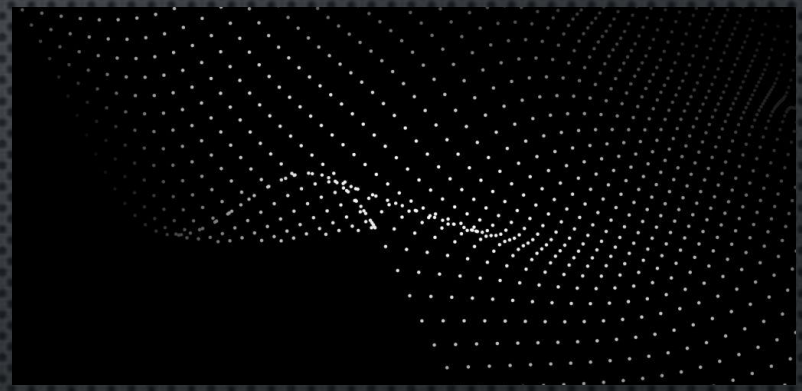
PREDICTIVE AND DISTINCTIVE LINGUISTIC FEATURES IN SCHIZOPHRENIA-BIPOLAR SPECTRUM DISORDERS



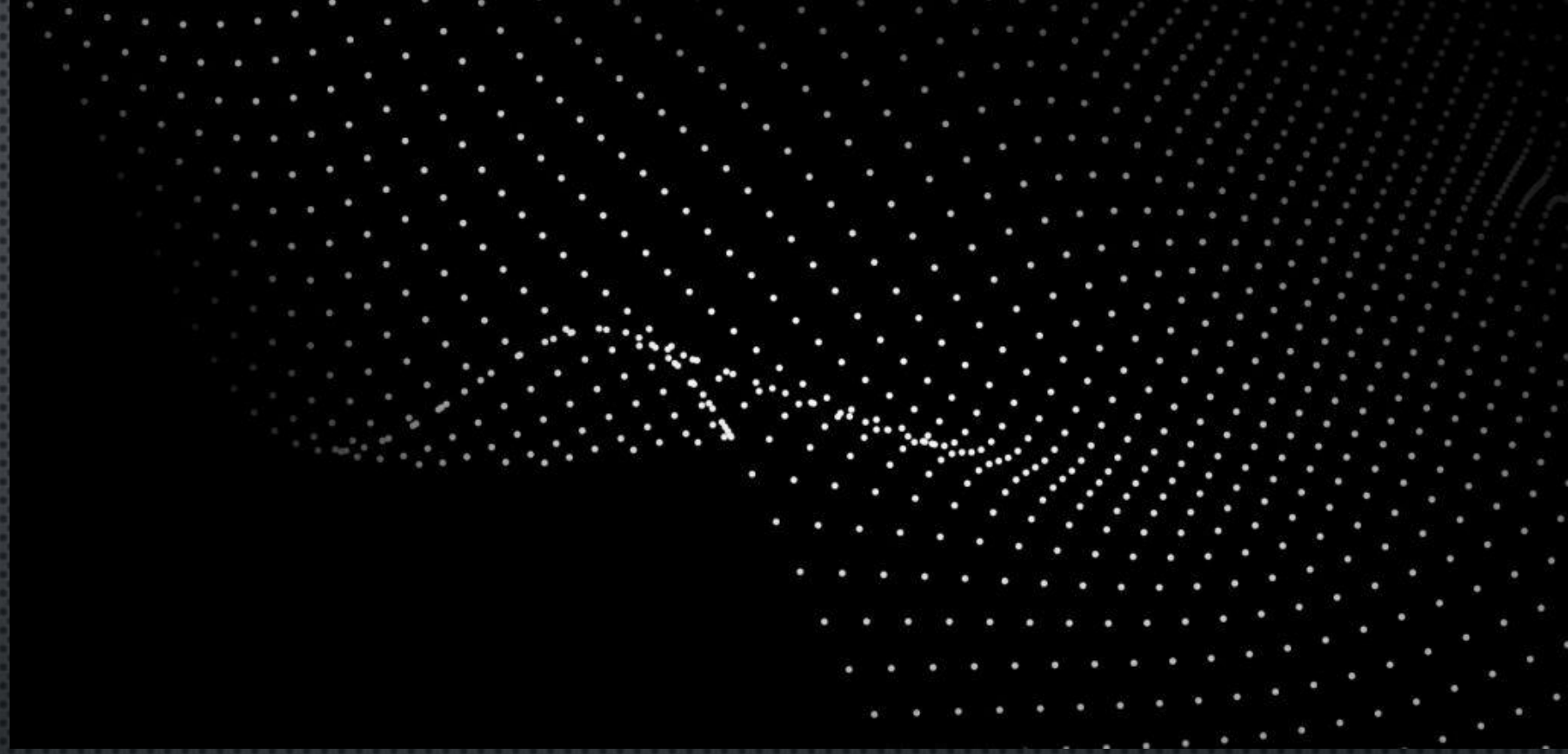
MARTINA KATALIN SZABÓ, VERONIKA VINCZE, BERNADETT DAM, CSENGE GUBA, ANITA BAGI, ISTVÁN SZENDI

University of Szeged, Hungary; HUN-REN Centre for Social Sciences, Hungary; Tokyo University of Foreign Studies, Japan; HUN-REN–SZTE Research Group on Artificial Intelligence, Hungary; Kiskunhalas Semmelweis Hospital, Teaching Hospital of the University of Szeged Psychiatry Department, Hungary

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

Introduction

- Spontaneous speech analysis of Hungarian patients with schizophrenia, schizoaffective, and bipolar disorders
- Comprehensive statistical analysis of linguistic parameters
- Goal: Identify distinctive linguistic features
- Method: Automatic differentiation among patient groups and controls using random forest algorithm
- Results: Effective distinction among SZ, SAD, BD, and controls, surpassing baseline results.



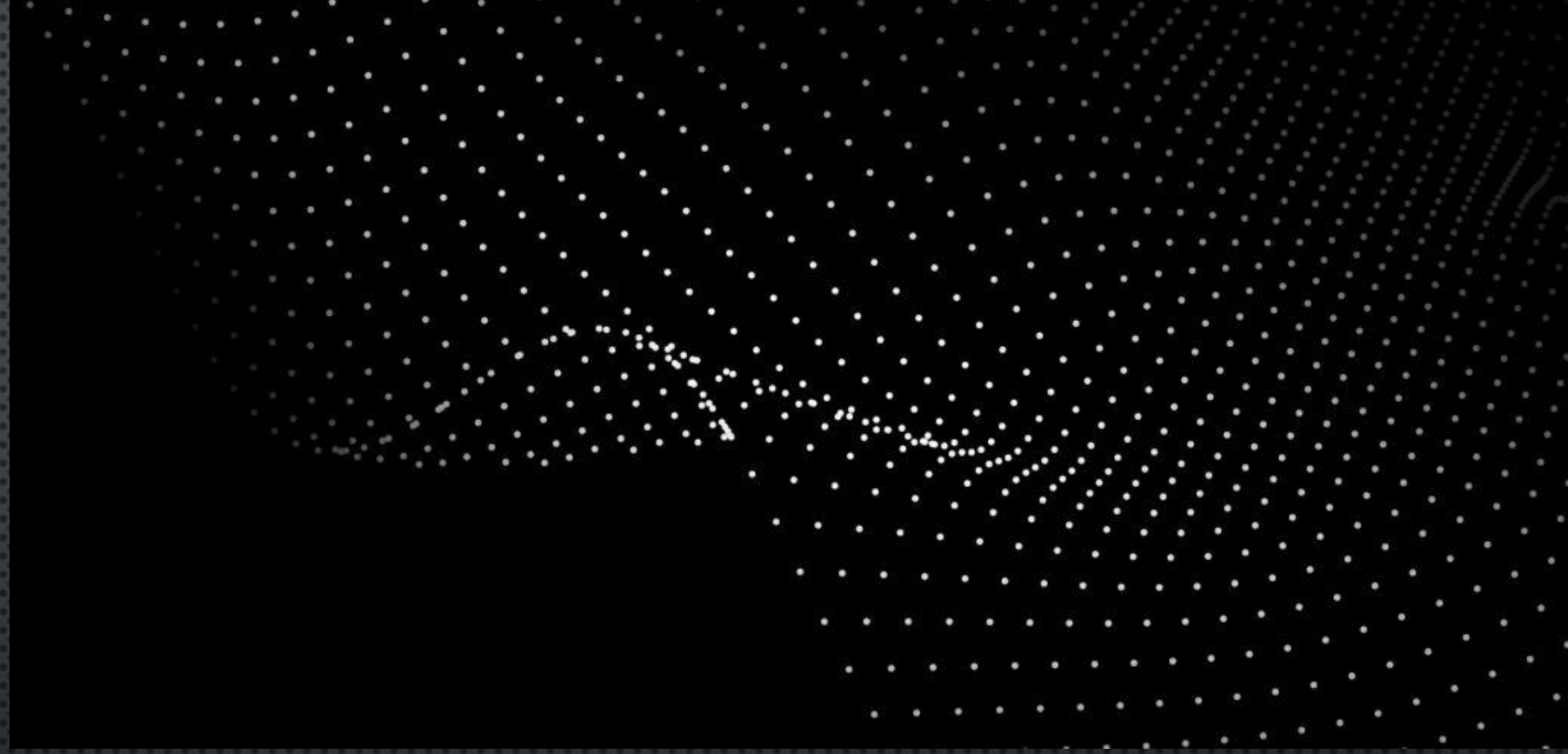
Understanding the relationship between psychosis spectrum disorders and language behavior

Understanding the relationship between psychosis spectrum disorders and language behavior

- Bipolar disorder (BD)
  - Recurrent episodes of mania, hypomania, and depression.
- Schizophrenia (SZ)
  - Symptoms: delusions, hallucinations, disorganized behavior.
- Schizoaffective disorder (SAD)
  - Mixed psychotic and affective symptoms.
- Cognitive impairment
  - Common in psychotic disorders.
  - May influence language use.
- Mental health and communication
  - Analysis of linguistic data provides insights into the relationship between linguistic factors and psychological aspects.

Research Gap

- **Limited Analysis:**
  - Only one research work compares linguistic features of text produced by SZ, SAD, and BD (Voleti et al., 2019).
  - However, this analysis does not distinguish between patients with schizophrenia or schizoaffective disorder.
- **Automatic Discrimination:**
  - Automatic discrimination among these patient groups based on linguistic features has not yet been addressed in the literature.
- **Hungarian Population Studies:**
  - Several papers have examined Hungarian patients with SZ, SAD, and BD (e.g., Kéri et al., 2001; Réthelyi et al., 2010; Inczédy et al., 2010; Kocsis et al., 2016; Döme et al., 2005; Kárpáti et al., 2018).
  - However, no study has systematically analyzed the linguistic features of these disorders in the Hungarian population.



Corpus compilation

Data collection

- Recorded by the Prevention of Mental Illnesses Interdisciplinary Research Group, University of Szeged, led by István Szendi
- Data collection approved by the Ethics Committee of the University of Szeged, following the Declaration of Helsinki
- Written informed consent and official permission obtained
- Spontaneous speech reflects language specificities more accurately than planned speech
- The database includes 458 monologues from 77 subjects

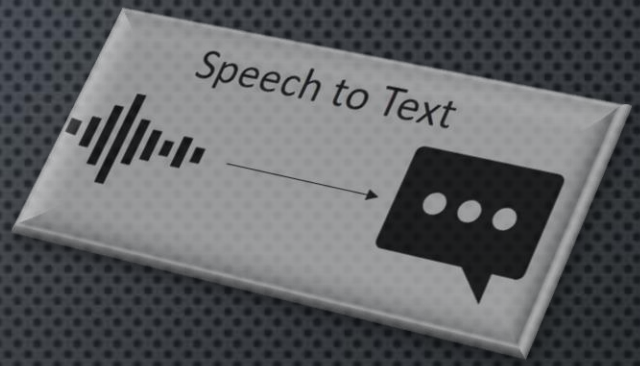
	Groups				
	Control	**SZ**	**SAD**	**BD**	**all**
Number of participants	21	27	14	15	77
Number of texts	126	162	84	86	458
Age; M(SD)	36.42(10.49)	38.80(10.17)	41.43(9.73)	49.08(8.67)	40.63(10.71)
Education; M(SD)	14.76(3.05)	14.19(2.87)	14.54(2.98)	15.83(3.62)	14.73(3.09)
Sex ratio; f:m	8:13	9:18	10:4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

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

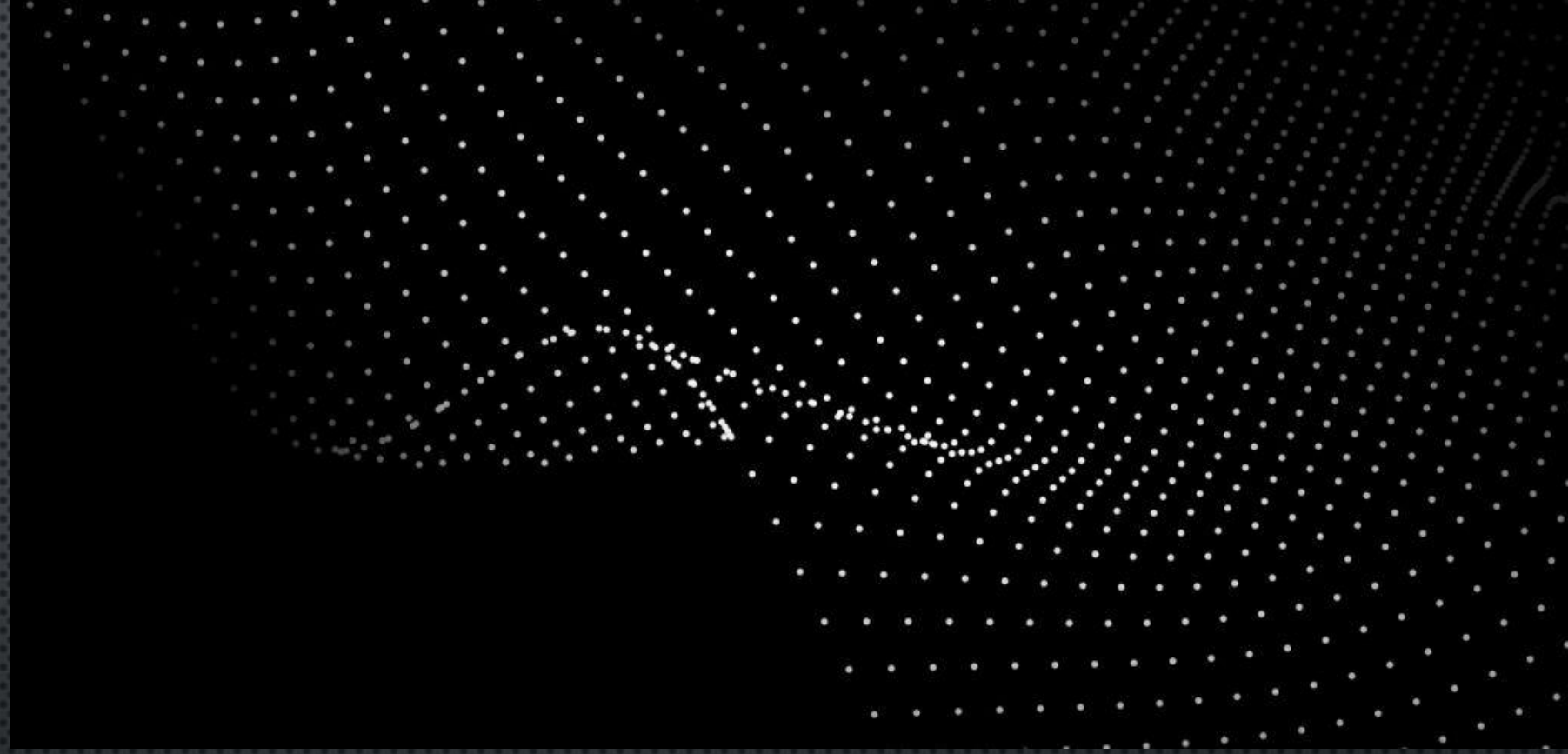
Speech tasks:

Description of Self (DescSelf), Mother (DescMother), and Father (DescFather), Younger Years (YoungSelf and YoungOther), Description of Previous Day (PrevDay)

Transcription process



- Manual transcription of recordings
- No specific software used
- Marked pauses, hesitations etc.
- Transcriptions stored in plain text format (UTF-8)
- Transcribers worked interchangeably on files
- Detailed guideline used for consistency
- Regular quality checks conducted



Corpus processing and analysis

Corpus processing

- We used the magyarlanc toolkit for automatic linguistic analysis, including sentence splitting, tokenization, lemmatization, part-of-speech tagging, and morphological tagging.
- Extracted: 17 basic statistical features and 10 speech-based, 87 morphosyntactic features
- Semantic and pragmatic linguistic features were analyzed, including sentiment and emotion words, discourse markers, and intensifiers.
- A total of 194 linguistic features

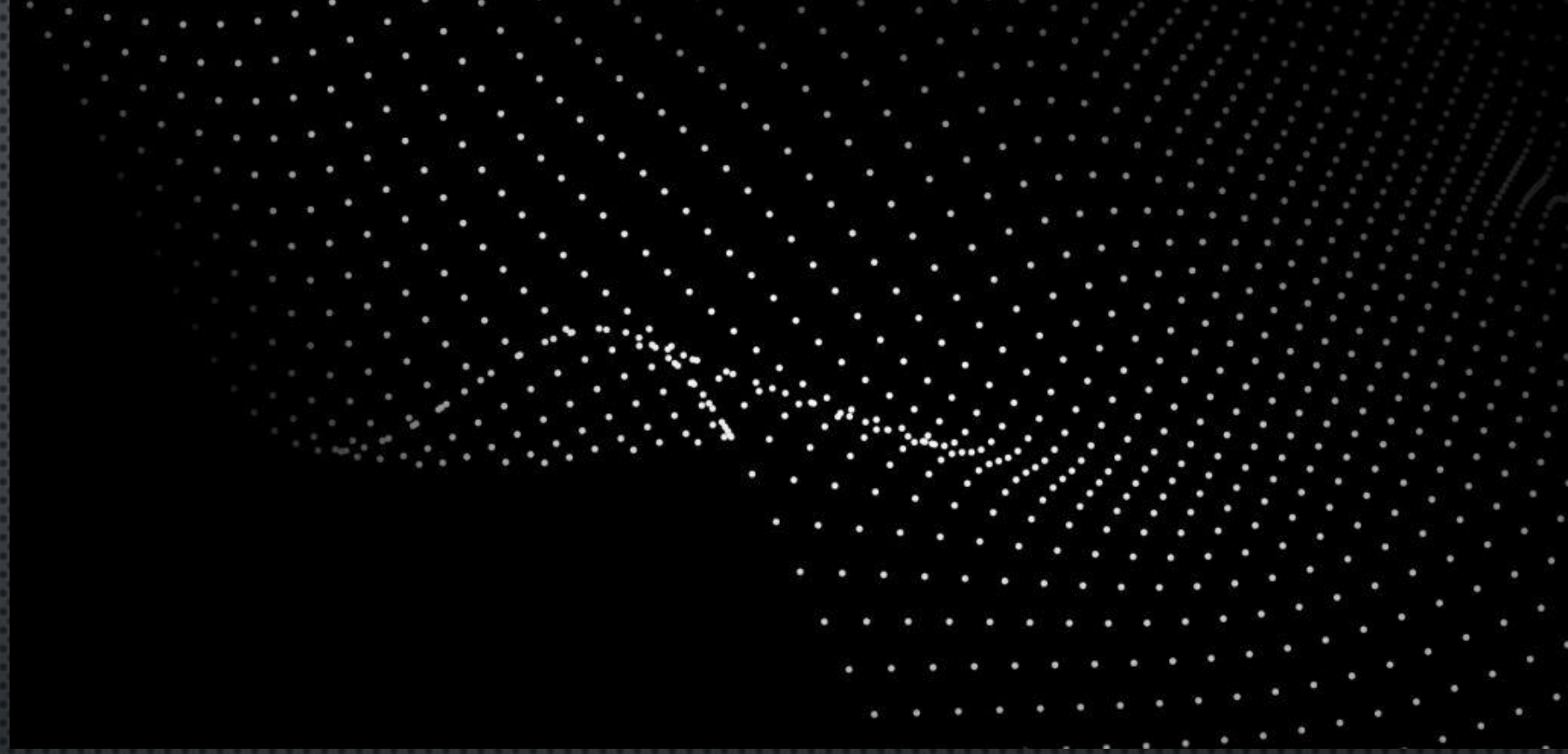
Example Features:

- Basic Statistical: Number of sentences, number and frequency of distinct lemmas compared to the number of words.
- Speech-Based: Number of pauses, hesitations, and filled pauses.
- Morphosyntactic: Count and occurrence rate of various parts-of-speech, and occurrences of third person singular verb forms.
- Semantic and Pragmatic: Number and frequency of sentiment words, emotion words, and discourse markers.

(For a detailed list of features, please refer to the appendix of our paper.)

Statistical analysis and machine learning experiments

- Pairwise t-tests -> evaluate feature usefulness
- Random forest classifier from the WEKA package -> automatic group discrimination
- Ten-fold cross-validation
- Results compared to majority classification baseline method



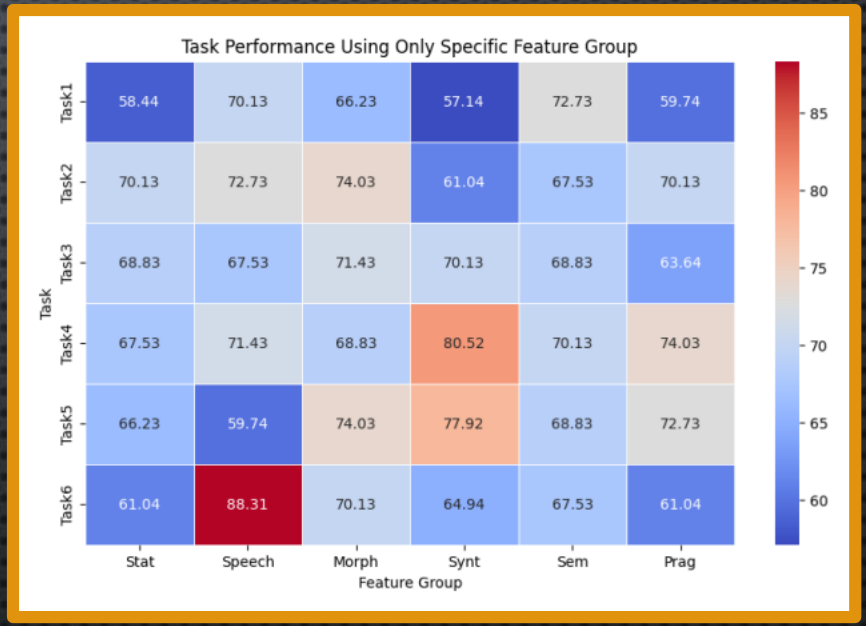
Corpus analysis results

Results of significance tests

- Several significant differences among the four groups:
  - The rate of **nouns and verbs** was significantly different in most sub-corpora.
  - Patients tended to use fewer **pronouns** compared to the control group.
  - Patients tended to use fewer **function** and more **content words** generally.
- When comparing each patient group with the control group, clear patterns of significant differences emerged:
  - The most apparent contrast was seen in BD patients, who had a lower rate of **filled and unfilled pauses** in several tasks.
  - SZ patients had a notably different **POS** distribution compared to the control group.
  - Patients used significantly more **positive words** when discussing someone close to them, but more **negative words** when discussing themselves.
  - BD speakers also used significantly more words related to **sorrow**
  - Healthy controls typically used the most **discourse markers** across all tasks.

Results of machine learning experiments I.

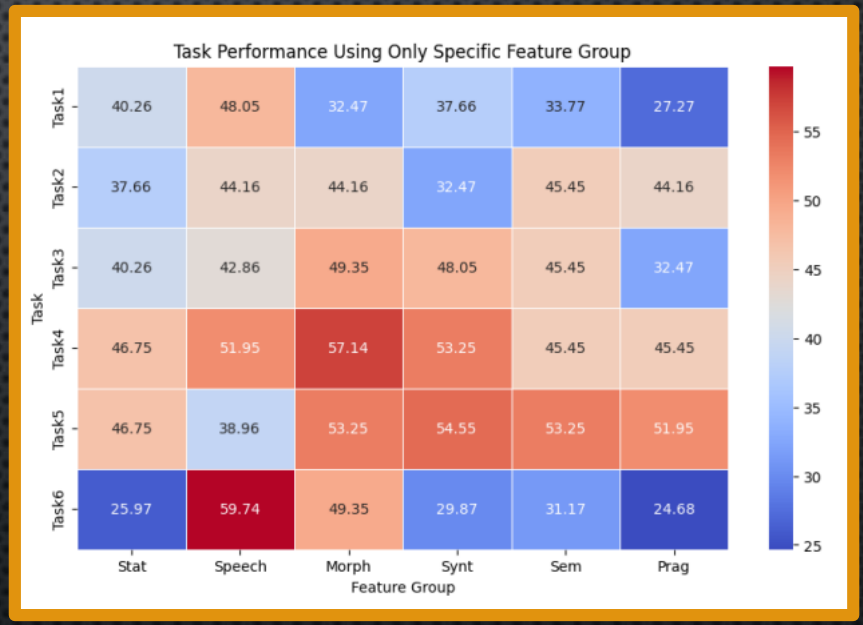
- Overall accuracy: 58.44%, outperforming the baseline.
- Dividing subjects into two groups yielded an accuracy of 72.73%
- Best performance: SZ subjects (F-score: 0.75).
- Speech-based features were the most effective (63.64%).



- Highest efficiencies by task and feature:
  - "PrevDay" with speech-based features (88.31%).
  - "YoungSelf" with syntactic features (80.52%).
  - "YoungOther" with syntactic features (77.92%).

Results of machine learning experiments II.

- When comparing the four groups: Excluding statistical or syntactic features improved accuracy
- Omitting some specific morphological and semantic features increased accuracy

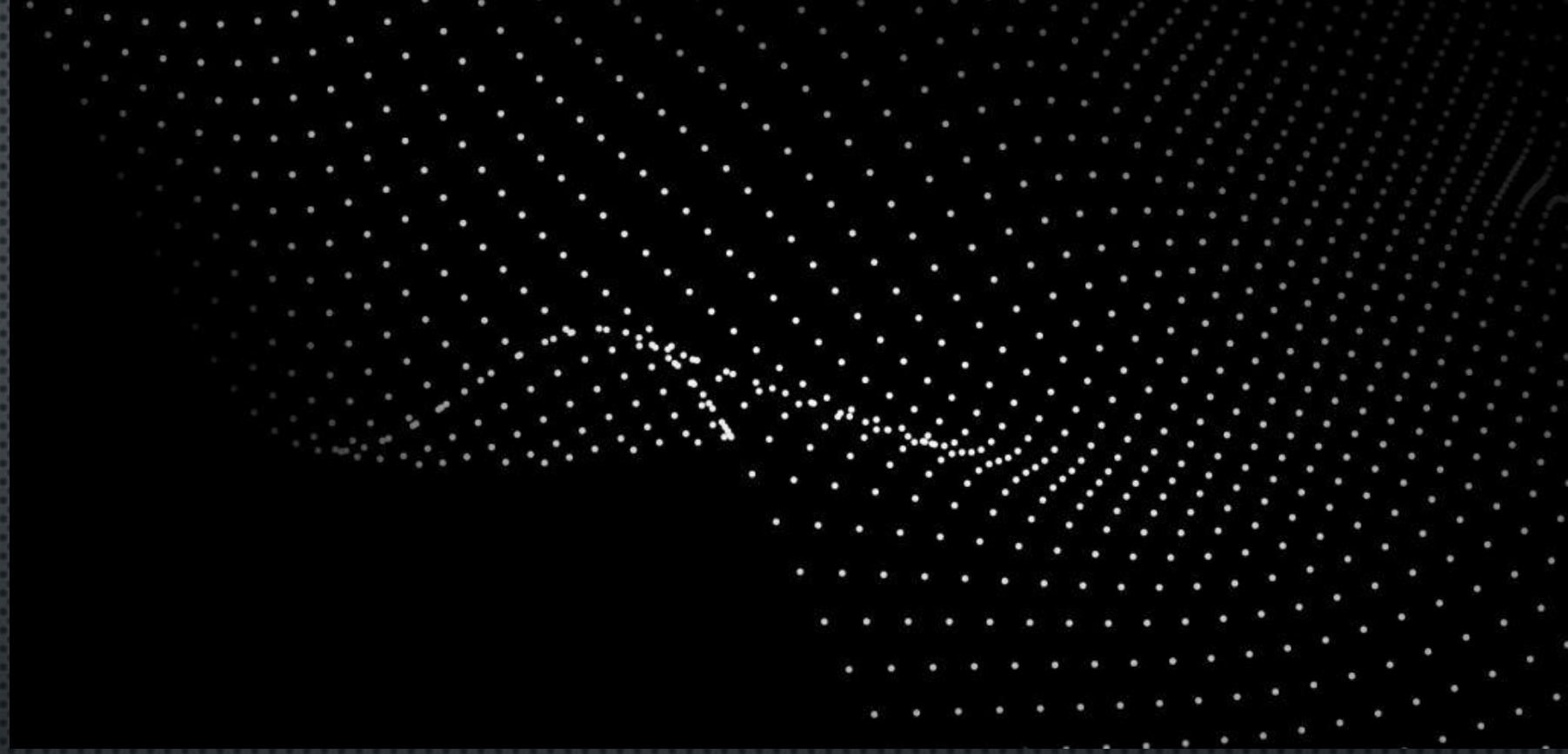


Results of machine learning experiments III.

- In our analysis and machine learning experiments, we also conducted experiments with intensifiers.
- Intensifiers are closely linked to emotion regulation (Athanasiadou, 2007; Strous et al., 2009); There are differences in the use of intensifiers among various mental disorders, such as schizophrenia (Strous et al., 2009, Szabó et al. 2023).
- Two types: Standard-register (non-emotive) intensifiers and Negative emotive intensifiers (NEI) (e.g. *awfully, crazy, damn, brutally* etc.)
- **Machine Learning Experiments:**
    - Excluding NEIs reduced overall accuracy but removing standard-register intensifiers increased accuracy
    - Including NEIs improved machine learning results

Importance of Feature Selection:

- These findings underscore the significance of feature selection.
- They highlight the complex interaction among different feature groups, significantly enhancing automatic classification performance.



Conclusions and future work

Conclusion and future work

- In this study, we analyzed spontaneous speech from Hungarian patients with SZ, SAD, and BD, aiming to identify distinctive linguistic features and improve automatic classification.
- Overall accuracy: 58.44%; Controls from patients: accuracy of 72.73%

Future work:

- Expand linguistic feature set; conduct deeper analysis
- Fine-tune machine learning methods
- Focus on most effective tasks and features
- Experiment with various machine learning algorithms
- Investigate data from other languages
- Release the corpus for research purposes after masking sensitive data.

REFERENCES

- Angeliki Athanasiadou. 2007. On the subjectivity of intensifiers. Language sciences, 29(4):554–565.
- P Döme, Z Rihmer, X Gonda, P Pestality, G Kovács, Z Teleki, and P Mandl. 2005. Cigarette smoking and psychiatric disorders in Hungary. International Journal of Psychiatry in Clinical Practice, 9(2):145–148.
- Gabriella Inczédy-Farkas, Judit Benkovits, Nóra Balogh, Péter Álmos, Beáta Scholtz, Gábor Zahuczky, Zsolt Török, Krisztián Nagy, János Réthelyi, Zoltán Makkos, et al. 2010. Schizobank – the Hungarian national schizophrenia biobank and its role in schizophrenia research. Orvosi Hetilap, 151(35):1403–1408.
- Eszter Kárpáti, Anita Bagi, István Szendi, Lujza Beatrix Tóth, Karolina Janacsek, and Ildikó Hoffmann. 2018. Rekurzió egy szkizoaffektív zavarral élő személy diskurzusaiban– esettanulmány. Iskolakultúra: Pedagógusok Szakmai-Tudományos Folyóirata, 28(5- 6):40–54.
- Szabolcs Kéri, O Kelemen, G Benedek, and Z Janka. 2001. Different trait markers for schizophrenia and bipolar disorder: a neurocognitive approach. Psychological medicine, 31(5):915–922.
- Krisztina Kocsis-Bogár, Zsófia Nemes, and Dóra Perczel-Forintos. 2016. Factorial structure of the Hungarian version of oxford-liverpool inventory of feelings and experiences and its applicability on the schizophrenia-schizotypy continuum. Personality and Individual Differences, 90:130–136.
- János M Réthelyi, Steven C Bakker, Patrícia Polgár, Pál Czobor, Eric Strengman, Péter I Pásztor, René S Kahn, and István Bitter. 2010. Association study of nrg1, dtnbp1, rgs4, g72/g30, and pip5k2a with schizophrenia and symptom severity in a Hungarian sample. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 153(3):792–801.
- Martina Katalin Szabó, Veronika Vincze, Csenge Guba, Bernadett Dam, Adrienn Solymos, Anita Bagi, and István Szendi. 2023. Fokozás szkizofréniában. In XIX. Magyar Számítógépes Nyelvészeti Konferencia, pages 17–32.
- Rohit Voleti, Stephanie Woolridge, Julie M Liss, Melissa Milanovic, Christopher R Bowie, and Visar Berisha. 2019. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. arXiv preprint arXiv:1904.10622.
- Rael D Strous, Moshe Koppel, Jonathan Fine, Smadar Nachliel, Ginette Shaked, and Ari Z Zivotofsky. 2009. Automated characterization and identification of schizophrenia in writing. The Journal of nervous and mental disease, 197(8):585–588.

ACKNOWLEDGE MENTS

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

THANK YOU FOR YOUR ATTENTION.