Key Linguistic Markers for Differentiating Schizophrenia, Schizoaffective, and Bipolar Disorders

Bernadett Dam (1), Martina Katalin Szabó (2,3,4), Veronika Vincze (5), Csenge Guba (1), Adrienn Solymos (6), Anita Bagi (6), István Szendi (7)

(1) SZTE Doctoral School of Linguistics
(2) Tokyo University of Foreign Studies,
(3) HUN-REN TK CSS-RECENS,
(4) SZTE Institute of Informatics,
(5) HUN-REN-SZTE Research Group on Artificial Intelligence
(6) SZTE Department of Hungarian Linguistics
(7) Psychiatry Unit, Kiskunhalas Semmelweis Hospital, University Teaching Hospital

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal

Társadalomtudományi Kutatóközpont tk

東京外国語大学 Tokyo University of Foreign Studies

TM Semmelweis Egyetem Transzlációs Medicina Központ

Topic and research question of the current work

- Analysis of spontaneous speech samples from patients with **schizophrenia** (SZ), **schizoaffective** (SAD) and **bipolar disorders** (BD), and healthy controls (HC)
- Which **linguistic features** are the most helpful in distinguishing between patient groups and controls?
- How effectively can we **automatically differentiate** between these groups based on a set of linguistic features?

Structure of the presentation

1. Brief description of the disorders in question
2. Corpus compilation and processing steps
    - Methods and analysis
3. Results: significance tests and machine learning experiments
4. Discussion of the results

Schizophrenia-bipolar spectrum disorders

- **Schizophrenia** (henceforth SZ): characterized by symptoms of delusions, hallucinations, disorganized speech or behavior, and impaired cognitive ability
- **Schizoaffective disorder** (henceforth SAD): mixed psychotic (hallucinations or delusions) and affective symptoms (mood episodes) → intermediate position between BD and SZ in the schizophrenia-bipolar spectrum
- **Bipolar disorder** (BD): characterized by episodes of mania, hypomania, and alternating or intertwining episodes of depression, possibly psychosis
- Common feature of the three disorders: cognitive deficits, impaired executive function → impaired verbal function

The corpus

- Hungarian database recorded by the Prevention of Mental Illnesses Interdisciplinary Research Group (University of Szeged) led by István Szendi
- **Spontaneous** speech recordings
- Task: describe the previous day
- Manually transcribed
- Part of the HuMenDisCo corpus (Szabó et al. 2023)

	SZ	SAD	BD	Control	All
Participants/Texts	27	14	15	21	77
Age; M(SD)	38.80(10.17)	41.43(9.73)	49.08(8.67)	36.42(10.49)	40.63(10.71)

Table 1: Basic demographic data of the corpus

Corpus processing steps

- Automatic linguistic analysis with magyarlanc (Zsibrita et al., 2013)
- The texts were split into sentences, tokenized, and the tokens were lemmatized and assigned a proper part-of-speech and morphological tag (lemmatization is especially important in the case of morphologically rich languages such as Hungarian)
- 17 basic **statistical features** (e.g., number of sentences, number and frequency of distinct lemmas compared to the number of words)
- 10 **speech-based features** (e.g., number of pauses and hesitations)
- 87 **morphosyntactic features** (e.g., parts-of-speech tags, number and frequency of superlative adjectives)
- Statistical analysis: pairwise t-tests for each feature and transcript
- Automatic classification: a random forest classifier of the WEKA package (Hall et al., 2009)

Results of the significance tests

- The ratio of **unknown words** (with an “unknown” POS tag) significantly higher in all patient groups than in the HC group
- SZ: fewest amount of sentences and tokens, BD: highest amount
- Unfilled **pauses** are significantly different in all patient groups compared to HC: SZ having a higher and BD and SAD having a lower rate
- SZ and SAD: somewhat more **function words** than HC and BD
- BD: highest rate of conjunctions → more complex sentences
- SZ: significantly lower rate of proper nouns than HC and SAD

Results of the machine learning experiments

1. Comparing all four groups

- Overall accuracy was **50.65%**, outperforming the baseline (35.06%)
- Most useful feature set: **speech-based** (59.74%)
- Excluding one feature set: best result without statistical features (54.55%)

Feature type	Only	Without
Statistical	25.974	55.8442
Speech-based	59.7403	44.1558
Morphological	49.3506	54.5455
Syntactic	29.8701	55.8442
All	50.6494	50.6494

Table 2: Results of ablation analysis

Results of the machine learning experiments

2. Comparing two groups: HC vs. patient groups

- Overall accuracy was **75.32%**
- Most valuable feature set: **speech-based** (88.31%)
- Excluding one feature set: best result without syntactic (77.92%)

Feature type	Only	Without
Statistical	61,039	76,6234
Speech-based	88,3117	68,8312
Morphological	70,1299	75,3247
Syntactic	64,9351	77,9221
All	75,32474	75,3247

Table 3: Results of ablation analysis (HC vs. patient groups)

Discussion I: Comparison between HCs and patients

- Overall accuracy of 75.32%
- **Speech-based** feature set: most valuable (accuracy of 88.31%); better than all the features together
- When omitting **syntactic** features → the model performed notably better than in any other cases (77,92%) → syntactic information may not be as crucial or relevant for the specific task; it underscores the importance of feature selection
- Importance of the **“unknown” POS tag** (significantly lower rate in HC) → neologisms or disfluencies

Discussion II: Comparison of all four groups

- Overall accuracy of 50.65%
- Effectiveness of **speech-based** features again (accuracy of 59.74%)
  - Similarities between SZ and HC (higher rate of pauses and hesitations) and between BD and SAD (lower rate of pauses and hesitations)
- When omitting **statistical** features, the result outperformed the overall accuracy (54.54%)

Conclusions and limitations

- Analysis of spontaneous speech samples from patients with **schizophrenia** (SZ), **schizoaffective** (SAD) and **bipolar disorders** (BD), as well as healthy **controls** (HC) in Hungarian
- Rich linguistic feature set
- **Automatic classification** between said groups based on linguistic differences
- **Speech-based** features proved to be the most effective in both the classification tasks
- We can get meaningful results even without deep linguistic analysis
- Some of the results of **the ablation analysis** underscored the importance of feature selection
- The algorithm is more successful at distinguishing patients from HC
- Limitation: smaller sample size

References

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Szabó, M. K., Vincze, V., Guba, Cs., Dam, B., Solymos, A., Bagi, A., & Szendi, I. (2023b). HuMenDisCo: A Hungarian speech corpus of schizophrenia, schizoaffective and bipolar disorders. *Language Resources and Evaluation*. Paper submitted.

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

Thank you for your attention!

The research work of Martina Katalin Szabó was funded by the Hungarian Research Fund of the National Research, Development and Innovation Office of Hungary (NKFIH, grant number: PD 132312) and the International Research Fellowship Program of Japan Society for the Promotion of Science (JSPS, Postdoctoral Fellowships for Research in Japan (Standard)) and by the OTKA Postdoctoral Research Grant. This research project was partly supported by the Centre for Social Sciences, the Eötvös Loránd Research Network, the Hungarian Academy of Sciences Centre of Excellence.