

Charles Torres, PhD

Machine Learning & Linguistics Research Scientist | Computational Linguist | Applied NLP Scientist

Research Profile

Machine learning and linguistics research scientist with a PhD in Language Science and six years of experience designing computational studies of language. Builds multilingual datasets and neural models, develops methods for testing what those models learn, and communicates the results through scientific software, papers, visualizations, presentations, and teaching. Research includes four peer-reviewed publications and a first-author preprint spanning computational linguistics, information theory, formal language theory, language change, and model evaluation.

Areas of Expertise

Machine Learning	Deep learning, natural language processing, model evaluation, controlled experiments, benchmark design, error analysis, interpretability through controlled comparisons
Models & Methods	Transformers, CNNs, LSTMs/RNNs, feedforward neural networks, multilayer perceptrons, logistic regression, sparsity regularization, prequential coding
Statistics & Data	pandas, scikit-learn, repeated-seed evaluation, permutation tests, t-tests, Tukey HSD, uncertainty estimates, data visualization
Language Science	Computational linguistics, multilingual data, phonology, IPA, orthography, language change, formal language theory, acoustic phonetics, cross-linguistic variation
Programming	Python (6 years), PyTorch (5 years), R (5 years), SQL (deployed app), regular expressions (regex), Hugging Face Transformers, Julia, $\LaTeX$
Research Software	HPC workflows, local GPU training, Git, GitHub Actions, Linux, Django, PostgreSQL, nginx, JavaScript, reproducibility documentation
Scientific Communication	Public speaking, presentation and figure design, technical writing, academic publishing, writing for specialist and general audiences, teaching, live question-and-answer discussion

Professional Experience

2025–Present	Postdoctoral Researcher , <i>UC Davis</i> , Davis, California. ▪ Research design: Develop research questions and study designs with the principal investigator, analyze results, resolve methodological problems, and write up findings. ▪ Computational modeling: Build Python and PyTorch simulations of language production and change using information-theoretic objectives; analyze and visualize behavior in R and matplotlib. ▪ Empirical research: Design experimental materials for an ongoing study of reduction; the study and analysis remain in progress. ▪ Scientific communication: Turn ongoing research into talks, slides, posters, figures, and manuscripts; delivered a lab lecture on the information-theory concepts used in the group's work.
2024–2025	ML Subject Matter Specialist , <i>Leading Frontier AI Lab</i> , Remote. ▪ LLM evaluation: Built prompts, response criteria, and scorecards for a new dataset used to train and benchmark language models with human feedback. ▪ Model-quality analysis: Reviewed model outputs and human judgments to identify failure patterns and revised tasks when the criteria did not produce useful distinctions. ▪ Technical collaboration: Met with the machine-learning team to compare effective and ineffective tasks, resolve ambiguous cases, and refine useful data characteristics; later reviewed tasks written by other researchers against project standards.

Davis, California

☎ (520) 238-1686 • ✉ torrescj92@gmail.com • 🌐 www.charles-torres.com

1/3

2020–2025 **Teaching Assistant**, *UC Irvine*, Irvine, California.

- **Teaching and public speaking:** Designed course materials and explained linguistics, statistics, and computational modeling across 13 university classes, adapting examples and technical detail to students with different backgrounds.
- **Audience adaptation:** Answered questions in class and changed the explanation when students needed a different route into the material.

2016–2020 **Tech Support Specialist**, *Banner Health*, Phoenix, Arizona.

- **Troubleshooting and communication:** Diagnosed software, hardware, and access problems for 30+ internal users per day and explained the problem and next steps in plain language.

Selected Research Projects

2024–Present **Clarifying Orthography: Orthographic Transparency as Compressibility**, *UC Irvine*. Project lead and first author.

- **Research question:** Developed a script-agnostic measure of how well spelling predicts pronunciation, and vice versa, using mutual compressibility and algorithmic information theory.
- **Multilingual data:** Built paired spelling-pronunciation datasets of 5,032 word types per language across 22 languages; implemented IPA cleanup, Unicode normalization, and Chinese, Japanese, and Korean character decomposition.
- **Modeling and algorithms:** Implemented sequence-to-sequence CNNs in PyTorch and replaced a quadratic prequential-coding procedure with an $O(kn)$ approximation, linear in dataset size for a fixed number of replay streams.
- **Evaluation:** Ran masked and unmasked comparisons across 40 paired model initializations for every language and prediction direction; analyzed results with t-tests, Tukey HSD, confidence intervals, and R visualizations.
- **Presentation and visualization:** Designed visualizations that made cross-language and cross-script comparisons legible, presented the project at CAMP, and released a first-author preprint with public research code.

2022–2024 **Simpler Neural Networks Prefer Subregular Languages**, *UC Irvine*. Project lead and first author.

- **Experimental design:** Built matched synthetic datasets to test which sequence patterns LSTMs learn most easily and whether smaller networks show stronger preferences.
- **Neural modeling:** Adapted continuous L0 regularization to remove unnecessary weights during training and trained hundreds of GPU-accelerated recurrent models across controlled sparsity settings.
- **Statistical analysis:** Compared accuracy, active parameter count, and generalization with permutation tests and R visualizations; found that smaller LSTMs learned structurally simpler, human-attested patterns with fewer active parameters.
- **Writing and visual explanation:** Designed accuracy-complexity figures showing how model preferences changed as networks became smaller and led the resulting first-author paper published in Findings of EMNLP 2023.

2021–2023 **Evaluating Adjective Ordering Across 32 Languages**, *UC Irvine*. Second-author collaboration.

- **Data engineering:** Combined participant ratings, parsed Universal Dependencies corpora, and aligned fastText embeddings into model-ready features across languages.
- **Modeling and evaluation:** Fit logistic regressions with leave-one-language-out evaluation and trained feedforward neural networks to compare individual and combined theories with random and distributional baselines.
- **Interdisciplinary communication:** Contributed multilingual modeling, exploratory analysis, and interpretation to a quantitative comparison for cognitive-science and computational-linguistics audiences, published in TACL.

2025–Present **Modeling Language Production and Change**, *UC Davis*. Ongoing postdoctoral research.

- **Agent-based simulation:** Built simple decision-making agents in Python and PyTorch that choose between two actions using a closed-form information-theoretic objective.
- **Iterative learning:** Modeled how one generation shapes the learning environment of the next; ran experiments across conditions and built figures in R and matplotlib.
- **Research presentation:** Presented the simulation as a scientific poster at HSP and delivered a lab lecture on its information-theory foundations; further analysis and a manuscript are in progress.

Davis, California

☎ (520) 238-1686 • ✉ torrescj92@gmail.com • 🌐 www.charles-torres.com

2/3

2024–2025 **Creolization versus Code-Switching**, *UC Irvine*. Project lead and first author.

- **Agent-based cognitive modeling:** Led a published simulation study of bilingual strategies in language contact and coordinated the work from research question through analysis, manuscript preparation, and peer-review revision.

Research Software

2025–Present **PreqTorch** – Open-source PyTorch library for reusable prequential-coding and replay-based evaluation workflows, published on PyPI.

- **Reusable components:** Built configurable encoders, custom data-loading support, masking, and loss functions so new evaluation methods can be added without rebuilding the entire pipeline.
- **Documentation and maintenance:** Document research methods and reproducibility guidance, review current literature, and add evaluation methods when they fit the library.

URL <https://github.com/cj-torres/preqtorch>

2025–Present **Eudaimoneion** – Personal work-in-progress scientific-collaboration web application built with Django, PostgreSQL/SQL, JavaScript, Ubuntu, and nginx.

- **Web development:** Designed database-backed features for accounts, research content, and social interaction and deployed the application on a remote Linux server.

URL <https://www.eudaimoneion.org>

Publications

- 2025 Charles John Torres, Weijie Xu, Yanting Li, and Richard Futrell. 2025. *Creolization versus code-switching: An agent-based cognitive model for bilingual strategies in language contact*. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 207–218, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- 2025 Charles J. Torres and Richard Futrell. 2025. *Clarifying orthography: Orthographic transparency as compressibility*. arXiv:2505.13657 [cs.CL]. Preprint.
- 2023 William Dyer, Charles Torres, Gregory Scontras, and Richard Futrell. 2023. *Evaluating a Century of Progress on the Cognitive Science of Adjective Ordering*. Transactions of the Association for Computational Linguistics, 11:1185–1200.
- 2023 Charles Torres and Richard Futrell. 2023. *Simpler neural networks prefer subregular languages*. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1651–1661, Singapore. Association for Computational Linguistics.
- 2023 Charles Torres, Kenneth Hanson, Thomas Graf, and Connor Mayer. 2023. *Modeling island effects with probabilistic tier-based strictly local grammars over trees*. In Proceedings of the Society for Computation in Linguistics 2023, pages 155–164, Amherst, MA. Association for Computational Linguistics.

Education

2020–2025 **PhD, Language Science**, *University of California, Irvine*.

Dissertation research on neural sequence models, multilingual data, information-theoretic evaluation, formal language theory, language change, and computational linguistics.

2017–2019 **Bachelor of Science, Neuroscience and Cognitive Science**, *University of Arizona, Tucson*.

Minor in Mathematics; Summa Cum Laude. Coursework included machine learning, natural language processing, databases and SQL, syntax, linear algebra, graph theory, and discrete mathematics.