Automating Paraphasia Identification in Discourse

Alexandra C. Salem[1], Robert C. Gale[1], Gerasimos Fergadiotis[2], Steven Bedrick[1]

[1]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR

[2]Department of Speech and Hearing Sciences, Portland State University, Portland, OR

INTRODUCTION

Previous work focused on automating scoring of picture-naming tests [1]. **Discourse**, however, is harder to analyze automatically because **paraphasias must be identified**.

Advancements in computer hardware (GPUs) have led to the development of **large language models (LLMs)**. Here, we automate paraphasia identification in Cinderella story retellings using a LLM we trained for use on speech-language pathology tasks, called **BORT** (Beyond Orthographically-Restricted Transformers) [3]. We had two research objectives:

1. Develop and demonstrate the utility of a **LLM** for automatically identifying **paraphasias in discourse**.
2. Explore the **impact of clinical characteristics** and **paraphasia type** on model performance.

METHOD

Data consisted of 353 Cinderella story retelling transcripts from 254 people with aphasia (PWA) from the English **AphasiaBank** database [6]. Demographic and clinical data are in Table 1. We filtered paraphasias [7] identified by AphasiaBank, leaving **3,107 paraphasias out of 93,842 total words** across all transcripts.

Table 1. Demographic data of 254 participants at their first session, where available.

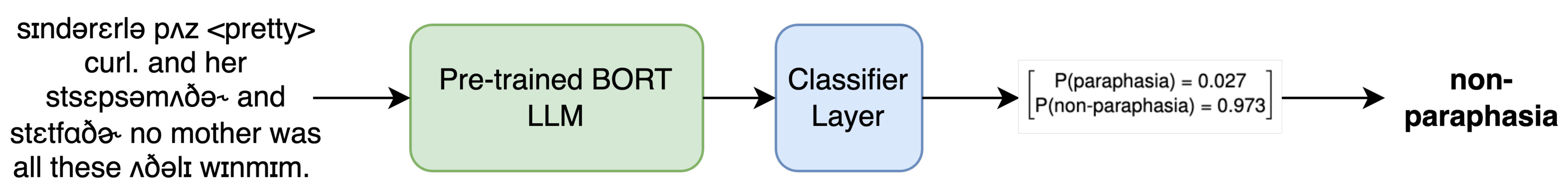
	Age	Years Post Onset	WAB-R AQ	BNT	VNT
***M* (*SD*)**	61.5 (12.4)	5.2 (4.7)	72.1 (17.9)	7.3 (4.5)	14.9 (6.3)
Min - Max	25.6 - 90.7	0.1 - 30.0	10.8 - 99.6	0.0 - 15.0	0.0 - 22.0
Missing (*N*)	3	3	8	13	11

Note. WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient [5]. BNT is the raw score from the Boston Naming Test-Short Form [4]. VNT is the raw score from the Verb Naming Test [2].

We **fine-tuned** BORT to **classify each word** as a paraphasia or non-paraphasia (Fig. 1). After fine-tuning, we used Receiver Operating Characteristic (**ROC**) analysis to determine the optimal threshold for final classification.

We evaluated the models' predictions against the known paraphasias by calculating **sensitivity**, **specificity**, **accuracy, and positive predictive value (PPV)**. We stratified our results by **error type**, aphasia **severity**, **fluent** vs **non-fluent** aphasia, and mean length of utterance in words (**MLUW**). We tested whether differences in accuracy for each stratification were **significant** using two-sided z-tests for independent proportions.

Figure 1. Classifying a sample transcript.



Using our LLM for clinical tasks, we can identify 86.7% of paraphasias in Cinderella story retellings

Take a picture to see a complete write-up and references!

Or go to this link:
alexandrasalem.com
My email: salem@ohsu.edu

Adapted from "Better" poster template: https://osf.io/ayjzg/

RESULTS

A comparison of performance using the **original classification threshold (0.5)** and the **optimal threshold (0.044)** determined from ROC analysis is in Table 2. By turning the threshold down, we were able to capture far more paraphasias and **increase sensitivity**, at the loss of some accuracy and PPV.

Table 2. Results using original classification threshold (0.5) and optimal threshold (0.044).

Test set	Threshold	Sens	Spec	PPV	Acc
All paraphasias	0.5	0.625	0.987	0.685	0.971
All paraphasias	0.044	0.867	0.923	0.278	0.921

Table 3. Breakdown LLM performance (with optimal threshold) by paraphasia type.

Paraphasia type	N paraphasias (%)	LLM Correct (%)
Non-real word (IPA)	1,554	1,547 (0.995)
Real word (orthographic)	1,553	1,147 (0.739)

RESULTS (CONTINUED)

Performance stratified by real words and non-real words is in Table 3. Non-real word paraphasias were more obvious, while **real word paraphasias were more challenging**.

Results stratified by clinical characteristics are in Table 4. **Sensitivity** was **higher** in **more severe** and **non-fluent** participants, and participants with **lower MLUW.**

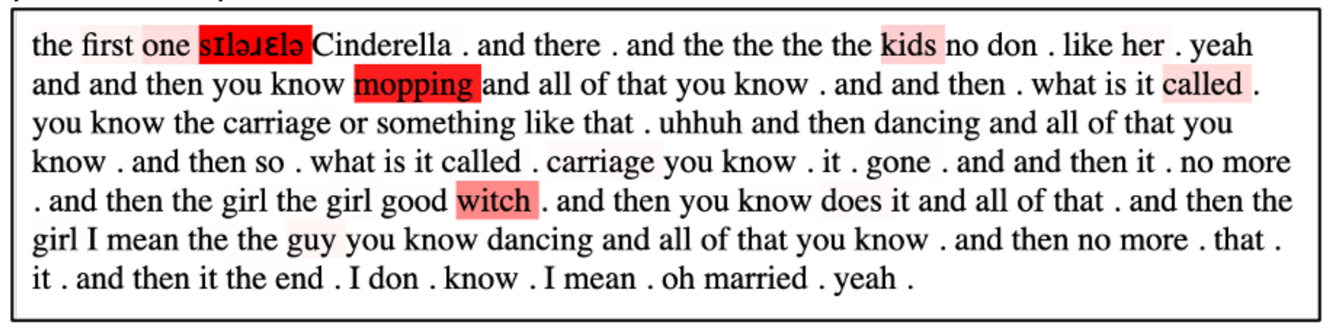
Table 4. Performance (with optimal threshold) across test set stratifications.

Test set	*N* sessions	*N* words	*N* paraphasias	Sens	Spec	PPV	Acc
All paraphasias	353	93,842	3,107	0.867	0.923	0.278	0.921
WAB-R AQ > median (74.05)	172	54,442	1,189	0.818	0.943	0.242	0.940
WAB-R AQ ≤ median (74.05)	172	36,911	1,857	0.896	0.892	0.305	0.892
Fluent participants	252	80,036	2,338	0.853	0.925	0.255	0.923
Non-fluent participants	92	11,317	708	0.907	0.903	0.384	0.903
MLUW > median (5.41)	177	62,633	1,793	0.852	0.928	0.258	0.926
MLUW ≤ median (5.41)	176	31,209	1,314	0.888	0.913	0.310	0.912

Note. 9 out of 353 total sessions had unavailable WAB-R results and were excluded just from analyses involving WAB-R scores. All differences in accuracy were significant (p < 0.001).

An example transcript is in Fig. 2. **Darker highlight** represents **higher prediction probability**. *first*, *one*, *sɪləɹɛlə*, *kids*, *mopping*, *called*, *witch* have prediction probabilities >0.044 and are classified as paraphasias. Actual paraphasias are ***sɪləɹɛlə*** and ***witch***.

Figure 2. Heat map showing prediction probability levels for each word in a sample transcript.



DISCUSSION

This work demonstrates the utility of developing a clinical tool for automatic identification of potential paraphasias in discourse. It is limited by **requiring transcription**, but advances in automatic speech recognition raise a solution to that problem. These findings take us closer to **automatic aphasic discourse analysis**.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of NIH/NIDCD award #R01DC015999 (PIs: Bedrick & Fergadiotis).

References

1. Casilio, M., Fergadiotis, G., Salem, A. C., Gale, R. C., McKinney-Bock, K., & Bedrick, S. (2023). ParAlg: A Paraphasia Algorithm for Multinomial Classification of Picture Naming Errors. *Journal of Speech, Language, and Hearing Research*, 1–21. https://doi.org/10.1044/2022_JSLHR-22-00255
2. Cho-Reyes, S., & Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). Aphasiology, 26(10), 1250– 1277. https://doi.org/10.1080/02687038.2012.693584
3. Gale, R., Salem, A., Fergadiotis, G., & Bedrick, S. (2023). Mixed Orthographic/Phonemic Language Modeling: Beyond Orthographically Restricted Transformers (BORT). *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, 212–225. https://doi.org/10.18653/v1/2023.repl4nlp-1.18
4. Kaplan, E., Goodglass, H., & Weintraub, S. (Eds.). (2001). Boston naming test (2. ed). Lippincott, Williams & Wilkins.
5. Kertesz, A. (2012). Western Aphasia Battery—Revised [Data set]. American Psychological Association. https://doi.org/10.1037/t15168-000
6. MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk (3rd ed.). Lawrence Erlbaum Associates.
7. Salem, A. C., Gale, R. C., Fleegle, M., Fergadiotis, G., & Bedrick, S. (2023). Automating Intended Target Identification for Paraphasias in Discourse Using a Large Language Model. *Journal of Speech, Language, and Hearing Research*, 1–18. https://doi.org/10.1044/2023_JSLHR-23-00121