Linguistic Markers of Cognitive Decline in Multiple Sclerosis

SUMMARY

While cognitive changes due to Multiple Sclerosis (MS) are typically very subtle and include decreased processing speed, memory challenges, and difficulties with communication, these changes can begin to appear before an official diagnosis. This research applies natural language processing to analyze patterns in naturalistic language communication and monitor cognitive changes over time. A public and anonymized Reddit dataset from the r/MultipleSclerosis community was used to compare language use between “patient” and “caregiver” users. Transformer-based computational models measured semantic density, coherence, vocabulary use, and emotional expression. While caregiver language stayed constant over time, patient language demonstrated quantifiable declines in meaning and structure and increased emotional variability. Complementary astrocyte gene expression changes in cognition-related genes suggest these linguistic shifts may also suggest molecular alterations in MS. This study introduces a novel tool in which social media discourse can offer new insights into cognitive function and MS progression.

KEYWORDS

Cognitive Decline, Linguistics, Multiple Sclerosis, Natural Language Processing, Social Media

INTRODUCTION

Early signs of cognitive impairment often go unnoticed, leading individuals to miss timely interventions that could mitigate disease progression. In fact, in Multiple Sclerosis (MS), an autoimmune disease, subtle cognitive symptoms can appear more than a decade before formal diagnosis[1]. However, these early impairments end up leading to rising healthcare burdens, with the United States’ MS mortality rate having a sharp rise since 2018[5]. Therefore, identifying accessible, early markers of cognitive decline is essential to prevent this misfortune.

Language research provides a potential solution. Cognitive function directly influences how people communicate. Social media offers an innovative resource to understand language usage, as there is an increase in individuals documenting their personal health and behavioral changes online[6]. Recent studies have revealed that specific language patterns and changes in cognitive social expression on the Reddit platform correlate with early symptoms of cognitive disorders[8].

At the same time, cognitive decline in MS is also associated with biological changes in the cortical grey matter[4]. Gene expression analysis of biomarkers involved could provide additional insight into cognitive decline progression.

This research aims to analyze language changes from users who self-report MS diagnosis on Reddit to identify linguistic features of cognitive decline using natural language processing techniques. Additionally, the study aims to test whether these behavioral findings are parallel to the biological changes observed in brain tissue. Ultimately, the goal is to understand a holistic picture of cognitive symptoms in Multiple Sclerosis through the potential of linguistic analysis of social media.

METHODS

Part 1: Computational Linguistic Analysis

Data Collection

Data was collected from the public r/MultipleSclerosis community on Reddit. Archived posts dated from 2010 to 2019. In total, 190,337 posts were collected. First, all user content was de-identified. Usernames were replaced with pseudonymous, hashed labels. By doing this, the dataset contained no direct personally identifiable information. Both content and posts with fewer than 50 characters were removed. After filtering, 189,919 posts remained.

Patient and Caregiver Classification

Users were classified with regex classification using 30 patterns. Phrases like “my wife has MS” indicated caregivers, while phrases like “I was diagnosed” indicated patients. Users were included if they had ≥10 posts, a ≥3-year posting span, and ≥8 posts with >50 words, resulting in 169 patients and 38 caregivers, averaging 4.2 years on Reddit each.

Feature Extraction
Sentence-BERT was used to convert each post into a vector. These vectors were compared with pre-set vectors that model cognitive impairment patterns. DistilRoBERTa classified emotional tone into the seven categories of fear, anger, disgust, surprise, neutrality, sadness, and joy. The spaCY library was used to analyze sentence structure and calculate vagueness based on the ratio of functional words to content words. Standard linguistic characteristics measured included type-token ratio, semantic density, entropy, sentiment scores, and post length.

Statistical Analysis
Paired t-tests were used for normally distributed features. Wilcoxon signed-rank tests were used for paired data. Mann-Whitney U tests were used to compare patients and caregivers.

Part 2: Gene Expression Analysis
Data Collection
Data was collected from dataset GSE131281 in the GEO database[2]. Six genes were selected based on their roles in MS progression. GFAP[7] and CHI3L1[9] for inflammation, SNAP25 for synaptic function, NEFL[7] for structural integrity, and VCAM1[10] and CXCL16[3] for vascular and immune processes.

Data Analysis
Data was normalized, and log2-fold change was calculated. Two-sample t-tests were utilized to compare the MS and control groups.

RESULTS AND DISCUSSION
MS patients showed significant declines in linguistic features over time, while caregivers from the same online community showed no changes.

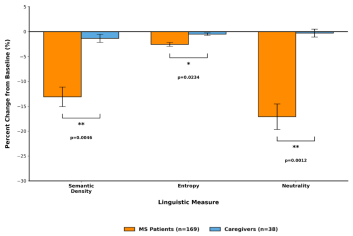

Figure 1. Comparative Analysis of Longitudinal Change Scores. Percent change in linguistic features was compared between MS patients (n=169) and caregiver controls (n=38) from Year 1 to Year 3+. MS patients exhibited significant declines in semantic density (-13.11%, Wilcoxon p=0.003), linguistic entropy (-2.58%, p<0.0001), and emotional neutrality (-17.1%, p<0.001) over a 3-year period. Caregivers showed no significant changes in any metric (semantic density: -1.34%, p>0.05; entropy: -0.5%, p>0.05; neutrality: -0.3%, p>0.05). Brackets indicate that the difference in change magnitude between the two groups is statistically significant (Mann-Whitney U), confirming that this linguistic decline is specific to the MS patient population. Error bars represent mean ± SD.

Reduced semantic density indicates a loss of meaningful content, and decreased entropy suggests reduced linguistic complexity, consistent with declining cognitive function. Emotional analysis showed decreased neutrality along with increases in both joy and sadness, indicating a shift toward extreme emotional expression. Clustering analysis revealed that emotionally focused posts had the highest vagueness, suggesting that cognitively demanding topics are associated with less precise language

Table 1. Differential expression of neuroinflammatory and synaptic markers in MS cortical grey matter. Expression values (log2 intensity) from the GEO dataset GSE131281 comparing MS (n=64) and control (n=42) cortical grey matter samples. Results represent independent two-sample t-tests.

Biomarker	Significance of Differential Expression (p-value)
GFAP (↑)	$p < 0.01$
SNAP25 (↑)	$p < 0.05$
NEFL (↓)	$p < 0.05$
VCAM1	n.s.
CXCL16	n.s.
CHI3L1	n.s.

The gene expression results contribute a biological foundation for these patterns. Increased GFAP suggests neuroinflammation. Decreased NEFL implies structural neuronal damage. Increased SNAP25 indicates possible compensatory synaptic activity. These findings support that Multiple Sclerosis progression causes damage in brain regions and biomarkers involved in cognition and language, which helps explain the linguistic decline. Limitations include that gene expression analyses are derived from cortical grey matter, which may not fully represent other brain systems affected by Multiple Sclerosis.

CONCLUSIONS

This study shows that cognitive decline in Multiple Sclerosis can be understood through changes in natural language over time on social media. Patients demonstrated reduced semantic clarity, linguistic complexity, and shifts in emotional expression. Caregivers from the same online space showed no comparable changes. Therefore, findings were specific to Multiple Sclerosis progression.

The addition of gene expression analysis strengthens these findings. With orthogonal validation, it is clear that neuroinflammation, structural damage, and synaptic changes in brain tissue are involved in cognition in Multiple Sclerosis and may be influencing language trends as well.

Accordingly, these results emphasize the potential of using social media as a passive and scalable method for early detection of cognitive impairment in Multiple Sclerosis. Since language analysis can help identify cognitive decline within three years, earlier than traditional clinical methods, this approach can allow for timely intervention. Future work will focus on comparing Multiple Sclerosis linguistic patterns with patterns of other neurodegenerative diseases across Reddit communities.

REFERENCES

1. Freeborn, Jessica. “Subtle Signs of Multiple Sclerosis May Appear Years before Onset.” Medicalnewstoday.com, Medical News Today, 13 Aug. 2025, www.medicalnewstoday.com/articles/subtle-signs-of-multiple-sclerosis-may-appear-years-before-onset#MS-may-start-earlier-than-previously-thought.
2. GSE131281: Cortical Grey Matter Samples from MS Patients and Healthy Controls. Gene Expression Omnibus, NCBI. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131281
3. Holmøy, T., Løken-Amsrud, K. I., Bakke, S. J., Beiske, A. G., Bjerve, K. S., Hovdal, H., Lilleås, F., Midgard, R., Pedersen, T., Šaltytė Benth, J., Torkildsen, Ø., Wergeland, S., Myhr, K.-M., Michelsen, A. E., Aukrust, P., & Ueland, T. (2013). Inflammation Markers in Multiple Sclerosis: CXCL16 Reflects and May Also Predict Disease Activity. PLoS ONE, 8(9), e75021. https://doi.org/10.1371/journal.pone.0075021
4. Messina, S., & Patti, F. (2014). Gray Matters in Multiple Sclerosis: Cognitive Impairment and Structural MRI. Multiple Sclerosis International, 2014, 1–9. https://doi.org/10.1155/2014/609694
5. Nadeem, Zain Ali, et al. “Trends in Mortality due to Multiple Sclerosis in the United States: A Retrospective Analysis from 1999 to 2020.” Multiple Sclerosis and Related Disorders, vol. 89, 8 July 2024, p. 105765, www.sciencedirect.com/science/article/abs/pii/S2211034824003420, https://doi.org/10.1016/j.msard.2024.105765.
6. Naslund, J. A., Bondre, A., Torous, J., & Aschbrenner, K. A. (2020). Social media and mental health: Benefits, risks, and opportunities for research and practice. Journal of Technology in Behavioral Science, 5(3), 245-257. https://doi.org/10.1007/s41347-020-00134-x
7. Petrou, P., Kassis, I., Levi, Y., Yaghmour, N., Epstein, T., Ginzberg, A., & Karussis, D. (2025). Kinetics of serum NFL and GFAP and changes in cognitive functions, in MS patients treated with repeated administrations of autologous mesenchymal stem cells (MSC-NG01). Journal of Neuroimmunology, 403, 578613. https://doi.org/10.1016/j.jneuroim.2025.578613
8. Seraj, S., Blackburn, K. G., & Pennebaker, J. W. (2021). Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. Proceedings of the National Academy of Sciences, 118(7), e2017154118. https://doi.org/10.1073/pnas.2017154118

9. Talaat, Foraysa, et al. “Chitinase-3-like 1-Protein in CSF: A Novel Biomarker for Progression in Patients with Multiple Sclerosis.” Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, vol. 44, no. 9, 1 Sept. 2023, pp. 3243–3252, pubmed.ncbi.nlm.nih.gov/36988727, https://doi.org/10.1007/s10072-023-06764-2.
10. Yong, H. Y. F., Batty, N. J., Tottenham, I., Koch, M., & Camara-Lemarroy, C. R. (2024). Soluble adhesion molecules: Cognitive worsening biomarkers in primary progressive multiple sclerosis? Journal of Neuroimmunology, 393, 578384. https://doi.org/10.1016/j.jneuroim.2024.578384