Brian Roark Curriculum Vitae

http://www.lanzaroark.org/brian-roark roarkbr@gmail.com
Google, Inc., 555 SW Morrison St., Ste. 500 Portland, OR 97204

Employment

Google	2019–present	Senior Staff Research Scientist
	2013–2019	Staff Research Scientist
Oregon Health & Science University	2008–2013	Associate Professor
	2004–2008	Assistant Professor
		Center for Spoken Language Understanding
		OGI School of Science & Engineering
AT&T Labs–Research	2001–2004	Senior Technical Staff Member

Education

Brown University	2001	Ph.D., Linguistics
		Cognitive and Linguistic Sciences Department
		Thesis: Robust Probabilistic Predictive Syntactic Processing.
		Committee: Mark Johnson (supervisor), Eugene Charniak, Julie Sedivy, Frederick Jelinek.
	2001	Sc.M., Applied Mathematics
Claremont Graduate School	1997	M.S., Information Science
University of California, Berkeley	1989	B.A., Double Major in Mathematics and Philosophy

Research interests

Computational linguistics (CL) and natural language processing (NLP): transliteration and text normalization; language modeling for automatic speech recognition, text entry and other applications; weighted transducers and grammars; supervised and unsupervised learning of language models; pronunciation modeling; text entry, accessibility and augmentative & alternative communication (AAC); syntactic parsing of text and speech; statistical models of human language processing; spoken language processing for diagnosis of neurodevelopmental and neurodegenerative disorders.

Publications

Books

B. Roark and R. Sproat. 2007. Computational Approaches to Morphology and Syntax. Oxford University Press, Oxford.

Refereed Journals

C. Kirov, C. Johny, A. Katanova, A. Gutkin and B. Roark. 2024. Context-aware transliteration of Romanized South Asian languages. *Computational Linguistics*, 50(2):475–534.

A.T. Suresh, B. Roark, M. Riley and V. Schogol. 2021. Approximating probabilistic models as weighted finite automata. *Computational Linguistics*, 47(2):221—254.

T. Pimentel, B. Roark, R. Cotterell. Phonotactic Complexity and Its Trade-offs. 2020. *Transactions of the Association for Computational Linguistics (TACL)*, 8:1–18.

H. Zhang, R. Sproat, A.H. Ng, F. Stahlberg, X. Peng, K. Gorman, B. Roark. 2019. Neural Models of Text Normalization for Speech Applications. *Computational Linguistics*, 45(2):293–337.

U. Orhan, H. Nezamfar, M. Akcakaya, D. Erdogmus, M. Higger, M. Moghadamfalahi, A. Fowler, B. Roark, B. Oken, M. Fried-Oken. 2016. Probabilistic Simulation Framework for EEG-Based BCI Design. *Brain Computer Interfaces*, 3(4):171-185.

E. Prud'hommeaux and B. Roark. 2015. Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549-578.

B. Roark, M. Fried-Oken and C. Gibbons. 2015. Huffman and Linear Scanning Methods with Statistical Language Models. *Augmentative and Alternative Communication*, 31(1):37-50.

R. Sproat, M. Yarmohammadi, I. Shafran and B. Roark. 2014. Applications of Lexicographic Semirings to Problems in Speech and Language Processing. *Computational Linguistics*, 40(4):733-761.

B.S. Oken, U. Orhan, B. Roark, D. Erdogmus, A. Fowler, A. Mooney, B. Peters, M. Miller and M. Fried-Oken. 2014. Brain-computer interface with language model-EEG fusion for locked-in syndrome. *Neurorehabilitation and Neural Repair*, 28(4):387–394.

U. Orhan, D. Erdogmus, B. Roark, B.S. Oken and M. Fried-Oken. 2013. Offline Analysis of Context Contribution to ERP-based Typing BCI Performance. *Journal of Neural Engineering*, 10(6):066003.

B. Roark, R. Beckley, C. Gibbons, M. Fried-Oken. 2013. Huffman scanning: using language models within fixed-grid keyboard emulation. *Computer Speech and Language*, 27(6): 1212–1234.

J. Higginbotham, B. Moulton, G. Lesher and B. Roark. 2012. The Application of Natural Language Processing to Augmentative and Alternative Communication. *Assistive Technology*, 24(1):14–24.

B. Roark, K. Hollingshead and N. Bodenstab. 2012. Finite-state chart constraints for reduced complexity context-free parsing pipelines. *Computational Linguistics*, 38(4):719–753.

E. Arisoy, M. Saraclar, B. Roark and I. Shafran. 2012. Discriminative Language Modeling with Linguistic and Statistically Derived Features. In *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):540-550.

B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead and J.A. Kaye. 2011. Spoken language derived measures for detecting Mild Cognitive Impairment. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7), pp. 2081-2090.

B. Roark, M. Saraclar and M.J. Collins. 2007. Discriminative n-gram language modeling. *Computer Speech and Language*, 21(2), pp. 373-392.

M. Saraclar and B. Roark. 2006. Utterance Classification with Discriminative Language Modeling. *Speech Communication*, 48(3-4), pp. 276-287.

M. Bacchiani, M. Riley, B. Roark and R. Sproat. 2006. MAP Adaptation of Stochastic Grammars. *Computer Speech and Language*, 20(1), pp. 41-68.

C. Allauzen, M. Mohri and B. Roark. 2005. The Design Principles and Algorithms of a Weighted Grammar Library. *International Journal of Foundations of Computer Science*, 16(3), pp. 403-421.

B. Roark. 2004. Robust garden path parsing. *Natural Language Engineering*, 10(1), pp. 1-24.

B. Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2), pp. 249-276.

Refereed Conference Publications

S. Ruder, J.H. Clark, A. Gutkin, M. Kale, M. Ma, M. Nicosia, S. Rijhwani, P. Riley, J.A. Sarr, X. Wang, J. Wieting, N. Gupta, A. Katanova, C. Kirov, D. L. Dickinson, B. Roark, B. Samanta, C. Tao, D.I. Adelani, V. Axelrod, I. Caswell, C. Cherry, D. Garrette, R. Ingle, M. Johnson, D. Panteleev and P. Talukdar. 2023. XTREME-UP: A User-Centric Scarce-Data Benchmark for Under-Represented Languages. In *Findings of EMNLP*, pp. 1856–1884

E. Nielsen, C. Kirov and B. Roark. 2023. Spelling convention sensitivity in neural language models. In *Findings of EACL*, pp. 1304–1316.

R. Doctor, A. Gutkin, C. Johny, B. Roark and R. Sproat. 2022. Graphemic Normalization of the Perso-Arabic Script. To appear in *Proceedings of Grapholinguistics in the 21st Century*.

I. Demirşahin, C. Johny, A. Gutkin and B. Roark. 2022. Criteria for Useful Automatic Romanization in South Asian Languages. In *Proceedings of LREC*, pp. 6662–6673.

A. Gutkin, C. Johny, R. Doctor, L. Wolf-Sonkin and B. Roark. 2022. Extensions to Brahmic script processing within the Nisaba library: new scripts, languages and utilities. In *Proceedings of LREC*, pp. 6450–6460.

K. Gorman, C. Kirov, B. Roark and R. Sproat. 2021. Structured abbreviation expansion in context. In *Findings of EMNLP*, pp. 995–1005.

T. Pimentel, B. Roark, S. Wichmann, R. Cotterell and D. Blasi. 2021. Finding Concept-specific Biases in Form–Meaning Associations. In *Proceedings of NAACL*, pp. 4416–4425.

T. Pimentel, R. Cotterell and B. Roark. 2021. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of EACL*, pp. 31–41.

B. Roark, L. Wolf-Sonkin, C. Kirov, S.J. Mielke, C. Johny, I. Demirşahin and K. Hall. 2020. Processing South Asian languages written in the Latin script: The Dakshina dataset. In *Proceedings of LREC*, pp. 2413–2423.

A. Datta, B. Ramabhadran, J. Emond, A. Kannan and B. Roark. 2020. Language-agnostic multilingual modelling. In *Proceedings of ICASSP*.

A.T. Suresh, B. Roark, M. Riley and V. Schogol. 2019. Distilling weighted finite automata from arbitrary probabilistic models. In *Proceedings of FSMNLP*, pp. 87–97.

L. Wolf-Sonkin, V. Schogol, B. Roark and M. Riley. 2019. Latin script keyboards for South Asian languages with finite-state normalization. In *Proceedings of FSMNLP*, pp. 108–117.

S.J. Mielke, R. Cotterell, K. Gorman, B. Roark and J. Eisner. 2019. What Kind of Language Is Hard to Language-Model? In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pp. 4975–4989.

T. Pimentel, A.D. McCarthy, D. Blasi, B. Roark and R. Cotterell. 2019. Meaning to Form: Measuring Systematicity as Information. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pp. 1751–1764.

R. Cotterell, S.J. Mielke, J. Eisner and B. Roark. 2018. Are All Languages Equally Hard to Language-Model?. In the *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Vol. 2 (Short Papers)*, pp. 536–541.

L. Hellsten, B. Roark, P. Goyal, C. Allauzen, F. Beaufays, T. Ouyang, M. Riley and D. Rybach. 2017. Transliterated mobile keyboard input via weighted finite-state transducers. In the *Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing (FSMNLP)*, pp. 10–19.

V. Kuznetsov, H. Liao, M. Mohri, M. Riley and B. Roark. 2016. Learning n-gram language models from uncertain data. In Proceedings of Interspeech, pp. 2323-2327.

Y. Halpern, K. Hall, V. Schogol, M. Riley, B. Roark, G. Skobeltsyn and M. Baeuml. 2016. Contextual prediction models for speech recognition. In Proceedings of Interspeech, pp. 2338-2342.

P. Aleksic, M. Ghodsi, A. Michaely, C. Allauzen, K. Hall, B. Roark, D. Rybach and P. Moreno. 2015. Bringing Contextual Information to Google Speech Recognition. In *Proceedings of Interspeech*, pp. 468-472.

K. Hall, E. Cho, C. Allauzen, F. Beaufays, N. Coccaro, K. Nakajima, M. Riley, B. Roark, D. Rybach and L. Zhang. 2015. Composition-based on-the-fly rescoring for salient n-gram biasing. In *Proceedings of Interspeech*, pp. 1418-1422.

E. Morley, A.E. Hallin and B. Roark. 2014. Data Driven Grammatical Error Detection in Transcripts of Children's Speech. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 980-989.

K. Wu, C. Allauzen, K. Hall, M. Riley and B. Roark. 2014. Encoding linear models as weighted finite-state transducers. In *Proceedings of Interspeech*, pp.1258-1262.

F. Biadsy, K. Hall, P. Moreno and B. Roark. 2014. Backoff inspired features for maximum entropy language models. In *Proceedings of Interspeech* pp. 2645-2649.

B. Roark and R. Sproat. 2014. Hippocratic Abbreviation Expansion. In *Proceedings of the ACL*, pp. 364-369.

M. Yarmohammadi, A. Dunlop and B. Roark. 2014. Transforming trees into hedges and parsing with "hedgebank" grammars. In *Proceedings of the ACL*, pp. 797-802.

R. Beckley and B. Roark. 2013. Pair Language Models for Deriving Alternative Pronunciations and Spellings from Pronunciation Dictionaries. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1584-1589.

A. Fowler, B. Roark, U. Orhan, D. Erdogmus and M. Fried-Oken. 2013. Improved inference and autotyping in EEG-based BCI typing systems. In *Proceedings of the 15th ACM SIGACCESS International Conference on Computers and Accessibility (ASSETS)*.

E. Dikici, E. Prud'hommeaux, B. Roark and M. Saraclar. 2013. Investigation of MT-based ASR Confusion Models for Semi-Supervised Discriminative Language Modeling. In *Proceedings of Interspeech*, pp. 1218-1222.

B. Roark, C. Allauzen and M. Riley. 2013. Smoothed marginal distribution constraints for language modeling. In *Proceedings of the ACL*.

M. Fried-Oken, U. Orhan, B. Roark, D. Erdogmus, A. Fowler, M. Miller, A. Mooney, B. Oken, B. Peters. 2013. The RSVP Keyboard: A Brain-Computer Interface for Communication by people with Locked-in Syndrome. In *Conference of the Rehabilitation Engineering and Assistive Technology Society of North America (RESNA)*.

M. Rouhizadeh, E. Prud'hommeaux, B. Roark and J. van Santen. 2013. Distributional semantic models for the evaluation of disordered speech. In *Proceedings of HLT-NAACL*, pp. 709–714.

M. Lehr, I. Shafran, E. Prud'hommeaux and B. Roark. 2013. Discriminative Joint Modeling of Lexical Variation and Acoustic Confusion for Automated Narrative Retelling Assessment. In *Proceedings of HLT-NAACL*, pp. 211–220.

U. Orhan, D. Erdogmus, B. Roark, B. Oken, S. Purwar, K. Hild II, A. Fowler, M. Fried-Oken. 2012. Improved Accuracy Using Recursive Bayesian Estimation Based Language Model Fusion in ERP-Based BCI Typing Systems. In *Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'12)*.

M. Lehr, E. Prud'hommeaux, I. Shafran and B. Roark. 2012. Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment. In *Proceedings of Interspeech*, pp. 1039-1042.

P. Xu, S. Khudanpur and B. Roark. 2012. Phrasal Cohort Based Unsupervised Discriminative Language Modeling. In *Proceedings of Interspeech*, pp. 198-201.

D. Karakos, B. Roark, I. Shafran, K. Sagae, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, S. Khudanpur, M. Saraclar, D. Bikel, M. Dredze, C. Callison-Burch, Y. Cao, K. Hall, E. Hasler, P. Koehn, A. Lopez, M. Post and D. Riley. 2012. Deriving conversation-based features from unlabeled speech for discriminative language modeling. In *Proceedings of Interspeech*, pp. 202-205.

A. Çelebi, H. Sak, E. Dikici, M. Saraçlar, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, D. Karakos, S. Khudanpur, B. Roark, K. Sagae, I. Shafran, D. Bikel, C. Callison-Burch, Y. Cao, K. Hall, E. Hasler, P. Koehn, A. Lopez, M. Post, D. Riley. 2012. Semi-supervised discriminative language modeling for Turkish ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5025-5028.

U. Orhan, K.E. Hild II, D. Erdogmus, B. Roark, B. Oken, M. Fried-Oken. 2012. RSVP Keyboard: an EEG based typing interface. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 645-648.

K. Sagae, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, D. Karakos, S. Khudanpur, B. Roark, M. Saraçlar, I. Shafran, D. Bikel, C. Callison-Burch, Y. Cao, K. Hall, E. Hasler, P. Koehn, A. Lopez, M. Post and D. Riley. 2012. Hallucinated n-best lists for discriminative language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5001-5004.

P. Xu, S. Khudanpur, M. Lehr, E. Prud'hommeaux, N. Glenn, D. Karakos, B. Roark, K. Sagae, M. Saraçlar, I. Shafran, D. Bikel, C. Callison-Burch, Y. Cao, K. Hall, E. Hasler, P. Koehn, A. Lopez, M. Post and D. Riley. 2012. Continuous space discriminative language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2129-2132.

A. Dunlop, N. Bodenstab and B. Roark. 2011. Efficient matrix-encoded grammars and low latency parallelization strategies for CYK. In *Proceedings of the 12th International Conference on Parsing Technologies (IWPT)*, pp. 163-174.

U. Orhan, D. Erdogmus, B. Roark, S. Purwar, K. Hild II, B. Oken, H. Nezamfar, M. Fried-Oken. 2011. Fusion with Language Models Improves Spelling Accuracy for ERP-based Brain Computer Interface Spellers. In *33nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'11)*.

E. Tucker Prud'hommeaux and B. Roark. 2011. Extraction of narrative recall patterns for neuropsychological assessment. In *Proceedings of Interspeech*, pp. 3021-3024.

Z. Li, Z. Wang, J. Eisner, S. Khudanpur and B. Roark. 2011. Minimum Imputed-Risk: Unsupervised Discriminative Training for Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 920-929.

N. Bodenstab, A. Dunlop, K. Hall and B. Roark. 2011. Beam-Width Prediction for Efficient CYK Parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 440-449.

N. Bodenstab, K. Hollingshead and B. Roark. 2011. Unary Constraints for Context-Free Parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, short papers, pp. 676-681.

B. Roark, R. Sproat and I. Shafran. 2011. Lexicographic Semirings for Exact Automata Encoding of Sequence Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, short papers, pp. 1-5.

M. Mitchell, A. Dunlop and B. Roark. 2011. Semi-supervised Modeling for Prenominal Modifier Ordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, short papers, pp. 236-241.

E. Tucker Prud'hommeaux, M. Mitchell and B. Roark. 2011. Using Patterns of Narrative Recall for Improved Detection of Mild Cognitive Impairment. In *Conference of the Rehabilitation Engineering and Assistive Technology Society of North America (RESNA) and 3rd International Conference on Technology and Aging (ICTA)*.

C. Whelan, B. Roark, and K. Sönmez. 2010. Designing Antimicrobial Peptides with Weighted Finite State Transducers. In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'10)*.

C. Monson, K. Hollingshead, and B. Roark. 2010. Simulating Morphological Analyzers with Stochastic Taggers for Confidence Estimation. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Revised Selected Papers. Lecture Notes in Computer Science 6241*, pp. 649-657. Springer.

T. Tchoukalov, C. Monson, and B. Roark. 2010. Morphological Analysis by Multiple Sequence Alignment. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Revised Selected Papers. Lecture Notes in Computer Science 6241*, pp. 666-673. Springer.

A. Dunlop, M. Mitchell and B. Roark. 2010. Prenominal Modifier Ordering via Multiple Sequence Alignment. In *Proceedings of the Human Language Technology Conference of the N.American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 600-608.

E. Arısoy, M. Saraçlar, B. Roark and I. Shafran. 2010. Syntactic and sub-lexical features for Turkish discriminative language models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5538-5541

B. Roark and A. Bachrach and C. Cardenas and C. Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 324-333.

B. Roark and K. Hollingshead. 2009. Linear complexity context-free parsing pipelines via chart constraints. In *Proceedings of the Human Language Technology Conference of the N.American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 647-655.

B. Roark and K. Hollingshead. 2008. Classifying chart cells for quadratic complexity context-free inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pp. 745-752.

E. Arisoy, B. Roark, I. Shafran and M. Saraclar. 2008. Discriminative N-gram Language Modeling for Turkish. In *Proceedings of Interspeech 2008*, pp. 825-828.

D. Vergyri, I. Shafran, A. Stolcke, R.R. Gadde, M. Akbacak, B. Roark and W. Wang. 2007. The SRI/OGI 2006 Spoken Term Detection System. In *Proceedings of Interspeech*, pp. 2393-2396.

K. Hollingshead and B. Roark. 2007. Pipeline Iteration. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 952-959.

S. Fisher and B. Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 488-495.

B. Roark, J.P. Hosom, M. Mitchell and J.A. Kaye. 2007. Automatically derived spoken language markers for detecting Mild Cognitive Impairment. In *Proceedings of the 2nd International Conference on Technology and Aging (ICTA)*.

M. Mohri and B. Roark. 2006. Probabilistic Context-Free Grammar Induction Based on Structural Zeros. In *Proceedings of the Human Language Technology Conference of the N.American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 312-319.

J. Hale, I. Shafran, L. Yung, B. Dorr, M. Harper, A. Krasnyanskaya, M. Lease, Y. Liu, B. Roark, M. Snover and R. Stewart. 2006. PCFGs with Syntactic and Prosodic Indicators of Speech Repairs. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 161-169.

B. Roark, M. Harper, E. Charniak, B. Dorr, M. Johnson, J. Kahn, Y. Liu, M. Ostendorf, J. Hale, A. Krasnyanskaya, M. Lease, I. Shafran, M. Snover, R. Stewart and L. Yung. 2006. SParseval: Evaluation Metrics for Parsing Speech. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya and L. Yung. 2006. Reranking for Sentence Boundary Detection in Conversational Speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

K. Hollingshead, S. Fisher and B. Roark. 2005. Comparing and Combining Finite-State and Context-Free Parsers. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pp. 787-794.

M.J. Collins, M. Saraclar and B. Roark. 2005. Discriminative Syntactic Language Modeling for Speech Recognition. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 507-514.

M. Saraclar and B. Roark. 2005. Joint Discriminative Language Modeling and Utterance Classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 561-564.

C. Allauzen, M. Mohri and B. Roark. 2004. A General Weighted Grammar Library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata (CIAA)*, volume 3317 of Lecture Notes in Computer Science, Springer-Verlag, pp. 23-34.

M.J. Collins and B. Roark. 2004. Incremental Parsing with the Perceptron Algorithm. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 111-118.

B. Roark, M. Saraclar, M.J. Collins and M. Johnson. 2004. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 47-54.

M. Bacchiani, B. Roark and M. Saraclar. 2004. Language model adaptation with MAP estimation and the perceptron algorithm. In *Proceedings of the Human Language Technology Conference of the N.American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 21-24.

B. Roark, M. Saraclar and M.J. Collins. 2004. Corrective language modeling for large vocabulary ASR with the perceptron algorithm. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 749-752.

M. Bacchiani and B. Roark. 2004. Meta-data conditional language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 241-244.

C. Allauzen, M. Mohri, M. Riley and B. Roark. 2004. A Generalized Construction of Speech Recognition Transducers. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 761-764.

S.R. Maskey, M. Bacchiani, B. Roark and R. Sproat. 2004. Improved name recognition with meta-data dependent name networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 789-792.

C. Allauzen, M. Mohri and B. Roark. 2003. Generalized algorithms for constructing language models. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 40-47.

B. Roark and M. Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. In *Proceedings of the Human Language Technology Conference of the N.American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 205-212.

M. Bacchiani and B. Roark. 2003. Unsupervised language model adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 224-227.

B. Roark. 2002. Markov parsing: lattice rescoring with a statistical parser. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 287-294.

B. Roark. 2001. Explaining vowel inventory tendencies via simulation: finding a role for quantal locations and formant normalization. In *Proceedings of the 31st Conference of the North East Linguistics Society (NELS 31)*, pp. 419-434.

T. Pepinsky, K. Demuth, and B. Roark. 2001. The status of 'filler syllables' in children's early speech. In *Proceedings of the 25th annual Boston University Conference on Language Development*, pp. 575-586.

M. Johnson and B. Roark. 2000. Compact non-left-recursive grammars using the selective left-corner transform and factoring. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pp. 355-361.

B. Roark and K. Demuth. 2000. Prosodic constraints and the learner's environment: a corpus study. In *Proceedings of the 24th annual Boston University Conference on Language Development*, pp. 597-608.

B. Roark and M. Johnson. 1999. Efficient probabilistic top-down and left-corner parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 421-428.

B. Roark and E. Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, pp. 1110-1116.

Refereed Workshop or System Demonstration Publications

K. Gorman and B. Roark. 2024. Abbreviation across the world's languages and scripts. In *Proceedings of the Second Workshop on Computation and Written Language (CAWL) at LREC/Coling*, pp. 36–42.

E. Nielsen, C. Kirov and B. Roark. 2023. Distinguishing Romanized Hindi from Romanized Urdu. In *Proceedings of the First Workshop on Computation and Written Language (CAWL) at ACL*, pp. 33–42.

A. Gutkin, C. Johny, R. Doctor, B. Roark and R. Sproat. 2022. Beyond Arabic: Software for Perso-Arabic Script Manipulation. In *Proceedings of the Arabic Natural Language Processing Workshop (WANLP)*, pp. 381–387.

B. Roark and A. Gutkin. 2022. Design principles of an open-source language modeling microservice package for AAC text-entry applications. In *Proceedings of the ACL Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pp. 1–16.

C. Johny, L. Wolf-Sonkin, A. Gutkin and B. Roark. 2021. Finite-state script normalization and processing utilities: The Nisaba Brahmic library. In *Proceedings of the EACL Demo Session*, pp. 14–23.

J. Emond, B. Ramabhadran, B. Roark, P. Moreno, M. Ma. 2018. Transliteration based approaches to improve code-switched speech recognition performance. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*.

C. Allauzen, M. Riley and B. Roark. 2016. Distributed representation and estimation of WFST-based n-gram models. In *Proceedings of the ACL Workshop on statistical NLP and weighted automata*, pp. 32-41.

E. Prud'hommeaux, E. Morley, M. Rouhizadeh, L. Silverman, J. van Santen, B. Roark, R. Sproat, S. Kauper, R. DeLaHunta. 2014. Computational Analysis of the Trajectories of Linguistic Development in Autism. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*.

E. Morley, A.E. Hallin and B. Roark. 2014. Challenges in Automating Maze Detection. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pp. 69-77.

E. Morley, B. Roark and J. van Santen. 2013. The Utility of Manual and Automatic Linguistic Error Codes for Identifying Neurodevelopmental Disorders. In *Proceedings of the NAACL-HLT 2013 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*, pp. 1–10.

E. Prud'hommeaux and B. Roark. 2012. Graph-based alignment of narratives for automated neurological assessment. In *Proceedings of the NAACL 2012 Workshop on Biomedical Natural Language Processing (BioNLP)*, pp. 1-10.

B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen and T. Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 Demo Session*, pp. 61-66.

Steven Bedrick, Russell Beckley, Brian Roark and Richard Sproat. 2012. Robust kaomoji detection in Twitter. In *Proceedings of the NAACL 2012 Workshop on Language in Social Media (LSM)*, pp. 56-64.

P.O. Kristensson, J. Clawson, M. Dunlop, P. Isokoski, B. Roark, K. Vertanen, A. Waller, J.O. Wobbrock. 2012. Designing and Evaluating Text Entry Methods. Workshop overview, CHI 2012, Austin, TX.

I. Shafran, R. Sproat, M. Yarmohammadi and B. Roark. 2011. Efficient Determinization of Tagged Word Lattices using Categorial and Lexicographic Semirings. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 283-288.

E. Prud'hommeaux and B. Roark. 2011. Alignment of spoken narratives for automated neuropsychological assessment. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 484-489.

B. Roark, A. Fowler, R. Sproat, C. Gibbons and M. Fried-Oken. 2011. Towards technology-assisted co-construction with communication partners. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pp. 22-31.

R. Beckley and B. Roark. 2011. Asynchronous fixed-grid scanning with dynamic codes. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pp. 43-51.

E. Prud'hommeaux, B. Roark, J. van Santen and L. Black. 2011. Classification of atypical language in autism. In *Proceedings of the ACL 2011 Workshop on Cognitive Modeling and Computational Linguistics*, pp. 88-96.

K. Hild, U. Orhan, D. Erdogmus, B. Roark, B. Oken, S. Purwar, H. Nezamfar and M. Fried-Oken. 2011. An ERP-based Brain-Computer Interface for text entry using Rapid Serial Visual Presentation and Language Modeling. In *Proceedings of the ACL 2011 Demo Session*, pp. 38-43.

B. Roark, J. de Villiers, C. Gibbons and M. Fried-Oken. 2010. Scanning methods and language modeling for binary switch typing. In *Proceedings of the NAACL-HLT Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pp. 28-36.

J.G. Kahn, M. Ostendorf and B. Roark. 2008. Automatic syntactic MT evaluation with expected dependency pair match. In *Proceedings of the NIST MetricsMATR Workshop*.

B. Roark, M. Mitchell and K. Hollingshead. 2007. Syntactic complexity measures for detecting Mild Cognitive Impairment. In *Proceedings of the ACL 2007 Workshop on Biomedical Natural Language Processing (BioNLP)*, pp. 1-8.

M. Riley, B. Roark and R. Sproat. 2003. Good-Turing estimation from word lattices for unsupervised language model adaptation. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 453-458.

B. Roark. 2002. Evaluating parser accuracy using edit distance. In *Proceedings of the LREC-2002 Workshop "Beyond PARSEVAL: Towards Improved Evaluation Measures for Parsing Systems"*, pp. 30-36.

B. Roark. 2001. Storing automatically generated treebanks in lattices of derivations. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pp. 210-218.

B. Roark and E. Charniak. 2000. Measuring efficiency in high-accuracy, broad-coverage statistical parsing. In *Proceedings of the COLING-2000 Workshop on Efficiency in Large-scale Parsing Systems*, pp. 29-36.

Refereed Book Chapters or Paper Collections

B. Roark. 2009. A survey of discriminative language modeling approaches for large vocabulary continuous speech recognition. In *Large Margin and Kernel Approaches to Speech and Speaker Recognition*, Joseph Keshet and Samy Bengio, Eds. Wiley, pp. 117-137.

Invited Overview Papers or Book Reviews

K. McCoy, J. Arnott, L. Ferres, M. Fried-Oken and B. Roark. 2013. Speech and Language Processing as Assistive Technologies: Introduction to the special issue on speech and language processing for assistive technologies. *Computer Speech and Language*, 27(6):1143–1146.

B. Roark. 2007. Review of Rayner et al. "Putting Linguistics into Speech Recognition". *Computational Linguistics*, , 33(2), pp. 271-273.

Other Publications

D. Bikel, C. Callison-Burch, Y. Cao, N. Glenn, K. Hall, E. Hasler, D. Karakos, S. Khudanpur, P. Koehn, M. Lehr, A. Lopez, M. Post, E. Prud'hommeaux, B. Roark, D. Riley, K. Sagae, M. Saracclar, I. Shafran and P. Xu. 2011. Confusion-based Statistical Language Modeling for Machine Translation and Speech Recognition. Final Report from 2011 CLSP Summer Workshop.

B. Roark. 2011. Expected surprisal and entropy. Technical Report CSLU-2011-04, Center for Spoken Language Processing, Oregon Health & Science University.

A. Dunlop, N. Bodenstab and B. Roark. 2010. Reducing the grammar constant: an analysis of CYK parsing efficiency. Technical Report CSLU-2010-02, Center for Spoken Language Processing, Oregon Health & Science University.

N. Bodenstab, A. Dunlop, B. Roark and K. Hall. 2010. Exponential Decay Pruning for Bottom-Up Beam-Search Parsing. Paper presented at NW-NLP Workshop, Microsoft Research, April.

C. Monson, K. Hollingshead, and B. Roark. 2009. Probabilistic ParaMor. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece.

T. Tchoukalov, C. Monson, and B. Roark. 2009. Multiple Sequence Alignment for Morphology Induction. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece.

B. Roark. 2009. Open vocabulary language modeling for binary response typing interfaces. Technical Report CSLU-09-001, Center for Spoken Language Processing, Oregon Health & Science University.

S. Fisher, A. Dunlop, B. Roark, Y. Chen and J. Burmeister. 2009. OHSU Summarization and Entity Linking Systems. In *Proceedings of the Text Analysis Conference (TAC)*.

K. Hollingshead and B. Roark. 2008. Reranking with baseline system scores and ranks as features. Technical Report CSLU-08-001, Center for Spoken Language Understanding, Oregon Health & Science University.

B. Roark. 2007. Structural Alignment for Finite-State Syntactic Processing. Technical Report CSLU-07-001, OGI School of Science & Engineering, Oregon Health & Science University.

A.M. Cohen, J. Yang, S. Fisher, B. Roark and W.R. Hersh. 2007. The OHSU Biomedical Question Answering System Framework. In *Proceedings of the Text Retrieval Conference (TREC)*.

S. Fisher and B. Roark. 2007. Feature expansion for query-focused supervised sentence ranking. In *Document Understanding (DUC-2007) Workshop Papers and Agenda*.

A.M. Cohen, J. Yang, S. Fisher, B. Roark and W.R. Hersh. 2006. Combining Lexicon Expansion, Information Retrieval, and Cluster-based Ranking for Biomedical Question Answering. In *Proceedings of the Text Retrieval Conference (TREC)*.

S. Fisher and B. Roark. 2006. Query-focused summarization by supervised sentence ranking and skewed word distributions. In *Document Understanding (DUC-2006) Workshop Papers and Agenda*.

S. Fisher, B. Roark, J. Yang and B. Hersh. 2005. OGI/OHSU baseline query-directed multi-document summarization system for DUC 2005. In *Document Understanding (DUC-2005) Workshop Papers and Agenda*, pp. 100-102.

M. Mohri and B. Roark. 2005. Structural Zeros versus Sampling Zeros. Technical Report CSEE-05-003, OGI School of Science & Engineering, Oregon Health & Science University.

Presentations

Refereed Conference Presentations (unpublished)

T. Pimentel, B. Roark and R. Cotterell. 2019. Rethinking Phonotactic Complexity. Poster at *Second annual meeting of the Society for Computation in Linguistics (SCiL)*, New York, January.

R. Cotterell, S.J. Mielke, J. Eisner and B. Roark. 2019. Are All Languages Equally Hard to Language-Model?. Poster at *Second annual meeting of the Society for Computation in Linguistics (SCiL)*, New York, January.

U. Orhan, M. Akcakaya, D. Erdogmus, B. Roark, M. Moghadamfalahi, M. Fried-Oken. 2013. Comparison of Adaptive Symbol Presentation Methods for RSVP Keyboard. Poster at *Fourth International BCI Meeting*, Pacific Grove, CA, June.

B. Peters, D. Erdogmus, A. Fowler, A. Mooney, B. Oken, U. Orhan, B. Roark, M. Fried-Oken. 2013. Effects of Varying Presentation Rate and Sequence Length on User Performance with the RSVP Keyboard BCI. Poster at *Fourth International BCI Meeting*, Pacific Grove, CA, June.

E. T. Prud'hommeaux, M. Rouhizadeh, B. Roark and J. van Santen. 2013. Identifying Unexpected and Inappropriate Words in ASD Language Samples. Poster at *The International Meeting for Autism Research (IMFAR)*, San Sebastián.

B. Roark, M. Fried-Oken and R. Sproat. 2012. Communication partner co-construction in speech generating devices. Presentation at *15th Biennial Conference of the International Society for Augmentative and Alternative Communication (ISAAC)*, Pittsburgh, July.

B. Oken, U. Orhan, K. Hild, D. Erdogmus, B. Roark, M. Miller, A. Mooney, S. Purwar, and M.B. Fried-Oken. 2013. EEG-based typing interface with language model for individuals who are functionally locked-in. 65th Annual Meeting of the American Academy of Neurology, San Diego.

E. T. Prud'hommeaux, B. Roark, L. M. Black and J. van Santen. 2012. Identifying Features of ASD Language Impairment in Narrative Retellings. Poster at *The International Meeting for Autism Research (IMFAR)*, Toronto.

U. Orhan, D. Erdogmus, K. E. Hild II, B. Roark, B. Oken, M. Fried-Oken. 2011. Context Information Significantly Improves Brain Computer Interface Performance – a Case Study on Text Entry Using a Language Model Assisted BCI. In *Proceedings of Asilomar SSC*, 2011.

R. W. Sproat, L. M. Black, E. T. Prud'hommeaux, J. van Santen and B. Roark. 2011. Automated Analysis of Natural Language Samples: Comparison of Children with Autism Spectrum Disorders, Developmental Language Disorders, and Typical Development. Poster at *The International Meeting for Autism Research (IMFAR)*, San Diego.

C. Whelan, B. Roark, C. Tamerler, M. Sarikaya and K. Sönmez. 2010. Design of Inorganic Binding Peptides by a Finite-State Transducer Based Transform Framework. Poster at *Intelligent Systems for Molecular Biology (ISMB)*, Boston, July.

B. Roark, C. Gibbons and M. Fried-Oken. 2010. Binary coding with language models for EEG-based access methods. Poster at *14th Biennial Conference of the International Society for Augmentative and Alternative Communication (ISAAC)*, Barcelona, July.

D. Erdogmus, K. Hild, B. Oken, B. Roark, M. Fried-Oken. 2010. Initial design of an AAC device with non-invasive BCI, adaptive language modeling and Rapid Serial Visual Presentation (RSVP). Poster at *Fourth International BCI Meeting*, Pacific Grove, CA, May/June.

E.T. Prud'hommeaux, J. van Santen, L. Black, and B. Roark. 2010. Automatic detection of idiosyncratic word use in autism spectrum disorders. Poster at *The International Meeting for Autism Research (IMFAR)*, Philadelphia, May.

A. Bachrach, B. Roark, C. Cardenas and C. Pallier. 2009. Parser derived surprisal and entropy measures: The effect of corpus and lexical information. Presentation given at *The Annual Conference on Architectures and Mechanisms for Language Processing*, Barcelona, September.

B. Roark and M. Johnson. 1999. Broad-coverage predictive parsing. Paper presented at *the 12th Annual CUNY Conference on Human Sentence Processing*, New York, March.

Invited Talks

Romanization, non-standard orthography and text entry.
- NAACL workshop on subword & character level models in NLP, New Orleans, June 2018.
- Center for Language and Speech Processing (CLSP), Johns Hopkins University, July, 2018.

Natural language modeling in AAC and BCI.
- The 7th International BCI Meeting, workshop on NLP & BCI, Pacific Grove, CA, May 2018.

Natural language modeling for efficient text entry.
- Facultad de Ingeniería. Universidad de la República, Montevideo, Uruguay, Nov. 2017

Pronunciation modeling. Oregon Health & Science University, Jan 2017 (NLP course guest lecture)

Good-Turing estimation from uncertain data for semi-supervised language model adaptation
- Center for Language and Speech Processing (CLSP), Johns Hopkins University, July, 2016

Distributed representation and estimation of WFST-based n-gram models
- Oregon Health & Science University, May 2016 (NLP course guest lecture)

MaxEnt, features and marginal distritubion constraints
- New York University, April 2015 (ASR course guest lecture)

Learning high precision text normalization systems from (mostly) unlabeled data
- Columbia University, Sept. 2014 ("From Data to Solutions" program)

Alignment of clinically-elicited spoken language samples for neuropsychological assessment
- Facultad de Ingeniería. Universidad de la República, Montevideo, Uruguay, Mar. 2013

Imposing marginal distribution constraints on language models
- Google Research, NYC, Jan., 2013

Hallucinating system outputs for discriminative language modeling
- Aalto University, Finland, Dec., 2012
- Nuance Communications Technical Speaker Series, Webinar, Nov., 2012
- Symposium on Machine Learning in Speech and Language Processing, Portland, Sept., 2012

Confusion-based Statistical Language Modeling
- NAACL-HLT Workshop on the Future of Language Modeling for HLT, Montreal, June, 2012

Improving text entry in augmentative and alternative communication devices with language models
- Universities of Texas (Dallas) and Oregon, May, 2012

Natural Language Processing for psycholinguistic modeling and neuropsychological assessment
- Universities of Dundee and Aberdeen (SICSA visit), Sept. 2011

Learning hard chart constraints for efficient context-free parsing
- Universities of Edinburgh and St. Andrews (SICSA visit), Sept. 2011; Cambridge, Oct. 2011

Language modeling and coding for fixed grid keyboard emulation interfaces
- Universities of Edinburgh, Dundee and Aberdeen (SICSA visit), Sept. 2011

Applications of Lexicographic Semirings in Speech and Language Processing
- University of Edinburgh (SICSA visit), Sept. 2011; Cambridge, Oct. 2011; Facultad de Ingeniería. Universidad de la República, Montevideo, Uruguay, Dec. 2011

Open vocabulary language modeling for binary switch typing interfaces
- Center for Language and Speech Processing (CLSP), Johns Hopkins University, Oct., 2010
- Courant Institute of Mathematical Sciences, New York University, Oct., 2010

Classifying chart cells for quadratic complexity context-free inference
- Signal, Speech and Language Interpretation Lab (SSLI), University of Washington, July, 2008

The Utility of Parse-derived Features for Automatic Discourse Segmentation
- Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay, Dec. 2007

Structural alignment for finite-state syntactic processing
- Center for Language and Speech Processing (CLSP), Johns Hopkins University, Sept., 2007
- Signal, Speech and Language Interpretation Lab (SSLI), University of Washington, Oct., 2007

Joint discriminative language modeling and utterance classification
- Microsoft Research, Redmond and CLSP, Johns Hopkins, September, 2004

Efficient incremental beam-search parsing with generative and discriminative models
- *ACL workshop on Incremental Parsing*, Barcelona, July, 2004

Supervised and unsupervised statistical grammar adaptation to novel domains
- OGI School of Science & Engineering, Oregon Health & Science University, October, 2003

Incremental stochastic parsing for speech recognition
- Departments of Linguistics at: Georgetown University; University of Texas, Austin; University of Washington; University of Michigan; and UCLA. February, 2002.

Lexical and Lexico-syntactic probabilistic language models. Brown University, IGERT Conference on Stochastic and Deterministic Approaches to Vision, Language, and Cognition, May, 2001.

Incremental Broad-coverage Predictive Parsing
- University of Minnesota, Departments of Psychology and CSE, January, 2001.
- University of Michigan, Department of Linguistics, January, 2001.
- San Diego State University, Department of Linguistics, December, 2000.
- *The 2nd Annual Graduate Conference of the Northeastern Cognitive Science Society*, University of Pennsylvania, IRCS, May, 1999.

Probabilistic top-down parsing and language modeling.
- AT&T Labs - Research, New Jersey, December, 2000.
- IBM Research, New York, October, 2000.
- Johns Hopkins, The Center for Language and Speech Processing, June, 2000.

Selected Honors, Fellowships and Awards

Best Paper Award for articles published in Computer Speech and Language (2007-2011), for "Discriminative n-gram language modeling" (with Murat Saraçlar and Michael Collins)

Excellence in Teaching Award, OHSU 2010-2011 (selected by graduate students)

Best Short Paper Award, ACL 2011, for "Lexicographic Semirings for Exact Automata Encoding of Sequence Models" (with Richard Sproat and Zak Shafran).

Scottish Informatics & Computer Science Alliance (SICSA) Distinguished Visiting Fellowship, 2011.

Google Faculty Research Award, 2009

Member, DARPA Computer Science Study Panel, 2008.

Best Paper Award, HLT-NAACL 2006, for "Probabilistic Context-Free Grammar Induction Based on Structural Zeros" (with Mehryar Mohri).

Patents

A. Datta, B. Ramabhadran, J. Emond, B. Roark. 2023. Language-agnostic multilingual modeling using effective script normalization. U.S. Patent Number 11,615,779

B. Ramabhadran, M. Ma, P.J. Moreno Mengibar, J. Emond, B. Roark. 2022. Transliteration for speech recognition training and scoring. U.S. Patent Number 11,417,322.

W.F.N. Quah, B. Horling, M. Garrett, B. Roark and R. Sproat. 2017. Generating output for presentation in response to user interface input, where the input and/or the output include chatspeak. U.S. Patent Number 10,268,683.

F. Biadsy and B. Roark. 2014. Enhanced maximum entropy models. U.S. Patent Number 9,412,365.

M. Bacchiani and B. Roark. 2004. System and method for using meta-data dependent language modeling for automatic speech recognition. U.S. Patent Number 7,752,046.

M. Bacchiani, S. Maskey, B. Roark and R. Sproat. 2003. System and method of using meta-data in speech processing. U.S. Patent Number 7,996,224.

Selected Grants

NIH 1R01DC012033-01A1. Computational characterization of language use in autism spectrum disorder (PI: Richard Sproat). Funded as co-PI (20%). 09/01/2011 - 08/31/2016.

NSF IIS-0964102. RI-Medium: Semi-supervised Discriminative Training of Language Models (PI: Brian Roark; Collaborative with PI Sanjeev Khudanpur, JHU). 07/01/2010 - 06/30/2013. $500,000.

NIH 1R01DC009834-01. Translational refinement of adaptive communication system for locked-in patients (PI: Melanie Fried-Oken). Funded as co-PI (25%), plus one student, 02/01/2009 - 01/31/2014.

NSF BCS-0826654. DHB: Measuring Spoken Language Variability in Elderly Individuals (PI: Brian Roark). 11/01/2008 - 10/31/2011. $750,000.

NSF IIS-0811745. RI-Small: Efficient hidden structure annotation via structural multiple-sequence alignments (PI: Brian Roark). 08/01/2008 - 07/31/2011. $400,000.

DARPA HR0011-09-1-0041. CSSG: Learning within NLP pipelines for scalable data mining and information extraction (PI: Brian Roark). 05/07/2009 – 05/06/2011.

NSF IIS-0447214. CAREER: Discriminative Syntactic Language Modeling: Automatic Feature Selection and Efficient Annotation (PI: Brian Roark). 04/01/2005 - 03/31/2010. $500,000.

Autism Speaks. Automated Measurement of Dialogue Structure in Autism (PI: Brian Roark). 12/01/2007 - 11/30/2009. $100,000.

Department of Defense. Computer Science Study Group Participant Proposal (PI: Brian Roark). 02/01/2008 - 01/31/2009. $100,000.

NSF IIS-0741585. SGER: RI: Text-Based Discriminative Language Modeling (PI: Brian Roark). 09/01/2007 - 08/31/2008. $100,000.

OHSU Medical Research Foundation. Differentiating between Autism Spectrum Disorder and Developmental Language Disorder via Story Recall (PI: Brian Roark). 12/01/2006 - 11/30/2007. $30,000.

Intel. OHSU BAIC: Technologies for behavioral assessment and intervention (PI: Tamara Hayes). Funded as co-PI (15%), 09/25/2006 - 09/24/2007.

Oregon Center for Aging and Technology. Automated Analysis of Spoken Story Recall Tests (PI: Brian Roark). 08/01/2005 - 07/31/2006. $50,000.

Oregon Partnership for Alzheimer's Research, Tax Checkoff Research Fund. Annotation and Automatic Approximation of Language-use Metrics for Detection of Mild Cognitive Impairment (PI: Brian Roark). 05/01/2005 - 04/30/2006. $25,000.

NSF IIS-0313383. Objective Methods for Predicting and Optimizing Synthetic Speech Quality (PI: Jan van Santen). Funded as co-PI (2 months), 09/01/2004 - 08/31/2005.

Teaching and Advising

Graduate courses taught, Program in Computer Science & Engineering, OGI/OHSU:
- CSE606 Topics in Information Retrieval: 2012(f) co-taught with S. Bedrick and E. Prud'hommeaux
- CSE606 Topics in Natural Language Processing: 2012(sp) co-taught with E. Prud'hommeaux
- CSE606 Speech and Language Processing for Augmentative and Alternative Communication: 2010(sp)
- CSE662 Natural Language Processing: 2005-2007(w), 2008(f), 2010(w) (co-taught w/ K. Hollingshead in 2010)
- CSE655 Biological and Linguistic Sequence Analysis: 2007(sp), 2008(w), 2011(w)
- CSE654 Text-Based Language Processing Systems: 2008(sp)
- CSE606 Computational Approaches to Speech and Language Disorders: 2006(su) co-taught with 3 others

Other courses taught
- Biological and Linguistic Sequence Analysis, intensive course, December 1-19, 2008
  Facultad de Ingeniería. Universidad de la República, Montevideo, Uruguay

Ph.D. students advised (OHSU Computer Science & Engineering program):

- Russ Beckley, 2010–2015. Went to Ernst & Young.
- Nate Bodenstab, 2005–2007 (M.S., 2006), 2009–2012. Ph.D., Aug., 2012. Went to Nuance.
  Now Senior Research Scientist Manager at Nuance.
- Yongshun Chen, 2008–2011 (M.S., 2011).
- Aaron Dunlop, 2008–2014. Ph.D., Jan., 2014. Research Scientist at Intel 2013–2016.
  Now Senior Research Scientist at Nuance.
- Seeger Fisher, 2004–2011.
- Andrew Fowler, 2011–2018. Ph.D. Mar., 2020 (Steven Bedrick, advisor).
  Software Engineer at Google 2015-2018. Now Senior Researcher at Nuance.
- Kristy Hollingshead, 2004–2010. Ph.D., Aug., 2010. Post-doc at UMD 2010–2012. At DoD, 2012-2014.
  Now Research Scientist at Florida Institute for Human & Machine Cognition (IHMC).
- Eric Morley, 2012–2016. Ph.D., April 2016. Senior Data Scientist at Red Owl Analytics 2015–2016.
  Goldman Sachs, 2016-2018. Now Software Engineer at Google.
- Emily T. Prud'hommeaux, 2009–2012. Ph.D., Aug., 2012. Post-doc at U. of Rochester 2013–2014.
  Assistant Professor at Rochester Institute of Technology 2014-2018.
  Now Assistant Professor at Boston College.
- Mahsa Yarmohammadi, 2009–2016. Ph.D., Sept. 2016. Software Architect at Intel 2015-2017.
  Now Assistant Research Scientist at Johns Hopkins University.

Post-doctoral fellows
- Steven Bedrick (Ph.D. OHSU, 2011), 2/2011–2/2013. Now faculty at OHSU.
- Christian Monson (Ph.D. CMU, 2008), 12/2008–11/2010. Senior Research Scientist at Nuance 2010–2013.
  Now Machine Learning Scientist at Amazon.

Visiting Ph.D. students advised:
- Ebru Arisoy (Bogaziçi Univ.), 3/2007–12/2007. Now Asst. Professor at MEF University, Istanbul.
- Erinç Dikici (Bogaziçi Univ.), 10/2012–1/2013. Now Speech Scientist at SAIL LABS Technology, Vienna.
- Meg Mitchell (Univ. of Aberdeen, Scotland), 8/2010–4/2012. Now Research Scientist at Hugging Face.

Member of completed Ph.D. Thesis committees
- Kyle Ambert, Oregon Health & Science University, Biomedical Informatics. Defended 2013.
- Meysam Asgari, Oregon Health & Science University, Computer Science & Engineering. Defended 2014.
  Shiran Dudy, Oregon Health & Science University, Computer Science & Engineering. Defended 2020.
- Tom Kalt, University of Massachusetts, Amherst, Computer Science. Defended 2005.
- Maider Lehr, Oregon Health & Science University, Computer Science & Engineering. Defended 2014.
- Chu-Cheng Lin, Johns Hopkins University. Defended 2022.
- Sabrina J. Mielke, Johns Hopkins University. Defended 2023.
- Ethan Selfridge, Oregon Health & Science University, Computer Science & Engineering. Defended 2013.
- Chris Whelan, Oregon Health & Science University, Computer Science & Engineering. Defended 2014.
- Fan Yang, Oregon Health & Science University, Computer Science & Engineering. Defended 2008.
- Jianji Yang, Oregon Health & Science University, Biomedical Informatics. Defended 2007.

Ph.D. Thesis reader/opponent/referee/examiner
- Luis Chiruzzo, Universidad de la República, Uruguay, 2020.
- Trevor Cohn, University of Melbourne, Australia, 2006.
- Felice Dell'Orletta, Università di Pisa, Italy, 2008.
- Kathleen Fraser, University of Toronto, Canada, 2016
- Guillermo Moncecchi, Universidad de la República, Uruguay, 2013.
- Ehsan Shareghi, Monash University, Australia, 2017.
- Sami Virpioja, Aalto University, Finland, 2012.
- Yufei Wang, MaquarieUniversity, Australia, 2022.

Summer intern host at Google:
- Elizabeth Nielsen, University of Edinburgh, 2022 (co-hosted with Christo Kirov).
- Sabrina J. Mielke, Johns Hopkins University, 2019.
- Felix Stahlberg, Cambridge University, 2017.
- Ryan Cotterell, Johns Hopkins University, 2017 (co-hosted with Vlad Schogol).

Undergraduate summer interns advisor at OHSU (mainly through NSF-REU program):
- 2008: Joseph Greer (Oregon); Adam Hesterberg (Princeton)
- 2009: Josh Burmeister (Northern Colorado); James Elwell (Lewis & Clark); Elijah Hamovitz
  (Oregon); Tzvetan Tchoukalov (Stanford)
- 2010: Kari Baker (Arizona); Jessica Ferguson (Pacific); Tzvetan Tchoukalov (Stanford)

Software and Data

Dakshina dataset. A collection of text in both Latin and native scripts for 12 South Asian languages.

OpenGrm library (with Cyril Allauzen, Michael Riley, Richard Sproat and Terry Tai). An open-source collection of libraries for constructing, combining, applying and searching formal grammars.

Incremental top-down parser. Open-source code for incremental statistical parsing and training of statistical parsing models.

BUBS parser: written by PhD students Nate Bodenstab and Aaron Dunlop.

GRM library (with Cyril Allauzen and Mehryar Mohri). A general software collection for constructing and modifying weighted automata and transducers representing grammars or language models. (AT&T library, no longer available.)

Professional activities and affiliations

CO-EDITOR IN CHIEF, Transactions of the Association for Computational Linguistics (TACL): 2018-2022.

ACTION EDITOR

ACL Rolling Review (ARR): 2024-present.

Transactions of the Association for Computational Linguistics (TACL): 2012-2018.

EDITORIAL BOARD MEMBER, Computational Linguistics (2003-2005)

GUEST EDITOR, special issue of Computer Speech and Language on Speech and Language Processing for Assistive Technologies, 27(6), September 2013

MEMBER, IEEE Signal Processing Society Speech and Language Technical Committee (2006-2008)

MEMBER, ACL working group on anonymity and reviewing policy, 2023.

BOARD MEMBER, ACL Special Interest Group in Speech & Language Processing for Assistive Technologies (SIG-SLPAT), founding (interim) Secretary-Treasurer, 2011-2012.

ADVISORY BOARD MEMBER, NIDILRR funded Rehabilitation Engineering Research Center on Augmentative and Alternative Communication (RERC on AAC), 2015-2019.

PROGRAM COMMITTEE AREA CHAIR

Annual Meeting of the Association for Computational Linguistics (ACL): 2005, 2014

Conference on Empirical Methods in Natural Language Processing (EMNLP): 2016, 2023, 2024

Human Language Technology / N.American Chapter of the ACL (HLT-NAACL): 2004, 2007

ORGANIZING COMMITTEE MEMBER

Co-organizer, Workshop on Computation and Written Language (CAWL) at ACL, 2023; and at LREC-COLING, 2024.

Co-organizer, Workshop on Designing and Evaluating Text Entry Methods at CHI, 2012

Local arrangements chair, 49th Annual Meeting of the ACL, 2011

Co-organizer, Workshop on Speech and Language Processing for Assistive Technologies (SLPAT) at HLT-NAACL, 2010; and at HLT-NAACL, 2012

Best paper selection committee, ACL 2010; 2014

Tutorials co-chair, HLT-NAACL, 2009

Student Research Workshop co-chair (Faculty advisor), ACL, 2009

Publications chair, IEEE/ACL Spoken Language Technology Workshop, 2006

Publications co-chair, HLT-NAACL, 2006

Internal sponsor of Google Faculty awards: Benjamin Snyder (2015); Adam Lopez (2015); Trevor Cohn (2016); Jason Eisner (2017); Tal Linzen (2019).

CLSP Summer Workshop teams, Johns Hopkins University

- Team leader, "Confusion-based Statistical Language Modeling for Machine Translation and Speech Recognition", 2011
- Member, "Parsing and Spoken Structural Event Detection.", Team leader: Mary Harper, 2005.

Panelist

HLT-NAACL Student Research Workshop, 2009

HLT-NAACL Doctoral Consortium, 2007

ACL Student Research Workshop, 2007

Academic Society Memberships

Association for Computational Linguistics (ACL).

Senior Member, IEEE Signal Processing Society (SPS).

International Speech Communication Association (ISCA).

Regular Reviewer

years:	200#	201#	202#
Journals	0123456789	0123456789	0123456789
Augmentative and Alternative Communication			xx
Computational Linguistics	xxxxxxxxx	xxx xx	
Computer Speech and Language	x x x	xx	
IEEE Transactions on Audio, Speech and Language Processing	x xx	xx xx x	
Natural Language Engineering	x	xx	
Speech Communication	xx x	x x	
Conferences			
Association for Computational Linguistics Rolling Review (ARR)			xx
Annual Mtg of the Assoc. for Computational Linguistics (ACL)	xxx xx xx	x xx xxx	
Annual Mtg of the European Chapter of the ACL (EACL)	x x	x x	
Annual Mtg of the N.American Chapter of the ACL (NAACL)	x x	x x xx x	
Annual Mtg of the Asia-Pacific Chapter of the ACL (AACL)			x
Conf. on Empirical Methods in Natural Language Proc. (EMNLP)	x xx xx	x x x x	
Conf. on Computational Natural Language Learning (CoNLL)	x	x xxxx	
Human Language Technology Conference (HLT)	x x x	x	
IEEE Int'l Conf. on Acoustics, Speech and Signal Proc. (ICASSP)	xxxx		
Int'l Conf. on Computational Linguistics (COLING)	x x x	x x x	
Int'l Joint Conf. on Natural Language Processing (IJCNLP)	x	x	x
Neural Information Processing Systems (NIPS)	x xx	xxx	
Workshops			
African Natural Language Processing (AfricaNLP)			xxx
Biomedical Natural Language Processing (BioNLP)		x x x	
Cognitive Modeling and Computational Linguistics at ACL		x x xx	
Computational Linguistics and Clinical Psychology at ACL		xxxxx	x x
IEEE/ACL, Spoken Language Technology (SLT)	x x		
IEEE, Automatic Speech Recognition & Understanding (ASRU)	x x x x		
SIGMORPHON wkshp on Comp.Morphology/Phonology/Phonetics			xx
Speech and Language Proc. for Assistive Technologies (SLPAT)		xxx	x
Student Research Workshop at *ACL conferences	x x	x	
Widening NLP (WiNLP) at *ACL conferences		x	xxxx
Granting agencies			
National Science Foundation	xx xxx	xxx xx	
Natural Sciences and Engineering Research Council of Canada		x x	

Single time Reviewer

Journals	2002	Int'l Journal of Computational Linguistics and Chinese Language Processing
	2003	Machine Learning
	2005	Journal of Machine Learning Research
	2006	IEEE Transactions on Pattern Analysis and Machine Intelligence
		Journal of Artificial Intelligence Research
	2008	Journal of Biomedical Informatics
	2011	Cognition
	2012	Language and Speech
		Topics in Cognitive Science
	2013	ACM Transactions on Asian Language Information Processing
	2016	Journal of Neural Engineering
Conferences	2010	International Conference on Language Resources and Evaluation (LREC)
	2012	Conference of the Int'l Speech Communication Association (Interspeech)
	2017	ACM CHI Conference on Human Factors in Computing Systems (CHI)
	2018	International Conference on Learning Representations (ICLR)
Workshops	2004	Spoken Language Understanding for Conversational Systems at HLT-NAACL
		Incremental Parsing Workshop at ACL
	2006	Computationally Hard Problems and Joint Inference for NLP at HLT-NAACL
	2010	Computational Neurolinguistics at HLT-NAACL
	2012	On the Future of Language Modeling for HLT at NAACL HLT
		Speech and Multimodal Interaction in Assistive Environments at ACL
		Designing and Evaluating Text Entry Methods at CHI
	2015	Workshop on Software at ACL
	2016	Statistical NLP and Weighted Automata at ACL
Granting agencies	2008	Alzheimer's Association
	2015	UK Engineering and Physical Sciences Research Council (EPSRC)