Yonatan Belinkov, Ph.D.

John A. Paulson School of Engineering and Applied Sciences (SEAS), Harvard University

Office: 33 Oxford Street, MD-208, Cambridge, MA 02138
Phone: +1-857-928-5869
Email: belinkov@seas.harvard.edu
Website: http://people.csail.mit.edu/belinkov

Education

2018 **Ph.D. in Electrical Engineering and Computer Science**, MIT, Cambridge, MA
Thesis: On Internal Language Representations in Deep Learning: An Analysis of Machine Translation and Speech Recognition
Advisor: James Glass, Senior Research Scientist, Computer Science and Artificial Intelligence Laboratory (CSAIL), and Faculty Member, Harvard-MIT Health Sciences & Technology

2014 **M.A. in Arabic and Islamic Studies** (*summa cum laude*), Tel Aviv University, Israel

2009 **B.Sc. in Mathematics** (*magna cum laude*) and **Arabic and Islamic Studies** (*summa cum laude*), Tel Aviv University, Israel

Professional Appointments

2020– **Senior Lecturer**, Computer Science Department, Technion, Haifa, Israel

2018–20 **Postdoctoral Fellow in Computer Science**, SEAS, Harvard University, Cambridge, MA
Faculty Host: Stuart Shieber, Professor of Computer Science

2018–20 **Postdoctoral Associate in Computer Science**, CSAIL, MIT, Cambridge, MA
Faculty Host: James Glass, Senior Research Scientist, CSAIL, and Faculty Member, Harvard-MIT Health Sciences & Technology

Publications

Journal Articles

[1] **Belinkov, Y.***, N. Durrani*, F. Dalvi, H. Sajjad, and J. Glass. On the Linguistic Representational Power of Neural Machine Translation Models. *Computational Linguistics*. Accepted.

[2] **Belinkov, Y.***, A. Magidow*, A. Barrón-Cedeño, A. Shmidman, and M. Romanov. Studying the History of the Arabic Language: Language Technology and a Large-Scale . *Language Resources and Evaluation*. 2019.

[3] **Belinkov, Y.** and J. Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics (TACL)*. 2019.

[4] Adi, Y., E. Kermany, **Y. Belinkov**, O. Lavi, and Y. Goldberg. Analysis of sentence embedding models using prediction tasks in natural language processing. *IBM Journal of Research and Development*. 2017.

[5] Romeo, S., G. Da San Martino, **Y. Belinkov**, A. Barrón-Cedeño, M. Eldesouki, K. Darwish, H. Mubarak, J. Glass, and A. Moschitti. Language processing and learning models for community question answering in Arabic. *Information Processing & Management (IPM)*. 2017.

[6] **Belinkov, Y.**, T. Lei, R. Barzilay, and A. Globerson. Exploring Compositional Architectures and Word Vector Representations for Prepositional Phrase Attachment. *Transactions of the Association for Computational Linguistics (TACL)*. 2014.

[7] Arts, T., **Y. Belinkov**, N. Habash, A. Kilgarriff, and V. Suchomel. arTenTen: Arabic Corpus and Word Sketches. *Journal of King Saud University – Computer and Information Sciences*. 2014.

Refereed Conference Papers

[8] **Belinkov, Y.**, A. Ali, and J. Glass. Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition. In: *Proceedings of Interspeech*, 2019.

[9] Hahn, M., F. Keller, Y. Bisk, and **Y. Belinkov**. Character-based Surprisal as a Model of Human Reading in the Presence of Errors. In: *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci, Oral presentation)*, 2019.

[10] **Belinkov, Y.***, A. Poliak*, S. M. Shieber, B. Van Durme, and A. M. Rush. Don?t Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[11] Luo, H., L. Jiang, **Y. Belinkov**, and J. Glass. Improving Neural Language Models by Segmenting, Attending, and Predicting the Future. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[12] Durrani, N., F. Dalvi, H. Sajjad, **Y. Belinkov**, and P. Nakov. One Size Does Not Fit All: Comparing NMT Representations of Different Granularities. In: *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[13] **Belinkov, Y.***, A. Poliak*, S. M. Shieber, B. Van Durme, and A. M. Rush. On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference. In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM, Oral presentation)*, 2019.

[14] Liu, N., M. Gardner, **Y. Belinkov**, M. Peters, and N. Smith. Linguistic Knowledge and Transferability of Contextual Representations. In: *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[15] Bau, A.*, **Y. Belinkov***, S. Sajjad, N. Durrani, F. Dalvi, and J. Glass. Identifying and Controlling Important Neurons in Neural Machine Translation. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[16] Dalvi, F., N. Durrani, S. Sajjad, **Y. Belinkov**, ,A. Bau, and J. Glass. What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[17] Dalvi, F., A. Nortonsmith, D. A. Bau, **Y. Belinkov**, H. Sajjad, N. Durrani, and J. Glass. NeuroX: A Toolkit for Analyzing Individual Neurons in Neural Networks. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI): Demonstrations Track*, 2019.

[18] Suzgun, M., **Y. Belinkov**, and S. M. Shieber. On Evaluating the Generalization of LSTM Models in Formal Languages. In: *Proceedings of the Society for Computation in Linguistics (SCiL)*, 2019.

[19] Poliak, A., **Y. Belinkov**, B. Van Durme, and J. Glass. On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference. In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018.

[20] **Belinkov, Y.*** and Y. Bisk*. Synthetic and Natural Noise Both Break Neural Machine Translation. In: *Proceedings of the International Conference on Learning Representations (ICLR, Oral presentation)*, 2018.

[21] **Belinkov, Y.** and J. Glass. Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[22] **Belinkov, Y.**, L. Màrquez, H. Sajjad, N. Durrani, F. Dalvi, and J. Glass. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. In: *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, 2017.

[23] Dalvi, F., N. Durrani, H. Sajjad, **Y. Belinkov**, and S. Vogel. Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder. In: *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, 2017.

[24] Khurana, S., M. Najafian, A. Ali, T. Al Hanai, **Y. Belinkov**, and J. Glass. QMDIS: QCRI-MIT Advanced Dialect Identification System. In: *Proceedings of Interspeech*, 2017.

[25] Sajjad, H., F. Dalvi, , N. Durrani, A. Abdelali, **Y. Belinkov**, and S. Vogel. Challenging Language-Dependent Segmentation for Arabic: An Application to Machine Translation and Part-of-Speech Tagging. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[26] **Belinkov, Y.**, N. Durrani, F. Dalvi, H. Sajjad, and J. Glass. What do Neural Machine Translation Models Learn about Morphology?. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[27] Adi, Y., E. Kermany, **Y. Belinkov**, O. Lavi, and Y. Goldberg. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[28] Romeo, S., G. Da San Martino, A. Barrón-Cedeño, A. Moschitti, **Y. Belinkov**,, W. Zhu, Y. Zhang, M. Mohtarami, and J. Glass. Neural Attention for Learning to Rank Questions in Community Question Answering. In: *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (Coling)*, 2016.

[29] **Belinkov, Y.** and J. Glass. Arabic Diacritization with Recurrent Neural Networks. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[30] Sajjad, H., K. Darwish, and **Y. Belinkov**. Translating Dialectal Arabic to English. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013.

Refereed Workshop Papers

[31] Suzgun, M., S. Gehrmann, **Y. Belinkov**, and S. M. Shieber. ALSTM Networks Can Perform Dynamic Counting. In: *Proceedings of the First Workshop on Deep Learning and Formal Languages: Building Bridges*, 2019.

[32] Vig, J. and **Y. Belinkov**. Analyzing the Structure of Attention in a Transformer Language Model. In: *Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP at ACL)*, 2019.

[33] Grand, G. and **Y. Belinkov**. Adversarial Regularization for Visual Question Answering: Strengths, Shortcomings, and Side Effects. In: *Proceedings of the 2nd Workshop on Shortcomings in Vision and Language (SiVL at NAACL-HLT, **Best paper award**)*, 2019.

[34] Sajjad, H., N. Durrani, F. Dalvi, **Y. Belinkov**, and S. Vogel. Neural Machine Translation Training in a Multi-Domain Scenario. In: *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2017.

[35] **Belinkov, Y.** and J. Glass. A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects. In: *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial at Coling)*, 2016.

[36] **Belinkov, Y.**, A. Magidow, M. Romanov, A. Shmidman, and M. Koppel. Shamela: A Large-Scale Historical Arabic Corpus. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH at Coling)*, 2016.

[37] **Belinkov, Y.** and J. Glass. Large-Scale Machine Translation between Arabic and Hebrew: Available Corpora and Initial Results. In: *Proceedings of the Workshop on Semitic Machine Translation (SeMaT at AMTA)*, 2016.

[38] Aharoni, R., Y. Goldberg, and **Y. Belinkov**. Improving Sequence to Sequence Learning for Morphological Inflection Generation: The BIU-MIT Systems for the SIGMORPHON 2016 Shared Task for Morphological Reinflection. In: *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology (SIGMORPHON at ACL)*, 2016.

[39] Mohtarami, M., **Y. Belinkov**, H. Wei-Ning, Y. Zhang, T. Lei, K. Bar, S. Cyphers, and J. Glass. SLS at SemEval-2016 Task 3: Neural-based Approaches for Ranking in Community Question Answering. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval),* 2016.

[40] **Belinkov, Y.**, A. Barrón-Cedeño, and H. Mubarak. Answer Selection in Arabic Community Question Answering: A Feature-Rich Approach. In: *Proceedings of the Second Workshop on Arabic Natural Language Processing (ANLP)*, 2015.

[41] **Belinkov, Y.**, M. Mohtarami, S. Cyphers, and J. Glass. VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval),* 2015.

Selected Talks

2019 Deep Learning Models for Language: What they learn, where they fail, and how to make them more robust – Hebrew University, Technion, Weizmann Institute, Carnegie Mellon University, University of Pennsilvenia

2018 Internal Representations in Neural Machine Translation – Amazon MT team, Pittsburgh

2018 Internal Representations in Deep Learning for Language and Speech Processing – Johns Hopkins University, University of Washington, Allen Institute for Artificial Intelligence, Toyota Technological Institute at Chicago, Radcliffe Institute for Advanced Study

2017 Understanding Internal Representations in Deep Learning Models for Language and Speech Processing – Machine Learning for Language, NYU, New York

2017 On Learning Form and Meaning in Neural Machine Translation Models – Computational Data Science Seminar, Technion; CompLang Discussion Group, MIT

2017 What do Neural Machine Translation Models Learn about Morphology? – Data Science Summit Europe, Jerusalem

2017 Language Technologies for Arabic: Historical Documents, Web Forums, and Machine Translation – Qatar Computing Research Institute, Doha

2016 A Computational Analysis of Judeo-Arabic Translations of the Passover Hagaddah – International Jewish Languages Conference, Hebrew University of Jerusalem, Jerusalem

2015 Deep Learning for Sentence Representation – IBM Research, Tel Aviv

2015 Exploring Compositional Architectures and Word Vector Representations for Prepositional Phrase Attachment – Tel Aviv University, Tel Aviv

Teaching & Advising Experience

Co-Instructor, Harvard University, Cambridge, MA (2019)

- Curricular Design for Computer Science: Computational Linguistics and Natural-language Processing

Lecturer, Tel Aviv University, Israel

- Fundamentals of Grammar (2009–2011)
- Arabic II (2009-2011)
- Grammar I (2010)

Teaching Assistant, MIT, Cambridge, MA (2015)

- Introduction to Computer Science and Programming in Python
- Introduction to Computational Thinking and Data Science

Guest Lecturer

- Automatic Speech Recognition, MIT, Cambridge, MA (2019)
- Machine Translation and Sequence-to-sequence Models, CMU, Pittsburgh, PA (2018)
- NLP and the Humanities, Hebrew University, Jerusalem, Israel (2015)

Pedagogical Training, MIT, Cambridge, MA (2015)
Kaufman Teaching Certificate Program, Teaching and Learning Laboratory

Advising Experience (2017–2019)

- Mentor for seven students in the Undergraduate Research Opportunities Program, MIT
- Mentor for two undergraduate students at Harvard SEAS

Fellowships & Awards

Fellowships

2018 Mind, Brain, and Behavior Postdoctoral Fellowship, Harvard University

2018 Moore-Sloan Data Science Fellow, NYU (*declined*)

Research Grants

2018–22 International Collaborator on Israel Science Foundation Grant no. 1191/18. Linguistic Analysis of Algerian Judeo-Arabic Corpora Assisted by Machine Learning. PI: Ofra Tirosh-Becker, Hebrew University. Grant amount: 520,000 NIS (approx. $143,000)

2019 Harvard Mind, Brain, Behavior Fellow Award. Language Representations in Humans and Machines ($5000)

2011 Gardner Fund Research Grant, Princeton University Arabic Dialectology Training Program ($2500)

Academic Recognition

2013 Elie Shaio Memorial Award, MIT
2012 Konard Adenauer Master's Thesis Scholarship, Tel Aviv University
2007–09 Honors list of the Dean of Exact Sciences, Tel Aviv University
2009 Excellence Scholarship, The Wolf Foundation
2009 Excellence Award, School of Mathematical Sciences, Tel Aviv University
2008 Honors list of the Dean of Humanities, Tel Aviv University

Travel Awards

2019 ICLR Travel Award, New Orleans, LA
2017 NeurIPS Travel Award, Long Beach, CA
2016 Coling Student Support Program, Osaka, Japan

Professional Service

Area Chair
Interpretability and Analysis of Models for NLP track at *ACL* (2020)

Reviewer
CL Journal, *ACL*, *EMNLP* (Best reviewer, 2018), *NAACL*, *NeurIPS*, *ICLR* (Outstanding reviewer, 2019), *Coling* (Outstanding reviewer, 2018), *CoNLL*, *IEEE TASL*, *CSL*, *SemEval*

Workshop Organizer
BlackboxNLP at *ACL* (2019), *Robustness Task* at *WMT* (2019)

Committee Work
Graduate Admissions Committee Member, EECS, MIT (2015–18)