same size (around 1200).

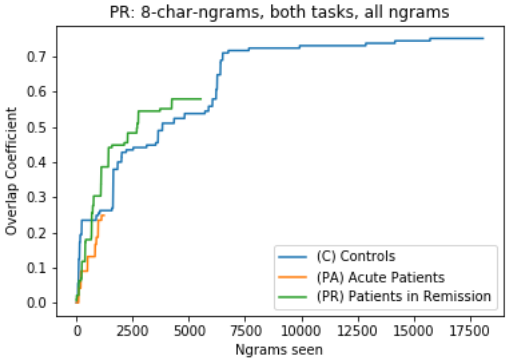

Figure 5: Raise of Overlap Coefficient by incremental increase of ngrams over all ngrams.

7 Word Graph Analysis

Mota and colleagues developed the SpeechGraph software[4] that represents texts as graphs with each word represented as a node:

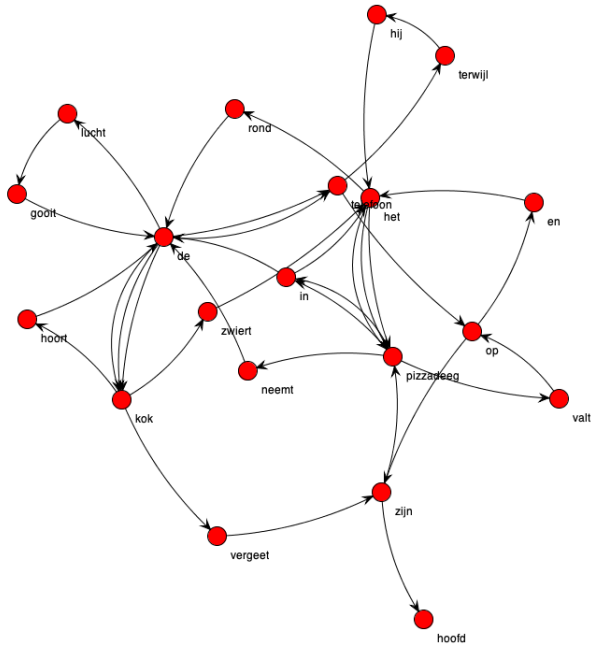

Figure 6: SpeechGraph example of a participant's language task 1.

Figure 6 shows an example of a SpeechGraph representation of the following sentence from the second language task of our data: "The cook swirls the pizza dough. The cook hears the telephone while he throws the pizza dough in the air. The cook forgets his pizza dough picks up the phone and the pizza dough falls on his head." (*NL: De kok zwiert het pizzadeeg in het rond. De kok hoort de telefoon terwijl hij het pizzadeeg in de lucht gooit. De kok vergeet zijn pizzadeeg neemt de telefoon op en het pizzadeeg valt op zijn hoofd.*) It extracts several measurements which Mota and colleagues called connectedness attributes (see Table 12 for further details). Those were used by Mota and colleagues as features for a Naïve Bayes classification (Mota et al., 2017). They interviewed 21 patients undergoing first clinical interview after a psychosis, which was followed by six months of clinical contact until a diagnosis was established. This interview was transcribed and fed into the SpeechGraph software to extract graphs and parameters. The parameters were used to classify negative symptoms and schizophrenia diagnosis. Mota and colleagues reached a high classification accuracy of 100% for negative symptom severity and 91.67% for schizophrenia diagnosis.

[4]SpeechGraphs: https://neuro.ufrn.br/softwares/speechgraphs, last accessed January 25th, 2020.

7.1 Implementation

The biggest difference between Mota and colleagues and our approach is that their main findings are based on dream reports of participants with schizophrenia compared to participants without schizophrenia while we are drawing from data based on two language tasks that do not involve a dream report. We used the free available SpeechGraph Software to extract the graphs and parameters from both language tasks. The data was divided into 30-word windows and their connectedness attributes were classified with Naïve Bayes, like Mota and colleagues proposed. Additionally we implemented a Support Vector Machine due to the sparsity of data and to compare to our other approaches.

7.2 Results

Tables 6 and 7 show that the Support Vector Machine achieves overall better classification results in regard to accuracy. It should again be noted that the first column shows the classification of three groups while the second and third column shows the classification of only two groups. Thus, the chance of the first column lies around 33.33% while it is around 50% for the latter two.

The classification for the Support Vector Machine in both language classes was not compilable, leaving gaps in the results shown in Table 4.

Task	Clf	C vs PR vs PA	C+PR vs PA	C vs PR+PA
1	NB	.53 (±.09)	.60 (±.07)	.64 (±.09)
	SVM	.53 (±.03)	.62 (±.04)	.59 (±.07)
2	NB	.60 (±.10)	.85 (±.05)	.66 (±.07)
	SVM	**.61 (±.09)**	**.89 (±.09)**	**.73 (±.05)**
both	NB	.48 (±.05)	.56 (±.05)	.61 (±.05)
	SVM	.59 (±.05)	.78 (±.04)	.61 (±.06)

Table 6: Accuracy (with STD) of the Word Graph Analysis including all features by SpeechGraph, no division into 30-word-windows, no lemmatization.

While accuracy seems very high, a look at the classification report (Table 12 shows that the Support Vector Machine misclassified all acute patients as controls and patients in remission. This holds across all trials of Support Vector Machine classification.

Task	Clf	C vs PR vs PA	C+PR vs PA	C vs PR+PA
1	NB	.68 (± .00)	.56 (±.02)	.56 (±.02)
	SVM	.74 (±.00)	.74 (±.00)	.73 (±.00)
2	NB	.52 (±.02)	.65 (±.03)	.58 (±.03)
	SVM	.79 (±.00)	**.96 (±.00)**	.79 (±.00)
both	NB	.76 (±.01)	.83 (±.05)	.76 (±.05)
	SVM	/	/	/

Table 7: Accuracy (with STD) of the Word Graph Analysis including all features by SpeechGraph, data was divided into windows of 30 words.

Clf	Class	Precision	Recall	F1-Score	Support
NB	C+PR	.97	.65	.78	620
	PA	.04	.36	.07	22
SVM	C+PR	.97	1.	.98	620
	PA	.00	.00	.00	22

Table 8: Classification report for Naïve Bayes and Support Vector Machine performing on C+PR vs PA in language task 2.

8 Discussion

Using lemmata instead of raw words worsened the accuracy in the bag-of-words baseline and the Ngram Tracing. The lemmatizer FROG was built with a corpus of Dutch newspapers while our participants were native speakers of a Flemish dialect of Dutch, which might lead to a inhibited performance of lemmatization.

The erratic behaviour as well as the strong bias in classification in all three classification methods is likely due to the sparsity of data.

Ngram Tracing only led to moderate results on our data (the highest being 81.65% in the 7-char ngrams in language task 2 for C+PR vs PA), but yet above chance for most participant and language task combinations. Similar as Grieve and colleagues, we found best results mainly in character ngrams of ngram order between 5 and 10. 2-word ngrams are performing on average not better than 1-word or 3-word ngrams, as found by Grieve and colleagues. Since Ngram Tracing was introduced for authorship attribution, motivated by the idea that the style of an author is reflected by their use of ngrams, this approach might not be suited for our classification task. The language output by people with schizophrenia is mainly characteristic in the inconsistency of thoughts. Since Ngram Tracing is solely based on the stylography of a document, it fails to capture the semantic ”jumps of thought” which are necessary for classification.

Approaches like the Word Graph Analysis fit this task better by mapping the semantic pattern of language output into nodes and edges. The division of items into 30-word-windows increases the input size, but Table 8 shows that still the same problems of bias occur as with our baseline which often leads to a misclassification of the patient groups as controls but rarely the other way around. This effect was stronger in the results of the Support Vector Machine compared to the Naïve Bayes classifier.

9 Conclusion

The small amount of data and the small size of the texts affected all three methods and were the main limitation for this exploratory study. In this regard, we have not found a better fitting approach to classify the data than Verachtert. Since the study was not controlled in age, gender or medication use of the schizophrenic patients, the findings can not be generalized for schizophrenia prediction.

In regard to the used models, we can exclude Ngram Tracing as a fruitful method for schizophrenia prediction. In order to compare the Ngram Tracing results to the other two methods, the f1-score would need to be computed. However, our baseline and the Word Graph Analysis yielded generally more promising results: Future research should combine a bag of words with Mota and colleagues' features in order to investigate possible improvement of classification on a larger dataset. Furthermore, there should be an extraction and analysis of the individual features of the classification methods in order to improve classification by deliberate selection of the parameters calculated by the SpeechGraph software.

10 Tables

	C vs PR vs PA	C+PR vs PA	C vs PR+PA
1-word	.35 (±.03) / .34 (±.03)	.55 (±.02) / .52 (±.04)	.52 (±.04) / .52 (±.04)
2-word	.30 (±.03) / .26 (±.03)	.59 (±.03) / .57 (±.05)	.61 (±.03) / .59 (±.04)
3-word	.37 (±.05) / **.49 (±.04)**	.60 (±.02) / .55 (±.03)	.58 (±.03) / .53 (±.03)
3-char	**.42 (±.03)** / .40 (±.04)	.59 (±.03) / .58 (±.04)	.62 (±.03) / .57 (±.02)
4-char	.38 (±.04) / .35 (±.04)	.61 (±.03) / .58 (±.02)	.61 (±.02) / .56 (±.04)
5-char	.31 (±.04) / .34 (±.05)	.64 (±.04) / .60 (±.03)	.61 (±.03) / .56 (±.02)
6-char	.30 (±.03) / .26 (±.04)	.65 (±.02) / .61 (±.03)	.66 (±.03) / .61 (±.03)
7-char	.27 (±.02) / .25 (±.02)	.65 (±.02) / **.63 (±.03)**	**.68 (±.03)** / .62 (±.02)
8-char	.27 (±.04) / .23 (±.03)	**.66 (±.02)** / .61 (±.03)	.67 (±.03) / **.63 (±.04)**
9-char	.29 (±.03) / .25 (±.02)	.65 (±.02) / .59 (±.03)	**.68 (±.02)** / .59 (±.03)
10-char	.32 (±.03) / .25 (±.03)	.64 (±.02) / .56 (±.03)	.65 (±.03) / .60 (±.02)
11-char	.33 (±.02) / .31 (±.04)	.63 (±.03) / .55 (±.02)	.66 (±.01) / .61 (±.03)
12-char	.37 (±.03) / .30 (±.04)	.62 (±.03) / .58 (±.02)	.65 (±.02) / .60 (±.02)
13-char	.37 (±.04) / .31 (±.03)	.63 (±.03) / .60 (±.02)	.66 (±.03) / .60 (±.02)
14-char	.35 (±.05) / .31 (±.04)	.65 (±.02) / .59 (±.02)	**.68 (±.02)** / .59 (±.03)
15-char	.36 (±.04) / .33 (±.04)	.63 (±.02) / .60 (±.02)	**.68 (±.02)** / .61 (±.01)
16-char	.33 (±.05) / .34 (±.04)	.64 (±.02) / .61 (±.02)	.67 (±.02) / .61 (±.02)

Table 9: Ngram Tracing yielding the accuracy averaged over 10 trials. The first number is raw words, the second lemmata. First language task.

	C vs PR vs PA	C+PR vs PA	C vs PR+PA
1-word	.42 (±.05) / .33 (±.03)	.70 (±.03) / .53 (±.06)	.56 (±.04) / .59 (±.04)
2-word	.45 (±.05) / .43 (±.05)	.79 (±.02) / **.71 (±.04)**	.52 (±.04) / .51 (±.05)
3-word	.47 (±.04) / .43 (±.05)	.75 (±.03) / .69 (±.04)	.62 (±.04) / **.67 (±.03)**
3-char	.50 (±.04) / .45 (±.06)	.78 (±.04) / .68 (±.04)	.62 (±.03) / .60 (±.04)
4-char	**.51 (±.04)** / .44 (±.04)	.73 (±.04) / .65 (±.03)	**.63 (±.02)** / .63 (±.02)
5-char	.45 (±.04) / .39 (±.03)	.75 (±.04) / .66 (±.03)	.62 (±.02) / .65 (±.03)
6-char	**.51 (±.03) / .47 (±.05)**	.81 (±.04) / .70 (±.03)	.62 (±.03) / .60 (±.03)
7-char	**.51 (±.05) / .47 (±.02)**	**.82 (±.03)** / .70 (±.02)	.59 (±.03) / .61 (±.01)
8-char	**.51 (±.04)** / .44 (±.03)	.79 (±.02) / .68 (±.03)	.58 (±.03) / .63 (±.03)
9-char	.49 (±.04) / .44 (±.03)	.77 (±.02) / .65 (±.02)	.61 (±.02) / .66 (±.04)
10-char	.49 (±.05) / .43 (±.03)	.73 (±.03) / .63 (±.04)	.60 (±.04) / .65 (±.04)
11-char	.45 (±.04) / .39 (±.03)	.71 (±.01) / .60 (±.02)	**.63 (±.02) / .67 (±.03)**
12-char	.42 (±.04) / .38 (±.05)	.69 (±.03) / .59 (±.02)	.62 (±.03) / .65 (±.03)
13-char	.44 (±.03) / .32 (±.04)	.71 (±.03) / .56 (±.03)	.58 (±.02) / .61 (±.04)
14-char	.40 (±.04) / .30 (±.04)	.68 (±.02) / .55 (±.03)	.56 (±.01) / .59 (±.02)
15-char	.41 (±.05) / .33 (±.03)	.67 (±.02) / .58 (±.03)	.55 (±.02) / .55 (±.02)
16-char	.41 (±.04) / .35 (±.04)	.66 (±.02) / .56 (±.03)	.52 (±.03) / .55 (±.02)

Table 10: Ngram Tracing yielding the accuracy averaged over 10 trials. The first number is raw words, the second lemmata. Second language task.

	C vs PR vs PA	C+PR vs PA	C vs PR+PA
1-word	.41 (±.02) / .31 (±.04)	.59 (±.02) / .55 (±.03)	.56 (±.03) / .55 (±.02)
2-word	.36 (±.03) / .35 (±.03)	.71 (±.02) / .63 (±.03)	.57 (±.02) / .55 (±.03)
3-word	.44 (±.03) / .42 (±.03)	.68 (±.02) / .63 (±.02)	.60 (±.02) / .62 (±.03)
3-char	**.47 (±.04) / .45 (±.04)**	.70 (±.03) / .63 (±.02)	.55 (±.02) / .58 (±.03)
4-char	.41 (±.04) / .41 (±.04)	.66 (±.03) / .60 (±.02)	.62 (±.03) / .61 (±.03)
5-char	.40 (±.02) / .35 (±.02)	.71 (±.02) / .61 (±.01)	**.64 (±.01)** / .62 (±.02)
6-char	.39 (±.03) / .35 (±.02)	.73 (±.02) / **.65 (±.03)**	.63 (±.01) / .61 (±.02)
7-char	.39 (±.02) / .36 (±.03)	**.74 (±.02) / .65 (±.02)**	.62 (±.02) / .61 (±.02)
8-char	.38 (±.02) / .35 (±.03)	**.74 (±.02)** / .64 (±.02)	.63 (±.02) / **.63 (±.02)**
9-char	.37 (±.03) / .33 (±.03)	.72 (±.02) / .63 (±.02)	.62 (±.02) / .60 (±.01)
10-char	.40 (±.01) / .33 (±.03)	.69 (±.02) / .63 (±.02)	.63 (±.02) / .61 (±.01)
11-char	.42 (±.02) / .34 (±.03)	.68 (±.01) / .62 (±.02)	.62 (±.01) / .61 (±.02)
12-char	.39 (±.02) / .34 (±.02)	.68 (±.02) / .60 (±.02)	.62 (±.02) / .61 (±.02)
13-char	.38 (±.03) / .33 (±.02)	.67 (±.02) / .58 (±.02)	.60 (±.01) / .60 (±.02)
14-char	.36 (±.02) / .32 (±.04)	.67 (±.02) / .60 (±.02)	.60 (±.02) / .60 (±.02)
15-char	.36 (±.03) / .35 (±.03)	.67 (±.01) / .59 (±.02)	.59 (±.01) / .58 (±.02)
16-char	.36 (±.03) / .36 (±.03)	.66 (±.02) / .62 (±.02)	.59 (±.01) / .60 (±.02)

Table 11: Ngram Tracing yielding the accuracy averaged over 10 trials. The first number is raw words, the second lemmata. Both language tasks combined.

Feature	Description
Edges (E)	amount of edges
Parallel Edges (PE)	sum of all parallel edges linking the same pair of nodes given that the source node of an edge could be the target node of the parallel edge.
Average Shortest Path (ASP)	average length of the shortest path between pairs of nodes of a network, calculated by the sum of all shortest path divided by number of paths. If the graph is not connected, this parameter is calculated based on the LCC of the graph.
Average Total Degree (ATD)	given a node n, the Total Degree is the sum of "in and out" edges. Average Total Degree is the sum of Total Degree of all nodes divided by the number of nodes.
Largest Strongly connected Component (LSC)	number of nodes in the maximal subgraph in which all pairs of nodes are reachable from one another in the directed subgraph (node a reaches node b, and b reaches a)
Largest Connected Component (LCC)	number of nodes in the maximal subgraph in which all pairs of nodes are reachable from one another in the underlying undirected subgraph.
Diameter (D)	length of the longest shortest path between the node pairs of a network. This parameter is based on the undirected graph. This parameter is based on the undirected graph. If the graph is not connected, this parameter is calculated based on the LCC of the graph.
Density (DI)	number of edges divided by possible edges. (D = 2*E/N*(N-1)), where E is the number of edges and N is the number of nodes. This parameter is based on the undirected graph.

Table 12: Word Graph features as described by Mota and colleagues.

References

A. Bosch, W. Daelemans, G. J. Busser, and S. Canisius. An efficient memory-based morphosyntactic tagger and parser for dutch. pages 99–114, 2007.

B. Elvevåg, P. Foltz, D. Weinberger, and T. Goldberg. Elvevåg b, foltz pw, weinberger dr, goldberg te. quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93:304–16, 08 2007.

J. Grieve, I. Clarke, E. Chiang, H. Gideon, A. Heini, A. Nini, and E. Waibel. Attributing the bixby letter using n-gram tracing. *Digital Scholarship in the Humanities*, 16, 2018.

S. D. Han, P. Nestor, M. Shenton, M. Niznikiewicz, G. Hannah, and R. McCarley. Associative memory in chronic schizophrenia: A computational model. *Schizophrenia research*, 61:255–63, 07 2003.

G. Kuperberg. Language in schizophrenia part 1: An introduction. *Language and linguistics compass*, 4:576–589, 2010.

T. C. Manschreck, B. A. Maher, J. E. Rosenthal, and J. Berner. Reduced primacy and related features in schizophrenia. *Schizophrenia Research*, 5:35–41, 1991.

N. Mota, N. Vasconcelos, N. Lemos, A. Pieretti, O. Kinouchi, G. Cecchi, M. Copelli, and S. Ribeiro. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7, 2012.

N. Mota, M. Copelli, and S. Ribeiro. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophrenia*, 3, 2017.

L. Sanders, J. Adams, H. Tager-Flusberg, and M. Shenton. A comparison of clinical and linguistic indices of deviance in the verbal discourse of schizophrenia. *Applied Psycholinguistics*, 16:325–338, 09 1995.

L. Verachtert. A stylometric approach to detecting schizophrenia from writing. Master's thesis, University of Antwerp, Belgium, 2017.