Jon Gauthier, Ph.D.

1651 4th Street, Room 674A
San Francisco, CA 94158 U.S.A.

email: jon@gauthiers.net
URL: http://www.foldl.me
phone: 480 788 8509

Current position

Postdoctoral Scholar
University of California, San Francisco
Department of Neurological Surgery

Affiliated with:
Chang Lab (PI: Edward F. Chang)

Education

2017–2023 Ph.D. in Cognitive Science, Massachusetts Institute of Technology (Cambridge, MA)
Thesis title: *Multi-level models of language comprehension in the mind and brain*
Advised by Roger P. Levy and Joshua B. Tenenbaum.

2013–2017 B.Sc. in Symbolic Systems, Stanford University (Palo Alto, CA)
Advised by Christopher D. Manning.

Work experience

2016 *Research Intern*, OpenAI (San Francisco, CA)
2015 *Research Intern*, Google Brain (Mountain View, CA)
2014–2017 *Research Assistant*, Stanford Natural Language Processing Group (Stanford, CA)
2012–2013 *Software Development Engineer*, Stremor Corp. (Phoenix, AZ)

Publications & talks

Refereed conference proceedings

2025 **Jon Gauthier**, Canaan Breiss, Matthew K. Leonard & Edward F. Chang. Emergent morpho-phonological representations in self-supervised speech models. Accepted at the *Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)* (Suzhou, China).
Jon Gauthier & Canaan Breiss. Emergent morpho-phonological patterning in a model of spoken word recognition. Accepted at the *Annual Meeting on Phonology (AMP 2025)* (Berkeley, California).
Canaan Breiss & **Jon Gauthier**. Emergent feature structure in self-supervised speech models' phone representations. Accepted at the *Annual Meeting on Phonology (AMP 2025)* (Berkeley, California).
2024 Catherine Wang, Deborah Levy, John Andrews, **Jon Gauthier**, & Edward Chang. The pSTG is

associated with prolonged language impairment following neurosurgical resection. Accepted at *Society for the Neurobiology of Language (SNL 2024)* (Brisbane, Australia).

2023 **Jon Gauthier** *&* Roger Levy. The neural dynamics of auditory word recognition and integration. Accepted at *Empirical Methods in Natural Language Processing (EMNLP 2023)* (Singapore).

Jon Gauthier *&* Roger Levy. The neural dynamics of auditory word recognition and integration. In *Proceedings of the Conference on Cognitive Computational Neuroscience (CCN 2023)* (Oxford, England).

Kinan Martin, **Jon Gauthier**, Canaan Breiss, *&* Roger Levy. Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration. In *Proceedings of INTERSPEECH 2023* (Dublin, Ireland).

{Koustuv Sinha, **Jon Gauthier**}, {Aaron Mueller, Kanishka Misra}, Keren Fuentes, Roger P. Levy, *&* Adina Williams. Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*. **Outstanding Paper Award.**

2020 Ethan Wilcox, **Jon Gauthier**, Jennifer Hu, Peng Qian, *&* Roger P. Levy. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020)*.

Jennifer Hu, **Jon Gauthier**, Peng Qian, Ethan Wilcox, *&* Roger P. Levy. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, *&* Roger P. Levy. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2020)*.

2019 **Jon Gauthier** *&* Roger P. Levy. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Hong Kong).

Jon Gauthier, Roger P. Levy, *&* Joshua B. Tenenbaum. A rational model of syntactic bootstrapping. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)* (Montreal, Canada).

2018 {**Jon Gauthier**, Anna Ivanova}. Does the brain represent words? An evaluation of brain decoding studies of language understanding. In *Proceedings of the 2nd Conference on Cognitive Computational Neuroscience (CCN 2018)* (Philadelphia, PA).

Jon Gauthier, Roger P. Levy, *&* Joshua B. Tenenbaum. Word learning and the acquisition of syntactic–semantic overhypotheses. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci 2018)* (Madison, WI).

2016 {Sam Bowman, **Jon Gauthier**}, Raghav Gupta, Abhinav Rastogi, Christopher D. Manning, *&* Christopher Potts. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)* (Berlin, Germany).

Other refereed content

2019 **Jon Gauthier**, João Loula, Eli Pollock, Tyler Brooke Wilson, *&* Catherine Wong. From mental representations to neural codes: A multilevel approach. *Behavioral and Brain Sciences, 42*, E228.

Book chapters

2022 Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, *&* Roger Levy. *Learning syntactic structures from string input.* In *Algebraic Structures in Natural Language.* CRC Press.

Presentations

2025 **Jon Gauthier**, Canaan Breiss, Matthew Leonard, & Edward F. Chang. What does word recognition look like?: Distributed representational signatures of word-level processing. Invited talk at Stanford GLySN Lab (Stanford, California).

Jon Gauthier, Canaan Breiss, Matthew Leonard, & Edward F. Chang. Emergent morpho-phonological representations in neural network models of speech. Invited talk at the Bay Area Language Processing Interest Group (Davis, California).

2023 **Jon Gauthier** & Roger Levy. The neural dynamics of auditory word recognition and integration. Invited talk at the NeuroCognition of Language Lab at Tufts University (Medford, MA).

Jon Gauthier. Multi-level models of language comprehension in the mind and brain. Thesis defense at MIT.

2022 **Jon Gauthier** & Roger Levy. Multi-level modeling for the cognitive neuroscience of language: two case studies. Invited talk at the Computation and Psycholinguistics Lab at New York University (New York, NY).

Jon Gauthier & Roger Levy. Multi-level modeling for the cognitive neuroscience of language: two case studies. Invited talk at UC San Francisco Speech Lab (San Francisco, CA).

Jon Gauthier & Roger Levy. Multi-level modeling for the cognitive neuroscience of language: two case studies. Invited talk at Meta Brain & AI (Paris, France).

2018 **Jon Gauthier**, Maxwell Nye, Roger Levy, & Joshua B. Tenenbaum. A scalable computational model for capturing the syntax–semantics link. Invited talk at Harvard Language and Cognition speaker series (Cambridge, MA).

Jon Gauthier, Maxwell Nye, Roger Levy, & Joshua B. Tenenbaum. A rational model of syntactic and semantic bootstrapping. Poster presentation at *Language Learning in Humans and Machines (L2HM 2018)* (Paris, France).

2017 **Jon Gauthier**. What does natural language processing tell us about language? Invited talk in the MIT Computation and Language talk series.

Li Lucy & **Jon Gauthier**. Are distributional representations ready for the real world? Poster presentation at the *ACL 2017 Workshop on Language Grounding for Robotics* (Vancouver, Canada).

2016 **Jon Gauthier** & Igor Mordatch. A paradigm for situated and goal-driven language learning. Oral presentation at the *NIPS 2016 Machine Intelligence Workshop* (Barcelona, Spain).

Jon Gauthier. Structured deep models for sentence representation. Invited talk at Google DeepMind (London, UK).

Preprints

2018 **Jon Gauthier**. Conceptual issues in AI safety: the paradigmatic gap.

2015 **Jon Gauthier**. Conditional generative adversarial networks for convolutional face generation.

2014 **Jon Gauthier**, Danqi Chen, and Christopher D. Manning. Exploiting long-distance context in transition-based dependency parsing with recurrent neural networks.

Awards

2018–2023 Open Philanthropy AI Fellow

2019 Angus MacDonald Award for Excellence in Undergraduate Teaching, MIT Department of Brain

2017 and Cognitive Sciences Stanford J.E. Wallace Sterling Award for Scholastic Achievement

Awarded to 25 undergraduates in the School of Humanities and Sciences in the class of 2017.

2017 Stanford Dean's Award for Academic Excellence

Awarded to ten undergraduates in the class of 2017 by faculty nomination.

2016 Phi Beta Kappa

2014 Stanford President's Award for Academic Excellence in the Freshman Year

2013 National Merit Scholar
2013 AP Scholar with Honor

Service

2023– Area chair for ACL Rolling Review
2018–2020 Co-founder of the MIT Brain and Cognitive Sciences Philosophy Circle
2015– Reviewer for NeurIPS, ICLR, CCN, AAAI, *ACL

Relevant coursework

This section lists graduate-level coursework and relevant research output. Full transcript available upon request.

2019 Pragmatics in Linguistic Theory — MIT 9.S913 (Roger Levy and Danny Fox)
Neural Mechanisms of Cognitive Computation — MIT 9.017 (Michael Halassa)
2018 Computational Psycholinguistics — MIT 9.012 (Roger Levy)
Project: "A rational model of syntactic and semantic bootstrapping." Presented at L2HM 2018.
Developmental Proseminar — Harvard PSY 2170 (Liz Spelke)
2017 Computational Cognitive Science — MIT 9.660 (Josh Tenenbaum)
Project: "Language is not ambiguous! An evolutionary simulation study of communication in grounded contexts."
2016 Developmental Psycholinguistics — Stanford LINGUIST 248 (Eve Clark)
Computational Cognitive Science — Stanford PSYCH 204 (Noah Goodman)
Project: "Online learning of compositional semantics in spatial reference games." With Sebastian Schuster.
2015 Probabilistic Graphical Models — Stanford CS 228 (Stefano Ermon)
Artificial Intelligence — Stanford CS 221 (Percy Liang)
Project: "Reinforcement learning pointer networks." With Ilya Sutskever & Oriol Vinyals. Submitted to ICML 2015.
Independent Research — Stanford CS 199 (Christopher Manning)
Project: "Just-in-time estimation of unknown word embeddings." Submitted to EMNLP 2015.
2014 Foundations of Psycholinguistics — Stanford LINGUIST 246 (Eve Clark)
Theoretical Neuroscience — Stanford APPPHYS 293 (Surya Ganguli)
Convolutional Neural Networks for Visual Recognition — Stanford CS 231N (Andrej Karpathy)
Project: "Conditional generative adversarial networks for convolutional face generation." Published as technical report, 2015.
Natural Language Processing — Stanford CS 224N (Christopher Manning)
Project: "Buffer-aware transition-based dependency parsing with recurrent neural networks."
Machine Learning — Stanford CS 229 (Andrew Ng)
Project: "Language identification and accent variation detection in spoken language recordings." With Shyamal Buch and Arthur Tsang.
Independent Research — Stanford CS 199 (Christopher Manning)
Project: "Improved data selection methods for low-resource machine translation applications." With Danqi Chen.
Project: "Exploiting long-distance context in transition-based dependency parsing with recurrent neural networks." With Danqi Chen. Submitted to ICLR 2015.

2013 Natural Language Understanding Stanford CS 224U (Christopher Potts)
Project: "Deep neural models for bilingual lexicon extraction." With Arthur Tsang. Submitted to ACL 2014.