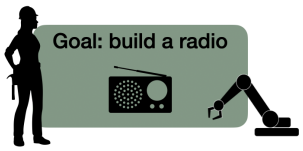
Alane Suhr: Research Statement

I build systems that use natural language to interact with human users. My research has two central goals: (a) to design models that reason about language within world and interaction context, and (b) to develop learning algorithms for acquiring language in interaction with users. Such systems have enormous potential impact; they will enable non-experts to access complex systems such as robots and databases, and will learn from and teach people new skills and concepts.

Natural language is intrinsically interactive and situated. We use language to communicate with each other about our shared world (e.g., by making references to our environment) and to relay our intents (e.g., by making requests to one another). For example, consider an interaction where a human and a robot collaborate to build a radio (Figure 1). Both participants use language to delegate tasks, make corrections, ask for clarifications, and exchange information about their progress. However, the majority of research in NLP focuses on static texts, at times with accompanying context (e.g., an image), but often in isolation. My work expands the perspective of the field to consider the full complexity of situated language interactions. This reveals both challenges and opportunities that significantly alter the language problem.



1	User	Can you build the amp while I add the volume control?
2	System	Sure, which speaker do you want me to use?
3	User	The smallest one in the red bin to the right.
4	System	*grasps a speaker*
5	System	Do you mean this one?
6	User	No, that's the largest one. I want the smallest.
7	System	*derives a learning signal and applies it to model parameters*
		...
8	System	Did you finish the volume control? I think I have the potentiometer you need.
9	User	Oh, there it is! Thanks.

Figure 1: Illustration of a natural language interaction between a user and a robot working together to build a radio.

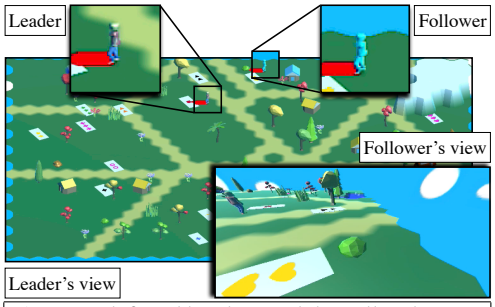
The first set of challenges is modeling how language depends on the context of the interaction. Participants should understand and generate references to the world around them (e.g., *"the smallest one in the red bin to the right"* in utterance 3 of Figure 1), and should be able to recall and refer to the interaction history (e.g., the system referring to the user's current task in utterance 8). Interactive systems must efficiently reason about interaction context, including visual observations of the world and the history of the interaction, both linguistic and non-linguistic. I designed several of the first systems that use interaction history to map language to expressive, executable representations, including SQL (Suhr et al., 2018) and low-level actions (Suhr and Artzi, 2018). My work laid the foundation for recent progress in vision-dependent language reasoning by creating benchmarks that require systems to understand how references to sets, counts, comparisons, and other phenomena arise in images (Suhr et al., 2017, 2019b; Chen et al., 2019).

Learning is of equal importance to modeling in this scenario, where it presents both open problems and new opportunities. The dynamic nature of interaction reveals a key learning challenge not addressed by contemporary NLP approaches. Existing approaches predominantly rely on learning from static datasets, e.g., recordings of interactions between two people. However, the language and observations a system encounters during interaction are highly dependent on the dynamics of that interaction. For example, if the system struggles to complete a high-level task requested by the user, the user may adapt by describing the task in a sequence of low-level steps, or may simply decide to do it themselves, and delegate other tasks to the system. This dependence between the type of language seen and the system's capabilities results in a constantly changing data distribution. Static data cannot provide the necessary learning signal to handle such dynamic data. However, the interaction itself provides natural opportunities for learning, allowing users to shape system behavior via feedback without disrupting the interaction. For example, in utterances 3–7 of Figure 1, the system misunderstands the meaning of *"smallest"*, and the user's correction provides a

learning signal that is used to update the model. I study this problem by building research platforms, such as CerealBar, a collaborative game environment where two players coordinate using natural language (Suhr et al., 2019a). We have used CerealBar to analyze language change in interactions (Effenberger et al., 2021), and build systems that follow instructions (Suhr et al., 2019a) and continually learn from interaction to generate instructions (Kojima et al., 2021).

1 Learning and Using Language in Collaborative Interactions

People use natural language as a collaborative instrument to coordinate acting together in the world. However, most NLP research does not account for this collaborative aspect, abstracting over important complexities of the problem and missing opportunities to aid both interpretation and learning. A significant obstacle to studying this problem is creating situated, task-oriented, and collaborative scenarios that emphasize natural language interaction. I addressed this by creating CerealBar, a collaborative language-based 3D game where two players complete shared tasks (e.g., collecting sets of cards), using language to coordinate their actions (Suhr et al., 2019a).



1	Turn left and head toward the yellow hearts, but don't pick them up yet. I'll get the next card first.
2	Okay, pick up yellow hearts and run past me toward the bush sticking out, on the opposite side is 3 green stars

Figure 2: A snapshot from an interaction in CerealBar. Players coordinate by writing and following instructions.

CerealBar is a full-fledged game implemented from scratch in a professional game development environment. This enables large-scale data collection of human interaction data via crowdsourcing, and deployment and training of language-using systems in live interactions with human users. In a series of papers, we showed how the CerealBar data allows us to study language in a way previously not possible, and how it provides an effective platform to train language-using models through interaction with users. For example, previous work found that as people form language conventions over time, their sentences become simpler (Clark and Wilkes-Gibbs, 1986). We showed that CerealBar reveals a more nuanced evolution: as conventions form, instructors adapt to relay more complex tasks by using more complex sentences (Effenberger et al., 2021). This illustrates the importance of developing more expressive scenarios like CerealBar.

With CerealBar, we studied instruction following in a way previously not explored. The collaborative nature of CerealBar brings about more flexible instruction giving: if a system executing an instruction fails, the human giving the instructions can rephrase, simplify the task, or even decide to do the task themselves and delegate simpler ones to the system. We showed that a neural model can effectively learn to follow instructions in such scenarios, operating alongside human users (Suhr et al., 2019a). A key challenge was using static data from recorded human-human interactions, which does not reflect the dynamics of system failure and recovery. While full-fledged reinforcement learning is not possible using such static data, we showed that there is an effective middle ground. We introduced a learning algorithm that combines supervised learning from recorded interactions with as much self-experience as possible without deviating too much from the recordings. This method uses the static data to learn between-instruction error recovery.

Most recently, together with collaborators, we used CerealBar to seamlessly incorporate learning into interaction (Kojima et al., 2021). Interactions are replete with implicit learning signals, which existing work does not take advantage of, relying instead on static datasets that are both expensive to collect and not fully representative of the interaction dynamics. We studied the problem of learning to generate instructions by observing human behavior. A system communicates an internal intent in the instructions it gives to human users. Comparing the user's execution of an instruction with this internal intent provides a signal about

1	show me flights from seattle to boston next monday `(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.from_airport IN ...(SELECT city.city_code FROM city WHERE city.city_name = 'SEATTLE'))) AND (flight.to_airport IN ...(SELECT city.city_code FROM city WHERE city.city_name = 'BOSTON'))) AND (flight.flight_days IN ...(SELECT date_day.day_name FROM date_day date_day.month_number = 2 AND date_day.day_number = 8))));`
2	on american airlines `(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.airline_code = 'AA') AND (flight.from_airport IN ...(SELECT city.city_code FROM city WHERE city.city_name = 'SEATTLE'))) AND (flight.to_airport IN ...(SELECT city.city_code FROM city WHERE city.city_name = 'BOSTON'))) AND (flight.flight_days IN ...(SELECT date_day.day_name FROM date_day WHERE AND date_day.month_number = 2 AND date_day.day_number = 8)));`

Figure 3: The first two utterances of an interaction between a user and a natural language travel planning system. Each user request is mapped to a SQL query. Highlighted SQL segments arise from implicit dependency on interaction history, rather than the current utterance.

the quality of the communication (i.e., the generated instruction). We designed a contextual bandit learning method that uses this signal to continually train a language generation system through interaction with users; as the system interacts with users, it becomes better and better at communicating its intent.

2 Learning to Reason about Interaction History to Resolve Meaning

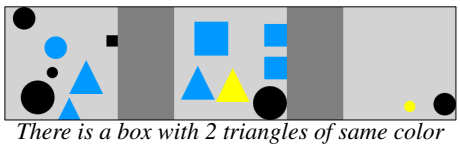
The common approach for mapping language to formal meaning representations, at the time I started studying this problem in 2017, was to use handcrafted, linguistically-inspired representations such as lambda calculus (e.g., Zettlemoyer and Collins, 2007). The rationale behind this approach is simple: targeting a representation that is relatively close to natural language, but still formal, simplifies the learning problem. However, there are multiple drawbacks to this approach. The design process is labor-intensive, and it often implicitly incorporates latent assumptions that limit coverage of language phenomena. Recovering these symbolic representations requires careful search over a large combinatorial space, and the output representations are often not executable, requiring an additional step to translate them to actual system representations or actions. This problem is further exacerbated in interactive scenarios, where hand-crafted representations must also account for all the different ways that an utterance's meaning can depend on the interaction history. In a series of papers, we showed that learning to directly map from language to system representations in interactive systems does not only obviate the design process, it also results in better systems. The repeating theme of this line of work is trading off representation design with learning challenges.

In Suhr et al. (2018), we built the first system that directly maps user requests to executable SQL queries in interaction, a task that was previously addressed by using non-executable linguistic representations (Zettlemoyer and Collins, 2009). A key challenge is that much of the recovered SQL comes from the interaction history, and not from the utterance itself. Figure 3 illustrates this. While the request *"on american airlines"* does not explicitly mention a desired city or date, its meaning implicitly depends on the previous utterance, which directly specifies these search parameters. We designed a neural network that does this in two ways: it maintains an implicit representation of the interaction as it progresses, and it explicitly copies segments from previously-generated queries. Our approach is not only effective in recovering history-dependent meaning, but also more efficient, as generation decisions from previous queries can be copied as complete segments. **This work received an Outstanding Paper Award at NAACL 2018**, and demonstrated the feasibility of using system-provided executable representations for mapping language to meaning in interaction. As a result, this area has received increasing attention in recent years, with the introduction of new datasets and models (e.g., Yu et al., 2019; Zhang et al., 2019). An exciting focus of this recent work is cross-database generalization, where text-to-SQL systems are tested on databases and domains that were not seen during training. In Suhr et al. (2020), I characterized the challenges this new scenario raises, and re-formulated existing benchmarks to better reflect them.

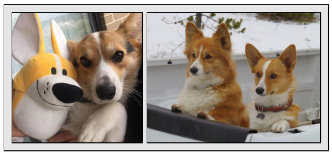
Of course, not all systems utilize a symbolic compositional representation like SQL. When systems operate with lower-level actions (e.g., a robotic agent that moves in the world), the gap between language and low-level actions is much larger. We showed that even then, it is better to map language directly to low-level actions without going through intermediate symbolic representations (Suhr and Artzi, 2018). We studied the problem of following a sequence of instructions in an interaction. We built a neural network that maps directly from language and interaction history to low-level actions, trading-off handcrafted representation design with learning challenges. We introduced a contextual bandit learning algorithm that uses weak goal-state annotations only, and overcomes exploration challenges that arise in large action spaces. Our method avoids the problem of representation design, and outperforms prior work dependent on symbolic representations work by up to 25 accuracy points. It is also competitive with supervised learning, which has stronger annotation requirements.

3 Reasoning about Language in Situated Environments

Language is an expressive medium for describing visual and spatial environments, including images. It can precisely describe complex and compositional structure in images, including sets, spatial relations, and quantities. These kind of reasoning skills are critical for intelligent systems that operate in the real world. However, datasets for language-and-vision traditionally focus on shallow types of reasoning that require little more than object recognition, and include spurious biases that make them unreliable for model evaluation (Goyal et al., 2017). A key challenge is collecting data that includes complex images with natural language annotations. In a series of papers, I defined procedures for collecting large amounts of high-quality language-and-vision data, which resulted in several datasets that require complex, compositional reasoning.



There is a box with 2 triangles of same color nearly touching each other.



All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.

Figure 4: Examples from NLVR (top) and NLVR2 (bottom). The task is to classify if the statement is true or false with respect to the image pair.

Natural Language for Visual Reasoning (NLVR; Suhr et al., 2017) includes synthetically generated images paired with natural language descriptions, where the task is to determine whether a sentence is true about an image (Figure 4, top). We defined a new data collection process that results in complex, compositional language that requires reasoning about counts, sets, spatial relations, comparisons, and more. The key idea is to generate images that require reasoning over groups of objects, and to ask crowdworkers to write sentences that implicitly contrast images. Our crowdsourcing procedure has influenced creation of several recent datasets (e.g., Gardner et al., 2020; Liu et al., 2021). **This work received the Best Resource Paper Award at ACL 2017.**

In Suhr et al. (2019b), we scaled the NLVR process to real photographs from the web to create the NLVR2 dataset (Figure 4, bottom). Without the ability to generate images, the key challenge was finding a large number of images that contain sets, quantities, and spatial relations to refer to. NLVR2 has been instrumental in recent progress towards stronger vision-and-language systems, especially as a testbed to evaluate the first large pre-trained multimodal models (Tan and Bansal, 2019; Chen et al., 2020). NLVR2 remains influential, as it is now used as a standard benchmark for evaluating such models.

We also scaled this line of work to systems that follow natural language instructions in visual environments. Together with collaborators, we developed the Touchdown task and dataset (Chen et al., 2019). Touchdown requires systems to follow long instructions, identify objects, and reason about complex spatial relations in a realistic urban environment based on the Google Street View platform.

4 Future work

My research goal is to build systems that use natural language to interact with human users, in a safe, transparent, and interpretable way. Dynamic interaction dramatically complicates the language problem, but also provides opportunities for learning from implicit and explicit feedback signals, opening the way for systems that continually improve and adapt. Below are some of the research directions I plan to pursue.

Learning from Interaction Feedback Interactions are rich in implicit and explicit feedback signals that we can use to train better models. Equally important is that human users expect systems to learn and adapt in interaction, because it is a critical part of their own interaction with others. I will take inspiration from theories of child language acquisition to design new learning algorithms that make the most of interaction feedback. Children learn language in interaction by observing and predicting interaction dynamics, with very little explicit feedback (Tomasello, 1992). This process eventually builds to a point where word learning accelerates rapidly (Goldfield and Reznick, 1990). In designing these algorithms, I will consider questions that have been debated at length in linguistics and cognitive science about innateness, action, and learning during language acquisition. Building systems that acquire language requires investigating similar questions, including: what inductive biases must our models have before they can effectively learn from interaction? How should they act so that they can get the most informative feedback, while remaining cooperative and useful? How should this feedback be transformed into a learning signal that can effectively update a model?

Rapid System Personalization People in natural language interactions show rapid partner-specific adaptation and convention formation (Clark and Wilkes-Gibbs, 1986), and the same is not only expected from systems, but required for efficient interaction. For example, in goal-directed interactions, systems should use different jargon depending on the domain expertise of the user. I will address this problem by maintaining and mixing models of both general and partner-specific interaction (Hawkins et al., 2021). A key challenge is sample complexity: it is infeasible to learn a model of each new user from scratch given only a short interaction with them. I will address this challenge by designing meta-learning methods that simultaneously learn both user-specific models and strategies for quickly adapting to users in general (Zhu et al., 2021).

Language and Beyond for Natural Interaction The complex facets of human language interaction – including bidirectional conversation, speech and gesture, and complex visual situations – support more natural and streamlined interaction, but have been widely understudied in interaction. Two barriers are responsible for this: the difficulty of mapping all the different signals to designed meaning representations and the lack of rich interaction platforms that display all facets. My work on representation learning within interactions and visual domains places me in an ideal position to incorporate further communication signals. I will address the challenge of limited language interaction data by incorporating pre-trained multimodal models into my approach. Drawing on my experience with CerealBar, I will design immersive, engaging, extensible, and scalable platforms to drive this research via large-scale studies.

Safe and Interpretable Systems Systems that participate in and learn from interaction must be transparent, interpretable, and safe to deploy. Because these systems learn from users, we must ensure that what they learn (and thus, what they act on) is not harmful. Systems should be able to provide verifiable and easy-to-understand explanations about how they learn and act. I will explore how modular neural networks (Andreas et al., 2016) and program synthesis (e.g., Wong et al., 2021) can support this by automatically learning libraries of executable, high-level, and domain-specific functions that have direct correspondence with language. This will support interpretability by providing users with derivations of how language is mapped to low-level actions via compositions of discrete library functions. In addition, users can quickly debug and adjust undesirable or incorrect model behavior by analyzing and modifying these learned functions.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *CVPR*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *ECCV*.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Anna Effenberger, Rhia Singh, Eva Yan, Alane Suhr, and Yoav Artzi. 2021. Analysis of language change in collaborative instruction following. In *Findings of the Association for Computational Linguistics: EMNLP*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP*.

Beverly A. Goldfield and J. Steven Reznick. 1990. Early lexical acquisition: rate, content, and the vocabulary spurt. *Journal of Child Language*, 17(1):171–83.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Robert D. Hawkins, Michael Franke, Michael C. Frank, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2021. From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*.

Noriyuki Kojima, Alane Suhr, and Yoav Artzi. 2021. Continual learning for grounded instruction generation by observing human following behavior. *TACL*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *EMNLP*.

Alane Suhr and Yoav Artzi. 2018. Situated mapping of sequential instructions to actions with single-step reward observation. In *ACL*.

Alane Suhr, Ming-Wei Chang, Peter Shaw, and Kenton Lee. 2020. Exploring unexplored generalization challenges for cross-database semantic parsing. In *ACL*.

Alane Suhr, Srinivasan Iyer, and Yoav Artzi. 2018. Learning to map context-dependent sentences to executable formal queries. In *NAACL-HLT*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *ACL*.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019a. Executing instructions in situated collaborative interactions. In *EMNLP*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019b. A corpus for reasoning about natural language grounded in photographs. In *ACL*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Michael Tomasello. 1992. The social bases of language acquisition. *Social development*, 1(1):67–87.

Catherine Wong, Kevin Ellis, Joshua B. Tenenbaum, and Jacob Andreas. 2021. Leveraging language to learn program abstractions and search heuristics. In *ICML*.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019. SParC: Cross-domain semantic parsing in context. In *ACL*.

Luke Zettlemoyer and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *EMNLP-CoNLL*.

Luke Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. In *ACL-AFNLP*.

Rui Zhang, Tao Yu, Heyang Er, Sungrok Shim, Eric Xue, Xi Victoria Lin, Tianze Shi, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019. Editing-based SQL query generation for cross-domain context-dependent questions. In *EMNLP-IJCNLP*.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *ICML*.