DAVID ARTHUR SMITH

Computer & Information Science
Northeastern University
340 Huntington Avenue
Boston, MA 02115, USA

dasmith@ccs.neu.edu
Office: +1 617 373 8526
Mobile: +1 410 900 6238
https://khoury.neu.edu/home/dasmith

Education

Johns Hopkins University 2010 **Ph.D. in Computer Science**
Advisor: Jason Eisner
- *National Science Foundation fellowship (2003–6)*
- *Wolman fellowship (2002–3)*

Harvard University 1994 **A.B. *summa cum laude* in Classics (Greek)**
- *Harvard National Scholar*

Employment

Northeastern University July 2018 – present
Associate Professor, Khoury College of Computer Sciences

Northeastern University September 2012 – June 2018
Assistant Professor, College of Computer and Information Science

University of Massachusetts Amherst September 2008 – August 2012
Research Assistant Professor, Department of Computer Science, Center for Intelligent Information Retrieval

Johns Hopkins University September 2002 – September 2008
Research Assistant, Department of Computer Science, Center for Language and Speech Processing
Machine learning for natural language processing: semi-supervised learning and efficient inference techniques; syntactic parsing; morphological disambiguation; machine translation and word alignment
Summer Research Workshop, 2003: Member of Syntax for Statistical Machine Translation team

Google, Inc. May 2005 – September 2005
Internship in Machine Translation group
Research on improved training and decoding for machine translation

Tufts University July 1994 – August 2002
Perseus Digital Library Project
Information retrieval and extraction, named-entity disambiguation, document alignment, morphological analysis

Publications

Book

[1] Ryan Cordell, David Smith, Abby Mullen, and Jonathan D. Fitzgerald. *Going the Rounds: Virality in Nineteenth-Century Newspapers*. University of Minnesota Press, 2024.

Refereed Conference & Journal Publications

[2] Jacob Murel and David A. Smith. Active learning with relevance feedback for handwriting detection in historical print. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2024.

[3] Jaydeep Borkar and David A. Smith. Mind the gap: Analyzing lacunae with transformer-based transcription. In *International Workshop on Computational Paleography*, 2024.

[4] Danlu Chen, Jacob Murel, Taimoor Shahid, Xiang Zhang, Jonathan Parkes Allen, Taylor Berg-Kirkpatrick, , and David A. Smith. MONSTERMASH: Multidirectional, overlapping, nested, spiral text extraction for recognition models of Arabic-script handwriting. In *International Workshop on Computational Paleography*, 2024.

[5] Jacob Murel and David A. Smith. Retrieving and analyzing translations of American newspaper comics with visual evidence. In *Workshop on coMics ANalysis, Processing and Understanding (MANPU)*, 2024.

[6] Liwen Hou and David A. Smith. Detecting syntactic change with pre-trained transformer models. In *Findings of EMNLP*, 2023.

[7] David A. Smith, Jacob Murel, Jonathan Parkes Allen, and Matthew Thomas Miller. Automatic collation for diversifying corpora: Commonly copied texts as distant supervision for handwritten text recognition. In *Computational Humanities Research Conference (CHR)*, 2023.

[8] Ryan Muther, Mathew Barber, and David A. Smith. Querying the past: Automatic source attribution with language models. In *Computational Humanities Research Conference (CHR)*, 2023.

[9] Caroline Craig, Kartik Goyal, Gregory R. Crane, Farnoosh Shamsian, and David A. Smith. Testing the limits of neural sentence alignment models on classical Greek and Latin texts and translations. In *Computational Humanities Research Conference (CHR)*, 2023.

[10] Ryan Muther and David A. Smith. Citations as queries: Source attribution using language models as rerankers. In *SIGIR Workshop on Retrieval-Enhanced Machine Learning (REML)*, 2023.

[11] Si Wu and David A. Smith. Composition and deformance: Measuring imageability with a text-to-image model. In *Proceedings of the Workshop on Narrative Understanding*, 2023.

[12] Ryan Muther, David A. Smith, and Sarah Bowen Savant. From networks to named entities and back again: Exploring classical Arabic *isnad* networks. *Journal of Historical Network Research*, 2022.

[13] Giulia Taurino and David A. Smith. Machine learning as an archival science: Narratives behind artificial intelligence, cultural data, and archival remediation. In *NeurIPS Workshop on AI Cultures*, 2022.

[14] Alejandro Toselli, Si Wu, and David A. Smith. Digital editions as distant supervision for layout analysis of printed books. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2021.

[15] Ansel MacLaughlin, Shaobin Xu, and David A. Smith. Recovering lexically and semantically reused texts. In *Proceedings of the Joint Conference on Lexical and Computational Semantics (*SEM)*, 2021.

[16] Helen O'Neill, Anne Welsh, David A. Smith, Glenn Roe, and Melissa Terras. Text mining Mill: Computationally detecting influence in the writings of John Stuart Mill from library records. *Digital Scholarship in the Humanities*, 2021.

[17] Rui Dong and David A. Smith. Structural encoding and pre-training matter: Adapting BERT for table-based fact verification. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

[18] Ansel MacLaughlin and David A. Smith. Content-based models of quotation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

[19] Liwen Hou and David A. Smith. Drivers of English syntactic change in the Canadian Parliament. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, 2021.

[20] Liwen Hou and David A. Smith. Emerging English transitives over the last two centuries. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, 2021.

[21] Ansel MacLaughlin, John Wihbey, Aleszu Bajak, and David A. Smith. Source attribution: Recovering the press releases behind science health news. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2020.

[22] Maha Alkhairy, Afshan Jafri, and David A. Smith. Finite state machine pattern-root Arabic morphological generator, analyzer and diacritizer. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2020.

[23] Shijia Liu and David A. Smith. Detecting *de minimis* code-switching in historical German books. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2020.

[24] Rui Dong, David A. Smith, Shiran Dudy, and Steven Bedrick. Noisy neural language modeling for typing prediction in BCI communication. In *Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 44–51, 2019.

[25] Rui Dong and David A. Smith. Multi-input attention for unsupervised OCR correction. In *Proceedings of the Association for Computational Linguistics*, 2018.

[26] Ansel MacLaughlin, John Wihbey, and David A. Smith. Predicting news coverage of scientific articles. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2018.

[27] Shiran Dudy, Steven Bedrick, Shaobin Xu, and David A. Smith. A multi-context character prediction model for a brain-computer interface. In *Proceedings of the Workshop on Subword and Character Level Models in NLP (SCLeM)*, 2018.

[28] Shaobin Xu and David A. Smith. Contrastive training for models of information cascades. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[29] Liwen Hou and David A. Smith. Modeling the decline in English passivization. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, 2018.

[30] Shaobin Xu and David A. Smith. Retrieving and combining repeated passages to improve OCR. In *Proceedings of the ACM+IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2017.

[31] Kriste Krstovski and David A. Smith. Bootstrapping translation detection and sentence extraction from comparable corpora. In *Proceedings of the Conference on Human Language Technology of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2016.

[32] Kriste Krstovski and David A. Smith. Online multilingual topic models with multi-level hyperpriors. In *Proceedings of the Conference on Human Language Technology of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2016.

[33] Kriste Krstovski, David A. Smith, and Michael Kurtz. Automatic construction of evaluation sets and evaluation of document similarity models in large scholarly retrieval systems. In *AAAI Workshop on Scholarly Big Data*, 2016.

[34] David A. Smith, Ryan Cordell, and Abigail Mullen. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, 27(3), 2015.

[35] Kriste Krstovski, David A. Smith, and Michael Kurtz. Evaluating retrieval models through histogram analysis. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

[36] John Wilkerson, David A. Smith, and Nick Stramp. Tracing the flow of policy ideas on legislatures: A text reuse approach. *American Journal of Political Science*, 59(4):943–956, 2015.

[37] David A. Smith, Ryan Cordell, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. Detecting and modeling local text reuse. In *Proceedings of the ACM+IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2014. **Nominated for best paper**.

[38] Youngho Kim, Jangwon Seo, W. Bruce Croft, and David A. Smith. Automatic suggestion of phrasal-concept queries for literature search. *Information Processing & Management*, 50(4):568–583, July 2014.

[39] Shaobin Xu, David Smith, Abigail Mullen, and Ryan Cordell. Detecting and evaluating local text reuse in social networks. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 2014.

[40] Xiaoxi Xu, Tom Murray, Beverly Park Woolf, and David A. Smith. Identifying social deliberative behavior from online communication—a cross-domain study. In *Proceedings of the Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 237–242, 2014.

[41] Xiaoxi Xu, Tom Murray, Beverly Park Woolf, and David A. Smith. Social network signatures of effective online communication. In *Intelligent Tutoring Systems*, pages 621–622, 2014.

[42] David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *IEEE Workshop on Big Data and the Humanities*, 2013.

[43] Kriste Krstovski, David A. Smith, Hanna M. Wallach, and Andrew McGregor. Efficient nearest-neighbor search in the probability simplex. In *Proceedings of the International Conference on the Theory of Information Retrieval (ICTIR)*, 2013.

[44] Kriste Krstovski and David A. Smith. Online polylingual topic models for fast document translation detection. In *Proceedings of the Workshop on Statistical Machine Translation*, 2013.

[45] Jacqueline L. Feild, Erik G. Learned-Miller, and David A. Smith. Using a probabilistic syllable model to improve scene text recognition. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2013.

[46] Xiaoxi Xu, Tom Murray, Beverly Park Woolf, and David A. Smith. Mining social deliberation in online communication: If you were me and I were you. In *International Conference on Educational Data Mining (EDM)*, 2013.

[47] Jason Naradowsky, Tim Vieira, and David A. Smith. Grammarless parsing for joint inference. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2012.

[48] Jason Naradowsky, Sebastian Riedel, and David A. Smith. Improving NLP through marginalization of hidden syntactic structure. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012.

[49] Sebastian Riedel, David A. Smith, and Andrew McCallum. Parse, price and cut—delayed column and row generation for graph based parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012.

[50] Yanchuan Sim, Noah A. Smith, and David A. Smith. Discovering factions in the computational linguistics community. In *ACL Workshop on Rediscovering 50 Years of Discoveries*, 2012.

[51] Michael Bendersky and David A. Smith. A dictionary of wisdom and wit: Learning to extract quotable phrases. In *NAACL Workshop on Computational Linguistics for Literature*, pages 69–77, 2012.

[52] David Bamman and David A. Smith. Extracting two thousand years of Latin from a million book library. *ACM Journal on Computing and Cultural Heritage*, 5(1), 2012.

[53] Jangwon Seo, W. Bruce Croft, and David A. Smith. Online community search using conversational structures. *Information Retrieval*, 14(6):547–571, 2011.

[54] Andrew Kae, David A. Smith, and Erik Learned-Miller. Learning on the fly: A font-free approach towards multilingual OCR. *International Journal on Document Analysis and Recognition*, 14(3):289–301, 2011.

[55] Jeffrey Dalton, James Allan, and David A. Smith. Passage retrieval for incorporating global dependencies in sequence labeling. In *Conference on Information and Knowledge Management (CIKM)*, pages 355–364, 2011.

[56] Jinyoung Kim, W. Bruce Croft, David A. Smith, and Anton Bakalov. Evaluating an associative browsing model for personal information. In *Conference on Information and Knowledge Management (CIKM)*, pages 647–652, 2011.

[57] Jae-Hyun Park, W. Bruce Croft, and David A. Smith. A quasi-synchronous dependence model for information retrieval. In *Conference on Information and Knowledge Management (CIKM)*, pages 17–26, 2011.

[58] David A. Smith, R. Manmatha, and James Allan. Mining relational structure from millions of books: Position paper. In *Proceedings of the CIKM BooksOnline Workshop*, pages 49–54, 2011.

[59] Kriste Krstovski and David A. Smith. A minimally supervised approach for detecting and ranking document translation pairs. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 207–216, 2011.

[60] Michael Bendersky, W. Bruce Croft, and David A. Smith. Joint annotation of search queries. In *Proceedings of the Association for Computational Linguistics*, pages 102–111, 2011.

[61] John S. Y. Lee, Jason Naradowsky, and David A. Smith. A discriminative model for joint morphological disambiguation and dependency parsing. In *Proceedings of the Association for Computational Linguistics*, pages 885–894, 2011.

[62] Elif Aktolga, James Allan, and David A. Smith. Passage reranking for question answering using syntactic structures and answer types. In *European Conference on Information Retrieval (ECIR)*, pages 617–628, 2011.

[63] Michael Bendersky, W. Bruce Croft, and David A. Smith. Structural annotation of search queries using pseudo-relevance feedback. In *Conference on Information and Knowledge Management (CIKM)*, pages 1537–1540, 2010.

[64] Jinyoung Kim, Anton Bakalov, David A. Smith, and W. Bruce Croft. Building and evaluating a semantic representation for personal information. In *Conference on Information and Knowledge Management (CIKM)*, pages 1741–1744, 2010.

[65] Xiaobing Xue, W. Bruce Croft, and David A. Smith. Query reformulation using query distributions. In *Conference on Information and Knowledge Management (CIKM)*, pages 1497–1500, 2010.

[66] Sebastian Riedel, David A. Smith, and Andrew McCallum. Inference by minimizing size, divergence, or their sum. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 227–234, 2010.

[67] Sebastian Riedel and David A. Smith. Relaxed marginal inference and its application to dependency parsing. In *Proceedings of the Conference on Human Language Technology of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 760–768, 2010.

[68] Jangwon Seo, W. Bruce Croft, and David A. Smith. Online community search using thread structure. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pages 1907–1910, 2009.

[69] David A. Smith and Jason Eisner. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 822–831, 2009.

[70] David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 880–889, 2009.

[71] Michael Bendersky, W. Bruce Croft, and David A. Smith. Two-stage query segmentation for information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 810–811, 2009.

[72] David A. Smith and Jason Eisner. Dependency parsing by belief propagation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 145–156, 2008.

[73] David A. Smith and Jason Eisner. Bootstrapping feature-rich dependency parsers with entropic priors. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 667–677, 2007.

[74] David A. Smith and Noah A. Smith. Probabilistic models of nonprojective dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 132–140, 2007.

[75] Keith Hall, Jiří Havelka, and David A. Smith. Log-linear models of non-projective trees, k-best MST parsing and tree-ranking. In *Proceedings of the CoNLL Shared Task*, pages 962–966, 2007.

[76] David A. Smith and Jason Eisner. Minimum risk annealing for training log-linear models. In *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics*, pages 787–794, 2006.

[77] David A. Smith and Jason Eisner. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 23–30, 2006.

[78] Markus Dreyer, David A. Smith, and Noah A. Smith. Vine parsing and minimum risk reranking for speed and precision. In *Proceedings of the CoNLL Shared Task*, pages 201–205, 2006.

[79] Noah A. Smith, David A. Smith, and Roy W. Tromble. Context-based morphological disambiguation with random fields. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 475–482, 2005.

[80] David A. Smith and Noah A. Smith. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 49–56, 2004.

[81] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. A smorgasbord of features for statistical machine translation. In *Proceedings of the Conference on Human Language Technology and the North American Association for Computational Linguistics*, pages 161–168, 2004.

[82] David A. Smith and Gideon S. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL Workshop on Analysis of Geographic References*, pages 45–49, 2003.

[83] Gregory Crane, Clifford E. Wulfman, Lisa M. Cerrato, Anne Mahoney, Thomas L. Milbank, David Mimno, Jeffrey A. Rydberg-Cox, David A. Smith, and Christopher York. Towards a cultural heritage digital library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2003*, pages 75–86, Houston, TX, June 2003.

[84] David A. Smith. Detecting and browsing events in unstructured text. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 73–80, Tampere, Finland, August 2002.

[85] David A. Smith. Detecting events with date and place information in unstructured text. In *Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, pages 191–196, Portland, OR, July 2002.

[86] David A. Smith, Anne Mahoney, and Gregory Crane. Integrating harvesting into digital library content. In *Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, pages 183–184, Portland, OR, July 2002.

[87] Gregory R. Crane, Robert F. Chavez, Anne Mahoney, Thomas L. Milbank, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman. Drudgery and deep thought: Designing a digital library for the humanities. *Communications of the Association for Computing Machinery*, 44(5):35–40, 2001.

[88] Gregory Crane, David A. Smith, and Clifford E. Wulfman. Building a hypertextual digital library in the humanities: A case study on London. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 426–434, Roanoke, VA, June 2001. **Best paper award**.

[89] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of the European Conference on Digital Libraries (ECDL)*, pages 127–136, Darmstadt, Germany, September 2001.

[90] David A. Smith, Jeffrey A. Rydberg-Cox, and Gregory R. Crane. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.

[91] David A. Smith, Anne Mahoney, and Jeffrey A. Rydberg-Cox. Management of XML documents in an integrated digital library. *Markup Languages: Theory and Practice*, 2(3):205–214, 2000.

[92] David A. Smith, Anne Mahoney, and Jeffrey A. Rydberg-Cox. Management of XML documents in an integrated digital library. In *Proceedings of Extreme Markup Languages 2000*, pages 219–224, Montreal, August 2000.

[93] David A. Smith. Textual variation and version control in the TEI. *Computers and the Humanities*, 33(1-2):103–112, 1999.

Other Publications

[94] David A. Smith. Modeling errors in estimating historical trends. In *Digital Humanities*, 2024.

[95] Shijia Liu and David A. Smith. Tracing accounts of racial terror in historical newspapers. In *New Directions in Analyzing Text as Data (TADA)*, 2023.

[96] Si Wu and David A. Smith. The language of US partisan newspapers from 1869 to 1925. In *New Directions in Analyzing Text as Data (TADA)*, 2023.

[97] Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet. Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292). *Dagstuhl Reports*, 12(7):112–179, 2023.

[98] Giulia Taurino, Si Wu, and David A. Smith. Archeologies of data in contemporary journalism: The digital afterlives of newspapers' photo morgues. In *Computation + Journalism*, New York, June 2022.

[99] Soumya Mohanty and David Smith. Alignment-based training for detecting reader annotations in printed books. In *Proceedings of Digital Access to Textual Cultural Heritage (DATeCH)*, 2019.

[100] David A. Smith and Ryan Cordell. A research agenda for historical and multilingual optical character recognition. Technical report, Northeastern University, 2018. `https://repository.library.northeastern.edu/files/neu:f1881m409`.

[101] Kriste Krstovski, Michael J. Kurtz, David A. Smith, and Alberto Accomazzi. Multilingual topic models for indexing scientific articles. Under review, 2019.

[102] Ryan Muther and David A. Smith. Charting the changes: Modeling edits in the lawmaking process. In *PoliInformatics*, Bainbridge Island, WA, August 2017.

[103] Ryan Cordell and David A. Smith. What news is new?: Ads, extras, and viral texts on the nineteenth-century newspaper page. In *Digital Humanities*, 2017.

[104] Ryan Cordell, David Smith, and Shaobin Xu. Aggregating exchange in the nineteenth-century newspaper. In *Society for the History of Authorship, Reading, and Publishing (SHARP)*, Victoria, BC, June 2017.

[105] David A. Smith, Anne Washington, and John Wilkerson. Attacking the code: A computational approach to discovering issue networks in congress. In *Political Networks*, Portland, OR, June 2015.

[106] Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler, editors. *Computational Humanities – Bridging the Gap between Computer Science and Digital Humanities (Dagstuhl Seminar 14301)*, volume 4 of *Dagstuhl Reports*, Dagstuhl, Germany, 2014. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[107] John Wilkerson, David A. Smith, and Nick Stramp. The inclusiveness of lawmaking: A text reuse approach to tracing the progress of policy ideas in legislation. In *MPSA Annual Meeting*. Midwest Political Science Association, April 2014.

[108] John Wilkerson, David A. Smith, and Nick Stramp. Tracing the flow of policy ideas in legislatures: A text reuse approach. In *New Directions in Analyzing Text as Data*. London School of Economics, September 2013.

[109] John Wilkerson, David A. Smith, Nick Stramp, and James Dashiell. Tracing the flow of policy ideas in legislatures: A computational approach. In *APSA Annual Meeting*. American Political Science Association, September 2013.

[110] Ryan Cordell, Elizabeth Maddock Dillon, and David A. Smith. Uncovering reprinting networks in nineteenth-century American newspapers. In *Digital Humanities*, 2013.

[111] Ryan Cordell and David A. Smith. Uncovering reprinting networks in nineteenth-century American newspapers. In *Chicago Colloquium on Digital Humanities & Computer Science*, November 2012.

[112] Xiaoye Wu and David A. Smith. Right-branching tree transformation for eager dependency parsing. Technical Report CIIR-776, University of Massachusetts, 2010.

[113] Jason Naradowsky, Joe Pater, David Smith, and Robert Staubs. Learning hidden metrical structure with a log-linear model of grammar. In *Computational Modelling of Sound Pattern Acquisition*, pages 59–60, Edmonton, February 2010. Department of Linguistics, University of Alberta.

[114] Joe Pater, David A. Smith, Robert Staubs, Karen Jesney, and Ramgopal Mettu. Learning hidden structure with a log-linear model of grammar. In *Linguistic Society of America (LSA)*, Baltimore, January 2010.

[115] Gregory Druck and David A. Smith. Computing conditional feature covariance under non-projective tree conditional random fields. Technical Report UM-CS-2009-060, University of Massachusetts, 2009.

[116] David A. Smith. Debabelizing libraries: Machine translation by and for digital collections. *D-Lib Magazine*, 12(3), March 2006.

[117] Anne Mahoney, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman. Generalizing the Perseus XML document manager. In *Linguistic Exploration: Workshop on Web-based Language Documentation and Description*, Philadelphia, December 2000.

Dissertation

[118] David A. Smith. *Efficient Inference for Trees and Alignments: Modeling Monolingual and Bilingual Syntax with Hard and Soft Constraints and Latent Variables*. PhD thesis, Johns Hopkins University, 2010.

Software

[119] Perseus Digital Library. `http://www.perseus.tufts.edu/`, 1999-2002. Head programmer for one of the largest, and most popular, humanities digital libraries, Perseus presents sources for language, literature, art, and archaeology for several periods from the ancient Mediterranean through 19th century North America. Users viewing documents receive automatically generated information on morphology, lexicon, translations, technical terms, and named entities, as well as temporal and spatial visualizations.

[120] Gregory Crane, editor. *Perseus 2.0: Interactive Sources and Studies on Ancient Greece*. Yale University Press, 1996-2000. Head programmer for two major releases.

Grants and Contracts

External

2024–2025 NEH Digital Humanities Advancement Geant: *Machine Learning for Large-Scale Journalism Collections* (PI; $150k)

2023–25 Mellon Foundation Higher Learning program: *The Virality of Racial Terror in US Newspapers, 1863–1921* (co-PI; PI, Ryan Cordell, University of Illinois at Urbana-Champaig; $500k)

2022–25 Mellon Foundation Scholarly Communications and Information Technology program: *Arabic-script OCR Catalyst Project Phase II* (co-PI; PI, Matthew Miller, University of Maryland; $1.75M)

2021–22 NEH Digital Humanities Advancement Grant: *Automatic Collation for Diversifying Corpora: Improving Handwritten Text Recognition (HTR) for Arabic-script Manuscripts* (co-PI; PI, Matthew Miller, University of Maryland; $325k)

2020–25 NIH NIDCD: *Optimizing BCI-FIT: Brain-Computer Interface – Functional Implementation Toolkit* (co-PI; PI, Melanie Fried-Oken, Oregon Health and Science University; $677k in Y1)

2020–22 Department of Energy Small Business Innovation Research: *A Multi-Task Learning Framework for Automating the Classification of Building Data* (co-PI; PI, Brian Simmons, Onboard Data Inc.; $127k)

2019–21 Mellon Foundation Scholarly Communications and Information Technology program: *Arabic-script OCR Catalyst Project* (co-PI; PI, Matthew Miller, University of Maryland; $800k)

2019–20 NEH Digital Humanities Advancement Grant: *Improving Optical Character Recognition and Tracking Reader Annotations in Printed Books by Collating and Transcribing Multiple Exemplars* (PI; $100k)

2019 Qatar National Library: *Project to Identify and Study the Sira in the Digital Age* (co-PI; PI, Sarah Savant, ISMC–AKU, London; $446k)

2018 Qatar National Library: *Pilot Project to Identify and Study the Sira in the Digital Age* (co-PI; PI, Sarah Savant, ISMC–AKU, London; $260k)

2018–19 Swiss National Science Foundation: *Research Visit of Prof. D. Smith to EPFL* (co-PI; PI, Frédéric Kaplan, EPFL; $19k)

2018–19 Google Faculty Research Award: *Finding Consensus: Improving OCR with Text-Reuse Detection, Multi-Input Encoders, and Lattice-Based Classification and Sequence Labeling* (PI; $44k)

2017–18 Mellon Foundation Scholarly Communications and Information Technology program: A Research Agenda for Historical and Multilingual OCR (PI; $50k)

2017–19 IMLS Transatlantic Platform: Digging into Data: *Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840–1914* (co-PI; PI, Ryan Cordell; Northeastern is prime institution with $200k out of $1.2M to all international partners)

2015–20 NIH NIDCD: *Clinic Interactions of a Brain-Computer Interface for Communication* (co-PI; PI, Melanie Fried-Oken, Oregon Health and Science University; $665,012 in Y1)

2014–16 Mellon Foundation Scholarly Communications and Information Technology program: *Proteus: Supporting Scholarly Information Seeking through Text-Reuse Analysis and Interactive Corpus Construction* (PI; $500k)

2013–14 Google Faculty Research Award: *Trees over Time: Diachronic Syntactic Analysis of Millions of Books with Unsupervised Domain Adaptation* (PI; $35k)

2013–14 Mellon Foundation Scholarly Communications and Information Technology program: *Proteus Infrastructure: Work Aggregation and Entity Extraction* (co-PI; PI, R. Manmatha, UMass; $205k)

2011–16 DARPA BOLT: *Effective Interactive Retrieval Combining Multiple Annotations and Representations* (co-PI; PI, James Allan, UMass; $2.8M)

2009–13 DARPA Machine Reading: *A Universal Machine Reading System* (co-PI; PI, Andrew McCallum, UMass; $2.5M)

2009–13 NSF Data-Intensive Computing: *Mining a Million Books: Linguistic and Structure Analysis, Fast Expanded Search, and Improved OCR* (co-PI; PI, James Allan, UMass; $2.3M)

2007–12 Army/MURI: *SUBTLE: Situation Understanding Bot through Language and Environment* (co-PI; PI, Andrew McCallum, UMass; $634k)

2010–15 NIH Clinical and Translational Science program (CIIR subcontract; PI, Bruce Croft, UMass; $40k)

2009-11 NSF CluE: *Learning Word Relationships Using TupleFlow* (sr. personnel; PI, James Allan, UMass; $450k)

2009-11 Yahoo!, Inc.: *Data-Intensive Processing for Better Search, Analysis, and OCR* (PI, in-kind access to Yahoo!'s Hadoop cluster)

2009–10 NEH Digital Humanities Start-up Grant: *OCRonym: Entity Extraction and Retrieval for Scanned Books* (co-PI; PI, James Allan, UMass; $50k)

Internal

2023–24 Northeastern University Tier 1 grant: *Archeologies of Data in Contemporary Journalism: Bootstrapping AI Models for Access to Photo Morgues* (PI; co-PI, Meg Heckman; $50k)

2023 Northeastern University Tier 2 grant to support development of NSF Expeditions in Computing proposal on *Narrative Interfaces* (PI; $40k)

2018–19 Northeastern University Tier 1 grant: *Building a Digital Archive of Cherokee: A Lexical Database* (co-PI; PI, Julia Flanders; $50k)

2014–15 Northeastern University Tier 1 grant: *Big Data and Text Curation: TEI and NLP Methods and Intersections* (co-PI; PI, Elizabeth Maddock Dillon; $50k)

2014–15 Northeastern University Tier 1 grant: *Extracting and Visualizing Multi-Dimensional Text Networks: Tools for Structural Reading and Text Exploration* (co-PI; PI, Christoph Riedl; $50k)

2013–14 Northeastern University Tier 1 grant: *Infectious Texts: Uncovering Reprinting Networks in 19th Century Newspapers* (co-PI; PI, Ryan Cordell; $50k)

Teaching and Advising

Courses

Special Topics in AI: Artificial Intelligence as an Archival Science (Northeastern CS7180) Spring 2024
Designer and instructor for graduate course; enrollment: 5

Natural Language Processing (Northeastern CS4120/CS6120) Spring 2013–2017, 2020–21
Designer and instructor for graduate, later undergrad/grad course; enrollment: 21 (2013), 31 (2014), 32 (2015), 37 (2016), 50 (2017), 64 (2020), 68 (2021)

Information Retrieval (Northeastern IS4200/CS6200) Fall 2012–2015, 2019, 2021–23, Spring 2022
Instructor for graduate/undergraduate course; enrollment: 55 (2012), 97 (2013), 111 (2014), 120 (2015: including organizing extra sections), 73 (2019), 60 (2022), 50 (2023)

Foundations of Artificial Intelligence (Northeastern CS4100/5100) Spring 2018, 2023
Instructor for undergrad/grad course; enrollment: 39 (2018), 94 (2023)

Special Topics in AI: Text Modeling for the Humanities and Social Sciences (Northeastern CS7180) Fall 2017
Designer and instructor for graduate course; enrollment: 11

Search Engines (UMass CMPSCI 446) Spring 2012
Instructor for undergraduate course on information retrieval; enrollment: 40

Freshman Computer Science Seminar (UMass CMPSCI 191a) Fall 2011
Co-instructor for Residential Academic Program Seminar; enrollment: 24

Introduction to Natural Language Processing (UMass CMPSCI 585) Fall 2009
Designer and instructor for advanced undergrad/grad course, students from CS and linguistics; enrollment: 17

Mining Text and Images in Digital Libraries Using Grid Computing (UMass CMPSCI 791MT) Spring 2009

Designer and Instructor, with James Allan and R. Manmatha
Graduate seminar with readings and final project; enrollment: 10

Empirical Research Methods in Computer Science (JHU 600.408) Fall 2005
Designer and Primary Instructor (with Noah Smith)
One-credit course for advanced undergraduates and graduate students on computer-intensive statistics and experimental design; enrollment: 18

An Overview of Statistical Machine Translation August 2006
Conference of the Association for Machine Translation in the Americas, Cambridge, MA
Designer and Primary Instructor (with Charles Schafer)
Tutorial on data, models, and algorithms in statistical MT for broad audience; enrollment: 12

Doctoral Supervision

- Jason Naradowsky. UMass Ph.D. student, 2008–2014. Dissertation defended on 30 June 2014, *Learning with Joint Inference and Latent Linguistic Structure in Graphical Models*. Postdoc, University College London and University of Cambridge.
- Kriste Krstovski. UMass Ph.D. student, 2009–2016. Dissertation defended on 4 February 2016, *Efficient Inference, Search and Evaluation for Latent Variable Models of Text with Applications to Information Retrieval and Machine Translation*. Postdoc, Columbia University.
- Shaobin Xu. Northeastern Ph.D. student, 2013–2019. Dissertation defended on 9 December 2019, *Modeling Text Embedded Information Cascades*. Google.
- Liwen Hou. Northeastern Ph.D. student, 2014–2022. Dissertation defense on 29 April 2022, *Detecting and Modeling Syntactic Change*.
- Ansel MacLaughlin. Northeastern Ph.D. student, 2016–2021. Dissertation defended on 13 April 2021, *Analyzing the Usage of Source Texts in New Documents*. Amazon.
- Rui Dong, Northeastern Ph.D. student, 2017–2021. Dissertation defended on 4 August 2021, *Natural Language Processing on Noisy Text*. Amazon.
- Ryan Muther. Northeastern Ph.D. student, 2016–2024. Dissertation defensed on 25 January 2024, *Citation-Augmented Text Reuse Detection*.
- Shijia Liu. Northeastern Ph.D. student, 2020–
- Si Wu. Northeastern Ph.D. student, 2020–
- Jaydeep Borkar. Northeasstern Ph.D. student, 2023–

Doctoral Committees

- Gregory Druck. Advisor, Andrew McCallum. 2009 – 2011 (defended)
- David Mimno. Advisor, Andrew McCallum. 2009 – 2011 (defended)
- Jangwon Seo. Advisor, Bruce Croft. 2010 – 2011 (defended)
- Xiaobing Xue. Advisor, Bruce Croft. 2010 – 2012 (defended)
- Michael Bendersky. Advisor, Bruce Croft. 2010 – 2012 (defended)
- Jinyoung Kim. Advisor, Bruce Croft. 2011 – 2012 (defended)
- Kedar Bellare. Advisor, Andrew McCallum. 2009 (proposal).
- Lisa Friedland. Advisor, David Jensen. 2010 – 2016 (defended)
- Elif Aktolga. Advisor, James Allan. 2012 – 2013 (defended)
- Xiaoxi Xu. Advisor, Beverly Park Woolf. 2012 – 2013 (defended)
- Marc-Allen Cartright. Advisor, James Allan. 2012 – 2013 (defended)
- Jeff Dalton. Advisor, James Allan. 2013 – 2014 (defended)
- Maryam Bashir. Advisor, Javed Aslam. 2013 – 2014 (defended)
- Pavel Metrikov. Advisor, Javed Aslam. 2013 – 2015 (defended)
- Karl Wiegand. Advisor, Rupal Patel. 2013 – 2014 (defended)
- Peter Golbus. Advisor, Javed Aslam. 2013 – 2014 (defended)
- Matthew Ekstrand-Abueg. Advisor, Javed Aslam. 2014 – 2016 (defended)
- Jesse Anderton. Advisor, Javed Aslam. 2017 – 2019 (defended)

- Xinyu Hua. Advisor, Lu Wang. 2020 – 2021 (defended)
- Benjamin Nye. Advisor, Byron Wallace. 2021 – 2022 (defended)
- Sarthak Jain. Advisor, Byron Wallace. 2021 – 2022 (defended)
- Ryan Gallagher. Advisor, Brooke Foucault Welles. 2020 – 2022 (defended)
- Wenjun Sun, La Rochelle Université. Advisor, Mickael Coustaty. 2023–
- Danlu Chen, University of California San Diego. Advisor, Taylor Berg-Kirkpatrick. 2024–

MS Thesis Supervision

- Samuel Scarano. Northeastern M.S. student, 2013–2014. Thesis defended on 17 December 2014, *Applying Unsupervised Grammar Induction to OCR Error Correction*.
- Poonam Bhide. Northeastern M.S. student, 2013–2015. Thesis defended 23 April 2015, *Exploiting Implicit and Explicit Network Structures for Text Classification*.
- Kunal Asarsa. Northeastern M.S. student, 2015–2016. Thesis defended on 12 August 2016, *Analysis of Named Entity Recognition and Entity Linking in Historical Text*.

Other Research Supervised

- Andrew Kae (UMass Ph.D. student). Qualifying synthesis project, with Erik Learned-Miller. 2009–10.
- Jacqueline Feild (UMass Ph.D. student). Qualifying synthesis project, with Erik Learned-Miller. 2009–10.
- David Goff (Cornell undergraduate). Summer REU Site advisee. 2010.
- Weize Kong (UMass Ph.D. student). Qualifying synthesis project, with James Allan. 2012–13.
- Zoe Winkworth. Northeastern B.S. thesis. 2015-2016.
- Maha Alkhairy. Northeastern B.S. independent study. 2016.

Service

Institutional service: Khoury College Area Chair for Artificial Intelligence and Machine Learning Research (2022–24); Elected Faculty Senator (Northeastern 2020–22, 2022–24); Faculty Senate Committee on Library Policies and Operations (Northeastern, 2017–18, 2022–23); Faculty hiring committee (Northeastern, 2014–15, 2016–18); Digital Humanities Search Committee (Northeastern, 2013, 2016); Digital Humanities certificate and curriculum committee (2016–18); Ph.D. committee (Northeastern, 2012–14); graduate program committee (UMass, 2010–11); ad-hoc committee for new institute for computational and experimental linguistics (UMass, 2009–10); curriculum (UMass, 2008–10); graduate student recruiting (JHU, 2003–7), system administration (JHU, 2003–8).

Conference organizing: A Research Agenda for Historical and Multilingual Optical Character Recognition, Northeastern, 2018; Text as Data, Northeastern, 2016: The primary conference at the intersection of computer science and the social sciences.

Journals: Associate Editor, *Harvard Data Science Review*; reviewing for *Computational Linguistics*, *Computers and the Humanities*, *Journal of Artificial Intelligence Research*, *Literary and Linguistic Computing*, *Proceedings of the National Academy of Sciences*.

Conference reviewing: ACL/IJCNLP (NLP Applications, area chair, 2021), EMNLP (Text Mining and Information Retrieval, area co-chair, 2018), ACH/ALLC, ACL, COLING (Machine Learning area chair, 2010), DH, ICML, HLT-NAACL (Information Retrieval area chair, 2015), EACL, EMNLP (Text Mining area chair, 2016), IJCNLP (Machine Learning area chair, 2011), NIPS, SIGIR.

Invited Presentations

- Freie Universität, Zuse Institute, Berlin, March 2024
- Emory University, Department of Quantitative Theory and Methods, October 2023
- Princeton University, Center for Digital Humanities, December 2022
- MIT, Digital Humanities Seminar, September 2022

- EPFL, Digital Humanities Seminar, February 2019
- New York University, Institute for the Study of the Ancient World, April 2018
- University of Richmond, CS Colloquium, February 2018
- Columbia University, NLP Seminar, February 2018
- Google, New York, NY, Tech Talk, February 2018
- Johns Hopkins University, CLSP Seminar, February 2018
- Carnegie Mellon University, LTI Colloquium, October 2017
- University of California, Berkeley, NLP Seminar, September 2017
- Google, Mountain View, CA, Tech Talk, September 2017
- University of Leipzig, Global Philology Workshop, July 2017
- University of South Carolina, Center for Digital Humanities, September 2016
- UCLA, Institute for Pure and Applied Mathematics, May 2016
- Princeton University, Center for Digital Humanities, February 2016
- University of Leipzig, December 2015
- Northwestern University, October 2015
- University of Notre Dame, October 2015
- Bloomberg, July 2015
- National Digital Newspaper Program, NEH, Washington, September 2013
- Wolfram Data Summit, Washington, September 2013
- IBM, T. J. Watson NLP Seminar, December 2012
- University of Leipzig, Computer Science Department, October 2012
- Brown University, Computer Science Department, April 2012
- Virginia Tech, Computer Science Department, March 2012
- CUNY, Graduate Center, March 2012
- Carnegie Mellon University, Language Technologies Institute, March 2012
- University of Chicago, Computer Science Department, February 2012
- Toyota Technical Institute, Chicago, February 2012
- Syracuse University, School of Information Studies, February 2012
- Yale University, Linguistics Department, February 2012
- Cornell University, AI Seminar, February 2012
- Northeastern University, College of Computer and Information Science, February 2012
- Harvard University, Institute for Quantitative Social Science, May 2011
- Princeton University, Computer Science Department, February 2011
- Carnegie Mellon University, Language Technologies Institute, February 2011
- MIT, Computer Science and Artificial Intelligence Laboratory, January 2011
- Humboldt University, Berlin, Institut für deutsche Sprache und Linguistik, January 2011
- UCLA, Institute for Pure and Applied Mathematics, August 2010
- University of Edinburgh, School of Informatics, March 2008
- University of Maryland, Computer Science Department, February 2008

Personal Details

Date of Birth: 27 October 1972

Citizenship: USA

Languages: English (native); ancient Greek, Latin, French, German (reading); Arabic (basic)