

# Catherine Arnett

[catherine.arnett@gmail.com](mailto:catherine.arnett@gmail.com)

[catherinearnett.github.io](https://catherinearnett.github.io)

## Education

- PhD Candidate, Linguistics with Computational Social Science**, UC San Diego 2019-2025  
Dissertation: "A Linguistic Approach to Crosslingual and Multilingual NLP"  
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

- EleutherAI**, *remote*  
NLP Researcher 2025 - *present*
- PleIAs**, *Paris, France (remote)*  
Research Scientist 2024-2025  
Research Intern Summer 2024

## Publications

### Refereed Journal Articles

1. Jesse Quinn, Matthew Goldrick, **Catherine Arnett**, Victor S. Ferreira, Tamar H. Gollan (2024). Syntax Drives Default Language Selection in Bilingual Connected Speech Production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

### Refereed Conference Proceedings

1. **Catherine Arnett** and Benjamin K. Bergen (2025). [Why do language models perform worse for morphologically complex languages?](#) The 31st International Conference on Computational Linguistics (COLING). Abu Dhabi, UAE and online. **Best Paper Award**.
2. Pavel Chizhov\*, **Catherine Arnett\***, Elizaveta Korotkova, Ivan P. Yamshchikov (2024). [BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training](#). Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, FL, USA. \*equal contribution.
3. Tyler A. Chang, **Catherine Arnett**, Zhuowen Tu, Benjamin K. Bergen (2024). [When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages](#). Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, FL, USA. **Outstanding Paper Award**.
4. 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, PA, USA.
5. **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
6. **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

7. 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
8. **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, USA.

## Manuscripts

1. **Catherine Arnett**, Tyler A. Chang, James A. Michaelov, and Benjamin K. Bergen (under review). [On the Acquisition of Shared Grammatical Representations in Bilingual Language Models](#).
2. **Catherine Arnett\***, Eliot Krzysztof 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](#)

## Conference Presentations

1. **Catherine Arnett** (2025). Toxic Commons: Toxicity of the Commons: Curating Open-Source Pre-Training Data. *The First Conference of the International Association for Safe & Ethical AI*. Official event of the AI Action Summit. Paris, France.
2. **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.
3. **Catherine Arnett** & Maho Takahashi (2022). Creating a Baseline to Evaluate Correlations Between Language and Environment. *Machine Learning and the Evolution of Language*. Kanagawa/online. [\[abstract\]](#) [\[poster pdf\]](#)
4. **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. [\[poster pdf\]](#)
5. **Catherine Arnett** & Eva Wittenberg (2020). Conceptual Effects of Verbal Reduplication in Mandarin Chinese. *North American Conference on Chinese Linguistics 32*. Storrs, CT/online.
6. **Catherine Arnett** & Eva Wittenberg (2019). Conceptual Effects of Verbal Reduplication in Mandarin Chinese. California Meeting on Psycholinguistics. Santa Cruz, California. [\[poster pdf\]](#)
7. **Catherine Arnett** (2018). Diachronic study of the typology of motion verbs in the Romance languages. Undergraduate Linguistics Association of Britain Conference, University of Edinburgh.

## Invited Talks

**Why do language models perform worse for morphologically complex languages?**

tinlab, Boston University

March 2025

**Characterizing Shared Multilingual Representations with Structural Priming**

Brown NLP, Brown University

March 2025

## Tools and Resources Created

### Primary Creator

**B-GPT**

February 2025

[\[model family\]](#)

Controlled bilingual models trained with different proportions of data from each language.

Model checkpoints released.

**MorphScore**

November 2024

<https://github.com/catherinarnett/morphscore>

Tokenizer evaluation dataset for assessing morphological validity of tokenization.  
Comparable datasets for 22 languages.

**Byte Premium Tool**

Feb 2024

<https://github.com/catherinarnett/byte-premium-tool>

Command-line tool to calculate dataset scaling needed to achieve cross-lingual training data equity.

**Contributor****Pleias 1.0 Models**

December 2024

[[model collection](#)][[release blog post](#)]

The first language models trained on exclusively permissively licensed data.

3 base models (300M, 1B, 3B) and 2 RAG models (300M, 1B).

**Common Corpus**

November 2024

[https://huggingface.co/datasets/PleIAs/common\\_corpus](https://huggingface.co/datasets/PleIAs/common_corpus) [[release blog post](#)]

2 trillion token corpus of permissively licensed data for language model training.

Multilingual corpus covering a range of domains, genres, and time periods.

**Goldfish Language Models**

August 2024

<https://huggingface.co/goldfish-models>

Small, Comparable Monolingual Language Models for 350 Languages.

**Media and Outreach****NLP Blog Posts**

“Small is Beautiful”

“Releasing the largest multilingual open pretraining dataset”

“Detoxifying the Commons”

“wHy DoNt YoU jUsT uSe ThE ILaMa ToKeNiZeR??”

**Online Publications**

Interviewed in [Code and Community: Computational Social Science Program Addresses Social Questions with Data](#)

**Author and Subject-Matter Expert,**[Cognitive Foundations, an open Textbook](#)

Update and rewrite portions of the *Language* chapter using bookdown in R.

**Teaching Experience****Graduate Teaching Consultant for International Instructors, UC San Diego**

2022-2024

Instruct incoming international graduate students in teaching skills and subject-specific English

**Instructor of Record, UC San Diego**

Summer 2023

LIGN 170: Psycholinguistics

[Course Website](#)**Teaching Assistant, UC San Diego**

2019-2022

LIGN 101: Introduction to Language, Fall 2020-Spring 2022

LIDS 19: Independent Language Study, Winter 2020-Spring 2020

LIGN 170: Psycholinguistics, Fall 2019

**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

Best Paper Award, COLING 2025	2025
Outstanding Paper Award, EMNLP 2024	2024
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, transformers  
 basic SQL, basic HTML

## Professional Activities

### Reviewing

International Conference on Computational Linguistics (COLING)	2024
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-