Automated Main Concept Analysis in Narrative Speech for Detecting Linguistic Impairments

Seung Hee Lee[1], Sharon Wang[1], Maria Varkanitsa[1], Swathi Kiran[1]

[1]Center for Brain Recovery, Boston University, Boston, MA, USA

1 INTRODUCTION

Discourse Analysis in Neurological Conditions

- Discourse analysis provides insight into language abilities beyond the sentence level, often assessed through storytelling.
- While traditional analyses often focus on microlinguistic features, like mean length of utterance and type-token ratio, they often do not assess macrolinguistic features, such as Main Concept Analysis (MCA)[1] and topic coherence, which are important for comprehensive evaluation of language impairments.
- Macrolinguistic analysis remains largely manual[2-3].

2 STUDY AIMS

Overarching goal: Automate microlinguistic and macrolinguistic discourse analysis for clinical detection of neurogenic communication disorders.

Aim 1: Develop an NLP-based pipeline that automatically extracts main concept (MC), coherence, and sequencing features from retellings of the Cinderella story.

Aim 2: Apply machine learning (ML) to evaluate the diagnostic potential of automated macrolinguistic features in distinguishing healthy controls (HC), people with dementia (PWD), and people with aphasia (PWA).

3 METHODS

Variable	HC (n=113)	PWD (n=94)	PWA (n=102)
Age	67.21 (33-88)	72.41 (58-91)	60.09 (26-88)
Sex (F)	84	53	51
WAB-R AQ (/100)	NA	NA	67.21 (40.5-99.6)
MOCA (/30)	NA	25.44 (14-30)	NA

1. DATA ACQUISITION

Extract Cinderella story retelling audio samples from the public AphasiaBank[4] and DementiaBank[5] corpora

2. AUTOMATED TRANSCRIPTION & SEGMENTATION

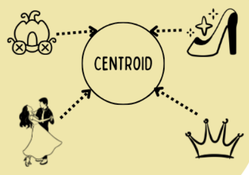
Transcribe audio with Amazon Transcribe and segment utterances by terminators and conjunctions.

3. SEMANTIC EMBEDDINGS

Generate vector representations of utterances and MCs[3] using the all-mpnet-base-v2 Sentence-Transformers model[6]

4. CENTROID & MC MATCH

MC semantic embeddings are averaged to get the centroid. Utterances within 1 SD of the centroid are matched as MCs.



5. AUTOMATED FEATURE EXTRACTION PIPELINE

Extract 23 features, including distance to centroid, coherence error types[7], MC metrics, and sequence score[8].

MACROLINGUISTIC FEATURES

6. ML CLASSIFICATION

Conduct a 3-way multinomial logistic regression classification among PWA, HC, and PWD, as well as one-vs-one pairwise classifications. Identify the most discriminative features based on coefficient magnitudes.

PWA HC PWD

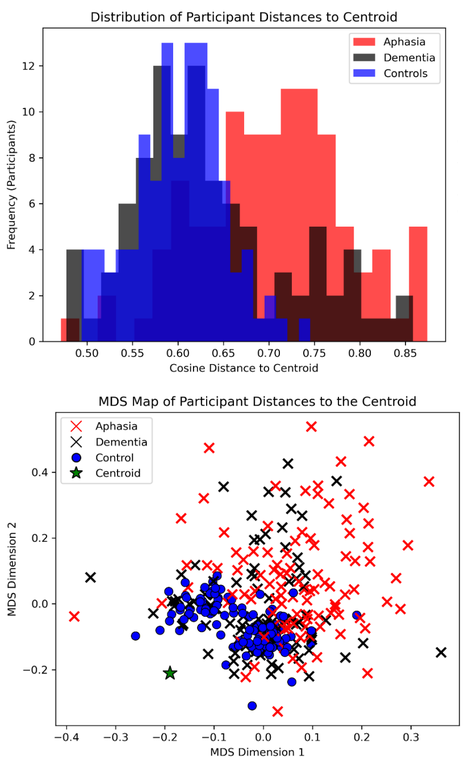
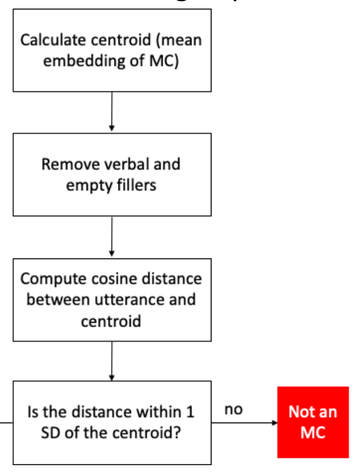
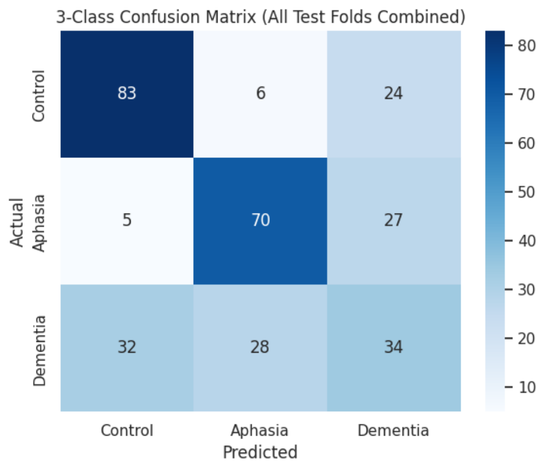


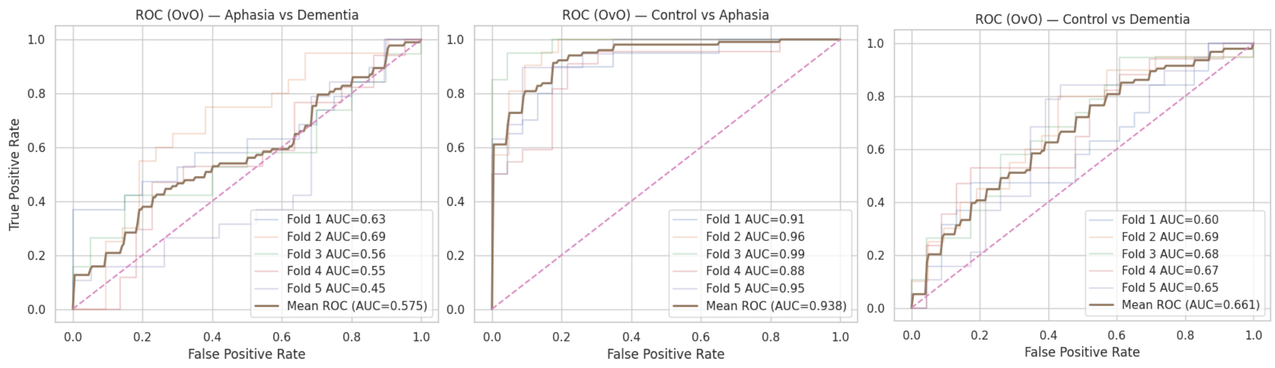
Figure 1. Increasing distance to centroid (representing semantic alignment to MCs) from controls → dementia → aphasia demonstrates its ability to capture narrative informativeness across groups



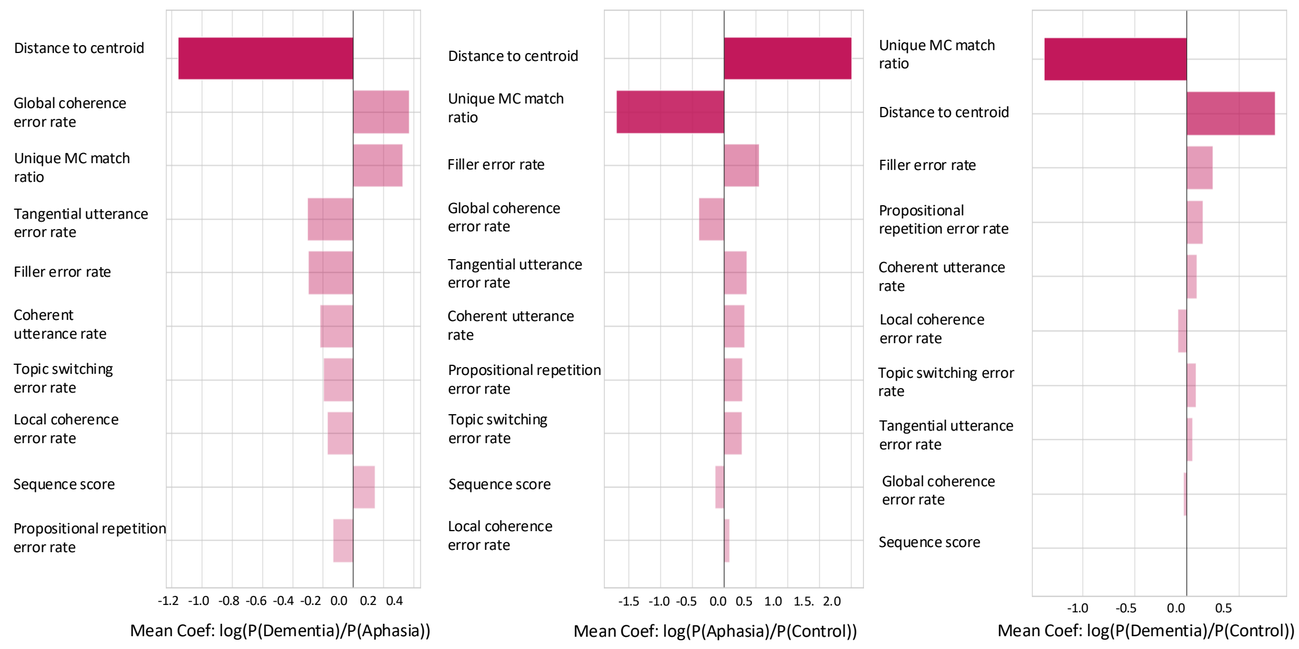
4 RESULTS



- Most accurate classification for HC, followed by PWA, then PWD.
- PWD are misclassified roughly equally as HC and PWA → dementia occupies an intermediate position in macrolinguistic abilities relative to other groups.



- Excellent PWA and HC discrimination (mean AUC = 0.938)
- Moderate performance for PWD vs. HC (mean AUC = 0.661) and poor for PWD vs PWA (mean AUC = 0.575) → macrolinguistic features alone insufficient for dementia differential diagnosis.



- Distance to centroid was the strongest predictor.
- Greater semantic deviation from narrative core characterized both clinical groups vs HC, and aphasia vs dementia → consistent, interpretable marker of discourse-level impairment

5 CONCLUSIONS

Fully automated extraction of macrolinguistic features from narrative speech is feasible.

The automated pipeline captured clinically meaningful group differences.

Automated discourse analysis is a scalable, objective, and clinically relevant approach for detecting linguistic impairment.

References

1 Nicholas, L. E., & Brookshire, R. H. (1995). Presence, Completeness, and Accuracy of Main Concepts in the Connected Speech of Non-Brain-Damaged Adults and Adults With Aphasia. *Journal of Speech, Language, and Hearing Research*, *38*(1), 145–156. https://doi.org/10.1044/jshr.3801.145

1. Leaman, M. C., & Edmonds, L. A. (2021). Measuring Global Coherence in People With Aphasia During Unstructured Conversation. *American Journal of Speech-Language Pathology*, *30*(1S), 359–375. https://doi.org/10.1044/2020_AJSLP-19-00104
2. Richardson, J. D., & Dalton, S. G. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. *Aphasiology*, *30*(1), 45–73. https://doi.org/10.1080/02687038.2015.1057891
3. Macwhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for Studying Discourse. *Aphasiology*, *25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893
4. Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., & Cohen, M. L. (2023). DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses. *American Journal of Speech-Language Pathology*, *32*(2), 426–438. https://doi.org/10.1044/2022_AJSLP-22-00281
5. Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084
6. Marini, A., Andreetta, S., del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology*, *25*(11), 1372–1392. (104641331). https://doi.org/10.1080/02687038.2011.584690
7. Richardson, J. D., Dalton, S. G., Greenslade, K. J., Jacks, A., Haley, K. L., & Adams, J. (2021). Main Concept, Sequencing, and Story Grammar Analyses of Cinderella Narratives in a Large Sample of Persons with Aphasia. *Brain Sciences*, *11*(1), 110. https://doi.org/10.3390/brainsci11010110

Funding

This study has been internally funded by the Boston University Center for Brain Recovery

Contact

Sharon Wang: sharonw2@bu.edu