Curriculum Vitae April, 2026

Nay San

Contact Information

rime labs inc.
911 Minna St
San Francisco, CA 94013, USA

E-mail: nay@rime.ai
Web: https://resunay.com
ORCID: 0000-0002-3533-5025

Research Interests

Applied spoken and natural language processing; development of robust speech technologies for under-served languages and speaker populations; documentation & revitalisation of endangered languages

Employment

Apr 2024-pres. **AI/ML Engineer**. rime labs inc. San Francisco, CA.
Jun-Sep 2023 **AI/ML Engineer Intern**. rime labs inc. San Francisco, CA.
Jun-Sep 2022 **Research Intern in AI/ML**. Apple. Cambridge, MA.
2017-2018 **Research Officer**. Australian National University. Canberra, Australia.
2014-2016 **Research Assistant**. Macquarie University. Sydney, Australia.

Education

2018–2024 **PhD, Linguistics**, Stanford University
Thesis title: Improving access to untranscribed speech corpora using AI
Supervisor: Dan Jurafsky

2015–2016 **M.Res., Linguistics**, Macquarie University (Sydney, Australia)
Thesis title: A quantitative analysis of vowel variation in Kaytetye
Supervisor: Michael Proctor

2013–2014 **B.A. (Hons. I), French**, University of Queensland (Brisbane, Australia)
Thesis title: Perceptual assimilation and categorical discrimination of French vowels by Australian-English learners of French
Supervisors: Barbara Hanna, Michael Tyler

2007–2011 **B.Sc., Mathematics/B.A., Philosophy; French**, University of Queensland

Other Academic Experience

Dec 2016 Summer School. ARC Centre of Excellence for the Dynamics of Language, Melbourne, Australia.
Dec 2015 Summer School. ARC Centre of Excellence for the Dynamics of Language, Sydney, Australia.
Feb 2015 Machine Learning Summer School. Sydney, Australia.
Feb 2012 Winter School on Multilingualism across the Lifespan. Fribourg, Switzerland.
Aug–Dec 2010 Exchange semester. University of Lausanne, Switzerland.

APPOINTMENTS

2019–pres. Visiting Fellow. College of Asia & the Pacific. Australian National University. Canberra, Australia.
2018–2022 Affiliate Member. ARC Centre of Excellence for the Dynamics of Language.
2018-2019 Visitor. College of Arts & Social Sciences. Australian National University. Canberra, Australia.

HONOURS AND AWARDS

2016	Summer Research Scholarship (Linguistics). Australian National University.	AUD 3,200
2016	Research Training Pathway Scholarship, Year 2. Macquarie University.	AUD 26,228
2015	Research Training Pathway Scholarship, Year 1. Macquarie University.	AUD 8,000
2014	G.M. Grassie Memorial Prize. University of Queensland.	AUD 800
2013–2014	Dean's Commendations for Academic Excellence. University of Queensland.	
2013	Summer Research Scholarship (Linguistics). University of Queensland.	AUD 2,100
2010	Exchange student allowance. University of Lausanne, Switzerland.	CHF 2,250
2010	Travel grant for exchange studies. University of Queensland.	AUD 1,000
2010	Summer Research Scholarship (Mathematics). University of Queensland.	AUD 2,100
2009	Summer Research Scholarship (Mathematics). University of Queensland.	AUD 2,100

GRANTS

2019	**San, N.**, Foley, B., Disbray, S., Turpin, M. and Simpson, J. *Towards an extensible, open-source picture dictionary template and processing system.* Transdisciplinary and Innovation Grant from the ARC Centre of Excellence for the Dynamics of Language.	AUD 19,809
2017	Harvey, M., Hercus, L., Carew, M. and **San, N.** *Metrical prominence and pre-stopping in Arabana.* Language Documentation Grant from the ARC Centre of Excellence for the Dynamics of Language.	AUD 10,992

RESEARCH ACTIVITY

Conference proceedings

San, N., Paraskevopoulos, G., Arora, A., He, X., Kaur, P., Adams, O. & Jurafsky, D. (2024). Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In *Proceedings of the Sixth Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP-2024)* (pp. 100–112). Association for Computational Linguistics. Available on https://aclanthology.org/2024.sigtyp-1.13

Field, A., Verma, P., **San, N.**, Eberhardt, J. & Jurafsky, D. (2023). Developing speech processing pipelines for police accountability. In *INTERSPEECH 2023*. International Speech Communication Association. doi:10.21437/Interspeech.2023-2109

Bartelds, M., **San, N.**, McDonnell, B., Jurafsky, D. & Wieling, M. (2023). Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the Association for Computational Linguistics conference 2023 (ACL2023)*. Association for Computational Linguistics. doi:10.18653/v1/2023.acl-long.42

San, N., Bartelds, M., Billings, B., De Falco, E., H., F., Safri, J., Sahrozi, W., Foley, B., McDonnell, B. & Jurafsky, D. (2023). Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-6)*. Association for Computational Linguistics. Available on https://aclanthology.org/2023.computel-1.1/

San, N., Bartelds, M., Ògúnrèmí, T., Mount, A., Thompson, R., Higgins, M., Barker, R., Simpson, J. & Jurafsky, D. (2022). Automated speech tools for helping communities process restricted-access corpora for language revival efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-5)* (pp. 41–51). Association for Computational Linguistics. Available on https://aclanthology.org/2022.computel-1.6

San, N. & Turpin, M. (2021). Text-setting in Kaytetye. In R. Bennett, R. Bibbs, M. Loren Brinkerhoff, M. J. Kaplan, S. Rich, A. Rysling, N. van Handel & M. Wax Cavallaro (Eds.), *Supplemental proceedings of the 2020 Annual Meeting on Phonology* (pp. 1–9). doi:10.3765/amp.v9i0.4911

***San, N.**, *Bartelds, M., Browne, M., Clifford, L., Gibson, F., Mansfield, J., Nash, D., Simpson, J., Turpin, M., Vollmer, M., Wilmoth, S. & Jurafsky, D. (2021). Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages. In *Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding workshop (ASRU)*. doi:10.1109/ASRU51503.2021.9688301. (* indicates co-first authorship)

van Esch, D., Foley, B. & **San, N.** (2019). Future directions in technological support for language documentation. In M. Silfverberg (Ed.), *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages (ComputEL-3)* (Vol. 1). Available on https://scholar.colorado.edu/scil-cmel/vol1/iss1/3

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D., Heath, S., Kratochvíl, F., Maxwell-Smith, Z., Nash, D., Olsson, O., Richards, M., **San, N.**, Stoakes, H., Thieberger, N. & Wiles, J. (2018). Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (Elpis). In S. S. Agrawal (Ed.), *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)* (pp. 200–204). doi:10.21437/SLTU.2018-43

Articles

Longpre, S., Biderman, S., Albalak, A., Schoelkopf, H., McDuff, D., Kapoor, S., Klyman, K., Lo, K., Ilharco, G., **San, N.**, Rauh, M., Skowron, A., Vidgen, B., Weidinger, L., Narayanan, A., Sanh, V., Adelani, D. I., Liang, P., Bommasani, R., Henderson, P., Luccioni, S., Jernite, Y. & Soldaini, L. (2024). The responsible foundation model development cheatsheet: A review of tools & resources. *Transactions on Machine Learning Research*.

Harvey, M., **San, N.**, Proctor, M., Panther, F. & Turpin, M. (2023). The Kaytetye segmental inventory. *Australian Journal of Linguistics*. doi:10.1080/07268602.2023.2218270

Foley, B., van Esch, D. & **San, N.** (2022). Managing transcription data for automatic speech recognition with Elpis. In A. Berez-Kroeker, B. McDonnell, E. Koller & L. Collister (Eds.), *The Open Handbook of Linguistic Data Management*. Cambridge, MA: MIT Press. doi:10.7551/mitpress/12200.003.0041

Harvey, M., **San, N.**, Carew, M., Strangways, S., Simpson, J. & Stockigt, C. (2019). Pre-stopping in Arabana. *Australian Journal of Linguistics, 39*(4). doi:10.1080/07268602.2019.1643290

Invited talks

San, N. (2023). Improving access to language documentation corpora using self-supervised models for speech. Phonetics, Phonology and Psycholinguistics Forum (Phorum), UC Berkeley (February, 24) and the Department of Computational Linguistics, University of Zurich (April, 18).

San, N. & van Esch, D. (2019). Accelerating transcription of fieldwork data using machine learning. Fieldwork Forum (FForum), UC Berkeley. October, 3.

San, N. (2018). Quantitative methods in linguistics: An overview. Department of Mathematics & Statistics, La Trobe University, Melbourne. February, 8.

Presentations

Turpin, M. & **San, N.** (2022). The prosodic word in literary traditions of variable line length. Paper presented at the conference of the Annual Meeting of the Linguistics Society of America.

San, N., Disbray, S., Foley, B. & Simpson, J. (2019). Towards an extensible, open-source picture dictionary template and processing system. Paper presented at the conference of the Australasian Association for Lexicography (AustraLex). Australian National University, Canberra, Australia. doi:10.5281/zenodo.3404266

San, N., Carne, M., Carew, M., Harvey, M., Hercus, L. & Simpson, J. (2018). An acoustic analysis of pre-stopping in Arabana. Paper presented at the 18th Australian Languages Workshop (ALW). Marysville, Victoria. doi:10.5281/zenodo.3404271

San, N. (2017). A corpus-based approach to vocalic contrasts in Kaytetye. Paper presented at the Corpus Workshop of the Centre of Excellence for the Dynamics of Language. Melbourne, Australia.

San, N., Proctor, M., Turpin, M., Harvey, M., Ringbauer, K., Ross, A. & Demuth, K. (2015). An acoustic analysis of Kaytetye vowel variability. Paper presented at the 46th Annual Conference of the Australian Linguistic Society, Western Sydney University, Australia.

San, N., Proctor, M., Turpin, M., Harvey, M., Ringbauer, K., Ross, A. & Demuth, K. (2015). Variability of vowels in Kaytetye words. Paper presented at the Arandic Phonetics & Phonology Workshop. Alice Springs, NT, Australia.

San, N. (2015). Visualising the articulatory characteristics of Kaytetye coronal consonants. Paper presented at the Building Corpora for Australian Languages workshop. Australian National University, Canberra.

San, N. & Turpin, M. (2014). Acoustic correlates of stress in Kaytetye words. Paper presented at the 45th Annual Conference of the Australian Linguistic Society (ALS). University of Newcastle, Australia.

San, N. & Turpin, M. (2014). Acoustic correlates of stress in Kaytetye words. Paper presented at the Workshop on Word Stress & Accent. Leiden University, the Netherlands.

Posters

San, N., Bartelds, M. & Jurafsky, D. (2021). Improving access to untranscribed speech by leveraging spoken term detection and self-supervised learning of speech representations. Invited non-archival extended abstract presented at SigTyp 2021.

San, N. & Turpin, M. (2020). Text-setting in Kaytetye. Poster presented at the 2020 Annual Meeting on Phonology (Online).

San, N. (2016). Using version control to facilitate a reproducible and collaborative workflow in acoustic phonetics. In C. Carignan & M. D. Tyler (Eds.), *Proceedings of the 16th Australasian Speech Science and Technology Association conference* (pp. 341–344). Available on http://www.assta.org/sst/2016/papers/San_SST2016.pdf

San, N. (2013). The perception of French vowels by Australian-English-speaking learners. Poster presented at the 21st Annual Conference of the Australian Society of French Studies (ASFS). The University of Queensland, Brisbane, Australia.

Research Assistantships

2017–2018 Warlpiri Lexicography Project. Australian National University. Supervisor: Jane Simpson.
2016–2017 Developing *Yerrampe*: A Kaytetye-to-English multimedia dictionary. Sydney Conservatorium of Music. Supervisor: Myfany Turpin.
2014–2015 Investigating tongue shaping of Kaytetye obstruents using ultrasound imaging. Macquarie University. Supervisors: Michael Proctor & Katherine Demuth.
Feb–Jun 2014 Examining the acoustic correlates of stress in Kaytetye words. University of Queensland. Supervisor: Myfany Turpin.

Software

- Yinarlingi: An R package for testing Warlpiri lexicon data (Tidylex configured with settings for Warlpiri). https://coedl.github.io/yinarlingi
- Tidylex: An R package for tidying and testing semi-structured lexicographical data. https://coedl.github.io/tidylex
- Phonpack: A package of fun(ctions) for doing phonetics in R. https://github.com/fauxneticien/phonpack

Teaching, as Teaching Assistant

Stanford University

2024 Phonetics, Winter Quarter (Instructor: Chelsea Sanker).
2023 Introduction to Psycholinguistics, Autumn Quarter (Instructor: Judith Degen).
2021 Phonetics, Winter Quarter (Instructor: Meghan Sumner).
2020 Introduction to Phonology, Winter Quarter (Instructor: Arto Anttila).

Macquarie University

2016 Speech Physiology, Semester 1 (Instructor: Michael Proctor).

University of Queensland

2009–2011 Calculus & Linear Algebra I (4 Semesters; Instructor: Phil Isaac).
2009–2011 Calculus & Linear Algebra II (4 Semesters; Instructor: Phil Isaac).
2010 Multivariate Calculus & Differential Equations, Semester 1 (Instructor: Phil Isaac).
2009 Discrete Mathematics I, Semester 2 (Instructor: Murray Elder).

Teaching, as Workshop Instructor

2023 *Introduction to Elpis for developing Automatic Speech Recognition for Field Data.* Co-presented with Daan van Esch at the Department of Linguistic and Cultural Evolution's workshop on Linguistic Analysis/Data Management at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. April, 4-5.
2021

Semi-automated transcription for Language Documentation with Elpis. Co-presented with Ben Foley, Nicholas Lambourne, and Daan van Esch at the 7th International Conference on Language Documentation and Conservation (Online). University of Hawai'i at Mānoa. March, 5-6.

2019 *Transcription Acceleration for Language Documentation with ELPIS.* Co-presented with Ben Foley and Daan van Esch at the 6th International Conference on Language Documentation and Conservation. University of Hawai'i at Mānoa. March, 1-2.

2018 *Docker: An overview for linguists.* Presented at the Transcription Acceleration Project workshop. Melbourne, Australia. February, 4.

2017 *Visualise your own vowels: A short introduction to Praat for beginners.* 1-day tutorial at the Summer School of the ARC Centre of Excellence for the Dynamics of Language. Canberra, Australia. November, 29.

2017 *Formalising data structures with the Nearley toolkit.* University of Queensland. September, 28.

2017 *Introduction to data wrangling using R.* University of Melbourne, Australia. May, 1.

2015 *A beginner's guide to programming in R and Matlab.* Child Language Lab, Macquarie University. May, 4.

2014 *LaTeX for Linguistics.* School of Languages & Comparative Cultural Studies. University of Queensland. Weekly tutorials, April–May.

GUEST LECTURES AND PUBLIC OUTREACH

2018 *The building blocks of spoken language.* Presentation to visiting final-year high schoolers at the National Youth Science Forum. Australian National University. Canberra, Australia. January, 10.

OTHER TEACHING EXPERIENCE

2012 Freelance English language instructor. Passmore College, Marburg, Germany.

2011-2012 Foreign language teaching assistant. Ministry of Education, Aurillac, France.

SERVICE

Stanford University

2021-2023 Corpus TA.

Jan-Mar 2021 Admissions Committee.

2019-2020 Colloquium Committee.

2018-2019 Social Committee.

University of Queensland

2014 Honours students representative, Semester 2. School of Languages and Cultures.

Elsewhere

2024 Reviewer, INTERSPEECH 2024.

2024 Reviewer, Third Annual Meeting of the ELRA-ISCA Special Interest Group on Under-resourced Languages (SIGUL 2024).

2024 Reviewer, Seventh Workshop on Computational Methods in the Study of Endangered Languages (ComputEL-7).

2023, 2024 Reviewer, Journal of Open Source Software.

2024 Reviewer, PeerJ Computer Science.
2016 Reviewer, Journal of the Canadian Acoustical Association.

Non-academic service

2021-2023 Volunteer tutor. Refugee and Immigrant Transitions. Oakland.
2020-2021 Volunteer tutor. Berkeley Public Schools Fund.
2017 Volunteer tutor. Migrant and Refugee Settlement Services. Canberra.
2016 Volunteer. Refugee Action Coalition. Sydney.

LANGUAGES

English (native), French (proficient, CEFR C1), German (conversational, CEFR B2), Burmese (heritage)