Dr Jonathan Dunn (he/him)

Department of Linguistics
University of Illinois at Urbana-Champaign

jedunn@illinois.edu
www.jdunn.name
www.github.com/jonathandunn

Specialization

Computational Linguistics; Cognitive Linguistics; Corpus Linguistics; Syntax & Semantics

Overview

I am a computational linguist working to model both (i) the emergence of grammatical structure within individuals and (ii) variation in that structure across populations. To carry out these kinds of natural experiments, I've had to first develop large multi-lingual corpora. In practical terms, my work has also focused on the impact that linguistic variation has on models in NLP and on improving corpora for low-resource contexts.

Appointments held

2024– Associate Professor in Linguistics, University of Illinois at Urbana-Champaign
2023-2023 Senior Lecturer Above the Bar in Linguistics, University of Canterbury (NZ)
2021-2022 Senior Lecturer in Linguistics, University of Canterbury (NZ)
2018-2020 Lecturer in Linguistics, University of Canterbury (NZ)
2015-2018 Research Assistant Professor in Computer Science, Illinois Institute of Technology
2015-2018 Visiting Scientist, National Geospatial-Intelligence Agency
2014-2015 Post-Doc in Computer Science, Illinois Institute of Technology

Education

2013 PhD, Purdue University
Dissertation: *Automatic Identification of Metaphoric Utterances*
Advisor: Victor Raskin

2010 MA, Purdue University
Thesis: *Towards a Computational Model of Metaphor*
Advisor: Victor Raskin

2008 BA, Classics, Hillsdale College

Books

Dunn, J. (2022). *Natural Language Processing for Corpus Linguistics*. Elements in Corpus Linguistics. Cambridge University Press. Interactive Labs

Reviews: *Natural Language Engineering*, *Corpus Pragmatics*, *Int'l Journal of Corpus Linguistics*

Dunn, J. (Draft). *Computational Construction Grammar: A Usage-Based Approach*. Elements in Cognitive Linguistics. Cambridge University Press. (Manuscript).

MOOCs on edX

Now over 14,000 students

Text Analytics 1: Introducing Natural Language Processing

Text Analytics 2: Visualizing Natural Language Processing

Visiting and Leadership Positions

2021-2023 Theme Leader at the New Zealand Institute for Language, Brain and Behaviour

2022 Visiting Scholar at the Center for Spatial Data Science, University of Chicago

2015 Faculty, LSA Linguistic Institute, University of Chicago

Major Funding

2024-2027 Marsden Fund Standard Grant (NZL)
Does Machine-Assisted Writing Erase Linguistic Diversity?
Co-PI: $660,000 (NZD)

2020-2022 Science for Technological Innovation, National Science Challenges, MBIE (NZL)
Domain Adaptation to Support Polynesian Language Technology
PI: $200,000 (NZD)

2015-2018 Visiting Scientist Research Fellowship, Oak Ridge Institute for Science and Education
Geolinguistics
PI: Approx. $500,000 (USD)

Awards and Other Funding

2023 Early Career Research Award in Humanities, Social Sciences or Creative Arts
Faculty of Arts, University of Canterbury
Award: $2,000 (NZD)

2021 Early Career Researcher Development Fund, University of Canterbury
Towards Equitable Language Models
Grant: $5,000 (NZD)

2020 Teaching Development Grant, University of Canterbury
Syntactic Annotations for Māori Corpora
Grant with Jeanette King: $5,000 (NZD)

2019 Support Grant, University of Canterbury
Automating Tourist Profiles By Integrating Spatial and Textual Artificial Intelligence
Grant with Ben Adams and Girish Prayag: $5,000

2017 Certificate of Merit, Oak Ridge Institute for Science and Education
Award: $3,000 (USD)

2014-2015 IC Postdoctoral Research Fellowship, Office of the Director of National Intelligence
Computational Cognitive Stylistics For Multi-Modal Identity Analytics

2013 Bilsland Dissertation Fellowship, Purdue University Graduate School
Automated Identification of Metaphoric Utterances

2011-2012 Bilsland Strategic Initiatives Grant, Purdue University Graduate School
Evaluating Cross-Listed Graduate Courses

Papers

My research is situated at intersection between linguistics and natural language processing. Thus, I have published widely in both fields. Here I am distinguishing between NLP proceedings and linguistics journals for convenience. Both represent peer-reviewed publications.

2023

1. Li, H.; **Dunn, J.***; & Nini, A. (2023). "Register Variation Remains Stable Across 60 Languages." *Corpus Linguistics and Linguistic Theory*, 19(3): 397-426. (Q1 Journal in Linguistics). Link. **Corresponding author*.

2. **Dunn, J.** (2023). "Syntactic variation across the grammar: Modelling a complex adaptive system." *Frontiers in Complex Systems*. DOI: 10.3389/fcpxs.2023.1273741.

3. **Dunn, J.** (2023). "Exploring the Constructicon: Linguistic Analysis of a Computational CxG." In *Proceedings of the First International Workshop on Construction Grammars and NLP*. Association for Computational Linguistics.

4. **Dunn, J.** (2023). "Variation and Instability in Dialect-Based Embedding Spaces." In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*. Association for Computational Linguistics. (Workshop @ Top 10 Venue in NLP). Link.

5. Wong, S. G.-J., Durward, M., Adams, B., & **Dunn, J.** (2023). "Cantnlp@ LT-EDI-2023: Homophobia/transphobia detection in social media comments using spatio-temporally retrained language models." In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*. Link.

2022

6. **Dunn, J.** (2022). "Exposure and Emergence in Usage-Based Grammar: Computational Experiments in 35 Languages." *Cognitive Linguistics*, 33(4): 659-699. (Q1 Journal in Linguistics). Data. Link.

7. **Dunn, J.** & Wong, S. G.-J. (2022). "Stability of Syntactic Dialect Classification Over Space and Time." In *Proceedings of the International Conference on Computational Linguistics*, COLING. 26-36. (Top 10 Venue in NLP). Link.

8. **Dunn, J.**; Li, H.; & Sastre, D. (2022). "Predicting Embedding Reliability in Low-Resource Settings Using Corpus Similarity Measures." In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC. European Language Resources Association. 6461-6470. (Top 10 Venue in NLP). Link.

9. **Dunn, J.** & Nijhof, W. (2022). "Language Identification for Austronesian Languages." In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC. European Language Resources Association. 6530-6539. (Top 10 Venue in NLP). Link.

10. Li, H. & **Dunn, J.*** (2022). "Corpus Similarity Measures Remain Robust Across Diverse Languages." *Lingua*, 275: 103377. (Q1 Journal in Linguistics). Link. ** Corresponding author*.

11. **Dunn, J.** (2022). "Cognitive Linguistics Meets Computational Linguistics: Construction Grammar, Dialectology, and Linguistic Diversity." In Tay, D. & Xie Pan, M. (eds.), *Data Analytics in Cognitive Linguistics: Methods and Insights*. 273-308. Berlin: De Gruyter. Link

2021

12. **Dunn, J.** & Tayyar Madabushi, H. (2021). "Learned Construction Grammars Converge Across Registers Given Increased Exposure." In *Proceedings of the Conference on Computational Natural Language Learning*, CoNLL. Association for Computational Linguistics. 268-278. (Top 10 Venue in NLP). Link.

13. **Dunn, J.** (2021). "Representations of Language Varieties Are Reliable Given Corpus Similarity Measures." In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects* @ EACL. Association for Computational Linguistics. 28-38. (Workshop @ Top 10 Venue in NLP). Link.

14. **Dunn, J.** & Nini, A. (2021). "Production vs Perception: The Role of Individuality in Usage-Based Grammar Induction." In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* @ NAACL. Association for Computational Linguistics. 149-159. (Workshop @ Top 10 Venue in NLP). Link.

2020

15. **Dunn, J.** (2020). "Mapping Languages: The Corpus of Global Language Use." *Language Resources and Evaluation*, 54(4): 999-1,018. (Q1 Journal in Linguistics). Link.

16. **Dunn, J.** & Adams, B. (2020). "Geographically-Balanced Gigaword Corpora for 50 Language Varieties." In *Proceedings of the Conference on Language Resources and Evaluation*, LREC. European Language Resources Association. 2,521-2,529. (Top 10 Venue in NLP). Link.

17. **Dunn, J.**; Coupe, T.; & Adams, B. (2020). "Measuring Linguistic Diversity During COVID-19." In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science* @ EMNLP. Association for Computational Linguistics. 1-10. (Workshop @ Top 10 Venue in NLP). Link.

18. **Dunn, J.** (2020). "Ontological and Grammatical Constraints on Metaphor Productivity." In Attardo, S. (ed.), *Script-Based Semantics: Foundations and Applications. Essays in Honor of Victor Raskin*. Berlin: De Gruyter. 55-76. Link

19. Mohammadhassan N., Mitrovic A., Neshatian K. & **Dunn J.** (2020). "Automatic Quality Assessment of Comments in Active Video Watching Using Machine Learning Techniques." In So H-J; Rodrigo M; Mason J; Mitrovic A (Eds). *Proceedings of the 28th International Conference on Computers in Education*. I: 1-10. Taiwan: Asia-Pacific Society for Computers in Education. Link

20. Mohammadhassan N., Mitrovic A., Neshatian K. & **Dunn J.** (2020) "Developing Personalized Nudges to Improve Quality of Comments in Active Video Watching." In *Proceedings of 28th International Conference on Computers in Education* 2: 766-769. Taiwan: Asia-Pacific Society for Computers in Education. Link

2019

21. **Dunn, J.** (2019). "Global Syntactic Variation in Seven Languages: Towards a Computational Dialectology." *Frontiers in Artificial Intelligence: Language and Computation.* Section on Computational Sociolinguistics. 10.3389/frai.2019.00015 (Altmetric: Top 10%). Link.

22. **Dunn, J.** (2019). "Modeling Global Syntactic Variation in English Using Dialect Classification." In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects* @ NAACL. Association for Computational Linguistics. 42-53. (Workshop @ Top 10 Venue in NLP). Link.

23. **Dunn, J.** (2019). "Frequency vs. Association for Constraint Selection in Usage-Based Construction Grammar." In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* @ NAACL. Association for Computational Linguistics. 117-128. (Workshop @ Top 10 Venue in NLP). Link.

24. **Dunn, J.** & Adams, B. (2019). "Mapping Languages and Demographics with Georeferenced Corpora." In *Proceedings of GeoComputation 19*. Link

2018

25. **Dunn, J.** (2018). "Multi-Unit Association Measures: Moving Beyond Pairs of Words." *International Journal of Corpus Linguistics*, 23(2): 183-215. (Q1 Journal in Linguistics). Link

26. **Dunn, J.** (2018). "Finding Variants for Construction-Based Dialectometry: A Corpus-Based Approach to Regional CxGs." *Cognitive Linguistics*, 29(2): 275-311. (Q1 Journal in Linguistics). Link.

27. **Dunn, J.** (2018). "Modeling the Complexity and Descriptive Adequacy of Construction Grammars." In *Proceedings of the Society for Computation in Linguistics*. Association for Computational Linguistics. 81-90. Link.

2017

28. **Dunn, J.** (2017). "Computational Learning of Construction Grammars." *Language and Cognition*, 9(2): 254-292. (Q1 Journal in Linguistics). Link.

2016

29. **Dunn, J.**; Argamon, S.; Rasooli, A.; & Kumar, G. (2016) "Profile-Based Authorship Analysis." *Digital Scholarship in the Humanities*, 31(4): 689-710. (Q1 Journal in Linguistics). Link.

2015

30. **Dunn, J.** (2015) "Modeling Abstractness and Metaphoricity." *Metaphor & Symbol*, 30(4): 259-289. (Q1 Journal in Linguistics). Link.

31. **Dunn, J.** (2015). "Three Types of Metaphoric Utterances That Can Synthesize Theories of Metaphor." *Metaphor & Symbol*, 30(1): 1-23. (Q1 Journal in Linguistics). Link.

2014

32. **Dunn, J.** (2014). "Measuring Metaphoricity." In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL. Association for Computational Linguistics. 745-751. (Top 10 Venue in NLP). Link.

33. **Dunn, J.** (2014). "Multi-Dimensional Abstractness in Cross-Domain Mappings." In *Proceedings of the Workshop on Metaphor in NLP* @ ACL. Association for Computational Linguistics. 27-32.
(Workshop @ Top 10 Venue in NLP). Link.

34. **Dunn, J.**; Beltran de Heredia, J.; Burke, M.; Gandy, L.; Kanareykin, S.; Kapah, O.; Taylor, M.; Hines, D.; Frieder, O.; Grossman, D.; Howard, N.; Koppel, M.; Morris, S.; Ortony, A.; & Argamon, S. (2014). "Language-Independent Ensemble Approaches to Metaphor Identification." In *Proceedings of the 28th Conference on Artificial Intelligence: Workshop on Cognitive Computing for Augmented Human Intelligence*. 6-12.
(Workshop @ Top 10 Venue in AI). Link.

2013

35. **Dunn, J.** (2013). "How Linguistic Structure Influences and Helps To Predict Metaphoric Meaning." *Cognitive Linguistics*, 24(1): 33-66.
(Q1 Journal in Linguistics). Link.

36. **Dunn, J.** (2013). "Evaluating the Premises and Results of Four Metaphor Identification Systems." In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, Vol. 1*. Heidelberg: Springer. 471-486. Link

37. **Dunn, J.** (2013). "What Metaphor Identification Systems Can Tell Us About Metaphor-in-Language." In *Proceedings of the Workshop on Metaphor in NLP* @ NAACL. Association for Computational Linguistics. 1-10.
(Workshop @ Top 10 Venue in NLP). Link.

2011

38. **Dunn, J.** (2011). "Gradient Semantic Intuitions of Metaphoric Expressions." *Metaphor & Symbol*, 26(1): 53-67.
(Q1 Journal in Linguistics). Link.

Teaching

Computational Linguistics

Natural Language Processing (COSC 441: Graduate-level)
University of Canterbury
Recent Teaching Evaluation: 4.75 (out of 5)

Text Analytics (LING 223: Undergraduate-level)
University of Canterbury
Recent Teaching Evaluation: 4.90 (out of 5)

Linguistics

Topics in Syntactic Theory (LING 306: Undergraduate-level)
University of Canterbury
Teaching Evaluation: 4.50 (out of 5)

Grammatical Structure (LING 217: Undergraduate-level)
University of Canterbury
Teaching Evaluation: 4.64 (out of 5)

Forensic Linguistics (LING 225: Undergraduate-level)
University of Canterbury
Teaching Evaluation: 4.70 (out of 5)

General Linguistics

English Structures (LING 400: Graduate-level Intro)
University of Canterbury
Teaching Evaluation: 4.90 (out of 5)

The English Language (LING 101: Undergraduate-level Intro)
University of Canterbury
Teaching Evaluation: 4.10 (out of 5)

Previous Teaching

Data-Driven Computational Pragmatics: LSA Linguistic Institute @ University of Chicago
Introduction to Linguistics (LING 227) @ Purdue University
Classroom Communication for International Graduate Students @ Purdue University
First Year Composition for International Students (ENGL 1011) @ Purdue University
First-Year Composition (ENGL 101) @ Purdue University
Residential Learning Community (ENGL 101) @ Purdue University

Teaching Awards and Recognition

2012 Quintilian Award (Top 10% of Instructor Evaluations), Purdue University, Department of English
2012 Graduate Teacher Certificate, Center for Instructional Excellence, Purdue University

Supervisions

2021-Present Sidney Wong (Ph.D. Student, Linguistics; With Ben Adams)
2021-Present Matthew Durward (Ph.D. Student, Linguistics; With Chris Thomson)
2020-2022 Negar Mohammadhassan (Ph.D. Student, Computer Science; Primary supervisor: Tanja Mitrovic)
2021-2022 Haipeng Li (Post-doc, New Zealand Institute for Language Brain and Behaviour)

Software and Data

earthLings: Computational Linguistic Atlas
A 440 billion word geo-referenced corpus from the web and social media
https://www.earthLings.io
`https://www.github.com/jonathandunn/earthLings`

C2xG: Computational Construction Grammar
Python package for learning, evaluating, and annotating CxGs
`https://www.github.com/jonathandunn/c2xg`

idNet: Identifying minority languages and regional varieties of majority languages
Python package for state-of-the-art LID and DID
https://www.github.com/jonathandunn/idNet

pacificLID: Identifying languages and code-switching with a focus on Austronesian languages
Python Version: https://github.com/jonathandunn/pacific_CodeSwitch
fastText Version: https://jdunn.name/corpora/

text_analytics: Package for teaching computational linguistics in Python
https://www.github.com/jonathandunn/text_analytics

corpus_similarity: Package for measuring corpus similarity and homogeneity
https://www.github.com/jonathandunn/corpus_similarity

Common Crawl Corpus: Corpus building from the common crawl data
https://www.github.com/jonathandunn/common_crawl_corpus

Service to the Discipline

EDITORIAL WORK

2018-Present Associate Editor, *Frontiers in Artificial Intelligence* (Language and Computation)

AD HOC REVIEWING

2024 Endeavour Fund, MBIE, New Zealand
Royal Society Open Science
2023 European Association for Computational Linguistics (EACL)
Annual Meeting of the Association for Computational Linguistics (ACL)
International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)
Empirical Methods in Natural Language Processing (EMNLP)
Conference on Computational Natural Language Learning (CoNLL)
Association for Computational Linguistics Rolling Review (ARR) (x1)
17th Linguistic Annotation Workshop (LAW-XVII)
Cognitive Science
Psychological Review
Austrian Science Fund
European Science Foundation
2022 Conference on Computational Natural Language Learning (CoNLL)
Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)
Asia-Pacific Association for Computational Linguistics (AACL)
Association for Computational Linguistics Rolling Review (ARR) (x2)
Workshop on Figurative Language Processing
Fund for Scientific Research (F.R.S.-FNRS), Belgium
2021 North American Association for Computational Linguistics (NAACL)

Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)
Empirical Methods in Natural Language Processing (EMNLP)
Conference on Computational Natural Language Learning (CoNLL)
Joint 15th Linguistic Annotation Workshop and 3rd Designing Meaning Representations
2020 Association for Computational Linguistics (ACL)
Conference of the European Chapter of the Association for Computational Linguistics (EACL)
Conference on Empirical Methods in Natural Language Processing (EMNLP)
Conference on Computational Natural Language Learning (CoNLL)
Asia-Pacific Association for Computational Linguistics (AACL)
Linguistic Annotation Workshop (LAW XIV)
Workshop on Figurative Language Processing
Workshop on NLP and Computational Social Science
Transactions on Audio, Speech and Language Processing
2019 *Journal of Cognitive Science*
IEEE Transactions on Audio, Speech and Language Processing
Conference on Computational Natural Language Learning (CoNLL)
Association for Computational Linguistics (ACL)
Linguistic Annotation Workshop (LAW XIII)
North American Association for Computational Linguistics (NAACL)
2018 Linguistic Society of America, 2019 Annual Meeting
Joint LAW-MWE-CxG Workshop (@ COLING)
IEEE Transactions on Audio, Speech and Language Processing (2x)
2017 Social Sciences and Humanities Research Council of Canada
Cognitive Linguistics
Language & Cognition
Linguistic Society of America, 2018 Annual Meeting
Corpus Linguistics and Linguistic Theory
The Computer Journal
2016 Linguistic Society of America, 2017 Annual Meeting
Fourth Workshop on Metaphor in NLP (@ NAACL)
2015 Linguistic Society of America, 2016 Annual Meeting
Synthesis Lectures in Human Language Technologies, Book Series (Morgan Claypool)
Journal of Artificial Intelligence Research
Language Resources and Evaluation
Third Workshop on Metaphor in NLP (@ NAACL)
2014 *IEEE Transactions on Audio, Speech and Language Processing*
Croatian Science Foundation
Second Workshop on Metaphor in NLP (@ ACL)
2012 International Conference on Computational Linguistics (COLING)

University Service

Committees

2023

Academic Programmes Committee, College of Arts, University of Canterbury
2023 Hiring Committee, Department of Linguistics, University of Canterbury
2023 Hiring Committee, NZILBB, University of Canterbury
2020-2021 Academic Programmes Committee, College of Arts, University of Canterbury
2018-2023 Academic Programmes Committee, College of Science, University of Canterbury
2020-2022 Steering Group for Bachelors of Data Science Degree, University of Canterbury
2019-2020 Research Committee, School of Language, Social & Political Sciences, University of Canterbury
2019 Hiring Committee, Department of Linguistics, University of Canterbury

Administrative Achievements

2020 Developed and delivered one of the flagship MOOCs for the University of Canterbury
2020 Developed the *Computational Linguistics* undergraduate major
2020 Helped develop the undergraduate degree, *Bachelors of Data Science*

Partially Peer-Reviewed Publications

2018 1. Dunn, J. (2018). "Recursively Emerging Structure: A Discovery-Device CxG." In *Proceedings of the Chicago Linguistics Society*, 53: 71-86.

2017 2. Dunn, J. (2017). "Learnability and Falsifiability of Construction Grammars: A Learning-Based Approach." *Proceedings of the Linguistic Society of America Annual Meeting*, 2(1):1-15.

3. Dunn, J. (2017). *Automating Human Geography With Dialectology*. Technical Report for National Geospatial-Intelligence Agency. Washington, D.C.

2015 4. Dunn, J. (2015). "Review of Frames and Constructions in Metaphoric Language (Constructional Approaches to Language, 14)." *Cognitive Linguistics*, 26(2): 371-375.

5. Dunn, J. (2015). "Review of The Semantic Representation of Natural Language." *Studies in Language*, 39(2): 492-500.

2013 6. Dunn, J. (2013). "Review of Converging Evidence: Methodological and Theoretical Issues for Linguistic Research. (Human Cognitive Processing, 33)." *Cognitive Linguistics*, 24(4): 711-717.

Invited Talks (Since 2022)

2024 1. "Teaching Computational Linguistics to Non-Computational Students: Lessons from a MOOC." University of La Rioja

2. "Linguistic Diversity in the Digital World: From Perception to Production." University of Galway

2023 3. "Emerging Structure in Computational Construction Grammar." Plenary at the Construction Grammars and NLP Workshop, Georgetown University (CxGs+NLP)

4. "Learnability and Variability in Computational Linguistics: Implications for Second Language Acquisition." Purdue University

2022

5. "Improving Corpus Resources for Low-Resource Languages." University of Galway
6. "Computational Linguistics for Socially-Responsible Technology." University of Wisconsin-Madison
7. "Linguistic Variation in NLP." Pacific Northwest National Lab