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
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, Biological Language Models.

6. Teaching Experience

Lecturer, Technion, Haifa, Israel

- CS236004: Introduction to Transformers (Fall 2023, Spring 2025)
- CS236299: Introduction to Natural Language Processing (Fall 2020, Spring 2022, Spring 2023, Fall 2024)
- CS236756: Introduction to Machine Learning (Fall 2021, Fall 2022, Fall 2023, Fall 2024)
- 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, *EMNLP* 2023, *EMNLP* 2024, and *EMNLP* 2025), *Robustness Task* (at *WMT* 2019 and *WMT* 2020), *RobustML* (at *ICLR* 2021), *NLP and DH: Hebrew and other languages* (2024)

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, 2025), *Interpretability and Analysis of Models for NLP track* at *ACL* (2022, 2025), *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), *ACL ARR* (May 2025), *NeurIPS* (2021, 2022), *COLM* (2024, 2025)

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, 2024), *ICLR* (2019 [outstanding reviewer], 2020, 2021 [outstanding reviewer], 2022, 2025), *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), ERC Starting (2025)

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

ArXiv moderator
Moderator of the cs.CL subject on ArXiv (2020–2025)

10. Membership in Professional Societies

- The Association for Computational Linguistics (since 2015)
- ELLIS Scholar (since 2021)

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

2025	Krill Prize for Excellence in Scientific Research, Wolf Foundation
2024	Best Paper Award, Conference on Empirical Methods in Natural Language Processing
2024	Henry Taub Prize for Academic Excellence, Technion
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, Wolf Foundation
2009	Excellence Award, School of Mathematical Sciences, Tel Aviv University

12. Graduate Students

Completed PhD theses

Completed MSc theses

- Ido Levy, MSc student, Technion, *Unsupervised Translation of Emergent Communication* (2023–2025)
- Yanay Soker, MSc student, Technion, *Predicting Success of Model Editing Through Intrinsic Features* (2023–2024)
- Rotem Ben-Zion, MSc student, Technion, *Semantics and Spatiality of Emergent Communication*, (2023–2024)
- Zachary Bamberger, MSc student, Technion, *DEPTH: Discourse Education through Pre-Training Hierarchically* (2022–2024)
- Shadi Iskander, MSc student, Technion, *Mitigating Social Bias in Language Models: Fairness Strategies in Labeled and Unlabeled Demographics Settings* (2022–2024) (Co-advisor with Kira Radinsky)
- Shachar Katz, MSc student, Technion *Visualizing and Interpreting the Semantic Information Flow of Transformers* (2022–2024)
- 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

- Tal Haklay, PhD student, Technion, *Interpretabilty of LLMs in Real World Settings* (2022–)
- Tomer Ashuach, PhD student, Technion, *Uncovering Mechanisms of Language Models' Capabilities* (2023–)
- Yaniv Nikankin, PhD student, Technion, (2024–)
- Michael Toker, PhD student, Technion, *Interpretable LLMs: Robustness & Alignment* (2022–; Azrieli PhD Fellowship)
- Dana Arad, PhD student, Technion, *Editing Text-to-Image Models* (2023–; Ariane de Rothschild PhD Fellowship)
- Adi Simhi, PhD student, Technion, *Promoting Trust in AI Outputs* (2023–, VATAT PhD Fellowship in AI and Data Science)
- 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–; Apple AI/ML PhD Fellowship)
- Michael Hanna, PhD student, University of Amsterdam (2022–) (Co-advisor with Sandro Pezzelle)

MSc theses in progress

- Gal Kesten, MSc student, Technion, *Interpretability of Protein Language Models* (2025–)

PhD thesis reader / committee member

- Zachary Bamberger, PhD student, Technion, *Generating and Evaluating Persuasive Arguments with Large Language Models* (2025) (PhD committee member)
- Liran Ringel, PhD student, Technion, *Reliable and Efficient Language Models* (2025) (PhD committee member)
- Yara Shamshoum, PhD student, Technion, *Efficiency and Reasoning in AI Models* (2025) (PhD committee member)
- Sahar Admoni, PhD student, Technion, *Leveraging Large Language Models to Interpret Reinforcement Learning Agents* (2025) (PhD committee member)
- Mati Shufan, PhD student, Technion, *Cognitive Traits of the Human Mind as an Intentional System in Interactions with AI Personas* (2025) (PhD committee member)
- Guy Azran, PhD student, Technion, *Decomposing Task and Motion Planning into Learnable Components: Towards Generalizable Robotic Intelligence* (2025) (PhD committee member)
- Zorik Gekhman, PhD student, Technion, *Towards Robust and Trustworthy NLP Models* (2024) (PhD committee member)
- Tomasz Limisiewicz, PhD student, Charles University, *Interpreting and Controlling Linguistic Features in Multilingual Language Models* (2024) (PhD thesis reader)
- Tsachi Blau, PhD student, Technion, *Threat-Model Agnostic Defenses* (2024) (PhD committee member)
- Eyal Cohen, PhD student, Technion, *Smarter Decoders for Automatic Speech Recognition* (2024) (PhD committee member)
- Alon Jacovi, PhD student, Bar Ilan University, *Explaining Artificial Intelligence: Foundations and Practice* (2024) (PhD thesis reader)
- Hanqi Yan, PhD student, University of Warwick, *Enhancing Robustness and Interpretability in Natural Language Processing through Representation Learning* (2024) (PhD thesis reader)
- 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. Fiacco, PhD student, Carnegie Melon 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

- Oz Huly, Master's student, Technion, *Predicting RAG Performance for Text Completion* (2025)
- Inbar Nachmani, Master's student, Technion, *A Proximity Aware Loss Function for Ordinal Classification* (2025)
- Niv Kook, Master's student, Technion, *Benchmarking Confounder Suggestion Capabilities of Large Language Models in Observational Studies* (2025)
- Daniela Gottesman, Master's student, Tel Aviv University, *Estimating Knowledge in Large Language Models Without Generating a Single Token* (2024)
- Niv Kook, Master's student, Technion, *Benchmarking Confounder Suggestion Capabilities of Large Language Models in Observational Studies* (2024)
- Daniel Weisberg Mitelman, Master's student, Reichman University, *Applying Language Models to Phylogenetic Linguistics and Transliteration* (2024)
- 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

- Anja Reausch, Azrieli Postdoctoral Fellow (2024–2026)
- Martin Tutek (2024–2025)
- Aaron Mueller, Zuckerman Postdoctoral Scholar (2023–2025) (joint supervision with David Bau)

14. Grants

Competitive

2024–29 European Research Council Starting Grant no. 101165402. *Controlling Large Language Models (Control-LM).* Grant amount: €1.5M (approx. $1.66M).

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

2025–26 EuroHPC AI and Data-Intensive Applications. *Model Editing for Generative Information Retrieval (EHPC-AI-2024A06-058).* Co-PI: Anja Reusch. Grant amount: 128,000 GPU hours.

2025–28 MAFAT Grant. *An LLM-based approach to predict the function of genes and design tailored pathways and genomic segments.* Co-PIs: Tal Pupko, Dudu Burstein, Eran Bacharach. Grant amount: 1.8M NIS.

2024–26 Open Philanthropy Grant. *Support research on interpretability of large language models.* Grant amount: $184,835.

2025–26 IDSAI inter-disciplinary, inter-center research projects. *Text2Protein: Generating Protein Sequences from Rich Textual Descriptions.* Co-PI: Tal Pupko. Grant amount: 100,000 NIS.

2024–25 Google Gift. *Full Circuit Finding.* Grant amount: $30K.

2024–25 Pfizer Sponsored Research. *Hallucination Detection in Biomedical Literature Summarization.* Grant amount: $155,297.

2024–25 Google Sponsored Research. *Personas and knowledge in Large Language Models.* Grant amount: $30K.

2024–25 NVIDIA. *Academic Gift Award.* Grant amount: 130,000 NIS (approx. $35,000).

2024–25 IBM Project Agreement. *Local Updates to Language Models.* Grant amount: 120,000 NIS (approx. $32,000).

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 Ǧisir 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.

[10] Itzhak, I., G. Stanovsky, N. Rosenfeld, and **Y. Belinkov**. Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias. *Transactions of the Association for Computational Linguistics (TACL)*. 2024.

[11] Dotan, E., G. Jaschek, T. Pupko, and **Y. Belinkov**. Effect of Tokenization on Transformers for Biological Sequences. *Bioinformatics*. 2024.

[12] Dotan, E., E. Wygoda, N. Ecker, M. Alburquerque, O. Avram, **Y. Belinkov**, and T. Pupko. BetaAlign: a deep learning approach for multiple sequence alignment. *Bioinformatics*. 2025.

[13] Hanna, M., **Y. Belinkov**, and S. Pezzelle. Are formal and functional linguistic mechanisms dissociated in language models?. *Computational Linguistics*. 2025.

Refereed Conference Papers

[14] 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.

[15] **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.

[16] 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.

[17] 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.

[18] **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.

[19] 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.

[20] 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.

[21] 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.

[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] **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.

[24] **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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] **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.

[32] 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.

[33] 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.

[34] **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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] Durrani, N., S. Sajjad, Dalvi, F., and **Y. Belinkov**. Analyzing Individual Neurons in Pretrained Language Models. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[44] 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.

[45] 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)*, 2020.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Antverg, O. and **Y. Belinkov**. On the Pitfalls of Analyzing Individual Neurons in Language Models. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[55] 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.

[56] 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.

[57] 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.

[58] Bansal, R., D. Pruthi, and **Y. Belinkov**. Measures of Information Reflect Memorization Patterns. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[59] 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.

[60] Carmeli, B., R. Meir, and **Y. Belinkov**. Emergent Quantized Communication. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] Toker, 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.

[77] Rahamim, A. and **Y. Belinkov**. ContraSim – Analyzing Neural Representations Based on Contrastive Learning. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024.

[78] Arad, D.*, H. Orgad*, and **Y. Belinkov**. ReFACT: Updating Text-to-Image Models by Editing the Text Encoder. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024.

[79] Iskander, Sh., K. Radinsky, and **Y. Belinkov**. Leveraging Prototypical Representations for Mitigating Social Bias without Demographic Information. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024.

[80] Carmeli, B., **Y. Belinkov**, and R. Meir. Concept-Best-Matching: Evaluating Compositionality in Emergent Communication. In: *Findings of the Association for Computational Linguistics (ACL)*, 2024.

[81] Toker, M., H. Orgad, M. Ventura, D. Arad, and **Y. Belinkov**. Diffusion Lens: Interpreting Text Encoders in Text-to-Image Pipelines. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[82] Hanna, M., S. Pezzelle, and **Y. Belinkov**. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanism. In: *Proceedings of the 2024 Conference on Language Models (COLM)*, 2024.

[83] Katz, Sh., **Y. Belinkov**, M. Geva, and L. Wolf. Backward Lens: Projecting Language Model Gradients into the Vocabulary Space. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP, **Best paper award**)*, 2024.

[84] Rahamim, A., N. Saphra, S. Kangaslahti, and **Y. Belinkov**. Fast Forwarding Low-Rank Training. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[85] Ben Zion, R., B. Carmeli, O. Paradise, and **Y. Belinkov**. Semantics and Spatiality of Emergent Communication. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[86] Stolfo, A., B. P. Wu, W. Gurnee, **Y. Belinkov**, X. Song, M. Sachan, and N. Nanda. Confidence Regulation Neurons in Language Models. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[87] Levy, I., O. Paradise, B. Carmeli, R. Meir, Sh. Goldwasser, and **Y. Belinkov**. Unsupervised Translation of Emergent Communication. In: *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[88] Toker, M., I. Galil, H. Orgad, R. Gal, Y. Tewel, G.Chechik, and **Y. Belinkov**. Padding Tone: A Mechanistic Analysis of Padding Tokens in T2I Models. In: *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2025.

[89] Marks, S., C. Rager, E.J. Michaud, **Y. Belinkov**, D. Bau, and A. Mueller. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In: *Proceedings of the International Conference on Learning Representations (ICLR; Oral)*, 2025.

[90] Carmeli, B., R. Meir, and **Y. Belinkov**. CtD: Composition through Decomposition in Emergent Communication. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[91] Lenz, B., O. Lieber, ..., **Y. Belinkov**, ..., and Y. Shoham. Jamba: Hybrid Transformer-Mamba Language Models. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[92] Nikankin, Y., A. Reusch, A. Mueller, and **Y. Belinkov**. Arithmetic Without Algorithms: Language Models Solve Math with a Bag of Heuristics. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[93] Orgad, H., M. Toker, Z. Gekhman, R. Reichart, I. Szpektor, H. Kotek, and **Y. Belinkov**. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[94] Wiegreffe, S., O. Tafjord, **Y. Belinkov**, H. Hajishirzi, and A. Sabharwal . Answer, Assemble, Ace: Understanding How LMs Answer Multiple Choice Questions. In: *Proceedings of the International Conference on Learning Representations (ICLR; Spotlight)*, 2025.

[95] Reusch, A. and **Y. Belinkov**. How Generative IR Retrieves Documents Mechanistically. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2025.

[96] Ashuach, T., M. Tutek, and **Y. Belinkov**. REVS: Unlearning Sensitive Information in Language Models via Rank Editing in the Vocabulary Space. In: *Findings of the Association for Computational Linguistics (ACL)*, 2025.

[97] Haklay, T., H. Orgad, A. Mueller, and **Y. Belinkov**. Position-aware Automatic Circuit Discovery. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[98] Mueller, A.*, A. Geiger*, S. Wiegreffe*, D. Arad, I. Arcuschin, A. Belfki, Y. S. Chan, J. F. Fiotto-Kaufman, T. Haklay, M. Hanna, J. Huang, R. Gupta, Y. Nikankin, H. Orgad, N. Prakash, A. Reusch, A. Sankaranarayanan, S. Shao, A. Stolfo, M. Tutek, A. Zur, D. Bau, and **Y. Belinkov**. MIB: A Mechanistic Interpretability Benchmark. In: *Proceedings of the Forty-Second International Conference on Machine Learning (ICML)*, 2025.

[99] Gekhman, Z., E. Ben David, H. Orgad, E. Ofek, **Y. Belinkov**, I. Szpektor, J. Herzig, and R. Reichart. Inside-Out: Hidden Factual Knowledge in LLMs. In: *Proceedings of the 2025 Conference on Language Models (COLM)*, 2025.

[100] Itzhak, I., **Y. Belinkov**, and G. Stanovsky. Planted in Pretraining, Swayed by Finetuning: A Case Study on the Origins of Cognitive Biases in LLMs. In: *Proceedings of the 2025 Conference on Language Models (COLM)*, 2025.

[101] Tutek, M., F. H. Chaleshtori, A. Marasović, and **Y. Belinkov**. Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[102] Arad, D., A. Mueller, and **Y. Belinkov**. SAEs Are Good for Steering – If You Select the Right Features. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[103] Simhi, A., I. Itzhak, F. Barez, G. Stanovsky, and **Y. Belinkov**. Trust Me, I'm Wrong: LLMs Hallucinate with Certainty Despite Knowing the Answer. In: *Findings of the Association for Computational Linguistics: EMNLP 2025 (EMNLP)*, 2025.

[104] Yu, Z., S. Ananiadou, and **Y. Belinkov**. Back Attention: Understanding and Enhancing Multi-Hop Reasoning in Large Language Models. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Refereed Workshop Papers

[105] **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.

[106] **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.

[107] 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.

[108] 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.

[109] **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.

[110] **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.

[111] **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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] Saleh, A., 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.

[117] Orgad, H. and **Y. Belinkov**. Choose Your Lenses: Flaws in Gender Bias Evaluation. In: *Proceedings of the Fourth Workshop on Gender Bias in NLP (GeBNLP)*, 2022.

[118] Antverg, O., 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.

[119] Igbaria, R. and **Y. Belinkov**. Learning from Others: Similarity-based Regularization for Mitigating Dataset Bias. In: *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP)*, 2024.

[120] Bamberger, Z., O. Glick, Ch. Baskin, and **Y. Belinkov**. DEPTH: Discourse Education through Pre-Training Hierarchically. In: *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP)*, 2025.

Edited Collections

[121] 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).

[122] 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).

[123] 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).

[124] Bastings, J., **Y. Belinkov**, Y. Elazar, D. Hupkes, N. Saphra, and S. Wiegreffe. Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (held in EMNLP 2022).

[125] **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).

[126] **Y. Belinkov**, N. Kim, J. Jumelet,, H. Mohebbi, A. Mueller, and H. Chen. Proceedings of the 7th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (held in EMNLP 2024).

Non-Refereed Conference Papers

[127] **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

2025 Interpreting and Fixing Text-to-Image Models – Bar Ilan University

2025 What do image generators see when they generate images? [In Hebrew] – TECHNOVATION Conference, Technion

2024 Recent Progress in Interpreting and Controlling Language Models – Tel Aviv University

2024 Interpreting Emergent Communication – Simons Institute

2024 Editing Knowledge in Text-to-Image Generation Models – NVIDIA Tel Aviv Lab, Berkeley University

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 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