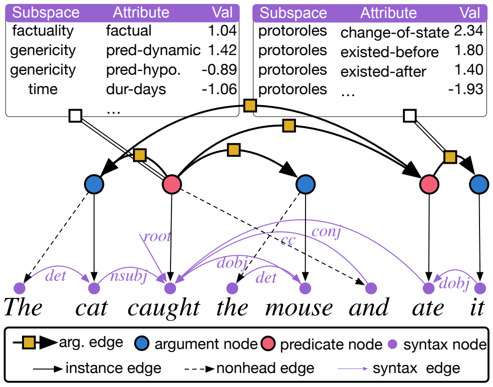
For both human and artificial agents, natural language represents an effective tool for communication and for reasoning rationally about the world. My research centers around this view of natural language and has three interconnected central goals:

1. Enabling humans to interact with digital agents via language.
2. Examining the way people use language to communicate with each other.
3. Leveraging language to improve agents' performance and efficiency.

There are two main themes which link these goals. The first involves **interaction, execution, and grounding**; this direction focuses on extracting structure from language as well as reasoning about and grounding that structure [1, 2, 3, 4, 5, 6, 7]. The second entails building an account of **implicit phenomena** inherent in natural language, such as ambiguity and vagueness [1, 2, 8, 9]. These phenomena are key to understanding language and acting intelligently but often pose a challenge for current systems, in part because they are not overtly expressed.

Consider asking a digital assistant to *"create meeting at noon tomorrow"*. The assistant might transform this utterance into a structured set of actions it can take in its domain (e.g. `CreateEvent` or `SetTime`); this structure needs to be bound to the environment (i.e. grounded) and executed. However, the language that people use with each other is not always so clear-cut. In conversation we rely on others' abilities to make inferences about implicit information. For example, we might assume our interlocutor can infer that *"let's meet around noon"* expresses an implicit preference for 12:00 (though 12:15 would be acceptable) or that *"let's have lunch on tuesday"* when uttered on a Monday might require a follow-up question (e.g. *"tomorrow or next week?"*). Handling such commonplace phenomena would allow us to communicate with digital agents more naturally while also revealing ways in which we might challenge and improve them.

Theme I: Interaction, Execution, and Grounding



Subspace	Attribute	Val
factuality	factual	1.04
genericity	pred-dynamic	1.42
genericity	pred-hypo.	-0.89
time	dur-days	-1.06
	...	

Subspace	Attribute	Val
protoroles	change-of-state	2.34
protoroles	existed-before	1.80
protoroles	existed-after	1.40
protoroles	...	-1.93

Figure 1: Predicting UDS graphs involves both structural semantic parsing *and* making implicit commonsense inferences (e.g. that the event *"caught"* likely did happen).

Parsing text into abstract structures of meaning is key to interacting with digital agents, communicating intent, executing commands, and reasoning, as well as to linguistic analysis. My work in this domain largely centers on parsing text into graph-based meaning representations, primarily using the transductive parsing paradigm [10], which recasts parsing as a sequence-to-graph problem. The resulting graphs can be efficiently processed using standard algorithms while remaining human-interpretable. When the graph is an abstract syntax tree or execution graph (e.g. [11, 4]) it can also be executed. Non-executable meaning representations, on the other hand, encode relevant linguistic properties which can be queried or used downstream.

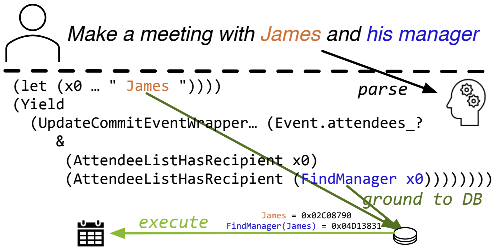
Descriptive Semantic Parsing Descriptive parsing aims to transform text into a structured representation describing relevant aspects of its meaning. Among the extant descriptive parsing formalisms, I have focused on Universal Decompositional Semantics (UDS) [12], a formalism which pairs/ a sentence with a Universal Dependencies (UD) syntactic parse, a semantic graph encoding its predicate-argument relations, and fine-grained, crowdsourced, scalar annotations over a variety of properties, from factuality (how likely it is that an event happened) to semantic proto-roles [13] such as volition, awareness, and change of state (see Fig. 1).

My work on UDS has been largely motivated by its utility in creating fine-grain descriptions of natural language: I have used UDS attributes associated with grammatical agents and patients to analyze the handling of subject and object control clauses in large pre-trained models [14], and in an ongoing study building on my work on ambiguous questions [9], where we are using UDS to analyze the factors making some "why" questions in English ambiguous.

I introduced the first-ever model to jointly predict a structured UDS graph and all of its attributes [1]. Parsing UDS graphs involves a unique multi-task challenge, as the model must predict a discrete and structured semantic graph as well as a set of over 50 distinct continuous attributes. In follow-up work [2], I also explored the role of the syntactic parse, developing an end-to-end model for simultaneously performing UD parsing, UDS graph parsing, and UDS attribute prediction, obtaining state-of-the-art results in UD and UDS, and testing for multilingual semantic transfer across 8 languages. Unlike previous work in joint syntactic-semantic parsing [15, 16], we were able to show bidirectional benefits between syntax and semantics. Additionally, we were among the first to introduce the popular Transformer neural architecture [17] – which has contributed to many of the recent successes in NLP and AI – to transductive parsing, where the relatively small sizes of semantic parsing datasets like UDS demanded model design and optimization changes.

Executable Semantic Parsing A long-standing goal in Artificial Intelligence (AI) is to transform text into an executable symbolic program [18, 19, 20]. This facilitates using language for instructing a digital agent as one would a person, and has become increasingly prevalent in the form of digital assistants (e.g. Siri, Alexa, Cortana, etc.) as well as in program synthesis tasks [21]. Furthermore, agents powered by pipelines of models often use executable domain-specific languages (DSLs) for communicating between elements in the pipeline; language- or code-like DSLs allow for interpretable communication and stand to benefit from large-scale unlabeled pre-training resources.

While executable parsing has seen remarkable advances, it is still brittle, especially to non-prototypical language; improvements in performance and robustness will broaden the domain in which it can be used. UDS parsing has proven useful for this sort of improvement: many of the methods we developed and refined for UDS transfer well to executable parsing. I have successfully modified UDS models for predicting programs for a neurosymbolic question-answering model [5] and for tackling the SMCalFlow benchmark [11, 4, 7]. SMCalFlow is a task-oriented parsing dataset mapping natural language commands to executable programs within Microsoft Outlook's calendar (cf. Fig. 2). It forms the basis for a digital assistant deployed to hundreds of millions of users. With minimal changes to a UDS model, we obtained state-of-the-art results in SMCalFlow. In ongoing work, we are examining how our model's confidence can help balance safety and usability [7].



Make a meeting with James and his manager

```
(let (x0 … " James "))))
(Yield
  (UpdateCommitEventWrapper… (Event.attendees_?
     &
      (AttendeeListHasRecipient x0)
      (AttendeeListHasRecipient (FindManager x0))))))))
```

```
James = 0x02C08790
FindManager(James) = 0x04D13831
```

Figure 2: An SMCalFlow program for a user utterance, containing an executable structure that can be grounded to referents and executed in a calendar.

Grounding A program's symbols must eventually be bound to referents in an environment in order to execute; for example, in Fig. 2 the function `FindManager` must return a valid pointer to a person in a database. Such *grounding* is key to any intelligent system, both for acting in the world and for learning world representations that are compatible with ours [22, 23, 24, 25] – a system that learns from massive amounts of ungrounded text is unlikely to represent the world the way a human (who learns from social and embodied interactions, etc.) does. I have been involved in grounding language to images [5, 6] and action [3]. Building on the expertise I acquired by implementing a custom Transformer architecture for UDS parsing [2], I developed a Transformer-based model for guiding a robot arm in a multi-step block manipulation task, trained on small amounts of raw image and text input [3]. In addition to achieving high performance – especially on compositional reasoning – and allowing for the re-use of pre-trained manipulation policies, the model represented one of the first uses of joint vision-and-language Transformers operating on text tokens and image patches (now a standard method), and was novel in the manipulation domain. Our method was able to handle both real and simulated images combined with templatic language (e.g. *"stack the red block on the blue block"*) and natural language from human annotators [26].

Theme II: Implicit Phenomena

The structured and discrete nature of semantic parses – which makes them well-suited for representing an utterance's logical content – also contributes to their rigidity and brittleness when faced with more scalar or implicit inferences. Much of what we communicate in language transcends what is said; this kind of implicit information, which rarely respects clear-cut boundaries, includes underspecification, ambiguity, vagueness, and commonsense.

Semantic Inferences and Commonsense In past work on UDS [1, 2] I have not only modeled discrete graph structures but also UDS attributes. These attributes represent flexible semantic inferences which are non-categorical in nature. For example, UDS attributes encode the fact that the event *"left"* is more likely to have happened in *"Jan knew that Chris left"* than *"Jan thought that Chris left"*, or that *"Dana broke the wishbone"* indicates more volition than does *"Dana broke her leg"*. These judgments are generally agreed upon by English speakers, despite being difficult to represent discretely. I have also explored commonsense inferences in a grounded setting; in [6] we probed models for commonsense inferences along a variety of visual attributes, including size (e.g. that an elephant is larger than a baseball) and color (e.g. that bananas are typically yellow but can be green). These inferences are subject to reporting bias [23]: the tendency for text to under- and over-report events. For example, bananas being yellow is under-reported, while plane crashes are over-reported. Using the dataset we introduced, we found that models trained on text alone were subject to more reporting bias than those trained on images and text. Our dataset and analysis have influenced lines of follow-up work on visual probing [27] and reporting bias [28, 29, 30].

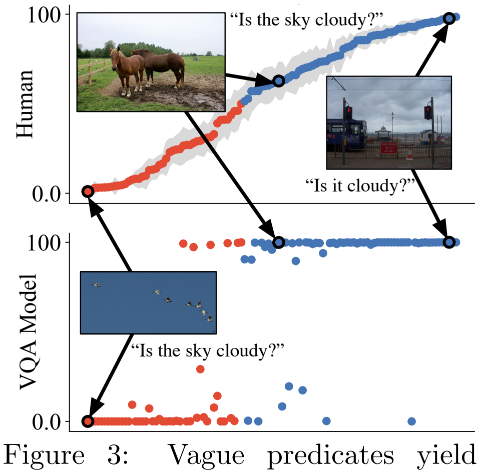


Figure 3: Vague predicates yield graded judgements in people but not in a common VQA model [8, 31]

Vagueness and Ambiguity Despite their nuance, commonsense inferences can often be coerced into a categorical format – given a forced binary choice, speakers generally agree on whether *Dana* in the above examples is volitional or not. However, for vague and ambiguous inputs, this is not the case; by definition, these phenomena defy a single categorical account. In [8] we explored vagueness in Visual Question Answering (VQA). We found that while people apply predicates like *"is sunny"* in a graded and continuous fashion (i.e. some skies are very sunny, some clearly cloudy, and some are in-between), models tend to give an all-or-nothing interpretation: 100% sunny or 100% cloudy (cf. Fig. 3). In VQA, questions can not only be vague but also ambiguous, simultaneously admitting several distinct interpretations. To explore this, we introduced a VQA dataset of ambiguous questions. Each question has multiple answers, which are grouped by the underlying question they address [9]. For example, answers to a question like *"What kind of flower is this?"* might be regrouped into species answers (e.g. *"lily"*, *"daisy"*) and those corresponding to a color (e.g. *"white"*, *"yellow"*). The question is rewritten for each group (e.g. *"What species of flower is this?"*). We then developed a question-disambiguation model able to recover implicit answer groups without supervision.

Future Work

Building on my past work [8, 14, 9], a core theme of my future research in interaction, execution, and grounding as well as in implicit phenomena is **documenting and utilizing the "long tail" of natural language phenomena**. Many NLP systems now work well on commonly-seen inputs. While frequent, these inputs represent a minority of the *types* of language phenomena. Despite recent progress, models often perform poorly on phenomena in the "tail" of the language distribution (i.e. the diverse majority of infrequently-seen types) which people grasp easily. Accounting for these non-prototypical phenomena is key to building robust natural language technology and revealing linguistic insights. I aim to explore the intersection of these phenomena and models, focusing on the models' representations and failure modes and using the rich body of linguistics and cognitive science research to guide model evaluation and development. Dovetailing with my work in multilingual parsing and cross-lingual word alignment [2, 32], my research will extend beyond English, as each language offers unique opportunities which are currently dramatically under-explored.

Working with various parsing formalisms [1, 2, 4] has drawn my attention to the challenges of designing abstract and structured meaning representations. In future work, I will explore how we can use the natural language's abstract nature to infer **domain-general representations** from data. Conversely, I also plan to examine how we can use existing structures to illustrate key differences between representations of the world learned from varying data sources. For example, an article and its corresponding video report might vary in their level of detail, focus, and style while expressing the same content; how can we characterize and quantify these differences? Broadening the scope of our past work on reporting bias [6], my research will address **data epistemology**: determining how factors of the input data contribute to the model's representations. I am especially keen to use meaning representations for describing the variation between models trained on different input modalities. Insights from this line of work could shed light on how meaning is represented in models, how that aligns with our concept space, and which differences lead to breakdowns.

Discrete, structured abstractions are key to **interpretability and trust**, a second major theme I plan to address in the future. In my past and ongoing work, I have built models for predicting interpretable meaning representations [1, 2, 5, 4, 7] and tackled some of the challenges of implicit phenomena like ambiguity and vagueness [8, 9]. My future work will combine these lines of research by exploring how agents can use structure for **reasoning under the uncertainty** induced by implicit language. Although implicit phenomena deal with things left unsaid, humans are able to reason about them and largely succeed in communication. I aim to explore the social, pragmatic, and psychological processes underlying reasoning about implicit concepts, and how we can better represent these processes in our agents and formalisms. I am especially keen to continue examining vagueness and ambiguity in interactive and grounded settings, as these have additional safety implications. Particularly in physical settings [3], where actions are often irreversible (e.g. robotic manipulation), a failure to correctly understand an interlocutor before acting can have disastrous consequences. This direction merges executable meaning representations, which will be central to acting on language instructions, with the ambiguity and vagueness endemic to natural language.

My work on implicit language has led to an ongoing line of research focused on **information-seeking agents**. Natural language comes "naturally" to us in part due to cognitive, experiential, and social commonalities allowing us to fill knowledge gaps. When we fail to fill these gaps, we rely on information-seeking behaviors. Acquiring information by asking questions is key to filling the gaps resulting from implicit language as well as to augmenting AI agents' reasoning capacity and facilitating human-AI coordination. To this end, in an ongoing collaboration with colleagues from Microsoft Research, I am examining the characteristics of questions and developing methods for augmenting agents with question-asking abilities.

References

[1] **Elias Stengel-Eskin**, Aaron Steven White, Sheng Zhang, and Benjamin Van Durme. "Universal Decompositional Semantic Parsing". *ACL* (2020).

[2] **Elias Stengel-Eskin**, Kenton Murray, Sheng Zhang, Aaron Steven White, and Benjamin Van Durme. "Joint Universal Syntactic and Semantic Parsing". *TACL* (2021).

[3] **Elias Stengel-Eskin***, Andrew Hundt*, Zhuohong He, Aditya Murali, Nakul Gopalan, Matthew Gombolay, and Gregory D. Hager. "Guiding Multi-Step Rearrangement Tasks with Natural Language Instructions". *CoRL* (2021).

[4] **Elias Stengel-Eskin**, Emmanouil Antonios Platanios, Adam Pauls, Sam Thomson, Hao Fang, Benjamin Van Durme, Jason Eisner, and Yu Su. "When More Data Hurts: A Troubling Quirk in Developing Broad-Coverage Natural Language Understanding Systems". *EMNLP* (2022).

[5] Zhuowan Li, **Elias Stengel-Eskin**, Yixiao Zhang, Cihang Xie, Quan Tran, Benjamin Van Durme, and Alan Yuille. "Calibrating Concepts and Operations: Towards Symbolic Reasoning on Real Images". *ICCV* (2021).

[6] Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and **Elias Stengel-Eskin**. "Visual Commonsense in Pretrained Unimodal and Multimodal Models". *NAACL* (2022).

[7] **Elias Stengel-Eskin** and Benjamin Van Durme. "Calibrated Interpretation: Confidence Estimation in Semantic Parsing". *In Submission* (2022).

[8] **Elias Stengel-Eskin**, Jimena Guallar-Blasco, and Benjamin Van Durme. "Human-Model Divergence in the Handling of Vagueness". *UnImplicit* (2021).

[9] **Elias Stengel-Eskin**, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. "Why Did the Chicken Cross the Road? Rephrasing and Analyzing Ambiguous Questions in VQA". *In Submission* (2022).

[10] Sheng Zhang. "Transductive Semantic Parsing". PhD thesis. The Johns Hopkins University, 2020.

[11] Semantic Machines et al. "Task-Oriented Dialogue as Dataflow Synthesis". *TACL* (2020).

[12] Aaron Steven White, **Elias Stengel-Eskin**, Siddharth Vashishtha, Venkata Subrahmanyan Govindarajan, et al. "The Universal Decompositional Semantics Dataset and Decomp Toolkit". *LREC* (2020).

[13] David Dowty. "Thematic proto-roles and argument selection". *Language* 67.3 (1991).

[14] **Elias Stengel-Eskin** and Benjamin Van Durme. "The Curious Case of Control". *EMNLP* (2022).

[15] Jayant Krishnamurthy and Tom Mitchell. "Joint syntactic and semantic parsing with combinatory categorial grammar". *ACL* (2014).

[16] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages". *CoNLL* (2009).

[17] Ashish Vaswani, Noam Shazeer, et al. "Attention is all you need". *NeurIPS* (2017).

[18] Terry Winograd. "Understanding natural language". *Cognitive psychology* (1972).

[19] John M Zelle and Raymond J Mooney. "Learning to parse database queries using inductive logic programming". *AAAI* (1996).

[20] Luke S Zettlemoyer and Michael Collins. "Learning to map sentences to logical form: structured classification with probabilistic categorial grammars". *UAI* (2005).

[21] Zohar Manna and Richard J Waldinger. "Toward automatic program synthesis". *Communications of the ACM* 14.3 (1971).

[22] Stevan Harnad. "The symbol grounding problem". *Physica D: Nonlinear Phenomena* (1990).

[23] Jonathan Gordon and Benjamin Van Durme. "Reporting bias and knowledge acquisition". *AKBC* (2013).

[24] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. "Experience Grounds Language". *EMNLP* (2020).

[25] Emily M Bender and Alexander Koller. "Climbing towards NLU: On meaning, form, and understanding in the age of data". *ACL* (2020).

[26] Yonatan Bisk, Deniz Yuret, and Daniel Marcu. "Natural language communication with robots". *NAACL* (2016).

[27] Tiancheng Zhao, Tianqi Zhang, et al. "VL-CheckList: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations". *arXiv* (2022).

[28] Yue Yang, Artemis Panagopoulou, Marianna Apidianaki, Mark Yatskar, and Chris Callison-Burch. "Visualizing the Obvious: A Concreteness-based Ensemble Model for Noun Property Prediction". *arXiv* (2020).

[29] Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. "Z-LaVI: Zero-Shot Language Solver Fueled by Visual Imagination". *arXiv* (2022).

[30] Fangyu Liu, Julian Martin Eisenschlos, Jeremy R. Cole, and Nigel Collier. "Do ever larger octopi still amplify reporting biases? Evidence from judgments of typical colour". *arXiv* (2022).

[31] Hao Tan and Mohit Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". *EMNLP-IJCNLP* (2019).

[32] **Elias Stengel-Eskin**, Tzu-Ray Su, Matt Post, and Benjamin Van Durme. "A Discriminative Neural Model for Cross-Lingual Word Alignment". *EMNLP-IJCNLP* (2019).