

# YONATAN BELINKOV, PH.D.

The Henry and Marilyn Taub Faculty of Computer Science  
Technion – Israel Institute of Technology

## 1. PERSONAL DETAILS

Office: CS Taub Building, Room 733, Technion, Haifa 3200003, Israel  
Phone: +972-52-8230230  
Email: [belinkov@technion.ac.il](mailto:belinkov@technion.ac.il)  
Website: <http://www.belinkov.com>  
ORCID iD: 0000-0002-6280-5045

## 2. ACADEMIC DEGREES

- 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

## 3. ACADEMIC APPOINTMENTS

- 2020– **Senior Lecturer** (Assistant Professor), Faculty of Computer Science, 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

## 4. PROFESSIONAL EXPERIENCE

- 2010–12 Software Engineer, Intuview Ltd., Israel

## 5. RESEARCH INTERESTS

Natural Language Processing, Machine Learning, Deep Learning, Interpretability, Robustness, Emergent Communication.

## 6. TEACHING EXPERIENCE

**Lecturer**, Technion, Haifa, Israel

- CS236004: Introduction to Transformers (Fall 2023)
- CS236299: Introduction to Natural Language Processing (Fall 2020, Spring 2022, Spring 2023)
- CS236756: Introduction to Machine Learning (Fall 2021, Fall 2022, Fall 2023)
- CS236817: Seminar in Natural Language Processing (Spring 2021, Fall 2021, Fall 2022)

**Co-Instructor**, MIT, Cambridge, MA (2020)

- Structure and Interpretation of Deep Networks

**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**

- Natural Language Processing, Princeton, Princeton, NJ (2021)
- Advanced Natural Language Processing, MIT, Cambridge, MA (2020)
- Language, Structure, and Cognition, Harvard, Cambridge, MA (2019)
- 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

## 7. TECHNION ACTIVITIES

## 8. DEPARTMENTAL ACTIVITIES

2023– Faculty Website, Computer Science, Technion

2022–23 Counselor for students with an unsatisfactory academic status, Computer Science, Technion

2021– Faculty Search Committee, Computer Science, Technion

2015–18 Graduate Admissions Committee Member, EECS, MIT

## 9. PUBLIC PROFESSIONAL ACTIVITIES

### Workshop Organizer

*BlackboxNLP* (at *ACL* 2019, *EMNLP* 2020, *ACL* 2021, *EMNLP* 2022, and *EMNLP* 2023), *Robustness Task* (at *WMT* 2019 and *WMT* 2020), *RobustML* (at *ICLR* 2021)

### Senior Area Chair

*Interpretability and Analysis of Models for NLP track* at *EMNLP* (2023), *Interpretability and Analysis of Models for NLP track* at *NAACL* (2021), *Interpretability and Analysis of Models for NLP track* at *ACL* (2022), *Ethical and Sustainable NLP track* at *EACL* (2023)

### Area Chair

*Interpretability and Analysis of Models for NLP track* at *ACL* (2020, 2021), *Interpretability and Analysis of Models for NLP track* at *EMNLP* (2020, 2021), *CoNLL* (2020), *NeurIPS* (2021, 2022)

### Reviewer

- **Journals:** *Computational Linguistics* (2021, 2022), *TACL* (2020–2022), *IEEE TASL* (2014, 2016, 2018), *Computer Speech and Language* (2017), *ACM Surveys* (2022)
- **Conferences:** *ACL Rolling Review* (2021), *ACL* (2018, 2019, 2023), *EMNLP* (2015, 2017, 2018 [best reviewer], 2019, 2022), *NAACL* (2018, 2019), *NeurIPS* (2019, 2020), *ICLR* (2019 [outstanding reviewer], 2020, 2021 [outstanding reviewer], 2022), *EACL* (2021), *Coling* (2018 [outstanding reviewer]), *CoNLL* (2016–2018, 2021), *IJCAI* (2019)
- **Workshops:** Various NLP workshops
- **Grant proposals:** Israeli Science Foundation (2021), Hasler Foundation (2021), Swiss National Science Foundation (2022), Czech Science Foundation (2022), Israeli Ministry of Science and Technology (2022), New Zealand Marsden Fund (2023)

### Tutorial Instructor

Tutorial on *Interpretability and Analysis in Neural NLP* at *ACL* (2020) (video)

## 10. MEMBERSHIP IN PROFESSIONAL SOCIETIES

- The Association for Computational Linguistics

## 11. FELLOWSHIPS, AWARDS, AND HONORS

### Fellowships

2020–23 Azrieli Early Career Faculty Fellowship  
2020–23 Viterbi Fellowship, Center for Computer Engineering, Technion  
2018–20 Mind, Brain, and Behavior Postdoctoral Fellowship, Harvard University  
2018 Moore-Sloan Data Science Fellow, NYU (*declined*)

## Awards and Honors

2023	Excellence in Teaching, Introduction to Machine Learning, Technion
2021	AAAI New Faculty Highlights Program, AAAI
2019	ICLR Travel Award, New Orleans, LA
2017	NeurIPS Travel Award, Long Beach, CA
2016	Coling Student Support Program, Osaka, Japan
2013	Elie Shaio Memorial Award, MIT
2012	Konard Adenauer Master’s Thesis Scholarship, Tel Aviv University
2010–11	Puzis Academic Achievements Award, Faculty of Humanities, Tel Aviv University
2010	The Rina Drori Excellence Scholarship, Faculty of Humanities, Tel Aviv University
2010	Excellence scholarship, Department of Arabic and Islamic Studies, Tel Aviv University
2009	Excellence Scholarship, The Wolf Foundation
2009	Excellence Award, School of Mathematical Sciences, Tel Aviv University

## 12. GRADUATE STUDENTS

### Completed PhD theses

#### Completed MSc theses

- Reda Igbaria, MSc student, Technion, *Debiasing Natural Language Understanding Models Through Biased Internal Components* (2021–2023)
- Omer Antverg, MSc student, Technion, *Analyzing Individual Neurons in Contextual Word Representations from Neural Language Models* (2021–2022)
- Michael Mendelson, MSc student, Technion, *How Debiasing Affects Internal Representations in Natural Language Understanding Models* (2020–2021)
- Yana Dranker, MSc student, Technion, *Invariant Risk Minimization for Natural Language Inference* (2020–2022)
- Dimion Asael, MSc student, Technion, *A Generative Approach for Mitigating Structural Biases in Natural Language Inference* (2020–2021)
- Michal Kessler, MSc student, Hebrew University, *Machine Learning for Judeo-Arabic* (2019–2021) (Co-advisor with Omri Abend)
- Rami Manna, MEng student, MIT, *Low Resource Speech-to-text Translation from Arabic to English* (2019–2021) (Co-advisor with James Glass)

#### PhD theses in progress

- Dana Arad, PhD student, Technion, *Editing Text-to-Image Models* (2023–)
- Adi Simhi, PhD student, Technion, *Promoting Trust in AI Outputs* (2023–)
- Adir Rahamim, PhD student, Technion, *Interpreting Language Models Across Time* (2023–)
- Edo Dotan, PhD student, Technion, *Deep Learning for Biology* (2023–) (Co-advisor with Tal Pupko)
- Itay Itzhak, PhD student, Technion, *Cognitive Biases of Language Models* (2023–) (Co-advisor with Gabi Stanovsky)
- Boaz Carmeli, PhD student, Technion, *Learning to Communicate* (2022–) (Co-advisor with Ron Meir)
- Hadas Orgad, PhD student, Technion, *Explaining, Improving and Evaluating Robustness in NLP Models* (2022–)
- Michael Hanna, PhD student, University of Amsterdam (2022–) (Co-advisor with Sandro Pezzelle)

**MSc theses in progress**

- Ido Levy, MSc student, Technion, *King Solomon - Understanding AI Agents' Emergent Communication* (2023–)
- Rotem Ben-Zion, MSc student, Technion, *Interpreting Emergent Communication by Translation to Natural Language*, (2023–)
- Yanay Soker, MSc student, Technion, *Early Identification of Difficulties in Model Editing* (2023–)
- Tal Haklay, MSc student, Technion, *Direct and Optimization-free Model Editing* (2022–)
- Shadi Iskander, MSc student, Technion, *Bias Mitigation with Bias-Adversarial Loss* (2022–) (Co-advisor with Kira Radinsky)
- Shachar Katz, MSc student, Technion *Noise Injection for Decoding Language Models* (2022–)
- Michael Toker, MSc student, Technion, *Metaphor Detection in Hebrew* (2022–) (Co-advisor with Benny Kimelfeld)
- Zachary Bamberger, MSc student, Technion, *Persuasive Language and Discourse* (2022–)
- Tomer Ashuach, MSc student, Technion, *Anonymizing LLMs: PII Deletion with Model Integrity* (2023–)

**PhD thesis reader / committee member**

- Adi Haviv, PhD student, Tel Aviv University (2024) (PhD committee member)
- Eilam Shapira, PhD student, Technion (2024) (PhD committee member)
- Aaron Mueller, PhD student, Johns Hopkins University (2023) (PhD committee member)
- Shalev Shaer, PhD student, Technion (2022) (PhD committee member)
- Yoav Levine, PhD student, Hebrew University, *Theoretical Insights on the Application of Deep Neural Networks in the Fields of Many-Body Quantum Physics and Natural Language Processing* (2022) (PhD thesis reader)
- Ido Galil, PhD student, Technion (2022) (PhD committee member)
- Damián Pascual Ortiz, PhD student, ETH Zurich, *Leveraging and Understanding Deep Learning Models from Brain Activity to Language Processing* (2022) (PhD thesis reader)
- James M. Fiocco, PhD student, Carnegie Mellon University, *Functional Components as a Paradigm for Neural Model Design and Explainability* (2022) (PhD committee member)
- Naomi Saphra, PhD student, University of Edinburgh, *Training Dynamics of Neural Language Models* (2021) (PhD thesis reader)

**Master's thesis reader**

- Safaa Shehadi, Master's student, Haifa University, *Triggers of Code-switching in Arabizi* (2024)
- Yair Gat, Master's student, Technion, *A Causal Framework for Model Explanations in NLP* (2023)
- Daniel Gilo, Master's student, Technion, *A General Search-based Framework for Generating Textual Counterfactual Explanations* (2023)
- Dave Makhervaks, Master's student, Technion, *Clinical Contradiction Detection* (2023)
- Yifan Jiang, Master's student, University of Washington, *The Weighted Möbius Score: A Unified Framework for Feature Attribution* (2023)
- Shaked Meirom, Master's student, Technion, *Geometric and Topological approaches for Natural Language Processing* (2023)
- Adi Simhi, Master's student, Technion, *Interpreting Embedding Spaces by Conceptualization* (2022)
- Ben Finkelshtein, Master's student, Technion, *Robustness and Rotation Equivariance in Geometric Deep Learning* (2022)

- Mohammed Dabbah, Master’s student, Technion, *Using Fictitious Class Representations to Boost Discriminative Zero-Shot Learners* (2022)
- Itay Itzhak, Master’s student, Tel Aviv University, *Models In a Spelling Bee: Language Models Implicitly Learn the Character Composition of Tokens* (2021)
- Daniel Rosenberg, Master’s student, Technion, *On the Robustness of Visual Question Answering Systems* (2021)
- Gal Sadeh-Kenigsfield, Master’s student, Technion, *Leveraging Auxiliary Text for Deep Recognition of Unseen Visual Relationships* (2021)
- Tomer Wullach, Master’s student, Haifa University, *Generalized Hate Speech Detection on Social Media* (2021)
- Ram Yazdi, Master’s student, Technion, *Perturbation Based Learning for Structured NLP Tasks with Application to Dependency Parsing* (2021)
- Shunit Haviv Hakimi, Master’s student, Technion, *Deep Neural Models for Jazz Improvisations* (2021)
- Elia Turner, Master’s student, Technion, *Charting and Navigating the Space of Solutions for Recurrent Neural Networks* (2021)
- Tom Beer, Master’s student, Technion, *Causal Inference with Mismeasured and Spurious Covariates* (2020)
- Muhammad Majadly, Master’s student, Haifa University, *Dynamic Ensembles in Named Entity Recognition for Historical Arabic Texts* (2020)

#### **Bachelor’s thesis reader**

- Mirac Suzgun, BA student, Harvard University, *Formal Language Theory as a Framework for Understanding the Limitations of Recurrent Neural Networks* (2020), Winner of the Hupes Prize
- Christine Jou, BA student, Harvard University, *Connecting Language Representations in Humans and Machines* (2020)
- Abdul Saleh, BA student, Harvard University, *Towards Social and Interpretable Neural Dialog Systems* (2020)

#### **Other advising experience**

- Mentor for seven undergraduate students at MIT (2017–2019)
- Mentor for six undergraduate students at Harvard SEAS (2018–2020)

### **13. SPONSORED LONG-TERM VISITORS AND POST-DOCTORAL ASSOCIATES**

#### **Post-Doctoral Associates**

- Aaron Mueller, Zuckerman Postdoctoral Scholar (2023–2025) (joint supervision with David Bau)
- Martin Tutek (2024–2025)

## 14. GRANTS

### Competitive

- 2023–27 U.S-Israel Binational Science Foundation Grant no. 2022330. *Emergent Communication in Artificial Agents: Knowledge Disparities and Language Decipherment*. Co-PIs: Shafi Goldwasser and Ron Meir. Grant amount: \$165,200.
- 2021–24 Ministry of Science and Technology Research Grant no. 0002215. *Automatic Detection of Figurative Language in Hebrew across the Eras*. Co-PIs: Benny Kimelfeld and Ophir Münz-Manor. Grant amount: 599,990 NIS (approx. \$189,000).
- 2020–24 Israel Science Foundation Personal Research Grant no. 448/20. *Interpretability and Robustness in Neural Natural Language Processing*. Grant amount: 920,000 NIS (approx. \$270,000).
- 2020–23 Israel Science Foundation New Faculty Equipment Grant no. 449/20. *Interpretability and Robustness in Neural Natural Language Processing*. Grant amount: 647,000 NIS (approx. \$200,000).

### Industrial and Other Sources

- 2022–24 Open Philanthropy Grant. *Initiative for the Interpretable Control of AI*. Co-PI: David Bau. Grant amount: \$1M.
- 2023–24 Google Sponsored Research. *Detecting generated hallucinations within Large Language Models (LLM) and intervening by locating them within the LLM architecture*. Grant amount: \$30K.
- 2020–23 Azrieli Faculty Fellowship Research Grant. *Information Storage in Models of Human Language*. Grant amount: \$209,440.
- 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).

## 15. PUBLICATIONS

### Theses

- 2018     PhD Thesis: On Internal Language Representations in Deep Learning: An Analysis of Machine Translation and Speech Recognition  
Electrical Engineering and Computer Science, MIT, Cambridge, MA  
Advisor: James Glass
- 2014     MA Thesis: The Arabic Dialect of Ġisir izZarga: Linguistic description and a preliminary classification, with sample texts  
Arabic and Islamic Studies, Tel Aviv University, Israel  
Advisor: Nasir Basal

### Journal Articles

- [1]     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.
- [2]     **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.
- [3]     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.
- [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]     **Belinkov, Y.** and J. Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics (TACL)*. 2019.
- [6]     **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.
- [7]     **Belinkov, Y.\***, N. Durrani\*, F. Dalvi, H. Sajjad, and J. Glass. On the Linguistic Representational Power of Neural Machine Translation Models. *Computational Linguistics*. 2020.
- [8]     **Belinkov, Y..** Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*. 2022.
- [9]     Tirosh-Becker, O.\*, M. Kessler\*, O. Becker, and **Y. Belinkov**. Part-of-Speech and Morphological Tagging of Algerian Judeo-Arabic. *Northern European Journal of Language Technology*. 2022.



## Refereed Conference Papers

- [10] 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.
- [11] **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.
- [12] 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 (Coling)*, 2016.
- [13] 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.
- [14] **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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] **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.
- [19] **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.
- [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] 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.
- [22] 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.
- [23] 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.

- [24] 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.
- [25] 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.
- [26] 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.
- [27] **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.
- [28] 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.
- [29] 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.
- [30] **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.
- [31] Li, X., P. Michel, A. Anastasopoulos, **Y. Belinkov**, N. Durrani, O. Firat, Ph. Koehn, G. Neubig, J. Pino, and H. Sajjad. Findings of the First Shared Task on Machine Translation Robustness. In: *Proceedings of the Fourth Conference on Machine Translation (WMT)*, 2019.
- [32] 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.
- [33] **Belinkov, Y.**, A. Ali, and J. Glass. Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition. In: *Proceedings of Interspeech*, 2019.
- [34] Rosenfeld, J., A. Rosenfeld, **Y. Belinkov**, and N. Shavit. A Constructive Prediction of the Generalization Error Across Scales. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- [35] Abdou, M., V. Ravishankar, M. Barrett, **Y. Belinkov**, D. Elliott, and A. Søgaard. The Sensitivity of Language Models and Humans to Winograd Schema Perturbations. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- [36] Mahabadi, R. K., **Y. Belinkov**, and J. Henderson. End-to-End Bias Mitigation by Modelling Biases in Corpora. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- [37] Wu, J.M.\*, **Y. Belinkov**\*, S. Sajjad, N. Durrani, F. Dalvi, and J. Glass. Similarity Analysis of Contextual Word Representation Models. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

- [38] Specia, L., Zh. Li, J. Pino, V. Chaudhary, F. Guzmán, G. Neubig, N. Durrani, **Y. Belinkov**, Ph. Koehn, H. Sajjad, P. Michel, And X. Li. Findings of the WMT 2020 Shared Task on Machine Translation Robustness. In: *Proceedings of the Fifth Conference on Machine Translation (WMT)*, 2020.
- [39] Durrani, N., S. Sajjad, Dalvi, F., and **Y. Belinkov**. Analyzing Individual Neurons in Pre-trained Language Models. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [40] Dalvi, F., S. Sajjad, N. Durrani, and **Y. Belinkov**. Analyzing Redundancy in Pretrained Transformer Models. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [41] Vig, J.\*, S. Gehrmann\*, **Y. Belinkov\***, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In: *Advances in Neural Information Processing Systems (NeurIPS, Spotlight presentation)*, 2020.
- [42] Ravichander, A., Y. Belinkov, and E. Hovy. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance?. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.
- [43] Mahabadi, R. K., **Y. Belinkov**, and J. Henderson. Variational Information Bottleneck for Effective Low-Resource Fine-Tuning. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [44] Sanh, V., Th. Wolf, **Y. Belinkov**, and A. M. Rush. Learning from others’ mistakes: Avoiding dataset biases without modeling them. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [45] Chung, Y., **Y. Belinkov**, and J. Glass. Similarity Analysis of Self-Supervised Speech Representations. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- [46] Finlayson, M.\*, A. Mueller\*, S. Gehrmann, S. Shieber, T. Linzen, and **Y. Belinkov**. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.
- [47] Mendelson, M. and **Y. Belinkov**. Debiasing Methods in Natural Language Understanding Make Bias More Accessible. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
- [48] Dranker, Y., H. He, and **Y. Belinkov**. IRM—when it works and when it doesn’t: A test case of natural language inference. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [49] Stacey, J., **Y. Belinkov**, and M. Rei. Supervising Model Attention with Human Explanations for Robust Natural Language Inference. In: *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2022.
- [50] Antverg, Omer and **Y. Belinkov**. On the Pitfalls of Analyzing Individual Neurons in Language Models. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.
- [51] Orgad, H., S. Goldfarb-Tarrant, and **Y. Belinkov**. How Gender Debiasing Affects Internal Model Representations, and Why It Matters. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2022.
- [52] Asael, D., Z. Ziegler, and **Y. Belinkov**. A Generative Approach for Mitigating Structural Biases in Natural Language Inference. In: *Proceedings of the Eleventh Joint Conference on Lexical and Computational Semantics (\*SEM)*, 2022.

- [53] Meng, K.\*, D. Bau\*, A. Andonian and **Y. Belinkov**. Locating and Editing Factual Associations in GPT. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [54] Bansal, R., D. Pruthi, and **Y. Belinkov**. Measures of Information Reflect Memorization Patterns. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [55] Zaman, K. and **Y. Belinkov**. A Multilingual Perspective Towards the Evaluation of Attribution Methods in Natural Language Inference. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
- [56] Carmeli, B., R. Meir, and **Y. Belinkov**. Emergent Quantized Communication. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.
- [57] Dotan, E., **Y. Belinkov**, O. Avram, E. Wygoda, N. Ecker, M. Alburquerque, O. Keren, G. Loewenthal, and T. Pupko. Multiple sequence alignment as a sequence-to-sequence learning problem. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [58] Meng, K., A.S. Sharma, A.J. Andonian, **Y. Belinkov**, and D. Bau. Mass-Editing Memory in a Transformer. In: *Proceedings of the International Conference on Learning Representations (ICLR, notable-top-25%)*, 2023.
- [59] Ratner, N., Y. Levine, **Y. Belinkov**, O. Ori, O. Abend, U. Karpas, A. Shashua, K. Leyton-Brown, and Y. Shoham. Parallel Context Windows Improve In-Context Learning. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.
- [60] Ram, O., L. Bezalel, A. Zicher, **Y. Belinkov**, J. Berant, and A. Globerson. What Are You Token About? Dense Retrieval as Distributions Over the Vocabulary. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.
- [61] Iskander, Sh., K. Radinsky, and **Y. Belinkov**. Shielded Representations: Protecting Sensitive Attributes Through Iterative Gradient-Based Projection. In: *Findings of the Association for Computational Linguistics : ACL 2023 (ACL)*, 2023.
- [62] Orgad, H. and **Y. Belinkov**. BLIND: Bias Removal With No Demographics. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.
- [63] Orgad, H., B. Kawar, and **Y. Belinkov**. Editing Implicit Assumptions in Text-to-Image Diffusion Models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [64] Stolfo, A., **Y. Belinkov**, and M. Sachan. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [65] Hanna, M., **Y. Belinkov**, and S. Pezzelle. When Language Models Fall in Love: Animacy Processing in Transformer Language Models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [66] Katz, Sh. and **Y. Belinkov**. VISIT: Visualizing and Interpreting the Semantic Information Flow of Transformers. In: *Findings of the Association for Computational Linguistics: EMNLP 2023 (EMNLP)*, 2023.
- [67] Gandikota, R., H. Orgad, **Y. Belinkov**, J. Materzynska, and D. Bau. Unified Concept Editing in Diffusion Models. In: *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2024.
- [68] Mor, A., **Y. Belinkov**, and B. Kimelfeld. Accelerating the Global Aggregation of Local Explanations. In: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

- [69] Hernandez\*, E., A. Sen Sharma\*, T. Haklay, K. Meng, M. Wattenberg, J. Andreas, **Y. Belinkov**, and D. Bau. Linearity of Relation Decoding in Transformer Language Models. In: *Proceedings of the International Conference on Learning Representations (ICLR, Spotlight)*, 2024.
- [70] Prakash, N., T. Rott Shaham, T. Haklay, **Y. Belinkov**, and D. Bau. Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- [71] Muhlgay, D., R. Ori, I. Magar, Y. Levine, N. Ratner, **Y. Belinkov**, O. Abend, K. Leyton-Brown, A. Shashua, and Y. Shoham. Generating Benchmarks for Factuality Evaluation of Language Models. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024.
- [72] Token, M., **Y. Belinkov**, O. Mishali, O. Münz-Manor, and B. Kimelfeld. A Dataset for Metaphor Detection in Early Medieval Hebrew Poetry. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024.

### Refereed Workshop Papers

- [73] **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.
- [74] **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.
- [75] 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.
- [76] 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.
- [77] **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.
- [78] **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.
- [79] **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.
- [80] 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.

- [81] 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.
- [82] 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.
- [83] Suzgun, M., S. Gehrmann, **Y. Belinkov**, and S. M. Shieber. LSTM Networks Can Perform Dynamic Counting. In: *Proceedings of the First Workshop on Deep Learning and Formal Languages: Building Bridges*, 2019.
- [84] Saleh, Abdelrhman, T. Deutsch, S. Casper, **Y. Belinkov**, and S. M. Shieber. Probing Neural Dialog Models for Conversational Understanding. In: *Proceedings of the Second Workshop on NLP for Conversational AI (NLP4ConvAI)*, 2020.
- [85] Orgad, Hadas and **Y. Belinkov**. Choose Your Lenses: Flaws in Gender Bias Evaluation. In: *Proceedings of the Fourth Workshop on Gender Bias in NLP (GeBNLP)*, 2022.
- [86] Antverg, Omer, E. Ben-David, and **Y. Belinkov**. IDANI: Inference-time Domain Adaptation via Neuron-level Interventions. In: *Proceedings of the Second Workshop on Deep Learning for Low-Resource NLP (DeepLoNLP)*, 2022.

#### Edited Collections

- [87] Linzen, T., G. Chrupała, **Y. Belinkov**, and D. Hupkes. Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (held in ACL 2019).
- [88] Alishai, A., **Y. Belinkov**, G. Chrupała, D. Hupkes, Y. Pinter, and H. Sajjad. Proceedings of the 2020 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (held in EMNLP 2020).
- [89] Bastings, J., **Y. Belinkov**, E. Dupoux, M. Giulianelli, D. Hupkes, Y. Pinter, and H. Sajjad. Proceedings of the fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (held in EMNLP 2021).
- [90] Bastings, J., **Y. Belinkov**, Y. Elazar, D. Hupkes, N. Saphra, and S. Wiegrefe. Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (held in EMNLP 2022).
- [91] **Y. Belinkov**, S. Hao, J. Jumelet, N. Kim, A. McCarthy, and H. Mohebbi. Proceedings of the Sixth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (held in EMNLP 2023).

#### Non-Refereed Conference Papers

- [92] **Belinkov, Y.** Large-Scale Electronic Corpora and the Study of Middle and Mixed Arabic. In: *Proceedings of the IVth AIMA International Conference (Emory University, Atlanta, GA, USA, 12–15 October 2013)*, 2021.

## 16. CONFERENCES

### Conference Co-Organizer

The Israeli Seminar on Computational Linguistics (*ISCOL* 2021)

## 17. SELECTED TALKS

- 2023      On Localization in Language Models – Google Research India, IBM Research Tel Aviv, Google Research Tel Aviv, TU Darmstadt, Düsseldorf University, Bar Ilan University, Simons Institute, Microsoft Research New England
- 2023      Editing Text-to-Image Generation Models – AI Summit, Israel
- 2022      Out-of-Distribution NLP – Hebrew University, Israeli Statistical Association
- 2021      Interpretability and Robustness in Natural Language Processing – AAAI New Faculty Highlights (video)
- 2020      Studying the History of the Arabic Language: Language Technology and a Large-Scale Historical Corpus – The Open University (video)
- 2020      Interpretability and Other Highlights from NLP – Workshop on Decoding Communication in Nonhuman Species, Simons Institute, UC Berkeley
- 2020–21   Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias – Stanford, UC Berkeley, UMass Amherst, Google, Salesforce, Amazon, NYU, Edingurgh
- 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 Pennsylvania
- 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