Byung-Doh Oh

oh.b@nyu.edu
Professional Website
GitHub Profile
Google Scholar

Research Interests

- Computational Linguistics; Psycholinguistics; Cognitive Modeling; Natural Language Processing; Large Language Models; Neural Network Interpretability
- My work aims to advance our understanding of real-time language processing in humans and machines by drawing on methods from computational psycholinguistics and machine learning.

Education

The Ohio State University — Columbus, OH
Ph.D. in Linguistics — 2024
- Dissertation: *Empirical Shortcomings of Transformer-Based Large Language Models as Expectation-Based Models of Human Sentence Processing*
- Advisor: William Schuler
- Committee: Michael White, Micha Elsner, Tal Linzen

Seoul National University — Seoul, Korea
M.A. in English Language Education — 2018
- Thesis: *Exploring English Online Research and Comprehension Strategies of Korean College Students*
- Advisor: Youngsoon So

Seoul National University — Seoul, Korea
B.A. in English Language Education, *summa cum laude* — 2016
- Thesis: *A Study on the Assessment Techniques of Language MOOCs*
- Advisor: Youngsoon So

Awards and Honors

CDS Faculty Fellowship — 2024 – 2026
Center for Data Science, New York University

Selection as DARPA Riser — 2022
Defense Advanced Research Projects Agency

Fulbright Graduate Study Award — 2018 – 2020
Bureau of Educational and Cultural Affairs

Chunjae Education Group Scholarship — 2015
Chunjae Education Group

College of Education Alumni Association Scholarship — 2011
Seoul National University

Employment

Assistant Professor/Faculty Fellow — Sept. 2024 –
Center for Data Science, New York University — New York, NY

Natural Language & Speech Processing Research Intern — May 2021 – Aug. 2021
Tencent AI Lab — Bellevue, WA (remote)
- Mentor: Lifeng Jin

Publications

In revision **Byung-Doh Oh** and William Schuler. The attenuation of frequency effects in large language models. *Journal of Memory and Language*.

2024 **Byung-Doh Oh** and William Schuler. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472.

2024 Christian Clark, **Byung-Doh Oh**, and William Schuler. Linear recency bias during training improves Transformers' fit to reading times. *arXiv preprint*, arXiv:2409.11250.

2024 **Byung-Doh Oh**, Shisen Yue, and William Schuler. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2644–2663.

2023 **Byung-Doh Oh** and William Schuler. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921.

2023 **Byung-Doh Oh** and William Schuler. Token-wise decomposition of autoregressive language model hidden states for analyzing model predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 10105–10117.

2023 **Byung-Doh Oh** and William Schuler. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*:336–350.

2022 **Byung-Doh Oh** and William Schuler. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334.

2022 **Byung-Doh Oh**, Christian Clark, and William Schuler. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*:777963.

2021 Lifeng Jin, **Byung-Doh Oh**, and William Schuler. Character-based PCFG induction for modeling the syntactic acquisition of morphologically rich languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4367–4378.

2021 Evan Jaffe, **Byung-Doh Oh**, and William Schuler. Coreference-aware surprisal predicts brain response. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3351–3356.

2021 **Byung-Doh Oh**, Christian Clark, and William Schuler. Surprisal estimators for human reading times need character models. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3746–3757.

2021 **Byung-Doh Oh** and William Schuler. Contributions of propositional content and syntactic category information in sentence processing. In *Proceedings of the 11th Workshop on Cognitive Modeling and Computational Linguistics*, pages 241–250.

2021 **Byung-Doh Oh**. Team Ohio State at CMCL 2021 shared task: Fine-tuned RoBERTa for eye-tracking data prediction. In *Proceedings of the 11th Workshop on Cognitive Modeling and Computational Linguistics*, pages 97–101.

2019 **Byung-Doh Oh***, Pranav Maneriker*, and Nanjiang Jiang*. THOMAS: The hegemonic OSU morphological analyzer using seq2seq. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 80–86.

2019 Micha Elsner, Andrea D. Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Shuan Karim, David L. King, Luana Lamberti Nunes, **Byung-Doh Oh**, Nathan Rasmussen, Cory Shain, Stephanie Antetomaso, Kendra V. Dickinson, Noah Diewald, Michelle McKenzie, and Symon Stevens-Guille. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, *7*(1):53–98.

2018 **Byung-Doh Oh** and Youngsoon So. Exploring English online research and comprehension strategies of Korean college students. *English Teaching*, *73*(3):53–76.

2017 **Byung-Doh Oh**. Predicting L2 writing proficiency with computational indices based on n-grams. *Foreign Language Education Research*, *21*:1–20.

Invited Talks

2024 **Byung-Doh Oh**. *What can linguistic data tell us about the predictions of large language models?* Invited talk at the Department of Computer Science and Engineering and Graduate School of Artificial Intelligence, POSTECH, Pohang, Korea.

2024 **Byung-Doh Oh**. *The bigger-is-worse effects of model size and training data of large language model surprisal on human reading times.* Invited talk in the Distinguished Speakers in Language Science Colloquium Series, Saarland University, Saarbrücken, Germany.

2023 **Byung-Doh Oh**. *The bigger-is-worse effects of model size and training data of large language model surprisal on human reading times.* Invited talk at the Center for Data Science, New York University, New York, NY.

2022 **Byung-Doh Oh**. *Computational models of sentence processing and syntactic acquisition.* Invited talk at the Department of English, Dongguk University, Seoul, Korea (online).

Conference Presentations

2023 **Byung-Doh Oh** and William Schuler. *On the bigger-is-worse nature of pre-trained language model surprisal.* Poster presented at the 36th Annual Conference on Human Sentence Processing, Pittsburgh, PA.

2023 **Byung-Doh Oh** and William Schuler. *Memory-based predictors from GPT-2 attention predict reading times over surprisal.* Poster presented at the 36th Annual Conference on Human Sentence Processing, Pittsburgh, PA.

2022 **Byung-Doh Oh**. *Unified unsupervised grammar induction for typologically diverse languages.* Poster presented at the DARPA Risers program, Columbus, OH.

2021 **Byung-Doh Oh**, Christian Clark, and William Schuler. *Comparison of structural and neural language models as surprisal estimators.* Short talk presented at the 34th Annual CUNY Conference on Human Sentence Processing, Philadelphia, PA (online).

2021 **Byung-Doh Oh** and William Schuler. *Contributions of propositional content and syntactic categories in sentence processing.* Short talk presented at the 34th Annual CUNY Conference on Human Sentence Processing, Philadelphia, PA (online).

2019 Evan Jaffe and **Byung-Doh Oh**. *The role of learnability in morphological change: A computational approach.* Talk presented at the Fourth American International Morphology Meeting, Stony Brook, NY.

Teaching

SP25	**Instructor of Record**, DS-GA 1015: Text-as-Data, New York University
2024	**Invited Lecture**, DS-GA 3001: Computational Linguistics and Cognitive Science, New York University (Instructor: Tal Linzen) *"Transformer-based language model surprisal predicts human reading times best with about two billion training tokens"*
2024	**Invited Lecture**, Metro Early College High School *"Guessing meaning and breaking it down (with NACLO problems)"*
AU20, SP21	**Instructor of Record**, LING 3801: Codes and Code Breaking, The Ohio State University Overall student evaluation: 4.69/5 (AU20), 4.86/5 (SP21)
SU17	**Teaching Assistant**, College English 1, Seoul National University

Service

Organizing

2024–2025	Workshop on Cognitive Modeling and Computational Linguistics

Reviewing

2024	ACL Rolling Review, BlackboxNLP, *Journal of Memory and Language*, *Frontiers in Communication*, *eLife*, *Nature Computational Science*
2023	ACL Rolling Review
2022	ACL Rolling Review
2021	ACL Rolling Review, 11th Workshop on Cognitive Modeling and Computational Linguistics

Departmental/University

2020, 2022	Student Volunteer, North American Computational Linguistics Olympiad (NACLO)
2019–2024	Laboratory/Computing Committee, Department of Linguistics, The Ohio State University
2019–2020	Treasurer, Student Linguistics Association, The Ohio State University
2018–2019	Speakers Committee, Department of Linguistics, The Ohio State University
2017–2018	Resident Advisor, Gwanak Residence Halls, Seoul National University
2015–2017	International Student Assistant, Gwanak Residence Halls, Seoul National University

Other

2015–2016	K-MOOC Monitoring Team, National Institute for Lifelong Learning, Seoul, Korea
2012–2014	Translator/Interpreter (ROK Air Force), ROK/US Combined Forces Command, Seoul, Korea

Skills

Languages	Korean (native), English (bilingual), Japanese (conversational)
Programming Languages	Python, C++, Bash, R, LaTeX, Prolog
Frameworks	PyTorch, TensorFlow, HuggingFace, pandas, Matplotlib, Git, Slurm, PBS