Title

Analysis of Spontaneous Speech in Parkinson's Disease by Natural Language Processing

Key Points

- Parkinson's disease (PD) patients have a variety of language problems, but there have been few reports on free speech text analysis and abnormalities.
- The number of parts-of-speech in a sentence spoken by PD patients was found to be lower than that of healthy controls, and PD patients had a higher percentage of verbs and a lower percentage of nouns and fillers compared to healthy controls.
- The extracted linguistic features discriminated the group of Parkinson's patients and healthy subjects with an accuracy of more than 80%.
- Natural language processing appears to be an effective tool to analyze the language of PD patients and to diagnose PD.

Summary

A research group headed by Prof. Masahisa Katsuno, Department of Neurology, Nagoya University Graduate School of Medicine, Principal Investigator, and Visiting Researcher Katsunori Yokoi, Department of Neurology, National Center for Geriatrics and Gerontology, in collaboration with Associate Professor Yurie Iribe, School of Information Science and Technology, Aichi Prefectural University, and Professor Norihide Kitaoka, Department of Computer Science and Engineering, Toyohashi University of Technology, conducted research on linguistic alteration in patients with Parkinson's disease (PD). Using natural language processing technology, we investigated the characteristics of PD patients' conversations and the possibility of diagnosing Parkinson's disease from free conversational texts. This work was published online in Parkinsonism & Related Disorders on May 12, 2023.

Patients with PD experience a variety of speech-related problems, including impaired speech production and language impairment. To elucidate the pathophysiological mechanisms of language changes in PD, Professor Katsuno and his research team used natural language processing technology to compare the speech of cognitively normal patients with that of healthy controls. We recorded the conversations of 53 cognitively normal PD patients and 53 healthy controls and used natural language processing to evaluate the

spontaneous conversational text. Then, we used machine learning algorithms to identify the characteristics of each group's conversation. For this analysis, 37 features focused on part-of-speech and syntactic complexity were used as evaluation items. Using a support vector machine (SVM), we used a 10-part cross-validation method to narrow down which of these features were effective in identifying PD patients' conversations, and we also tested the identification rate of each group.

The analysis revealed, first, that PD patients had fewer parts-of-speech per sentence than the healthy control group. Second, we found that PD patients had a higher percentage of verbs in total conversational sentences, a higher variance, a measure of data variability, for case particles, a higher percentage of verbs per sentence, and a lower percentage of common nouns, proper nouns, and fillers per sentence compared to the healthy control group. When we attempted to identify PD patients or healthy controls based on these conversational changes, the identification rate was more than 80%.

The results of this study suggest that even in the absence of cognitive decline, the conversations of PD patients are altered from those of healthy subjects. The results also suggest that natural language processing can be used to understand the characteristics of PD patients' conversations and to identify PD patients.

Research Background

PD is the second most common neurodegenerative disease after Alzheimer's disease. It is a relatively common disease, and its prevalence increases with age. PD is characterized by motor symptoms, such as slow movements, muscle rigidity, and resting tremor, and non-motor symptoms, such as cognitive impairment, mental impairment, sleep disturbance, autonomic dysfunction, and sensory disturbance.

Communicative changes are also common in PD. These are attributed to a variety of factors, including speech, prosodic changes, pronunciation and articulation changes, language changes, discourse management and pragmatics, and psychosocial influences. Studies have shown that more than 90% of PD patients experience some form of language impairment.

The purpose of this study was to analyze conversational sentences using natural language processing to clarify the pathophysiological mechanisms of speech transformation in PD patients without cognitive dysfunction.

Research Results

A total of 73 PD patients and 54 healthy subjects were recruited between April 2012 and March 2020. Of these, 17 PD patients and 1 healthy subject were

excluded due to insufficient data, as were 3 patients who were ultimately diagnosed with Lewy body dementia. As a result, 53 PD patients (24 males and 39 females) and 53 healthy subjects (24 males and 39 females) were included in the analysis. No significant differences were found between the groups with respect to age, sex, years of education, and MoCA-J scores.

There were no significant differences between PD and healthy subjects on the verbal fluency task, saying as many words beginning with "ka" as possible in one minute, and the semantic fluency task, saying as many animal names as possible in one minute, but the number of parts of speech was significantly lower in the PD group than in the healthy control group (Table 1). There were no significant group differences on the number of sentences.

Table 1. Linguistic characteristics of each group

	Healthy control (n = 53)	PD patients with normal cognitive (n = 53)	p-value
Phonemic fluency task	12.40 ± 4.08 (3 - 22)	11.61 ± 3.60 (5 - 21)	0.55
Semantic fluency task	19.88 ± 5.39 (9 - 33)	19.02 ± 4.72 (8 - 30)	0.678
Number of morphemes	342.41 ± 290.38 (135 - 2056)	225.73 ± 112.54 (94 - 602)	0.007
Number of sentences	14.98 ± 9.13 (6 - 54)	13.28 ± 6.45 (2 - 34)	0.271

Next, we conducted four trials using the Wrapper method to select features to discriminate between the PD and healthy subject groups. Ten Crossover results showed that the third trial had the highest F value. Sensitivity, specificity, positive predictive value, and negative predictive value all exceeded 0.83 for the third trial (Table 2).
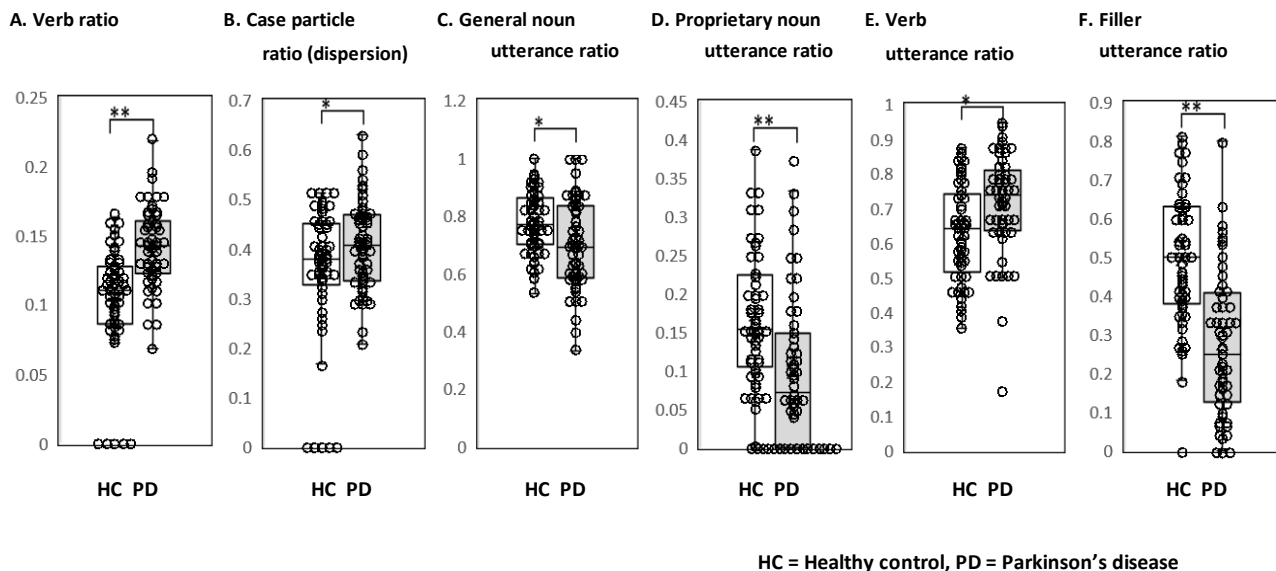
Table 2. Diagnostic accuracy of each group identification

	Disease Positive(PD)	Disease Negative(HC)	Sum of judgments of NLP	Predictive Values
NLP Positive(PD)	44	8	52	PPV = 0.833
NLP Negative(HC)	9	45	54	NPV = 0.846
Total diseases	53	53		
	Sensitivity = 0.830	Specificity = 0.849		

NLP = natural language processing, PPV = positive predictive value, NPV = negative predictive value

In the above analysis, six features were selected as important cues to distinguish PD patients from healthy controls: verb rate, case particle variance, common nouns, proper nouns, verbs, and fillers per sentence. Analysis of the selected features revealed significant differences between the PD group and the healthy group on all six items (Figure 1).

Figure 1. Differences between PD patients and healthy controls in each characteristic

A. Verb ratio B. Case particle ratio (dispersion) C. General noun utterance ratio D. Proprietary noun utterance ratio E. Verb utterance ratio F. Filler utterance ratio
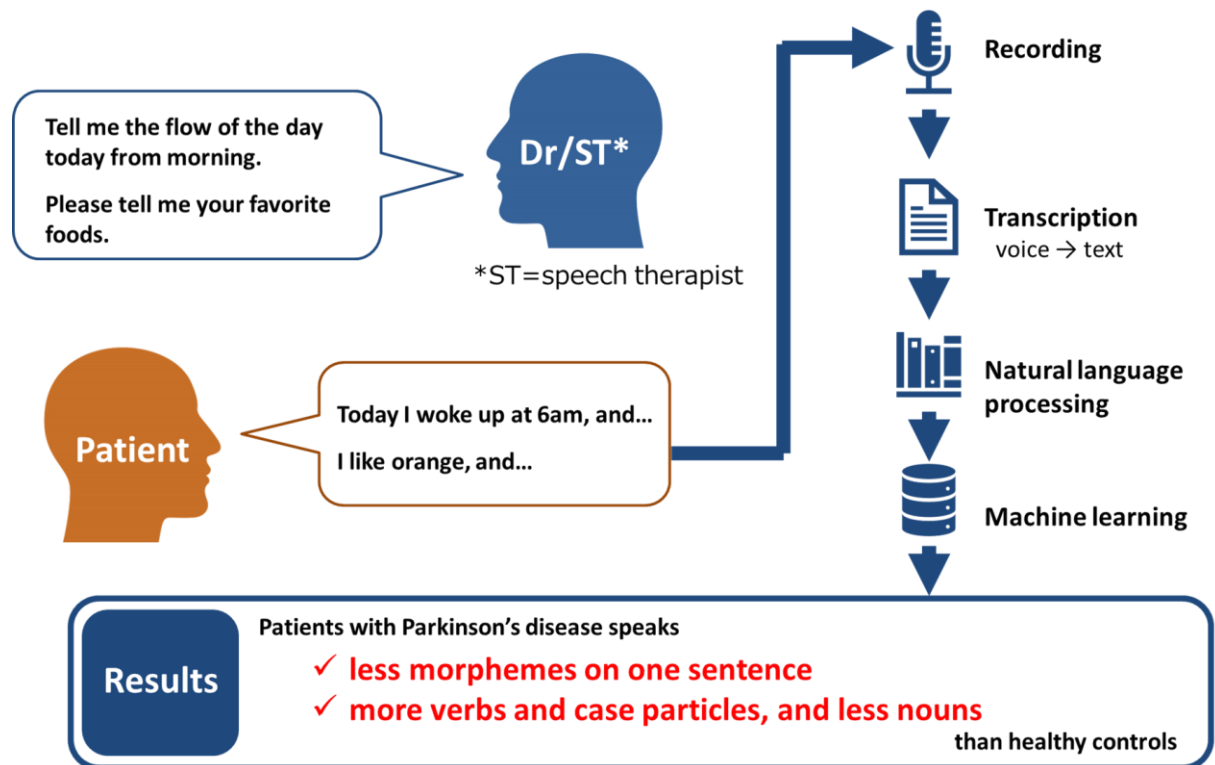
HC = Healthy control, PD = Parkinson's disease

Research Summary and Future Perspective

The results of this study indicate that there are differences in the content of conversations between PD patients without cognitive impairment and healthy controls. Specifically, it was found that compared to healthy subjects, PD patients' conversations consisted of fewer parts of speech and shorter sentences spoken in a single sentence, and more dispersion of verbs and case particles and fewer nouns and fillers (Figure 2). By applying this conversational variation, we confirmed that the support vector machine can discriminate between PD and healthy controls with more than 80% accuracy. This result suggests the possibility of language analysis using natural language processing, which could be used to diagnose PD.

In the future, we plan to use natural language processing to analyze the conversations of PD patients with cognitive decline and patients with neurodegenerative diseases other than PD, especially Alzheimer's disease.

Figure 2. Natural Language Processing of Conversations of Parkinson's Disease Patients