ROBERT PUGH

Email: <last><first[:3]>@iu.edu
Website: robertpugh.site

EDUCATION

PhD University of Indiana, Computational Linguistics — August 2021 - Present
Minor: Latin American and Caribbean Studies
Advisor: Dr. Francis Tyers

MA San Jose State University, Linguistics — May 2015
Graduate Certificate, Computational Linguistics

BA University of California, Santa Cruz, Linguistics — May 2011
Minored in Latin-American and Latino Studies

HONORS AND AWARDS

Foreign Language and Area Studies Award — 2022 Summer
Received complete tuition reimbursement and stipend for the summer to study intermediate Maya at Indiana University.

Foreign Language and Area Studies Award — 2021-2022 Academic Year
Received complete tuition reimbursement and stipend for the academic year to study intermediate Maya at Indiana University.

RESEARCH EXPERIENCE

Indiana University, Bloomington — 2021 to present
Research Assistant, Inclusive Technologies for Marginalised Languages

- Building computational models and resources for Nahuatl morphosyntax
- Developing automated speech recognition (ASR) systems for Nahuatl
- Exploring approaches to variation-robust natural language processing

Educational Testing Service, San Francisco, CA — 2017 to 2019
Research Engineer, NLP & Speech group

- Developed and researched automated written-response scoring engines
- Performed dialogue systems research focused on code-switching in spoken and written language

PROFESSIONAL EXPERIENCE

Transcendant Endeavors, New York, NY — 2023 to Present

Language Technology Consultant

- Advise on data and modeling strategy for text-to-speech (TTS) system for monolingual and code-switched Anishinaabemowin and English.
- Train models and prototype approaches to low-resource code-swithing TTS.

Mozilla Foundation, Mountain View, CA 2022 to 2023
Spanish Language Research Associate, Common Voice

- Engaged in community-building both online and in-person in Puebla, Mexico, with the goal of creating a more representative Spanish-language Common Voice dataset.
- Collected over 900 hours of Mexican Spanish speech data

Course Hero, Redwood City, CA 2019 to 2021
Staff Machine Learning Engineer

- Designed and built Machine Learning and NLP systems to understand unstructured student- and educator-created study documents.
- Wrote specialized parsers for segmenting domain-specific documents (e.g. homework, lecture notes, syllabi, etc.).
- Managed annotation projects and applied novel techniques for maximizing labeled data while minimizing time and cost.

NetBase Solutions, Santa Clara, CA 2014 to 2017
Data Science Engineer

- Owned NLP testing and evaluation, including regression testing, data annotation projects, and research for further linguistic development.
- Managed and maintained a MySQL database of annotated linguistic corpora.
- Performed acquisition, cleaning, and analysis of social media data (text, author statistics, and other metadata).

CafePress, Foster City, CA 2013 to 2014
Computational Linguisics Intern

- Worked primarily with the Search Engine Marketing team to identify, select, and generate optimal keywords.
- Developed language analytics tools, including custom part-of-speech tagger and statistical language generation models, using Python.

TEACHING EXPERIENCE

Syntactic annotation for Mesoamerican languages, Mexico City November 2023
Organizer/Instructor

- Organized (with two others) a week-long course in morphosyntactic linguistic annotation to speakers of Mesoamerican languages from Mexico and Guatemala.
- Taught fundamental linguistics concepts around syntax and morphology, in Spanish, focusing on the linguistic structures observed in the Mesoamerican linguistic area.
- Part of the project “Syntactically-annotated corpora for endangered languages in areal contact” (NSF)

Taller de anotación sintáctica, San Cristobal de las Casas, Chiapas November 2021
Instructor

- Helped plan and develop workshop materials for introducing indigenous mexican linguists to the Universal Dependencies framework
- Taught session on morphology and morphological annotation.

San Jose State University, San José, CA Spring 2014
Instructional Assistant, Linguistics and Language Development

- Acted as designated departmental tutor for undergraduate linguistics students (all courses)
- Occasionally lectured for undergraduate Syntax course focused on Role & Reference Grammar.

Multimundos Language School, Santa Cruz, CA Spring 2009
Spanish Language Instructor

- Instructed two classes of between 6-12 third-grade students in beginnning Spanish.
- Designed lesson plans and organized two "Showcases," performances where the students demonstrated to their parents and the community what they had learned.

PUBLICATIONS

Journal Publications

2020

- Rios, J. A., Ling, G., Pugh, R., Becker, D., & Bacall, A. (2020). Identifying critical 21st-century skills for workplace success: A content analysis of job advertisements. *Educational Researcher, 49*(2), 80-89.

Conference & Workshop Papers (Peer-reviewed)

2024

- Pugh, R., Sreedhar, V., and Tyers, F. M. (2024) "Wav2pos: Exploring syntactic analysis from audio for Highland Puebla Nahuatl". *AmericasNLP 2024*.
- Pugh, R., and Tyers, F. M. (2024) "Experiments in Multivariant Natural Language Processing for Nahuatl". *VarDial 2024*.
- Pugh, R., and Tyers, F. M. (2024) "A Universal Dependencies Treebank for Highland Puebla Nahuatl". In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1393–1403, Mexico City, Mexico. Association for Computational Linguistics.

2023

- Tyers, F., Pugh, R., & Berthoud, V. (2023, July). Codex to corpus: Exploring annotation and processing for an open and extensible machine-readable edition of the

Florentine Codex. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)* (pp. 19-29).

- Tyers, F. M., & Pugh, R. (2023, July). A finite-state morphological analyser for Highland Puebla Nahuatl. In *Third Workshop on Natural Language Processing for Indigenous Languages of the Americas* (p. 103).
- Pugh, R., Tyers, F., & Castañeda, Q. (2023, July). Developing finite-state language technology for Maya. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)* (pp. 30-39).
- O'Neil, A., Swanson, D., Pugh, R., Tyers, F., & Um, E. N. (2023, May). Comparing methods of orthographic conversion for Bàsàá, a language of Cameroon. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)* (pp. 97-105).
- Pugh, R., & Tyers, F. (2023). The ITML Submission to the IberLEF2023 Shared Task on Guarani-Spanish Code Switching Analysis. *Procesamiento del Lenguaje Natural*, *71*.

2022

- Pugh, R., Huerta Mendez, M., Sasaki, M., and Tyers, F.M. (2022). Universal Dependencies for Western Sierra Puebla Nahuatl. *LREC 2022 Conference Proceedings* pp.5011-5020.

2021

- Pugh, R. and Tyers, F. M. (2021). Investigating variation in written forms of Nahuatl using character-based language models. *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)* pp.21-27.
- Pugh, R., Tyers, F. M., and Huerta Mendez, M. (2021). Towards an Open Source Finite-State Morphological Analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. *Proceedings of ComputEL4* pp.80-85.

2019

- Riordan, B., Flor, M., & Pugh, R. (2019). How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* pp. 116-126.
- Qian, Y., Lange, P., Evanini, K., Pugh, R., Ubale, R., Mulholland, M., & Wang, X. (2019). Neural approaches to automated speech scoring of monologue and dialogue responses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 8112-8116. IEEE.

- Ramanarayanan, V., Pugh, R., Qian, Y., & Suendermann-Oeft, D. (2019). Automatic Turn-Level Language Identification for Code-Switched Spanish–English Dialog. In *9th International Workshop on Spoken Dialogue System Technology* pp. 51-61. Springer, Singapore.

2018

- Ramanarayanan, V., & Pugh, R. (2018, July). Automatic token and turn level language identification for code-switched text dialog: An analysis across language pairs and corpora. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue* pp. 80-8.
- Evanini, K., Mulholland, M., Ubale, R., Qian, Y., Pugh, R. A., Ramanarayanan, V., & Cahill, A. (2018). Improvements to an Automated Content Scoring System for Spoken CALL Responses: the ETS Submission to the Second Spoken CALL Shared Task. In *INTERSPEECH* pp. 2379-2383.

2017

- Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., ... & Qian, Y. (2017, September). A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* pp. 62-75.

PATENTS

Riordan, B. W., Steimel, K., Flor, M., & Pugh, R. (2023). *"Systems and methods for neural content scoring." U.S. Patent No. 11,790,227*. Washington, DC: U.S. Patent and Trademark Office.

Ramanarayanan, V., Pugh, R., Qian, Y., & Suendermann-Oeft, D. (2022). "Automatic turn-level language identification for code-switched dialog." *U.S. Patent No. 11,238,844*. Washington, DC: U.S. Patent and Trademark Office.

Qian, Y., Evanini, K., Lange, P., Pugh, R. A., & Ubale, R. (2020). "Native language identification with time delay deep neural networks trained separately on native and non-native english corpora." *U.S. Patent No. 10,783,873*. Washington, DC: U.S. Patent and Trademark Office.

PRESENTATIONS

Poster Presentation "Wav2Pos: Exploring syntactic analysis from audio for Highland Puebla Nahuatl." 4th Workshop on NLP for Indigenous Languages of the Americas. June, 2024.

Oral Presentation "Experiments in Multivariant NLP for Nahuatl." The Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). June, 2024.

Oral Presentation "Universal Dependencies for Highland Puebla Nahuatl." 2024 Conference of the North American Chapter of the Association for Computational Linguistics. June, 2024.

Seminar "Indigenous voices past, present and future: Digital language technology for Nahuatl and Maya" *Towards a National Collection*. May, 2024

Poster Presentation "Universal Dependencies for Western Sierra Puebla Nahuatl" Language Resources and Evaluation Conference (LREC), Marseille, France, June 2022.

Poster Presentation "A computational model of Maya morphology" Form and Analysis in Mayan Linguistics (FAMLi), San Cristóbal de las Casas, Chiapas, México, November 2021.

Paper Presentation "Language identification for indigenous languages of México" El Segundo Congreso Internacional de Procesamiento de Lenguaje Natural para Lenguas Indígenas, Online, November 2021.

Poster Presentation "Investigating variation in written forms of Nahuatl using character-based language models" *First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, Online, June 2021.

Paper Presentation "Towards a morphological analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl" 4th Workshop on Computational Methods for Endangered Languages, Online, February 2021.

Workshop "Applied Natural Language Processing for Educational Technology", Open Data Science Conference West, San Francisco, CA, October 2019.

Presentation, "Variable voicing assimilation of Spanish /s/" SJSU Linguistics & TESOL Syposium, San Jose, CA, September 2013.

PROFESSIONAL SERVICE

Workshop Co-Organizer, Jornada de tecnologías de voz, Mexico City, November 2021

Peer-Review for: Language Resources and Evaluation (Journal), NAACL and ACL (Conferences), AmericasNLP (Workshop), SIGDial (Workshop), BEA (Workshop)

LANGUAGES

English: Native Language

Spanish: Native-level fluency

Nahuatl: Intermediate

Maya: Intermediate

COMPUTER SKILLS

Programming: Python, Unix & Linux shell scripting, Regular Expressions, PyTorch, TensorFlow, Keras, Transformers, Scikit-learn, AllenNLP, Natural Language Toolkit (NLTK), Gensim, Stanford NLP Toolkit, SpaCy, Apache Spark, Apache Storm, Hadoop, Web development with HTML+CSS, Javascript, SQL

Applications: ELAN, Praat, Annotatrix

Platforms: Linux, MacOS, Windows

REFERENCES

Dr. Francis Tyers, Assistant Professor in Computational Linguistics
Department of Linguistics, Indiana Unviersity, Bloomington
Email: ftyers@iu.edu

Dr. Vikram Ramanarayanan, Chief Science Officer
Modality.AI
Email: vikram.ramanarayanan@modality.ai