Research Statement

I design, build, and analyze methods for semantic understanding of text — the process by which we derive *meaning* from a string of characters. This requires interpreting complex, often interacting phenomena, such as:

- **Lexical disambiguation:** "the *pitcher* went up to the mound" necessitates understanding that a pitcher is a baseball player (and not a liquid container).
- **Sentence-level semantics:** "John forgot that he *locked* the door" and "John forgot to *lock* the door" describe different situations, despite reading very similarly.
- **Common-sense reasoning:** "the *doctor* asked the nurse to help her in the operation" implies the doctor in this scenario is likely female.

Thanks to recent advances in artificial neural networks, natural language processing (NLP) reaches close to human performance on some *datasets* (such as textual entailment or sentiment analysis) while their performance often severely deteriorates on real-world examples, showing that the *tasks* underlying these datasets are far from being solved. Training deeper networks on ever-larger datasets does not solve the problem, as these heavily-parameterized models often prove stubbornly capable of overfitting artifacts in the data, rather than modeling meaningful aspects of the task. Instead, I use semantic understanding to approach this challenge from complementary directions:

Finding biases in state-of-the-art models [29, 27, 4]. I use semantic analysis to highlight and quantify bias and brittleness in NLP models. For example, I found that translation models are significantly gender biased (e.g., they assume doctors are men, regardless of context) and that word disambiguation models often opt for the most common sense of a word (e.g., associating *bats* with nocturnal animals rather than baseball instruments). These efforts drive the development of more robust models and highlight key performance bottlenecks (Section 1).

Parsing meaning representations [23, 31, 34, 32, 24, 6, 22, 36, 28, 25, 14, 30, 12]. One of the main reasons for this brittleness in state-of-the-art NLP models is their prevalent reliance on *end-to-end* neural networks, which learn associations between a sequence of word representations and the task label (e.g., sentiment, entailment or translation). This association may not be enough for these models to observe abstract patterns in language. To address this issue, I design models, datasets, and formalisms for broad coverage semantic representations of text. These aim to represent the meaning of sentences in formal discrete data structures, such as trees or graphs, which can serve as attractive, more abstract inputs for NLP models (Section 2).

Improving state of the art in NLP applications using semantic understanding [26, 33, 20, 5, 10]. Finally, I build models which rely on semantic features for *downstream applications* – these produce output with extrinsic value for a wide range of users in real-life applications. For example, I worked on automatically answering user queries over long paragraphs about history and sports, and identifying mentions of adverse drug reactions on social media. The natural, unrestricted nature of the texts in such applications requires effective modeling of the semantic phenomena I detailed above (Section 3).

Taken together, my combined expertise in core semantics, meaning representations, and algorithmic model design put me at a position to make significant progress by building algorithms for classic semantic challenges, introducing novel qualitative performance analyses, and driving future model improvements.

1 Exposing and Evaluating Biases in NLP Models

A central challenge in NLP is the brittleness of neural networks [8, 11]. This occurs when highly-parameterized models pick up on artifacts in their training data (e.g., societal or annotator biases) rather than more meaningful signal, resulting in a significant drop in performance when presented with data which does not exhibit these biases. For example, NLP models were shown to associate cooking or homemaking with women, and construction work or programming with men [3, 37, 18]. I expose these biases through specialized *challenge sets*: samples tailored to exhibit phenomena which may be under-represented in the training data, but crucial for particular end applications (e.g., texts describing male receptionists or female CEOs). Measuring model performance on such challenge sets gauges its ability to capture deeper semantic phenomena.

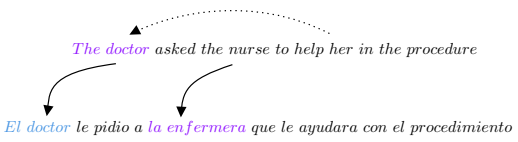

Figure 1: We identified that all popular machine translation models are gender biased. In the English source sentence (top), "her" identifies the doctor as female while the Spanish sentence (bottom) stereotypically changes the doctor gender: "*el* doctor" (male doctor) [29].

We presented the first evaluation of gender bias in machine translation [29]. Our approach built on existing datasets for coreference resolution, in which two human participants were cast into gender stereotypical and anti-stereotypical roles, such as male nurses and female doctors (see Figure 1) [38, 19]. We observed that when translated

to a language with grammatical noun gender (such as Spanish, French, or Hebrew), translation models need to rely on context to assign the right gender inflection (e.g., "her" in Figure 1). Following, we constructed an automatic evaluation method for 8 diverse target languages, and tested 6 state-of-the-art translation models, including the most popular online services, Google Translate and Microsoft Translator. **We found that all translation models were significantly gender biased on all tested languages**, tending to assign gender based on colocations in their training data (e.g., female receptionists and male CEOs), rather than cues from the input sentence. Our method can serve as an automatic measure for progress on this issue in future translation models. **This work was nominated for best paper at ACL 2019 (top 1% of all submissions).**

Another prominent instance of model bias can be found when models prefer the most common sense of a word with multiple meanings, rather than relying on context for disambiguation. To evaluate this kind of bias, we created the Odd-Man-Out challenge set. In a semi-automatic process, we collected roughly 500K Odd-Man-Out puzzles, exemplified in Figure 2, where a target answer word differs from the other four. Each puzzle also contains a *distractor* – a polysemous word (e.g., "crane" in the first example) chosen such that one of its meanings (a type of bird) shares a category with the odd word ("pelican" in this example), while another of its meanings (construction equipment) shares a category with the other words in the puzzle. While humans are able to home in on the correct sense of the distractor, popular word representation models are quite poor at doing so and often get confused by the distractor word. We designed a novel method to obtain *sense* representations by clustering contextual word representations in a large corpus, **leading to** 15% **absolute improvement over the state of the art** providing a method for improving out-of-domain performance for a wide range of other NLP models.

Category	Puzzle
construction	**crane**, *pelican*, excavator, hoist, upraise
guitar part	**fret**, *crying*, inlays, truss rod, neck
alcoholic drinks	**gin**, *poker*, vodka, tequila, wine

Figure 2: Examples from our crowdsourced Odd-Man-Out dataset, which probes the capability of neural models to disambiguate words with multiple meanings. Distractors appear in bold, the correct answer in italics. [27].

2 Broad-Coverage Semantics

One of the main reasons for the biases I found in NLP models is their sole dependence on correlations between input word representations (either pretrained, fine-tuned, or learned from scratch) and the task label, which may not be abstract enough for generalization. I address this issue by designing methods which model the *meaning* conveyed by an input sentence, either with a complete sentence representation or through discrete features for certain semantic aspects. These can then serve as a viable input for more robust NLP models, as I show in Section 3.

Sentence-level meaning representations aim to capture the meaning of a sentence in a formal data structure, such as a tree or a graph. Various meaning representations were suggested and developed in the literature at varying levels of abstraction [1], and were traditionally annotated by experts in an expensive and involved process.

I worked extensively within the novel QA-SRL paradigm [9], which, in contrast, annotates the meaning of a sentence with a list of questions and answers, shown to be intuitive enough for non-expert annotation (see Figure 3 for an example). We extended the paradigm to model relations between any two words in the sentence, and automatically inferred sentence-level meaning representations from question-answer pairs, as exemplified in Figure 4 [14]. **This annotation opens the door for a structured meaning representation created solely by non experts**, which can be used to label texts in low-resource domains and languages. Furthermore, we observed that QA-SRL subsumes the information contained within the popular Open IE representation [2], which represents the meaning of a sentence with a list of stand-alone propositions centered around an event and its participants (e.g., [Biden] *decided to run* [for president]). Following, we automatically converted QA-SRL annotations into the first large-scale annotated corpus for Open IE [24]. **This work has set the standard machine-learning setting for Open IE**, allowing various

In 2019, Biden and Warren announced their candidacy.

Argument	Relation
Biden / Warren	Who announced something?
their candidacy	What did someone announce?
in 2019	When did someone announce something?

Figure 3: Example QA-SRL annotation. The meaning of a sentence is represented using a list of intuitive and interpretable question-answer pairs.

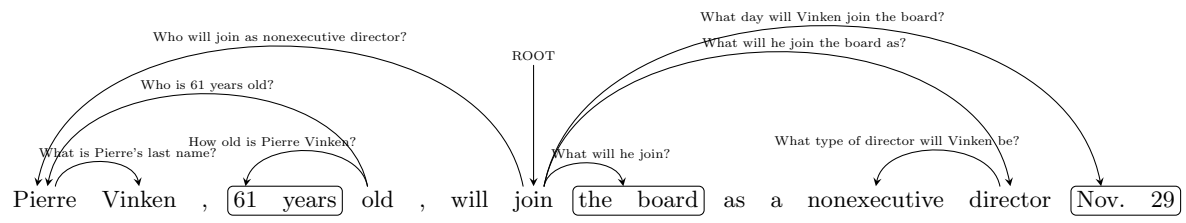

Figure 4: We construct a meaning representation graph from simple question-answer pairs, annotated by non-experts [14].

follow up works to train and evaluate neural models for the task. Using our corpus, we trained a sequence transducer which modeled Open IE as a word-level tagging task and achieved state-of-the-art results [28]. Since its publication, this model was introduced into the popular AllenNLP suite[1] and is being used extensively by developers in industry and academia [21, 7].

We also developed a model which leverages signal from *several* meaning representations by framing the task as a machine translation problem, where different representations are treated as foreign languages [25]. A prerequisite for doing so was the development of a linearization technique which "flattens" arbitrary structured representations, such as the graph shown in Figure 4, into a text sequence which can be directly fed to neural translation models. Our architecture **improves parsing accuracy** by training different representations *in parallel* using a single shared model, and **provides novel insights** by translating *between* meaning representations. For instance, we found equivalences and inherent differences between representations collected in the semantic dependency corpus [16]. These analyses can inform model developers when choosing between different meaning representations.

Modeling semantic features. I develop models that directly predict discrete features which can serve as additional signal for NLP models. For instance, we developed models for *factuality detection* [31], identifying whether an event described in text occurred or not. Given the sentence: "John forgot that he *locked* the door" we understand that John locked the door, while in "John forgot to *lock* the door" we infer that he did not. We normalized various factuality annotations onto a continuous scale, showing that training on the unified dataset achieves state-of-the-art results. **Our corpus have since become the standard training and evaluation setup for factuality**, and our method is the common baseline to compare against in this research area.

In this line of work, we also presented the first model trained on a crowdsourced corpus for the task of detecting *restrictive modifications* [23]. This model discriminates between two types of noun modifiers: *restrictive modifiers*, which constitute an integral part of the entity referenced by a noun phrase, e.g., the italicized modifier in "She wore the necklace *that her mother gave her*" versus *non-restrictive* modifiers, which provide additional information on an already definite entity, e.g., "The speaker thanked president Obama *who just came into the room*". Leveraging the QA-SRL corpus, we reduced the annotation task to a simple binary question, reaching good inter-annotator agreement. Our model relies on features such as enclosing commas, which formally introduce non-restrictive clauses, or the preposition used to introduce the modifier (e.g., "that" is traditionally considered as introducing a restrictive modifier). **These linguistically-motivated features achieve state-of-the-art results for this task**. Our model is used in applications such as text simplification to remove non-restrictive modifiers without changing the meaning of the sentence [13].

3 Semantic Understanding in Downstream Applications

Finally, I design models which integrate semantics and external knowledge in downstream applications – tasks which have extrinsic benefits for a wide range of users in real-life scenarios.

We used Open IE as an intermediate representation for *reading comprehension* [26], a task which involves answering questions over a a paragraph or a document, such as "When did the D-Day invasion start?" given an history document. **We demonstrated that Open IE is useful for several benchmarks as it focuses on the core proposition and participants in the text**. For example, "[The D-Day invasion; *began*; in June, 1944]" is a stand-alone Open IE extraction from a sentence which may contain additional details. By comparing various input representations, we showed that certain design choices in meaning representation (e.g., what constitutes a participant in an event) have profound implications for downstream tasks. In another work, we crowdsourced DROP [5] – a large-scale corpus of 96K questions and answers requiring discrete reasoning, such as mathematical operations over quantities or dates. For example, "When did the Six-Day War end?" given a paragraph describing that the "Six-Day War *started* on June 5th, 1967". We experimented with semantic parsing algorithms over various intermediate representations and found that all of them performed quite poorly, showing that **DROP is a challenging test-bed for future question-answering algorithms over longer, more complicated texts**.

We also studied the problem of identifying *mentions of adverse drug reactions* on Twitter [33]. Effective models for this task can aid pharmaceutical companies in sifting through social media posts to find adverse reactions for new drugs. We observed that the task is highly context dependent, where two opposing reports can have high lexical overlap. For example: "Aspirin gave me a headache" is an adverse reaction, while "Aspirin took away my headache" is a positive one. To cope with the small amounts of training data available for the task, we integrated *knowledge graph embeddings* into our neural network to help generalize across different drugs, their active ingredients, and known side effects. We found that adding this external knowledge significantly improved over the state of the art. Follow up work have adopted our setup and methods as the standard baseline for the task. In general, **we showed that external resources, such as Wikipedia, can readily aid in improving accuracy for tasks with limited data, even though they were not specifically created for this purpose**.

[1] Our online demo receives over a thousand requests a month: https://demo.allennlp.org/open-information-extraction

Future Directions

I plan to push the state of the art to new directions in all of the research areas I described above, expanding meaning representation to more languages, pursuing a novel learning setting which allows models to build predictions interactively, and applying NLP towards socially-conscious use-cases in real-world scenarios. In all of these projects, I intend to collaborate with the wide network of researchers and engineers with whom I worked during my academic career, including NLP practitioners from the top institutions in the world.

Multilingual meaning bank. Existing semantic representations are annotated predominantly in English. I plan to build a *multilingual*, parallel semantic representation bank which will **inspire research into new machine translation algorithms between semantic structures in different languages**, thus generating more semantically-coherent output. Our graph linearization technique can readily model multi-lingual representations [25].

The QA-SRL paradigm is a good candidate representation for multilingual annotation, thanks to its intuitiveness and our full sentence graph derivation (see Figures 3 and 4) [14]. In contrast to other large-scale annotation efforts, such as universal dependencies [15] or PropBank [17], the distance between representation definition and a fully annotated meaning bank is relatively short: the English QA-SRL bank, which consists of 1.2M words in more than 60K sentences, took only 9 days to annotate on Amazon Mechanical Turk. A natural first step towards a meaning representation annotated in a wide array of languages would be to design a Hebrew equivalent of QA-SRL, where one of the main challenges is its rich morphology requiring sub-word boundaries for both questions and answers.

Interactive semantics. The landscape of NLP is dominated by single-input, single-output supervision for tasks which imitate human intelligence (e.g., translation, summarization, and code generation). In all of these, a model has a "single shot": given an input instance, it needs to predict an output on which its accuracy is judged.

People, however, do not interact in this manner. For example, consider a reporter who is asked by their editor to *summarize* a recent U.N. agreement document. It is highly unlikely that either editor or reporter will contend with a single iteration on the output document. Instead, they will iterate over the story until reaching a version with which they are both reasonably happy. Similarly, a manager at a software company will give an initial requirement document to a programmer and through iterative cycles they will modify both output code and input document, which often deviates significantly from the its original version. Common to both of these scenarios, as well as many other human interactions, is the *mutual training* of both requester and worker. Both sides get acquainted to the other's expertise, lexicon and inherent limitations. **Introducing interactivity to the NLP framework will allow models iterative tries to reach a solution with a human counterpart**, correcting both model errors (e.g., getting the wrong word sense for an ambiguous word or incorrectly translating a person's gender), as well as informing users about possible mistakes in their requests. This setting poses many challenges in terms of training, evaluation and model design. An interpretable meaning representation, like those I have been developing [14, 25], will be useful in this scenario as it can point to misunderstandings between model and user at each iteration.

NLP for informed decision-making. I am excited to develop models which inform policy makers in a range of social issues. For example, we tracked the authorship trends in computer science and medicine literature in millions of papers over the last 50 years. We automatically estimated author gender based on first name distributions, and found that while more women are entering computer science research, they are doing so at a slow rate, projecting gender equality is at least a century away, and that men tend to collaborate and cite more male colleagues than expected [35]. These measurements and analyses can inform future decisions towards improving this state of affairs. **This work was covered in the New York Times, Science, Fortune, and more.**

In addition to measuring gender balance in academia, NLP can help inform a wide array of other much-debated topics. For example, combing news reports for gun-related incidents on a national scale is a laborious process, which unfortunately requires constant attention. The fruits of such effort, however, can have long-term positive implications. Imagine a system which provides law-makers easy access to all gun-related news stories, including automatic aggregates such as whether the depicted event was an accident or malicious, how many people were harmed, who were the perpetrators, and how the gun was obtained, all of which are often conveyed in news stories using natural language.

References

[1] Omri Abend and Ari Rappoport. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[2] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2670–2676, 2007.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, 2016.

[4] Anthony Chen, **Gabriel Stanovsky**, Sameer Singh, and Matt Gardner. Evaluating question answering evaluation. In *Machine Reading for Question Answering (MRQA) Workshop at EMNLP 2019*, Hong Kong, 2019 **(Received the Best Paper award)**.

[5] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, **Gabriel Stanovsky**, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*, 2019.

[6] Tobias Falke, **Gabriel Stanovsky**, Iryna Gurevych, and Ido Dagan. Porting an open information extraction system from english to german. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, November 2016. Association for Computational Linguistics.

[7] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. 2019.

[8] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

[9] Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[10] Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, **Gabriel Stanovsky**, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. SemEval-2019 task 10: Math question answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 893–899, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[11] Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? A critical investigation of popular benchmarks. *CoRR*, abs/1808.04926, 2018.

[12] Omri Koshorek, **Gabriel Stanovsky**, Yichu Zhou, Vivek Srikumar, and Jonathan Berant. On the limits of learning to actively learn semantic representations. In *Computational Natural Language Learning (CoNLL)*, 2019.

[13] John Lee and J Buddhika K Pathirage Don. Splitting complex english sentences. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 50–55, 2017.

[14] Julian Michael, **Gabriel Stanovsky**, Luheng He, Ido Dagan, and Luke S. Zettlemoyer. Crowdsourcing question-answer meaning representations. In *NAACL-HLT*, 2018.

[15] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal dependencies v1: A multilingual treebank collection. In *LREC*, 2016.

[16] Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[17] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

[18] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *EthNLP@EACL*, 2017.

[19] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *NAACL-HLT*, 2018.

[20] Vered Shwartz, **Gabriel Stanovsky**, and Ido Dagan. Acquiring predicate paraphrases from news tweets. In **SEM*, Vancouver, Canada, August 2017.

[21] Evangelia Spiliopoulou and Eduard Hovy. Definition frames: Using definitions for hybrid concept representations. *arXiv preprint arXiv:1909.04793*, 2019.

[22] **Gabriel Stanovsky**, Meni Adler, and Ido Dagan. Specifying and annotating reduced argument span via qa-srl. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

[23] **Gabriel Stanovsky** and Ido Dagan. Annotating and predicting non-restrictive noun phrase modifications. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

[24] **Gabriel Stanovsky** and Ido Dagan. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, November 2016. Association for Computational Linguistics.

[25] **Gabriel Stanovsky** and Ido Dagan. Semantics as a foreign langauge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[26] **Gabriel Stanovsky**, Ido Dagan, and Mausam. Open ie as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, 2015.

[27] **Gabriel Stanovsky** and Mark Hopkins. Spot the odd man out: Exploring the associative power of lexical resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[28] **Gabriel Stanovsky**, Julian Michael, Luke Zettlemoyer, and Ido Dagan. Supervised open information extraction. In *Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[29] **Gabriel Stanovsky**, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Association for Computational Linguistics (ACL)*, Florence, Italy **(Nominated for Best Paper award, top 1% of submissions)**, June 2019.

[30] **Gabriel Stanovsky** and Ronen Tamari. Y'all should read this! identifying plurality in second-person personal pronouns in english texts. In *Workshop on Noisy User-Generated Text at EMNLP 2019*, Hong Kong, 2019.

[31] **Gabriel Stanovsky**, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, Canada, August 2017.

[32] **Gabriel Stanovsky**, Jessica Ficler, Ido Dagan, and Yoav Goldberg. Intermediary semantic representation through proposition structures. *Semantic Parsing Workshop, ACL 2014*, page 66, 2014.

[33] **Gabriel Stanovsky**, Daniel Gruhl, and Pablo N. Mendes. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *Proceedings of the 2017 Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, April 2017. Association for Computational Linguistics.

[34] **Gabriel Stanovsky**, Omer Levy, and Ido Dagan. Proposition knowledge graphs. *Proceedings of the First AHA Workshop, Dublin, 2014*, pages 19–24, 2014.

[35] Lucy Lu Wang, **Gabriel Stanovsky**, Luca Weihs, and Oren Etzioni. Gender trends in computer science authorship, 2019 **(Covered in the New York Times, Science, Fortune, Axios, and more)**.

[36] Rachel Wities, Vered Shwartz, **Gabriel Stanovsky**, Meni Adler, Ori Shapira, Shyam Upadhyay, Dan Roth, Eugenio Martinez Camara, Iryna Gurevych, and Ido Dagan. A consolidated open knowledge representation for multiple texts. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem), in EACL*, Valencia, Spain, April 2017.

[37] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[38] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL-HLT*, 2018.