

Catherine Arnett

ccarnett@ucsd.edu

<https://catherinearnett.github.io/>

Education

- PhD Candidate, Linguistics with Computational Social Science**, UC San Diego 2019-
Committee: Farrell Ackerman, Benjamin Bergen, Leon Bergen, Victor Ferreira
- MA Chinese and Linguistics (Hons)**, University of Edinburgh 2014-2018
Including one year exchange at Zhejiang University, Hangzhou, P.R. China

Selected Professional Experience

- Research Intern**, PleIAs Apr - Sep 2024
Researching issues related to multilingual language modeling, including tokenization and toxicity detection.

Publications

Refereed Journal Articles

1. Jesse Quinn, Matthew Goldrick, **Catherine Arnett**, Victor S. Ferreira, Tamar H. Gollan (accepted). Syntax Drives Default Language Selection in Bilingual Connected Speech Production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Refereed Conference Proceedings

1. James Michaelov, **Catherine Arnett**, Benjamin K. Bergen (2024). [Revenge of the Fallen? Recurrent Models Match Transformers at Predicting Human Language Comprehension Metrics](#). The First Conference on Language Modeling (COLM). Philadelphia, USA.
2. **Catherine Arnett***, Pamela D. Rivière*, Tyler A. Chang, and Sean Trott (2024). [Different Tokenization Schemes Lead to Comparable Performance in Spanish Number Agreement](#). Special Interest Group on Computational Morphology and Phonology (SIGMORPHON) co-located at NAACL. Mexico City, Mexico. *equal contribution
3. **Catherine Arnett***, Tyler A. Chang*, Benjamin K. Bergen (2024). [A Bit of a Problem: Measurement Disparities in Dataset Sizes Across Languages](#). 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL), co-located at LREC-COLING. Torino, Italy. *equal contribution
4. James A. Michaelov*, **Catherine Arnett***, Tyler A. Chang, Benjamin K. Bergen (2023). [Structural priming demonstrates abstract grammatical representations in multilingual language models](#). Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore. *equal contribution
5. **Catherine Arnett** (2019). [Pathways of Change in Romance Motion Events: A Corpus-based Comparison](#). *Proceedings of the Thirtieth Western Conference on Linguistics (WeCOL)*. Vol 23. Fresno, CA.

Manuscripts

1. Catherine Arnett and Benjamin K. Bergen. Morphological Typology is Predictive of Language Modeling Performance, But Why?
2. **Catherine Arnett***, Eliot Jones*, Ivan P. Yamshchikov, Pierre-Carl Langlais (under review). Toxicity of the Commons: Curating Open-Source Pre-Training Data. *equal contribution
3. Tyler A. Chang, **Catherine Arnett**, Zhuowen Tu, Benjamin K. Bergen (under review). [Goldfish: Monolingual Language Models for 350 Languages](#)

4. Pavel Chizhov*, **Catherine Arnett***, Elizaveta Korotkova, Ivan P. Yamshchikov (under review). BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training. *equal contribution.
5. Tyler A. Chang, **Catherine Arnett**, Zhuowen Tu, Benjamin K. Bergen (under review). [When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages](#). arXiv preprint arXiv:2311.09205.

Conference Presentations

1. **Catherine Arnett**, Tyler A. Chang, James A. Michaelov, and Benjamin K. Bergen (2023). [Crosslingual Structural Priming and the Pre-Training Dynamics of Bilingual Language Models](#). *The 3rd Workshop on Multilingual Representation Learning co-located with EMNLP 2023*. Singapore.
2. **Catherine Arnett** & Maho Takahashi (2022). Creating a Baseline to Evaluate Correlations Between Language and Environment. *Machine Learning and the Evolution of Language*. Kanagawa/online. [[abstract](#)] [[poster pdf](#)]
3. **Catherine Arnett** & Eva Wittenberg (2020). Multiple Meanings of Doubling Up: Mandarin Verbal Reduplication. *The 26th Architectures And Mechanisms for Language Processing Conference (AMLaP)*. Potsdam/online. [[poster pdf](#)]
4. **Catherine Arnett** & Eva Wittenberg (2020). Conceptual Effects of Verbal Reduplication in Mandarin Chinese. *North American Conference on Chinese Linguistics 32*. Storrs, CT/online.
5. **Catherine Arnett** & Eva Wittenberg (2019). Conceptual Effects of Verbal Reduplication in Mandarin Chinese. California Meeting on Psycholinguistics. Santa Cruz, California. [[poster pdf](#)]
6. **Catherine Arnett** (2018). Diachronic study of the typology of motion verbs in the Romance languages. Undergraduate Linguistics Association of Britain Conference, University of Edinburgh.

Tools and Resources Created

Goldfish Language Models https://huggingface.co/goldfish-models Small, Comparable Monolingual Language Models for 350 Languages.	August 2024
Byte Premium Tool https://github.com/catherinearnett/byte-premium-tool Command-line tool to calculate dataset scaling needed to achieve cross-lingual training data equity.	Feb 2024

Media and Outreach

Interview Code and Community: Computational Social Science Program Addresses Social Questions with Data	Jan 2024
Author and Subject-Matter Expert , Cognitive Foundations, an open Textbook Update and rewrite portions of the <i>Language</i> chapter using bookdown in R.	Sep 2022

Teaching Experience

Graduate Teaching Consultant for International Instructors , UC San Diego Instruct incoming international graduate students in teaching skills and subject-specific English	2022-2024
Instructor , UC San Diego LIGN 170: Psycholinguistics Course Website	Summer 2023
Teaching Assistant , UC San Diego LIGN 101: Introduction to Language, Fall 2020-Spring 2022 LIDS 19: Independent Language Study, Winter 2020-Spring 2020 LIGN 170: Psycholinguistics, Fall 2019	2019-2022

Scholarships, Fellowships, and Grants

Yankelovich Graduate Research Funding, UC San Diego, \$3,500	2024
GPSA Travel Grant Award, UC San Diego, \$300	2023
Linguistics Department Anti-Racist Pedagogy Micro-Fellowship, UC San Diego Linguistics, \$500	2023
Summer Graduate Teaching Scholar, UC San Diego, \$1,000	2023
LaVerne Noyes Foundation Endowed Fellowship, UC San Diego, \$6,000	2022
Academic Senate Grant, UC San Diego, \$9,000	2022
Friends of the International Center Fellowship, UC San Diego, \$2000	2020
Research and Travel Grant, UC San Diego Linguistics, \$350	2019
GPSA Travel Grant Award, UC San Diego, \$500	2019
Student Opportunity Fund, University of Edinburgh, £250	2018
Eric Liddell China Saltire Scholarship, University of Edinburgh, £5,000	2016
St. Andrews Society of North Carolina Scholarship, \$37,000	2014-2018

Awards

Teaching Assistant Excellence Award, UCSD Linguistics	2023-2024
Edinburgh Award, Peer Learning and Support	2018

Skills and Languages

Languages

English - Native
Mandarin - Fluent
Spanish - Fluent
French - Conversational
Latin - Structural Knowledge
San Juan Piñas Mixtec - Fieldwork Experience

Coding and Software

R/RStudio tidyverse, lme4, bookdown
Python pandas, matplotlib, seaborn, numpy, BeautifulSoup
 basic SQL, basic HTML

Academic Service

Ad Hoc Reviewing

Multilingual Representation Learning (MRL) Workshop, EMNLP	2024
EMNLP Ethics Reviewer	2024
Undergraduate Linguistics Association of Britain	2020

Mentoring

Masters Students

Adelle Engmann, Computational Social Science	2023-2024
Xinyao Yi, Computational Social Science	2022-2023

Undergraduate Honors Theses Committee

Yuhan Fu, Cognitive Science B.S. Honors Thesis	2024
--	------

“Is There Evidence for Embodied Simulation During Language Translation?”

Undergraduate Research Assistants

Preeti Phan	Spring 2024
Sara Rojas	Spring 2024
Shani Spector	Spring 2024
Stephanie Huang	Winter 2024 - Spring 2024
Gwen O'brien	Winter 2024 - Spring 2024
Nattida Neal	Winter 2024 - Spring 2024
Sidney Ma	Winter 2024
Sarah Yu	Fall 2023 - Winter 2024
Livia Zhou	Fall 2023
Yoyo Wu	Fall 2023
Jiawei Lyu	Spring 2023
Jason Tran	Winter 2023 - Spring 2023
Emily Xu	Fall 2022, Spring 2023 - Fall 2023
Tiffany Wu	Fall 2022 - Winter 2023
Fiona Tang	Fall 2022 - Spring 2023
Jane Yang, now Psychology PhD student at UC San Diego	Winter 2020 - Spring 2022
Zhiyi Li	Winter 2020 - Winter 2021
Xutong Zhang	Fall 2020 - Winter 2021
Annabelle Chang	Fall 2019 - Winter 2021
Jessica Luo, now Linguistics PhD Student at University of Washington	Fall 2019 - Winter 2021
Ruoqi Wei	Fall 2019 - Winter 2021
Rebecca Xu	Fall 2019 - Winter 2020

Professional Memberships

Association for Computational Linguistics

2023-