Automating Intended Target Prediction for Paraphasias in Discourse Using a Large Language Model

Alexandra Salem, Mikala Fleegle, Robert Gale, Gerasimos Fergadiotis, Steven Bedrick

INTRODUCTION

Previous work focused on automating scoring of picture-naming tests [2], [9], [12]. **Discourse**, however, is harder to analyze because we do not know the intended target words.

Advancements in computer hardware (GPUs) have led to the development of **large language models (LLMs)**. Here, we automate predicting the intended targets of paraphasias in Cinderella story retellings using a LLM called Big Bird [13], [14].

We had two research objectives:

1. Assess the **feasibility** of applying a modern LLM to this task and establish a performance baseline
2. Explore the impact of **clinical factors** and intended **target ambiguity** on model performance.

METHOD

Data consisted of 332 Cinderella story retelling transcripts from 240 people with aphasia (PWA) from the English **AphasiaBank** database [7]. These sessions contained 2,489 paraphasias for which annotators obtained **76.8% average agreement** on target identification. Demographic and clinical data are shown in Table 1.

To prepare the transcripts, we replaced paraphasias with a "blank" token:

> ... and then and and she put her foot in the. and she rode off with the **[MASK]**. Cinderella was pretty girl. ...

We **fine-tuned** the model to **fill in the blank**. We compared this performance with the **pre-trained** LLM without fine-tuning. We used cross-validation to prevent overfitting.

We tested the models' predictions against our human-identified paraphasia targets by calculating **accuracy**. We stratified our results by Western Aphasia Battery-Revised (WAB-R) [6] **severity**, **fluent** vs **non-fluent** aphasia, whether humans had perfect **agreement** in target identification, and human **confidence** in target identification.

Table 1. Demographic data of 240 participants at their first session, where available.

	Age	Years Post Onset	WAB-R AQ	BNT	VNT
***M* (*SD*)**	61.5 (12.5)	5.4 (4.7)	72.8 (17.7)	7.5 (4.5)	15.2 (6.1)
Min - Max	25.6 - 91.7	0.1 - 30.0	10.8 - 99.6	0.0 - 15.0	0.0 - 22.0
Missing (*N*)	23	23	11	31	31

Note. WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient [6]. BNT is the raw score from the Boston Naming Test-Short Form [5]. VNT is the raw score from the Verb Naming Test [1].

Paraphasias in discourse are hard to analyze automatically because the ground truth targets are not readily accessible.

Here, our goal was to **predict intended target words** for paraphasias using a **large language model.**

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

Or go to this link:
alexandrasalem.com/talk/cac-2023/
My email: salem@ohsu.edu

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

ACKNOWLEDGEMENTS

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

RESULTS

Figure 1. Accuracy of pre-trained and fine-tuned LLMs matching the human-identified target within top 1-20 model predictions.

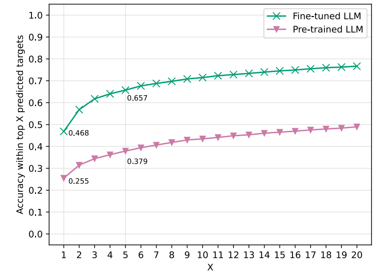

Table 2. Accuracy of pre-trained and fine-tuned LLMs matching the human-identified target, across test sets.

Test set	*N* paraphasias	Pre-trained: Accuracy exact match	Pre-trained: Accuracy within 5	Fine-tuned: Accuracy exact match	Fine-tuned: Accuracy within 5
All paraphasias	2489	0.255	0.379	0.468	0.657
Human agreement = 100%	1244	0.309	0.405	0.595	0.767
Human agreement < 100%	1245	0.201	0.353	0.342	0.548
Human confidence > median (3.3)	1089	0.319	0.419	0.605	0.768
Human confidence ≤ median (3.3)	1400	0.206	0.348	0.362	0.571
WAB-R AQ > median (74.6)	1039	0.294	0.410	0.527	0.703
WAB-R AQ ≤ median (74.6)	1076	0.204	0.325	0.416	0.621
Fluent participants	1666	0.261	0.385	0.487	0.670
Non-fluent participants	449	0.198	0.301	0.412	0.626

Note. 46 out of 332 total sessions had unavailable WAB-R results and were excluded just from analyses involving WAB-R scores.

DISCUSSION

We were able to automatically identify intended targets about half of the time. Performance was significantly **higher** on targets for which **humans had less difficulty**, and on participants with **fluent** or **less severe** aphasia.

These findings take us a step closer to **automatic aphasic discourse analysis**, and open up possibilities for applications that extend beyond assessment (e.g., AAC). In future work, we will incorporate **phonological information**.

References

1. 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
2. Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic Classification of Five Characteristic Types of Paraphasias. American Journal of Speech-Language Pathology, 25(4S). https://doi.org/10.1044/2016_AJSLP-15-0147
3. Goodglass, H., & Wingfield, A. (Eds.). (1997). Anomia: Neuroanatomical and cognitive correlates. Academic Press.
4. Hickin, J., Best, W., Herbert, R., Howard, D., & Osborne, F. (2001). Treatment of Word Retrieval in Aphasia: Generalisation to Conversational Speech. International Journal of Language & Communication Disorders, 36(s1), 13–18. https://doi.org/10.3109/13682820109177851
5. Kaplan, E., Goodglass, H., & Weintraub, S. (Eds.). (2001). *Boston naming test* (2. ed). Lippincott, Williams & Wilkins.
6. Kertesz, A. (2012). Western Aphasia Battery—Revised [Data set]. American Psychological Association. https://doi.org/10.1037/t15168-000
7. MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk (3rd ed.). Lawrence Erlbaum Associates.
8. Mayer, J., & Murray, L. (2003). Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. Aphasiology, 17(5), 481–497. https://doi.org/10.1080/02687030344000148
9. McKinney-Bock, K., & Bedrick, S. (2019). Classification of Semantic Paraphasias: Optimization of a Word Embedding Model. Proceedings of the 3rd Workshop on Evaluating Vector Space Representations For, 52–62. https://doi.org/10.18653/v1/W19-2007
10. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996
11. Pashek, G. V., & Tompkins, C. A. (2002). Context and word class influences on lexical retrieval in aphasia. Aphasiology, 16(3), 261–286. https://doi.org/10.1080/02687040143000573
12. Salem, A. C., Gale, R., Casilio, M., Fleegle, M., Fergadiotis, G., & Bedrick, S. (2022). Refining Semantic Similarity of Paraphasias Using a Contextual Language Model. Journal of Speech, Language, and Hearing Research, 1–15. https://doi.org/10.1044/2022_JSLHR-22-00277
13. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6
14. Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for longer sequences. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.