# Learning Language through Grounding

**Freda Shi**[1,2] **Ziqiao Ma**[3] **Jiayuan Mao**[4] **Parisa Kordjamshidi**[5] **Joyce Chai**[3]
[1]University of Waterloo [2]Vector Institute & Canada CIFAR AI Chair [3]University of Michigan
[4]Massachusetts Institute of Technology [5]Michigan State University
fhs@uwaterloo.ca, {marstin,chaijy}@umich.edu
jiayuanm@mit.edu, kordjams@msu.edu

## Abstract

*Grounding* has been a long-standing concept in natural language processing (NLP) and computational linguistics (CL). This tutorial provides a historical overview and introduces recent advances in learning language through grounding, with a particular emphasis on the latter. We will begin by tracing the history of grounding and presenting a unified perspective on the term. In Parts II to IV, we will delve into recent progress in learning lexical semantics, syntax, and complex meanings through various forms of grounding. We will conclude by discussing future directions and open challenges, particularly those related to the growing trend of large language models and scaling.

## 1 Introduction

Humans can learn and use language efficiently and effectively. Importantly, human language acquisition, learning, and use are *grounded* to contexts in the world, drawing on information from various sensory modalities such as visual and auditory signals, as well as interactions with other people and the surrounding physical environment. To better understand human language learning and build intelligent language agents, modeling language acquisition and learning has been a long-term interest in the fields of computational linguistics (CL) and natural language processing (NLP), drawing cross-disciplinary interest from fields such as cognitive science and robotics (Steels et al., 2000; Steels, 2003). Recently, we have all witnessed that advances in machine learning, especially those driven by large-scale multimodal data, have significantly advanced the field of grounded language learning (Kiros et al., 2014; Tan and Bansal, 2019; Radford et al., 2021, *inter alia*).

Although language grounding has emerged as a promising direction in language understanding research and is drawing growing attention from the CL and NLP community (Bisk et al., 2020), the usage of the term *grounding* has diverged (Chai et al., 2018). Two major lines of work on grounding started almost contemporaneously in different areas: [1; semantic grounding] in artificial intelligence (AI), which refers to connecting language units to the world that gives language meanings (Harnad, 1990), and [2; communicative grounding] in psychology, which means finding common ground between agents in a communication setting (Clark and Brennan, 1991). Given the interdisciplinary nature and rapid advancement of CL and NLP, especially driven by the emerging trend of language models and scaling, the convergence of various fields has made it imperative to re-examine the concept of grounding in the current context.

This tutorial will start by reviewing the evolution of the concept of grounding in the literature and introduce a unified view of grounding (Shi, 2024). After that, we will introduce recent work on computational approaches that model the acquisition and learning of natural language by grounding natural language to the world, including learning lexical semantics, syntax, complex forms of semantics, and pragmatics in various grounded settings. The tutorial will conclude with a discussion of future directions and open problems in the field, particularly regarding the emerging trend of language models and scaling.

**Tutorial type.** This tutorial should be classified as *cutting-edge* in CL/NLP, as most of the content was introduced within the last ten years; however, it will also include introductory content that familiarizes the audience with the necessary backgrounds.

**Related tutorials.** There have been two *ACL tutorials focusing on communicative grounding, i.e., building common ground through dialogue communications (Alikhani and Stone, 2020; Krishnaswamy and Pustejovsky, 2022); in contrast, this tutorial will (1) place greater emphasis on semantic grounding when introducing exemplar work and (2) delve deeper into the pragmatic aspects of



*Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 38–43

language use in communicative grounding, which has not been explicitly mentioned in previous tutorials. Additionally, while a related NAACL tutorial by a member of the instructor team introduces spatial and temporal language understanding (Kordjamshidi et al., 2024), this tutorial will offer a broader overview of the grounded language learning problems in general.

**Target audience.** This tutorial aims to attract a broad audience. We expect that the tutorial will be of interest to researchers and students in CL/NLP, particularly those who are interested in multimodal machine learning and grounded language learning. We expect the audience to understand basic concepts in calculus (derivatives and calculus), linear algebra (matrix, linear transformation), probability (probability distribution, expectation, and variance), and statistics (sampling from a distribution). Familiarity with neural networks, common practices in deep learning, and basic linguistic concepts is desired but not required.

## 2 Outline of Tutorial Content

We anticipate that the tutorial will be about 3 hours long. After an introduction to grounding, we break down the tutorial with respect to particular learning objectives.

- **Part I** (20 minutes): Introduction to grounding. We will review the history of grounding (Harnad, 1990; Clark and Brennan, 1991; Chai et al., 2018), and introduce the unified definition of grounding (Shi, 2024). In particular, grounding, in this tutorial, refers to processing the primary data with supervision from another source, where the two sources of data have positive mutual information. We will exemplify the definition through connection to existing work such as visual grounding (e.g., Kiros et al., 2014), acoustic grounding (e.g., Settle et al., 2019), factual grounding (e.g., Ghazvininejad et al., 2018), and cross-lingual grounding (e.g., Chen et al., 2021). We refer to Alikhani and Stone (2020) for building common ground through communication.
- **Part II** (30 minutes): Learning lexicons through grounding. Word acquisition has been a fundamental problem in language acquisition concerned by both cognitive sciences (Siskind, 1996; Roy and Pentland, 2002; Xu and Tenenbaum, 2007; Qu and Chai, 2010, *inter alia*) and robotics (Steels et al., 2000; Steels, 2003). With the advancement of neural networks and multimodal machine learning, there has been work on learning the meanings of written (Forbes and Choi, 2017; Mao et al., 2019; Ma et al., 2023) or spoken (Peng and Harwath, 2022) words by grounding language to visual signals. Particularly, there has been work focusing on grounding verb semantics to the change of the physical world (Gao et al., 2016; She and Chai, 2016, 2017). Another line of work on learning lexicons through cross-lingual grounding can be exemplified by Chen et al. (2021).

  In the first 10 minutes, we will introduce the background and focus on recent advances in the remaining time. Work on learning lexical semantics through interaction (Wang et al., 2016) or learning lexicon to compose sentence-level meanings (Mao et al., 2021) will be deferred to Part IV.
- **Part III** (30 minutes): Learning syntax through grounding. As shown by Shi et al. (2019), constituency parses of sentences can be learned by grounding to visual signals. Follow-up work has demonstrated the effectiveness of such visually grounded systems on learning variants of constituency (Kojima et al., 2020; Zhao and Titov, 2020; Zhang et al., 2021; Lai et al., 2023; Portelance et al., 2024) and dependency (Su et al., 2021) grammars. On another line, word alignment–based cross-lingual transfer can also be considered as an instantiation of learning syntax through cross-lingual grounding (Ma and Xia, 2014; Rasooli et al., 2021; Kurniawan et al., 2021, *inter alia*), where the text in the target language(s) is grounded to existing knowledge in the source language(s).

  A brief introduction of related syntactic knowledge, such as constituency (Chomsky, 1957), dependency (Tesnière, 1959), and combinatory categorial grammars (CCGs; Steedman, 2000), will be presented in the first 10 minutes of this part to help the audience better understand the content. We will focus on recent approaches to learning syntax through visual grounding and cross-lingual grounding in the rest of the time. Efforts on joint learning of syntax and semantics (e.g., Mao et al., 2021) will be delivered in Part IV.
- **Part IV** (60 minutes): Learning complex meanings, e.g., semantics and pragmatics, through grounding. It has attracted significant interest in learning (Kiros et al., 2014; Tan and Bansal,

2020; Mao et al., 2021, *inter alia*) and evaluating (Antol et al., 2015; Suhr et al., 2017; Kordjamshidi et al., 2017; Lake and Baroni, 2018; Mirzaee et al., 2021, *inter alia*) meaning acquisition in visually grounded settings. In addition to visual grounding, interaction is also a common source of supervision, where considerations regarding pragmatics and theory of mind are often taken into account (Wang et al., 2016; Yu et al., 2020; Zhu et al., 2022; Liu et al., 2023; Ou et al., 2023). Similarly to what has been mentioned in Part II, cross-lingual transfer on sentence or document-level meanings, particularly transferring knowledge from high-resource to low-resource languages, should also be considered as instantiations of cross-lingual grounding (Artetxe et al., 2020; Tran et al., 2020; Shi et al., 2023, *inter alia*).

This part will cover three topics for 20 minutes each: learning semantics through grounding, learning pragmatics through grounded interaction, and learning cross-lingual text representations through cross-lingual grounding.

- **Part V** (15 minutes): Discussion on future directions and open problems. A key discussion for future directions centers around whether grounding should emerge naturally from scaling models or whether we should enforce grounded supervision to achieve more efficient learning. Additionally, the scope of grounding can be broadened beyond traditional modalities, incorporating touch, olfaction, non-human sensors, video and temporal data, 3D environments, proprioception, episodic experiences, and even other forms of meta-cognition.

There will be a 5-minute Q&A session at the end of each part to address questions from the audience. A 30-minute coffee break will be held after Part III. We anticipate that 30% of the tutorial content will be work by at least one of the instructors, with the rest coming from other research groups.

### 2.1 Recommended Reading List

All readings below are optional; however, reading the materials listed below in advance will help the audience understand the tutorial content better.

We recommend Harnad (1990) and Clark and Brennan (1991) as background knowledge, as well as Chai et al. (2018) and Chapter 2 of Shi (2024) for an introduction to grounded language learning in a recent context. For audiences interested in more comprehensive surveys of recent work on language grounding, we recommend Bisk et al. (2020) for a roadmap and Chandu et al. (2021) for why communicative grounding has been overlooked.

Regarding background knowledge about linguistic structures, we recommend Chapters 6, 18, 19, and 21 of Jurafsky and Martin (2024), as well as Steedman (2000) for audiences interested in gaining a deeper understanding of CCGs.

For parts II-IV, we will refer the audience to the cited articles above in the tutorial outline. We encourage the audience to use the papers mentioned above as “pointers” to more comprehensive literature, which can be accessed through papers citing or cited by these aforementioned ones.

## 3 Specification of the Tutorial

**Diversity considerations.** This tutorial promotes diversity from the following aspects:

- **Topic diversity.** Language grounding is inherently a multidisciplinary field that intersects linguistics, machine learning, computer vision, robotics, cognitive science, and social science.
- **Linguistic diversity.** This tutorial will promote linguistic diversity in the NLP community. A key aspect of grounded language learning is cross-lingual grounding, which has been extensively explored in recent research, and the tutorial will cover a significant amount of work on cross-lingual grounding. In particular, much of the covered work in this tutorial has significantly contributed to low-resource NLP.
- **Diverse representation in the instructor team.** Our instructor team represents a diverse range of backgrounds, including various seniority levels, demographic backgrounds, home institutions, and geolocations.
- **Effort on participation diversity.** We are committed to making the tutorial content accessible to a broad audience with diverse academic backgrounds by ensuring that the fundamental concepts are thoroughly covered in the released tutorial material. Upon acceptance, we will invest efforts on encouraging everyone in the CL/NLP community to participate, especially those from a historically underrepresented group.

**Estimated audience size.** Given that cross-lingual NLP and cross-modal machine learning have been popular research topics in recent years and that the tutorial will cover a wide range of topics in grounded language learning, we expect around 200 in-person attendees.

**Technical equipment.** We have no major technical requirements beyond the basic settings of conference tutorials: we will need a projector, a screen, and a microphone. Connection to the Internet is strongly preferred but not required.

**Ethics statement.** Learning language through grounding involves the use of human language and multimodal data, which may raise ethical concerns in terms of data privacy and fairness. We will emphasize related ethical considerations in the tutorial, including the importance of data privacy, the potential biases in data used in relevant research, and the common practices in the field to mitigate negative impacts. All tutorial material will be made publicly available after the tutorial.

## 4 Tutorial Instructors

**Freda Shi** is an Assistant Professor at the University of Waterloo and a Faculty Member at the Vector Institute, where she holds a Canada CIFAR AI Chair position. Her research interests are in the intersection of natural language processing, with a focus on grounded language learning. Her work has been recognized with a Google PhD Fellowship, and several nominations for the Best Paper Award at ACL. Her doctoral dissertation entitled *Learning Language Structures through Grounding* focuses on learning syntactic and semantic structures through grounding language to various modalities that represent the real world. She has given invited talks on grounded language learning, syntactic parsing, and multilingual NLP at more than ten institutions in multiple countries.

**Ziqiao Ma** is a Ph.D. candidate at the University of Michigan. His research stands on the intersection of language, interaction, and embodiment from a cognitive perspective, with the goal of grounding and aligning language agents to non-linguistic modalities and rich interactive contexts. He received the Weinberg Cognitive Science Fellowship, an Outstanding Paper Award at ACL 2023, and the Amazon Alexa Prize Simbot Challenge Award. He co-organized the SpLU-RoboNLP workshop at ACL 2024, Bi-Align workshop at ICLR 2025. He has instructed NLP/AI courses three times, and has delivered guest lectures and invited talks on situated language processing and grounded and interactive language learning at multiple institutions worldwide.

**Jiayuan Mao** is a Ph.D. student at MIT, advised by Professors Josh Tenenbaum and Leslie Kaelbling. Her research agenda is to build machines that can continually learn concepts (e.g., properties, relations, rules, and skills) from their experiences and apply them for reasoning and planning in the physical world. Her research topics include visual reasoning, robotic manipulation, scene and activity understanding, and language acquisition. Her work is supported by an MIT presidential fellowship. She has co-organized the Workshop on Planning in the Era of LLMs at AAAI 2024, the Workshop on Learning Effective Abstractions for Planning at CoRL 2024, the Workshop on Visual Concepts at ECCV 2024, the workshop on Visually Grounded Interaction and Language (ViGIL) at NAACL 2021, and the Neuro-Symbolic Visual Reasoning and Program Synthesis tutorial at CVPR 2020.

**Parisa Kordjamshidi** is an Associate Professor at the CSE Department, Michigan State University. She has worked on spatial language understanding, language grounding, and a variety of vision and language problems. She organized shared tasks on Spatial role labeling, SpRL-2012, SpRL-2013, the Space Evaluation workshop, SpaceEval-2015 in SemEval Series, Multimodal spatial role labeling workshop (MSpRL) at CLEF-2017. She co-organized SpLU-2018, SpLU-RoboNLP-2019, SpLU-2020, SpLU-2021, and SpLU-RoboNLP-2024 workshops collocated with NAACL, EMNLP and ACL. She has given tutorials on spatial and temporal language understanding at NAACL 2024 and COLING 2022.

**Joyce Chai** is a Professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. Her research interests span NLP and embodied AI to human-AI collaboration. Her current work explores the intersection between language, perception, and action to enable situated communication with embodied agents. She served on the executive board of NAACL and as a Program Co-Chair for multiple conferences, most recently ACL 2020. She is a recipient of the NSF CAREER Award and multiple paper awards with her students (e.g., Best Long Paper Award at ACL 2010, Outstanding Paper Awards at EMNLP 2021 and ACL 2023). Her lab (the SEAGULL team) won the first prize in Amazon Alexa Simbot Challenge in 2023. She is a Fellow of ACL.

## References

Malihe Alikhani and Matthew Stone. 2020. Achieving common ground in multi-modal dialogue. In *ACL*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *ACL*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *EMNLP*.

Joyce Y. Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *IJCAI*.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W. Black. 2021. Grounding 'grounding' in nlp. In *Findings of ACL*.

Muhao Chen, Weijia Shi, Ben Zhou, and Dan Roth. 2021. Cross-lingual entity alignment with incidental supervision. In *EACL*.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton & Co.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition*, pages 127–149.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *ACL*.

Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Chai. 2016. Physical causality of action verbs in grounded language understanding. In *ACL*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing (3rd ed.)*. Prentice Hall.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Noriyuki Kojima, Hadar Averbuch-Elor, Alexander Rush, and Yoav Artzi. 2020. What is learned in visually grounded neural syntax acquisition. In *ACL*.

Parisa Kordjamshidi, Qiang Ning, James Pustejovsky, and Marie-Francine Moens. 2024. Spatial and temporal language understanding: Representation, reasoning, and grounding. In *NAACL*.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2017. Spatial role labeling annotation scheme. *Handbook of linguistic annotation*, pages 1025–1052.

Nikhil Krishnaswamy and James Pustejovsky. 2022. Grounding meaning representation for situated reasoning. In *AACL*.

Kemal Kurniawan, Lea Frermann, Philip Schulz, and Trevor Cohn. 2021. PPT: Parsimonious parser transfer for unsupervised cross-lingual adaptation. In *EACL*.

Cheng-I Jeff Lai, Freda Shi, Puyuan Peng, Yoon Kim, Kevin Gimpel, Shiyu Chang, Yung-Sung Chuang, Saurabhchand Bhati, David Cox, David Harwath, Yang Zhang, Karen Livescu, and James Glass. 2023. Audio-visual neural syntax acquisition. In *ASRU*.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *ICML*.

Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. 2023. Computational language acquisition with theory of mind. In *ICLR*.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *ACL*.

Ziqiao Ma, Jiayi Pan, and Joyce Chai. 2023. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *ACL*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *ICLR*.

Jiayuan Mao, Freda Shi, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. 2021. Grammar-based grounded lexicon learning. In *NeurIPS*.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. Spartqa: A textual question answering benchmark for spatial reasoning. In *NAACL-HLT*.

Jiefu Ou, Benno Krojer, and Daniel Fried. 2023. Pragmatic inference with a clip listener for contrastive captioning. In *Findings of ACL*.

Puyuan Peng and David Harwath. 2022. Word discovery in visually grounded, self-supervised speech models. In *Interspeech*.

Eva Portelance, Siva Reddy, and Timothy J O'Donnell. 2024. Reframing linguistic bootstrapping as joint inference using visually-grounded grammar induction models. *arXiv:2406.11977*.

Shaolin Qu and Joyce Yue Chai. 2010. Context-based word acquisition for situated dialogue in a virtual world. *Journal of Artificial Intelligence Research*, 37:247–277.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. "wikily" supervised neural translation tailored to cross-lingual tasks. In *EMNLP*.

Deb K Roy and Alex P Pentland. 2002. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146.

Shane Settle, Kartik Audhkhasi, Karen Livescu, and Michael Picheny. 2019. Acoustically grounded word embeddings for improved acoustics-to-word speech recognition. In *ICASSP*.

Lanbo She and Joyce Chai. 2016. Incremental acquisition of verb hypothesis space towards physical world interaction. In *ACL*.

Lanbo She and Joyce Chai. 2017. Interactive learning of grounded verb semantics towards human-robot communication. In *ACL*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *ICLR*.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *ACL*.

Haoyue Freda Shi. 2024. *Learning Language Structures through Grounding*. Ph.D. thesis, Toyota Technological Institute at Chicago.

Jeffrey Mark Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91.

Mark Steedman. 2000. *The Syntactic Process*. Cambridge, MA: MIT Press.

Luc Steels. 2003. Evolving grounded communication for robots. *Trends in cognitive sciences*, 7(7):308–312.

Luc Steels, Frédéric Kaplan, Angus Mcintyre, and Joris Van Looveren. 2000. Crucial factors in the origins of word meaning. In *In: Proceedings of the 3rd Conference on the Evolution of Language.*

Ruisi Su, Shruti Rijhwani, Hao Zhu, Junxian He, Xinyu Wang, Yonatan Bisk, and Graham Neubig. 2021. Dependency induction through the lens of visual perception. In *CoNLL*.

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

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

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *EMNLP*.

Lucien Tesnière. 1959. *Éléments de Syntaxe Structurale*. Librairie C. Klincksieck, Paris.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *NeurIPS*.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *ACL*.

Fei Xu and Joshua B Tenenbaum. 2007. Word learning as bayesian inference. *Psychological review*, 114(2):245.

Lili Yu, Howard Chen, Sida I. Wang, Tao Lei, and Yoav Artzi. 2020. Interactive classification by asking informative questions. In *ACL*.

Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. 2021. Video-aided unsupervised grammar induction. In *NAACL-HLT*.

Yanpeng Zhao and Ivan Titov. 2020. Visually grounded compound PCFGs. In *EMNLP*.

Hao Zhu, Yonatan Bisk, and Graham Neubig. 2022. Language learning from communicative goals and linguistic input. In *CogSci*.