Shomir Wilson

The Pennsylvania State University
College of Information Sciences and Technology
E310 Westgate Building
University Park, PA 16802

shomir@psu.edu
https://shomir.net
https://hltlab.org

EDUCATION

University of Maryland	2011 **Ph.D.**, Computer Science; 2008 **M.S.**, Computer Science
Virginia Tech	2005 **B.S.**, Computer Science, **B.S.**, Mathematics, **B.A.**, Philosophy

EXPERIENCE

Primary Affiliations

Penn State University College of Information Sciences and Technology (University Park, PA)	8/2018 – now	**Assistant Professor** Director, Human Language Technologies Lab Faculty Affiliate, Institute for CyberScience
University of Cincinnati EECS Department (Cincinnati, OH)	8/2016 – 8/2018	**Assistant Professor** Director, Human Language Technologies Lab Member, Institute for Analytics Innovation
Carnegie Mellon University School of Computer Science (Pittsburgh, PA)	8/2015 – 8/2016	**Project Scientist** *Supervisor: Norman Sadeh*
	1/2015 – 5/2015	**Lecturer** (Natural Language Processing)
	8/2013 – 7/2015	**NSF International Research Fellow** *Supervisor: Alan W Black*
	9/2011 – 7/2013	**Postdoctoral Fellow** *Supervisor: Norman Sadeh*
University of Maryland Computer Science Department (College Park, MD)	6/2006 – 5/2011	Graduate Research Assistant *Ph.D. Advisor: Donald Perlis*
	8/2005 – 5/2006	Graduate Teaching Assistant

Visiting Fellowships

University of Edinburgh School of Informatics (Edinburgh, United Kingdom)	8/2013 – 7/2014	**NSF International Research Fellow** *Host: Jon Oberlander*
Nat'l Univ. of Singapore School of Computing (Kent Ridge, Singapore)	6/2010 – 8/2010	**NSF EAPSI Fellow** *Host: Min-Yen Kan*
Macquarie University School of Computing (Sydney, Australia)	6/2009 – 8/2009	**NSF EAPSI Fellow** *Host: Robert Dale*

GRANTS AND RESEARCH SUPPORT

National Science Foundation $13,792 2018-09-01 – 2019-02-28
$76,249 2016-09-01 – 2018-07-31
#CNS-1330596 TWC SBE: Option: Frontier: Collaborative: Towards Effective Web Privacy Notice and Choice: A Multi-Disciplinary Prospective ("The Usable Privacy Policy Project")
(Structured as subcontracts from Carnegie Mellon University)

University of Cincinnati $2,350 2017-06-08 – 2018-11-17
College of Engineering and Applied Sciences Faculty Development Grant

Ohio Supercomputer Center (9,000 Resource Units) 2016-2018
#PES0731-1 Text Analysis for Online Privacy
#PES0746-1 Grant for Classroom Use: Natural Language Processing Term Projects

National Science Foundation $98,100 2013-08-01 – 2015-03-31
#OISE-1159236 IRFP: Metalanguage Identification for Interactive Language Technologies

National Science Foundation $5,000 2010-06-13 – 2010-08-06
#OISE-1015666 EAPSI (Singapore): Parsing Metalanguage and the Use-Mention Distinction

National Science Foundation $5,000 2009-06-22 – 2010-08-16
#OISE-0914091 EAPSI (Australia): Distinguishing Use and Mention in Natural Language

AWARDS, RECOGNITION, AND PRESS

Penn State News: "Outside of IST: Faculty and staff making a difference", October 17, 2018.

National Science Foundation News from the Field: "Carnegie Mellon Researchers Create an AI to Help Us Make Sense of Privacy Policies", March 1, 2018.

Television news interview about FaceID privacy, WCPO Cincinnati. November 14, 2017.

Best Paper Finalist at the 25th World Wide Web Conference, 2016.

National Science Foundation SEE Innovation Research Highlight: "Refining a Computer's Understanding of Language", 2012.

University of Maryland International Conference Student Support Award, 2011.

University of Maryland Block Grant Fellowship, 2005–2007.

SERVICE

Program Committee for NAACL 2019.

College of IST Security/Privacy Faculty Hiring Committee Member, 2018-2019.

ICDCIT Program Committee Member, 2017.

Lead Organizer of the AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies, 2019. (`https://sites.google.com/view/pal2019/`).

Session Chair for Social Applications II, EMNLP 2018.

Editorial Board, Journal of Intelligent and Information Systems, 2018-present.

NSF proposal review panelist: 2016, 2017, 2018.

College of Engineering and Applied Sciences eLearning Task Force, 2018.

Member of the EECS Graduate Admissions Committee, 2017-2018.

Grand Awards Judge for the Intel International Science and Engineering Fair: 2012, 2015, 2018.

Faculty Marshal for the College of Engineering and Applied Sciences: May 2017, May 2018.

Program Committee Member, LREC Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS), 2018.

Lead Organizer of the AAAI Fall Symposium on Privacy and Language Technologies, 2016. (`https://sites.google.com/site/fsplt2016/`)

Program Committee Member, Workshop on Privacy in Collaborative & Social Computing, 2016.

Judge for the Pittsburgh Regional Science and Engineering Fair, 2016.

Judge for Pittsburgh Regional FIRST Lego League, 2011 and 2013.

Proposal reviewer for the Portuguese Foundation of Science and Technology (FCT), 2012.

Graduate student representative to the Computer Science Department Council, 2008-2009.

Session Chair for Logic Programming and Knowledge Engineering at AAAI 2008.

Panel participant in forums on "Meta-Level Control" and "Models of Self", during the Metareasoning Workshop at the Twenty-Third AAAI Conference on Artificial Intelligence, 2008.

TEACHING

Semester Courses

Natural Language Processing: Fall 2017, University of Cincinnati; Spring 2015, Carnegie Mellon University (co-taught at CMU with Chris Dyer and Alan W Black).

Advanced Topics in Natural Language Processing: Spring 2017 and Spring 2018, University of Cincinnati.

Introduction to Low-Level Programming Concepts: (Teaching Assistant) Fall 2005 and Spring 2006, University of Maryland.

Condensed Courses

Natural Language Processing Methods and Applications: Workshop for faculty from Ming Chi University of Taiwan, July 11-13, 2018, University of Cincinnati (Cincinnati, Ohio).

Ethics of Artificial Intelligence: March 11-15, 2018, Future University (Cairo, Egypt).

ADVISEES

Abhijith Athreya, M.S., July 2018. (Chief Engineer, Samsung R&D)

Baradwaj Aryasomayajula, M.S., July 2018. (Software Developer, Verizon)

Ph.D. thesis committee memberships: Maha Aljohani (2018; external examiner at Dalhousie University)

M.S. thesis committee memberships: Sampurna Ravindranathan (2018), Abhro Mondal (2017)

PUBLICATIONS AND WRITINGS

Selected titles are prefixed with a caret (ˆ).

Peer-Reviewed Conference Proceedings

1. Supervised and Unsupervised Methods for Robust Separation of Section Titles and Prose Text in Web Documents. Abhijith Athreya Mysore Gopinath, Shomir Wilson, and Norman Sadeh. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018.
2. Identifying the provision of choices in privacy policy text. Kanthashree Sathyendra, Shomir Wilson, Florian Schaub, Norman Sadeh, and Sebastian Zimmeck. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 2017.
3. ˆ**Automated analysis of privacy requirements for mobile apps**. Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. In *Proceedings of the Network and Distributed System Security Symposium*, San Diego, California, March 2017.
4. ˆ**The creation and analysis of a website privacy policy corpus**. Shomir Wilson, Florian

Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, August 2016.

5. ˆ**Crowdsourcing annotations of websites' privacy policies: Can it really work?** Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah Smith and Frederick Liu. In *Proceedings of the 25th International World Wide Web Conference*, Montreal, Canada, April 2016. **Best Paper Finalist.**
6. This table is different: A WordNet-based approach to identifying references to document entities. Shomir Wilson, Alan W Black, and Jon Oberlander. In *Proceedings of the 8th International Global WordNet Conference*, Bucharest, Romania, January 2016.
7. Identifying relevant text fragments to help crowdsource privacy policy annotations. Rohan Ramanath, Florian Schaub, Shomir Wilson, Fei Liu, Norman Sadeh, and Noah Smith. In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing*, works-in-progress track, Pittsburgh, PA, November 2014.
8. Determiner-established deixis to communicative artifacts in pedagogical text. Shomir Wilson and Jon Oberlander. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, June 2014.
9. Toward automatic processing of English metalanguage. Shomir Wilson. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Nagoya, Japan, October 2013.
10. ˆ**Privacy manipulation and acclimation in a location sharing application**. Shomir Wilson, Justin Cranshaw, Norman Sadeh, Alessandro Acquisti, Lorrie Cranor, Jay Springfield, Sae Young Jeong, and Arun Balasubramanian. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Zurich, Switzerland, September 2013.
11. ˆ**Tweets are forever: A large-scale quantitative analysis of deleted tweets**. Hazim Almuhimedi, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. In *Proceedings of the 2013 ACM Conference on Computer Supported Cooperative Work*, San Antonio, TX, February 2013.
12. ˆ**The creation of a corpus of English metalanguage**. Shomir Wilson. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, South Korea, July 2012.
13. Application of MCL in a dialog agent. Darsana Josyula, Scott Fults, Michael L. Anderson, Shomir Wilson, and Don Perlis. In *Papers from the Third Language and Technology Conference*, Poznań, Poland, October 2007.

Peer-Reviewed Symposium Proceedings

1. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. Frederick Liu, Shomir Wilson, Florian Schaub and Norman Sadeh. In *Proceedings of the AAAI Fall Symposium on Privacy and Language Technologies*, Arlington, VA, November 2016.

2. Automatic extraction of opt-out choices from privacy policies. Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson and Norman Sadeh. In *Proceedings of the AAAI Fall Symposium on Privacy and Language Technologies*, Arlington, VA, November 2016.
3. Analyzing and predicting privacy law compliance of mobile apps. Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin and Joel Reidenberg. In *Proceedings of the AAAI Fall Symposium on Privacy and Language Technologies*, Arlington, VA, November 2016.
4. The Metacognitive Loop: An architecture for building robust intelligent systems. Hamid Haidarian, Wikum Dinalankara, Scott Fults, Shomir Wilson, Don Perlis, Matt Schmill, Tim Oates, Darsana Josyula, and Michael Anderson. In *Proceedings of the AAAI Fall Symposium on Commonsense Knowledge*, Arlington, VA, November 2010.
5. Toward domain-neutral human-level metacognition. Michael L. Anderson, Matt Schmill, Tim Oates, Don Perlis, Darsana Josyula, Dean Wright, and Shomir Wilson. In *Proceedings of the 2007 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Palo Alto, CA, March 2007.

Peer-Reviewed Journal Articles

1. ˆ(accepted, publication pending) **Analyzing privacy policies at scale: From crowdsourcing to automated annotations**. Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Sebsatian Zimmeck, Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah Smith. To appear in *ACM Transactions on the Web*.
2. **ˆPrivOnto: A semantic framework for the analysis of privacy policies**. Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B. Norton, N. Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. In *Semantic Web Journal*, May 2017.
3. **ˆNudges for privacy and security: Understanding and assisting users' choices online**. Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. ACM Computing surveys 50(3), August 2017.
4. In search of the use-mention distinction and its impact on language processing tasks. Shomir Wilson. *The International Journal of Computational Linguistics and Applications* 2(1-2), pp. 139-154, 2011.

Book Chapters

1. A bridge from the use-mention distinction to natural language processing. Shomir Wilson. In Saka, P., Johnson, M. (Ed.), *The Semantics and Pragmatics of Quotation*. Springer, 2017.
2. The metacognitive loop and reasoning about anomalies. Matthew Schmill, Michael L. Anderson, Scott Fults, Darsana Josyula, Tim Oates, Donald Perlis, Hamid Haidarian Shahri, Shomir Wilson, and Dean Wright. In Cox, M., Raja, A. (Ed.), *Metareasoning: Thinking About Thinking*. MIT Press, MA, 2010.

Magazine Article

- A self-help guide for autonomous systems. Michael L. Anderson, Scott Fults, Darsana P. Josyula, Tim Oates, Don Perlis, Matthew D. Schmill, Shomir Wilson, and Dean Wright. *AI Magazine*, Summer 2008.

Peer-Reviewed Conference Poster Abstracts

1. Increasing the salience of data use opt-outs online. Namita Nisal, Sushain K. Cherivirala, Kanthashree M. Sathyendra, Margaret Hagan, Florian Schaub, Shomir Wilson, Lorrie Faith Cranor, and Norman Sadeh. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security*, Santa Clara, CA, June 2017.
2. Mobile app privacy compliance: Automated technology to help regulators, app stores and developers. Sebastian Zimmeck, Lieyong Zou, Bin Liu, Shomir Wilson, Steven M. Bellovin, Ziqi Wang, Roger Iyengar, Florian Schaub, Norman Sadeh, and Joel Reidenberg. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security*, Santa Clara, CA, June 2017.
3. Visualization and interactive exploration of data practices in privacy policies. Sushain K. Cherivirala, Florian Schaub, Mads Schaarup Andersen, Shomir Wilson, Norman Sadeh, and Joel R. Reidenberg. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security*, Denver, CO, 2016.
4. Towards usable privacy policies: Semi-automatically extracting data practices from websites' privacy policies. Norman Sadeh, Alessandro Acquisti, Travis Breaux, Lorrie Cranor, Aleecia McDonald, Joel Reidenberg, Noah Smith, Fei Liu, N. Cameron Russell, Florian Schaub, Shomir Wilson, James Graves, Pedro Leon, Rohan Ramanath, and Ashwini Rao. In *Proceedings of the Tenth Symposium on Usable Privacy and Security*, Palo Alto, CA, July 2014.

Peer-Reviewed Workshop Proceedings

1. Demystifying Privacy Policies with Language Technologies: Progress and Challenges. Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy, and Norman Sadeh. In *Proceedings of the Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS) at LREC*, Portoroz, Solvenia, May 2016.
2. Distinguishing use and mention in natural language. Shomir Wilson. In *Proceedings of the NAACL HLT Student Research Workshop*, Los Angeles, CA, June 2010.
3. The role of metacognition in robust AI systems. Matt Schmill, Tim Oates, Michael L. Anderson, Darsana Josyula, Don Perlis, Shomir Wilson, and Scott Fults. In *Papers from the Workshop on Metareasoning at the 23rd AAAI Conference on Artificial Intelligence*, Chicago, IL, July 2008.
4. Ontologies for reasoning about failures in AI systems. Michael L. Anderson, Scott Fults, Darsana Josyula, Tim Oates, Don Perlis, Matt Schmill, and Shomir Wilson. In *Proceedings of the First International Workshop on Metareasoning in Agent-Based Systems*, Honolulu, HI, 2007.

Dissertation

- *A Computational Theory of the Use-Mention Distinction in Natural Language.* Shomir Wilson. University of Maryland, 2011.

Technical Reports

1. Towards Automatic Classification of Privacy Policy Text. Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. Technical Report CMU-LTI-17-010, Carnegie Mellon University, 2017.
2. The Usable Privacy Policy Project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about. Norman Sadeh, Alessandro Acquisti, Travis Breaux, Lorrie Cranor, Aleecia McDonald, Joel Reidenberg, Noah Smith, Fei Liu, N. Cameron Russell, Florian Schaub, and Shomir Wilson. Technical Report CMU-ISR-13-119, Carnegie Mellon University, 2013.
3. Automatic categorization of privacy policies: A pilot study. Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A. Smith. Technical Report CMU-LTI-12-019 / CMU-ISR-12-114, Carnegie Mellon University, 2012.

INVITED TALKS

"Human Language Technologies for Understanding Online Privacy", Computer Science Seminar, Dalhousie University, 5 December 2018.

Data Sciences Seminar, Penn State College of Information Sciences and Technology, 27 November 2018.

Northern Kentucky University College of Informatics, 24 March 2017.

"Text Analysis to Support the Privacy of Internet Users", Hutton Lecture Series, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 17 February 2017.

University of Dayton Computer Science Department Colloquium Series, 3 February 2017.

Georgetown University Computer Science Colloquium Series. 21 November 2016.

"Crowdsourcing Annotations of Websites' Privacy Policies: Can It Really Work?", **Encore Track** at the Fourth AAAI Conference on Human Computation and Crowdsourcing, Austin, Texas. 1 November 2016.

"Introspective Users and Introspective Text: Some Recent Results", CHIME Text Seminar, National University of Singapore. 5 January 2016.

"Identifying Deixis to Communicative Artifacts in Text", NLIP Seminar Series, Cambridge University. 9 May 2014.

"An Empirical Approach to Metalanguage", ILCC/HCRC Seminar Series, University of Edinburgh. 6 September 2013.

"A Computational Approach to Metalanguage and the Use-Mention Distinction", CL+NLP Lunch, Carnegie Mellon University. Pittsburgh, PA, 23 April 2013.

MISCELLANY

Professional Memberships: AAAI, ACL, ACM.

Safer People, Safer Places training, 2018.

Bias Busters training, 2015.

Licensed amateur radio operator, technician class.