Sebastin Santy

PhD student, UNIVERSITY *of* WASHINGTON

sebast.in @ ssanty@uw.edu github.com/sebastinsanty Google Scholar

CSE2 230, Bill & Melinda Gates Center, 3800 E Stevens Way NE, Seattle, WA 98195

Education

University of Washington, *Seattle, USA* 09/2021 - Present
Ph.D. in Computer Science (ongoing)
Advisor. Ranjay Krishna

Birla Insitute of Technology and Science, *Goa, India* 08/2015 - 07/2019
B.E. in Electronics & Instrumentation Engineering

Experience

Allen Institute for Artificial Intelligence, *Seattle, USA* 06/2022 - 09/2022
Research Intern with Maarten Sap, Ronan Le Bras

Microsoft Research, *Bangalore, India* 07/2019 - 07/2021
AI Center Fellow with Kalika Bali, Monojit Choudhury, Tanuja Ganu

Microsoft Research, *Bangalore, India* 01/2019 - 07/2019
Research Intern with Kalika Bali, Monojit Choudhury

Carnegie Mellon University, *Pittsburgh, USA* 07/2018 - 12/2018
Research Intern with David S. Touretzky

University College London, *London, UK* 05/2018 - 07/2018
Research Intern with Emine Yilmaz, Rishabh Mehrotra

Julia Language, *Remote* 05/2018 - 08/2018
Google Summer of Code Intern with Lyndon White

Mozilla, *Remote / Austin, USA* 05/2017 - 08/2017
Google Summer of Code Intern with Dylan Hardison

Publications

S=In Submission, C=Conference, W=Workshop, P=Poster/Demo, J=Journal

C.7 **Measuring Design Biases and Positionality of NLP Datasets and Models.**
Sebastin Santy*, Jenny Liang*, Ronan Le Bras, Katharina Reinecke, Maarten Sap (* = Equal Contribution)
Annual Conference of the Association for Computational Linguistics, Toronto, Canada
★ **Outstanding Paper** **[ACL 2023]**

C.6 **Language Translation as a Socio-Technical System: Case-Studies of Mixed-Initiative Interactions.**
Sebastin Santy, Kalika Bali, Monojit Choudhury, Sandipan Dandapat, Tanuja Ganu, Anurag Shukla, Jahanvi Shah, Vivek Seshadri
ACM SIGCAS Conference on Computing and Sustainable Societies (Virtual) **[COMPASS 2021]**

C.5 **Use of Formal Ethical Reviews in NLP Literature: Historical Trends and Current Practices.**
Sebastin Santy, Anku Rani, Monojit Choudhury
Annual Conference of the Association for Computational Linguistics (Virtual) **[ACL Findings 2021]**

C.4 **The State and Fate of Linguistic Diversity and Inclusion in the NLP World.**
Sebastin Santy*, Pratik Joshi*, Amar Budhiraja*, Kalika Bali, Monojit Choudhury (* = Equal Contribution)
Annual Conference of the Association for Computational Linguistics (Virtual) **[ACL 2020]**

C.3 **Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi.**
Sebastin Santy*, Devansh Mehta*, Ramaravind Mothilal, Brij Mohan Lal Srivastava, Alok Sharma, Anurag Shukla, Vishnu Prasad, Venkanna U, Amit Sharma and Kalika Bali (* = Equal Contribution)
International Conference on Language Resources and Evaluation (Virtual) **[LREC 2020]**

C.2 **Unsung Challenges of Building and Deploying Language Technologies for LRL Communities.**
Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, Kalika Bali
16th International Conference on Natural Language Processing, Hyderabad, India **[ICON 2019]**

C.1 **BITS Darshini: A Modular, Concurrent Protocol Analyzer Workbench.**
Prasad Talasila, Mihir Kakrambe, Anurag Rai, Sebastin Santy, Neena Goveas, Bharat Deshpande
19th ACM International Conference on Distributed Computing and Networking, Varanasi, India **[ICDCN 2018]**

P.3 **Deploying Language Technologies for Underserved Communities.** [Invited Poster]
Kalika Bali, Monojit Choudhury, Sunayana Sitaram, Sebastin Santy
UNESCO International Conference on Language Technologies for All, Paris, France **[LT4All 2020]**

P.2 **INMT: Interactive Neural Machine Translation Prediction.**
Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, Kalika Bali
Conference on Empirical Methods in Natural Language Processing, Hong Kong [Systems Demo] **[Demo, EMNLP 2019]**

P.1 **Towards Task Understanding in Visual Settings.**
Sebastin Santy, Wazeer Zulfikar, Rishabh Mehrotra, Emine Yilmaz.
33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA [Student Abstract] **[Abstract, AAAI 2019]**

W.2 **BERTologiCoMix: How does Code-Mixing interact with Multilingual BERT?**
Sebastin Santy, Anirudh Srinivasan, Monojit Choudhury
Workshop on Domain Adaptation for NLP, EACL (Virtual) **[AdaptNLP, EACL 2021]**

W.1 **CoSSAT: Code-Switched Speech Annotation Tool.**
Sanket Shah, Pratik Joshi, Sebastin Santy, Sunayana Sitaram
Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP, EMNLP, Hong Kong **[AnnoNLP, EMNLP 2019]**

J.1 **DataDepsGenerators.jl: Automatic generation of DataDeps.jl registration code.**
Lyndon White, Sebastin Santy
Journal for Open Source Software **[JOSS 2018]**

Work Mentions and Media Coverage

Center for Democracy & Technology on *"Lost in Translation: Large Language Models in Non-English [...]"* 05/2023
The Economic Times on *"Microsoft Research India is creating tools to help preserve fast disappearing languages"* 03/2023
Indian Express on *"How Microsoft's Project ELLORA is helping small languages like Gondi become [...]"* 01/2023
Microsoft Stories on *"Microsoft Research project helps languages survive — and thrive"* 01/2023
Slate on *"An A.I. Translation Tool Can Help Save Dying Languages. But at What Cost?"* 01/2023
U.S. Federal Trade Commission (FTC) on *"Who is being left behind? Enforcement Priorities for a Tech [...]"* 08/2022
The Gradient on *"New Technology, Old Problems: The Missing Voices in Natural Language Processing"* 04/2022
Emily Bender [Blog] on *"On Academic Freedom and Ethics Review"* 06/2021
Quartz on *"Data scientists are trying to make the internet accessible in every language"* 10/2020
Sebastian Ruder [Blog] on *"Why You Should Do NLP Beyond English"* 08/2020
Mint Lounge on *"Now a unique machine translation tool from Hindi to Gondi"* 08/2020
Underrated ML on *"Language Independence and Material Properties"* 06/2020
NLP Newsletter on *"Reviewing, Taking stock, Theme papers, Poisoning and stealing models, multimodal [...]"* 06/2020
SIGTYP on *"Recent Developments in Computational Typology & Multilingual NLP"* 04/2020
Times of India on *"Chhatisgarh: Now, tribals can listen to news in their own Gondi Language"* 08/2019
The Caravan on *"Mind Our Language: The Koitur community is reclaiming their linguistic identity despite [...]"* 08/2019
Hindustan Times on *"Gonds in Chhattisgarh get app for news in their language"* 08/2019
ETV on *"Now tribals can hear news and stories in their own language [Translated]"* 08/2019

Awards and Honors

Outstanding Paper Award @ ACL 2023 *"NLPositionality: Characterizing Design Biases of Datasets and Models"*

People's Choice Award @ UW CSE Annual Research Showcase 2022 [📰] *"Beyond WEIRDness of NLP"*

Outstanding Reviewer Award EMNLP 2021, CHI 2022

Talks and Tutorials

EMNLP 2023 Tutorial on *"Designing, Evaluating, and Learning from Humans Interacting with NLP Models"*
w/ Sherry Tongshuang Wu (CMU) and Diyi Yang (Stanford) — Singapore, 12/2023

UW CSE 512 Data Visualization Tutorial on *"D3.js (Data-Driven Documents)"* — Seattle, 04/2023

UW CSE 517 NLP Guest Lecture on *"Human-NLP Interaction"* — Seattle, 03/2023

UW CSE 440 HCI Guest Lecture on *"Societal Implications of Design and Technology"* — Seattle, 11/2022

NLP with Friends on *"The State and Fate of Linguistic Diversity and Inclusion in the NLP World"* — Remote, 09/2020

PyData 2018 on *"Repeatable Data Setup for Repeatable Science using Julia"* — NYC, 10/2018

In conv. with Dr. APJ Abdul Kalam (President of India) on *"Technology progress in India"*
televised on the National Geographic Channel — New Delhi, 06/2010

Teaching and Leadership Roles

Data Visualization, CSE 512 @ UW. *Teaching Assistant w/ Jeffrey Heer* — Spring 2023

Natural Language Processing, CSE 517 @ UW. *Teaching Assistant w/ Noah A. Smith* — Winter 2023

Introduction to Human-Computer Interaction, CSE 440 @ UW. *Teaching Assistant w/ Amy X. Zhang* — Fall 2022

Research Seminar on Human-Computer Interaction, CSE 590 @ UW. *Co-Organizer* — Winter 2023

Speech & NLP Reading Group, Microsoft Research India. *Organizer* — 2019 - 2021

Open Source Development (OSD) Labs, BITS Goa. *Founding Member* — 2016 - 2018

Introduction to Computer Programming, CS F111 @ BITS. *Teaching Assistant* — Spring 2018

Introduction to Programming, Center for Technical Education (CTE), BITS Goa. *Co-Instructor* — Spring 2017

Academic Service

PC/Reviewer	*Conferences:* FAccT'23, ACL'23, WWW'23, CHI'23, AAAI'23, EMNLP'22, NAACL'22, ACL'22, **CHI'22***, ARR'21, **EMNLP'21***, ACL'21, EACL'21, MLADS'20, ICON'20 *Conference Demos:* NAACL'22, ACL'22, EMNLP'21, NAACL'21, EMNLP'20 *Others:* UW CSE PhD Applications 2022 & 2023 (Area Chair for NLP, HCI) **(* = Outstanding Reviewer)**
Sub-Reviewer	EMNLP'20, ACL'20, LREC'20, CODS-COMAD'20, CoNLL'19, Interspeech'19, ICON'19
Volunteer	ACL'20, AAAI'19, Panini Linguistics Olympiad (PLO) '19/'20
Tech	ACL Rolling Review, ACL'20 Virtual Conference

Junior Collaborators

Andre Ye, Undergrad, UW CSE. *Cultural cognition differences reflected in vision-language systems* — 01/2023 - Present

Ayana Bharadwaj, Undergrad, UW CSE. *Curating machine learning data through games* — 04/2023 - Present

Software and Open Source Contributions

Microsoft. Interactive Neural Machine Translation - git,
Mozilla. Bugzilla - git, Firefox (Gecko Engine) - git, TreeHerder - git
Others. scikit-learn - git, DDGenerators.jl - git, Virtual ACL 2020 - git, UserContext -
BITS Pilani. BITS-Darshini - git, SWD - git, ARC - git, AutolabJS - git, Quark 2017 - git, Waves 2016 - git