Grounding Spoken Language

Cynthia Matuszek

July 14, 2023

1 Introduction

When we imagine interacting with robots in human environments, we imagine speech and language as a core modality. Human-robot interaction is a key area of study for using robots in human environments such as schools, homes, and care facilities; in order to usefully engage with people in such settings, robots will need to be able to gracefully interact with the people around them, and natural language represents an intuitive, comfortable mechanism for such interactions. This chapter will discuss the role of natural language processing (NLP) in modern robotics and human-robot interaction, and specifically, how grounded language learning is a critical modality for robots to understand the world. While it is crucial to study language from an auditory perspective, understanding the underlying semantics—the linguistic meaning and intent of those speech acts—is necessary for smooth interaction with robots in human-centric environments.

Robots can use language as a mechanism of interaction, but also as a tool for learning about the world around them. They can follow instructions [79, 32], respond to interactions using language (for example, by acknowledging commands or seeking clarification) [53], and use language to repair or reshape interactions [18, 28]. Frequently, these interactions are scripted: the human has a fixed set of possible commands, which may be provided, or which they may need to learn from experimentation. However, for truly flexible agents, it is preferable to learn from interactions what words and instructions mean. Grounded language learning is specifically the process of learning language from interactions with the world, and learning about the world from language used to describe it [87]. The core idea is that treating language learning as a problem grounded in the physical world via robotic agents will improve the effectiveness of both robotic interaction and natural language understanding [66].

Speech is already a key mechanism for embodied devices such as home assistants, phones, and even game consoles. While automated speech recognition (ASR) has become a relatively mainstream technology and continues to improve, there are still substantial difficulties with using those technologies for robotics [62], including environmental noise, latency, and a lack of datasets and models specific to real-world environment interaction. Some of the complexities of managing speech in robotics have been discussed in other chapters. However, even when the difficulties of accessing speech are disregarded, there are substantial NLP-based challenges with using language as it pertains to robots.

In this chapter, language is considered as a mechanism of referencing concepts in the physical world, and how natural language processing dovetails with work on using speech and making sense of language in a robot's environment is explored. Grounded language acquisition as a research area is examined, and a case study is presented of learning grounded language from speech by examining the question from three distinct-but-related angles: the need for complex perceptual data (and a resulting corpus), the need to learn to interact directly from speech without using a textual intermediary, and the problem of learning grounded language from richly multimodal data.

2 Grounded Language Acquisition

Language is comprised of symbols, and understanding those symbols and their underlying meaning—the *symbol grounding problem*—is a core aspect of artificial intelligence as a field [36, 14]. The fundamental idea underlying grounded (or embodied) language is that language does not exist in a vacuum: it derives from, and refers to, objects and actions in the physical environment in which robots operate [87]. Accordingly, this language can be learned by connecting co-occurrences of language with physical percepts perceived by a robot. While a substantial body of this work is related to what is frequently referred to as "Vision-and-Language," in practice, this terminology often refers to tasks such as Visual Question Answering [3] and Visual Commonsense Reasoning [96], where no literal physical agent is necessarily involved. This is a richly studied area, with connections to language modeling, automatic speech recognition, human-robot interaction, learning in simulation, and vision-and-language navigation and manipulation, among others. This section provides a necessarily partial overview of recent work.

One way to examine the current state of the art in grounded language learning is to consider the related and overlapping sub-tasks which people use as testbeds. Grounded language can refer to language about a wide range of aspects of the environment: objects and their attributes [16, 75], actions and tasks [41, 57, 20]—notably including vision-and-language navigation, a special case that has been studied since very early in the history of embodied language understanding (surveys: [35, 93, 69])—or referring expression groundings [89, 61], to name a few.

One significant body of physically situated language work revolves around the use of large pretrained vision-and-language models (VLMs). Contrastive language image pretraining (CLIP) encoders [72] have been successfully used for embodied navigation and vision-and-language navigation-related tasks [51, 80] and tabletop manipulation-based instruction following [81]. Other grounded language learning-adjacent works depend on such large language models (LLMs) as BERT [30], including ViLBERT [58] and Embodied BERT [85], which focuses on object-centric navigation in the ALFRED benchmark [82]. In particular, LLMs are frequently used to help derive plans for following natural language instructions for completing tasks [45]. SayCan [1] uses a combination of LLMs to extract possible useful actions in the world given a goal, and uses affordances of a physical robot to determine, of those actions, which are feasible to perform. In [43], an LLM is used to generate steps towards a goal, incorporating perceptual feedback about the environment in order to improve long-horizon planning.

Other approaches focus on the use of smaller but more task-specific knowledge bases. In [16], few-shot learning of object groundings is accomplished by adding to a database of examples of simulated objects overlaid on real environments; [68] learn to interpret task instructions by probabilistically connecting instructions to background knowledge found in part in relational and taxonomic databases. In the class of interactive language grounding in robotics, a physical agent can follow instructions interactively [59] or can learn to improve its performance on a task based on communication with a person (e.g., [20]). Other work has focused on collections of instructions that pertain to a specific environment [4, 63] and do not incorporate non-perceptual background knowledge.

Despite this preponderance of language-based approaches, the space of robots learning about the world via actual speech is comparatively nascent. Despite early work in learning the grounded semantics of spoken utterances [95, 76], most recent work on language grounding has focused on textual content, generally obtained via crowdsourcing [4, 70] or from web-scale data such as image captions or tags [54]. In some cases, text is transcribed from spoken language, either using automatic speech recognition (ASR) [9] or manually [90]. However, when interacting with embodied agents such as robots, speech is a more appropriate modality than typed text. This leads to the core technical aspect of this chapter: a discussion of grounding language via speech.

3 Learning Grounded Language about Objects

This section describes a case study of attempting to understand unconstrained spoken language about objects. While object recognition based on vision is an extremely active area of research in both 2D and 3D contexts [7, 71], using grounded language with robots in human environments introduces additional problems. First, although large pre-trained visual models contain extensive coverage of some object classes, they suffer from long tail problems when encountering rare objects in the environment or unusual exemplars of common objects [84]. Second, language about grounded concepts frequently evolves over the course of an interaction: people create new terminology and repurpose terminology on the fly during interaction [17, 44], meaning that a robot may need to learn new and remapped terms in real time as interactions unfold. Finally, existing large vision-and-language models tend to be Western-centric [78, 5], potentially limiting the usefulness of deployed systems in other cultural settings.

Learning to understand unconstrained language about objects may entail learning class names, but may also require learning to understand a variety of perceptually meaningful descriptors—for example, people may choose to describe objects based on color and shape, or on the material they are made from, e.g., ceramic or aluminum [75]. It is therefore necessary to learn the semantics of a variety of grounded terms above and beyond simple object names. This necessitates addressing a number of subproblems, including not only the collection of an appropriate dataset in order to benchmark the success of our efforts, but also the development of mechanisms for learning from speech without a textual intermediary and learning from rich multimodal sensor data.

3.1 A Dataset for Multimodal Language Understanding

While there has been extensive work on understanding language using large-scale pretrained natural language models such as BERT [30], in practice, grounding requires that such language-focused models be augmented with perceptual data from robotic sensors [12], typically in the form of visual data (survey: [26]). In our use case of allowing people to use speech directly with robots, there are two significant difficulties with this approach: First, robots tend to have perceptual capabilities beyond vision, such as depth perception (potentially along with more exotic sensory capabilities such as thermal sensing); and second, HRI contexts provide a comparatively small amount of spoken training data, as distinct from the very large amounts of textual data that are available (for example, in the form of image captions).

Nonetheless, when interacting with robots and other embodied agents, it is natural for people to want to speak to them, and many deployed systems use speech [88]. Spoken language is critical for interactions in physical contexts, despite the inherent difficulties: spoken sentences tend to be less well framed than written text, with more disfluencies and grammatical flaws [74]. However, despite these differences, text is commonly used for grounded language learning, presumably due to its wide availability and comparative ease of computational processing (with a few exceptions, *inter alia* [10, 34, 48]). In this section, the development of the Grounded Language Dataset (GoLD)[1] is described. GoLD is a dataset of object descriptions for speech-based grounded language learning [49]. This dataset contains visual- and depth-based images of objects aligned with spoken and written descriptions of those objects col-

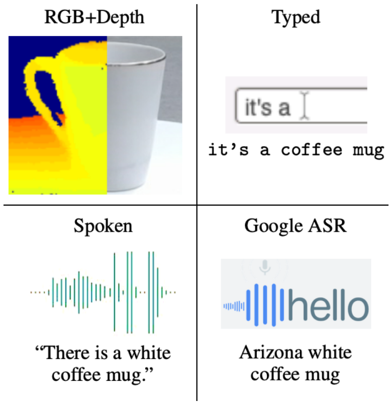

Figure 1: The GoLD dataset has RGB and depth point cloud images of 207 objects in 47 categories. It includes 8250 text and 4045 speech descriptions; all spoken descriptions include automatic transcriptions.

[1] https://github.com/iral-lab/gold

lected from Amazon Mechanical Turk.

GoLD contains perceptual and linguistic data in five high-level categories: *food*, *household*, *medical*, *office*, and *tools*. In these groups, 47 object classes contain 207 individual object instances. These categories were selected to represent objects that might be found in common human-centric environments such as homes and offices, and contain multiple examples of objects that are typical for such contexts, such as apples, dishes, analgesics, staplers, and pliers. For each of these objects, approximately 450 vision+depth images were collected as the object was rotated on a turntable in front of a Kinect RGB+D camera. For each object, four images from distinct angles were chosen to represent object 'keyframes.' Using rotational visual data helps to avoid a known problem with many image datasets, namely their tendency to show pictures of objects from a limited set of angles [8, 91, 89], whereas a robot in a home might see an object from any angle.

Three distinct types of language data were collected/created for each object in this dataset. For each keyframe, approximately twenty spoken descriptions were collected, leading to a dataset of 16,500 spoken descriptions. For comparison purposes, transcriptions of these descriptions were generated using Google Speech to Text. Manual evaluation of a subset of these transcriptions suggests that approximately 80% were good enough for grounded language understanding. Another 16,500 textual descriptions were separately collected; these were not associated with the spoken descriptions or provided by the same set of Mechanical Turk workers (although some workers did work on both problems, they were not given aligned examples to label). The types of data present in GoLD are shown in fig. 1.

GoLD fits into a landscape of datasets used for grounded or embodied language learning, and extends that landscape by providing a very rich dataset, in which each object is associated with a large number of images, depth images, and spoken descriptions. While there are many spoken-language datasets in existence, they are frequently handcrafted for the specific task that the research seeks to accomplish, often leading to narrower applications, for example, question answering [47]. Meanwhile, recent large-scale datasets that include speech typically incorporate synthetically generated speech [38, 27, 39, 29], use generated spoken descriptions from the text captions [56, 94], or ask crowdsourced workers to read captions [37, 42]. This may remove agrammatical constructs, disfluencies, and speech repair, effectively gating the complexities of speech through written language. Other larger datasets exist that contain (real or virtual) scenarios in which embodied vision+depth sensor data can be extracted, such as the ALFRED benchmark [83]; however, aligned, unconstrained spoken language is rarely included.

3.2 Learning Multimodal Groundings from Spoken Language

Grounded language learning offers a way for robots to learn about dynamic environments directly from individual end users. However, as described above, much of the current work in this area uses text as the linguistic input, rather than speech. Grounded language learning that does incorporate speech frequently relies on automatic speech recognition (ASR). Off-the-shelf ASR systems often have substantial drawbacks when used in robotic settings [62]: they introduce latency to the system, work poorly in noisy environments (including the noise produced by the robot itself), and cannot use perceptual information about the environment to improve on recognition. In addition, current ASR systems work inconsistently across demographics such as gender, race, and native language [2, 13, 33, 86, 92], which can lead to failures of inclusive design.

Despite this gap, in many cases, the learning methods applied to acquiring grounded language are relatively agnostic to the type of input, relying on broad approaches such as manifold alignment [67]. As a result, text can be replaced with appropriately featurized speech as an input to the joint language learning model [50]. A number of existing pre-trained speech representation models are available to encode speech into appropriate featurizations, making it possible to encode the spoken descriptions in GoLD and treat those encodings as input to a combined language learning model. This section describes a learning method that performs object grounding directly from speech, without relying on a textual intermediary; that work is then extended to a model that is capable of handling complex, multimodal input, even in cases where some sensory data becomes unavailable.

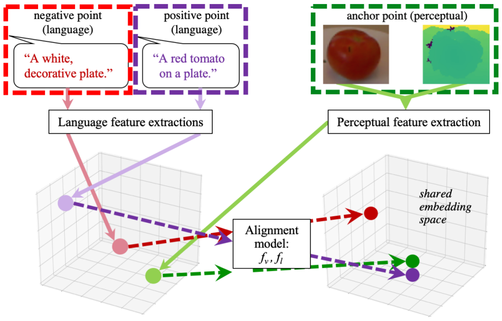
negative point
(language)
"A white, decorative plate."
positive point
(language)
"A red tomato on a plate."
anchor point (perceptual)
Language feature extractions
Perceptual feature extraction
Alignment model: f_v, f_l
shared embedding space

Figure 2: Triplet loss tries to minimize the distance between an anchor point, here the perceptual inputs of an object (boxed in green); an associated sample in another modality, here the language describing that object (boxed in purple); while maximizing the distance between the anchor point and a negative point in the other modality, such as the description of an unrelated thing (boxed in red). The learned embedding then should embed language "near" the things it describes and "far" from other things.

In order to learn from the data in GoLD, it is first necessary to featurize the disparate data types. Different featurizations are used for each of the vision, depth, and spoken language modalities. Visual features (image and depth) are extracted using ResNet152 pre-trained on ImageNet [40], and the last fully connected layer is removed to obtain 2048-dimensional features. RGB images are processed directly, while depth images are colorized before processing [75]. For the simple manifold alignment-based learning case, these vectors are concatenated to make a single visual vector. Speech is featurized using wav2vec 2.0 [6], in which audio is encoded via a convolutional neural network, then masked spans of the resulting speech representations in the latent space are input to a transformer network. Wav2vec 2.0 uses a two-stage training process: the first stage of training focuses on learning local patterns in the audio, such as phonemes, while the second stage focuses on learning patterns such as sentence structure. In our case, a pre-trained model was used, which was fine-tuned for speech recognition.

Manifold alignment: In the simple case, language groundings are learned from these encodings using a procedure known as manifold alignment, where featurized language and sensor data are treated as projections of some underlying manifold in a shared, non-observable latent space, and the goal is to find the functions that approximate the manifold. In this approach, groundings are learned by attempting to capture a manifold between speech and perceptual inputs. The goal is to find functions that make projections from both domains 'closer,' in feature space, to other projections of the same class.

For textual language, the approach used is triplet loss, a geometric approach that has shown success in learning metric embeddings [31]. Triplet loss is a form of contrastive learning, which learns representations of data by comparing points to representationally similar or dissimilar points. The goal is to learn an embedding which 'pulls' similar points closer together in the feature space and 'pushes' dissimilar points further apart. Our manifold alignment learning uses triplets of the form (a, p, n), where a is an 'anchor' point, p is a positive instance of the same class as the anchor (for example, tomato), and n is a negative instance from a different class (for example, plate). The goal is to maximize for each triplet the distance between a and n while minimizing the distance between a and p. This is achieved through the loss function $\mathcal{L} = \max(0, d(f(a) - f(p)) - d(f(a) - f(n)) + \alpha)$, where f is the relevant embedding function for the input domain, d is a distance metric, and α is a margin imposed between positive and negative instances. Given our multimodal data, the embedding function f is the encoder that projects instances of a given modality into the shared manifold (see fig. 2 for an example).

This approach to learning language groundings directly from speech has proven to be quite successful [50] on a downstream object retrieval task, in which the system is presented with a query where a robot must choose an object to retrieve from a set of alternatives (where the queries are spoken natural language inputs such as "The black and gold University mug"). This speech-based model outperforms text-based approaches to the same retrieval task, despite the fact that the model was initially designed for textual input [67].

Multimodal learning: As mentioned, although vision is a key way of perceiving the environment, it is not the only sensor available to robot platforms. The approach described above handles depth by concatenating the depth image to a scaled RGB image, and cannot incorporate additional modalities. Furthermore, it is still able to handle only a single communication modality (speech *or* text). However, people may wish to communicate in a multimodal fashion, in which case multiple interactive modalities should be taken into account simultaneously: while speech is an obvious mechanism for embodied interaction, there are cases when it is preferable to convey complex commands from a computer or via an interface on a phone. This section discusses an extension of the contrastive learning described above that begins to address these requirements.

In addition to handling multiple sensory and communication modalities, such a learning mechanism should be robust to missing modalities. A robot's percepts may be only partially available when handling a learning problem—sensors may fail, may be occluded at key moments, or may be missing from certain platforms; people may communicate via speech, gesture, text, or some subset of those. All of these desiderata taken together suggest the need for a broader learning mechanism that is (1) capable of handling arbitrary numbers of modalities, (2) robust in the face of modality dropouts, and (3) able to learn from relatively small-scale, human-provided inputs.

Broadly speaking, our approach is to extend the idea of geometric loss by combining it with a cross-entropy based supervised contrastive loss function [52], in which labels are used to allow points belonging to the same class to be pulled to the same area in embedding space, while points belonging to other classes are pushed apart. It is a general version of multiple contrastive loss functions including triplet loss, as well as general contrastive loss [24]. A distance-based loss function is defined that can be used for an arbitrary number of modalities.

Standard triplet loss, as described above, can be applied to only two modalities, and is not robust to sensor ablation. To address these issues, pairwise distance optimization is used for all data points. During training, two different instances are sampled and their corresponding representations from all modalities are split into two sets—one positive set (referring to a specific object) and one negative set (referring to some randomly-selected different object). In our setting, every item in the positive set becomes an anchor once, and the distance is minimized between that item and other items in the positive set, while minimizing the distance between that item and all items in the negative set. This can be seen as an one-to-many relationship instead of the one-to-two relationship in the triplet loss formulation.

This approach is tested over the four main modalities of the GoLD dataset: RGB, depth, speech, and text. Encoding mechanisms appropriate to each modality are selected. BERT [30] embeddings are used to featurize textual input, and wav2vec2 [6] to extract audio embeddings from speech. To process images, ResNet152 [40] is used for both RGB and depth images, producing a 2048-dimensional embedding vector. The objective is then to first minimize the distance between each pair of positive points from heterogeneous modalities, and second, maximize the distance between each pair of positive and negative points from all modalities. This combined loss function results in a learning mechanism that outperforms supervised contrastive loss in both the speed of convergence during training, and number of data points required to build a model capable of performing a downstream object retrieval task.

4 Open Challenges

Although learning to understand and learn from grounded language is an active and successful field of research, a number of challenges remain to be addressed. Discussing open questions in a fast-moving field such as language grounding carries an element of risk. There has recently been a surge of rapid development in applications of NL technology and robotics, enabled by new technologies and very large-scale data sources, that would have been difficult to predict a small number of years ago. Nonetheless, and despite this promising uptick in progress, there remain significant barriers to deploying robots that understand, learn from, and interact using language in a physical context. In this section, some open challenges are briefly discussed, as well as some characteristics problems may have that make them difficult to address using currently popular approaches.

Some of the problem characteristics of note in this space are familiar from machine learning and robotics more generally, although grounded language offers its own unique difficulties in solving those problems. Some of these include *scalability*, or how learned models of grounded language can scale to a wide range of objects, tasks, and modalities; *generalization*, how such models can generalize to new examples of learned concepts and generalize across different robot platforms, including via few-shot and zero-shot learning; *multi-modality*, how robots using multiple complex sensors can interact with people using a variety of communication modalities; and *common sense reasoning*, in which systems can use an understanding of the broader world to solve otherwise under-specified problems.

First, despite the progress described above, there remain substantial problems involved in using actual speech with robots. [62] provides an overview of these difficulties, sorted into eight categories. These categories cover human-robot interaction questions (such as improving the modeling of social components of language), systems-level questions (such as the timing and latency difficulties of performing speech-based interaction in real time and developing improved learning models), and infrastructure-level suggestions for improving the context in which speech for robotics is studied. Despite the progress described above on using speech directly, challenges such as disambiguating speech in noisy environments remain.

A broad class of problems in this space includes developing models that can learn from a small amount of data or in unsupervised settings. While there is extensive work on learning from a small number of examples based on pre-training [77, 60, 55], the complexity of human spaces and robotic sensing make performing few-shot learning in idiosyncratic real-world settings a distinct challenge. There is a long tail of potentially out-of-distribution objects that may be encountered, sensors may give partial information, and people interacting with a robot will be understandably reluctant to provide a significant number of training examples. This ties into another difficulty, that of dealing with low-resource settings. For example, while there are a tremendous array of resources available for English and a few other major languages, the same is not always true of smaller languages or dialects of the sort that may be spoken in human-centric settings, or of idiosyncratic or ambiguous language.

Given their current popularity and effectiveness, it is particularly worth discussing the strengths and drawbacks of applying large language models and large vision-and-language models to grounded language. As described above, LLMs have demonstrated tremendous success on a wide variety of NL applications, including some language grounding problems. Nonetheless, while they are broadly good at producing output that seems correct at first glance, they do not necessarily fully grasp the semantics of complex grounded language [12]—such models have been trained on large amounts of (typically) textual data, but lack data to understand and reason about the physical world, making it difficult for them to understand contextual language about physical settings. These models have shown some success in planning tasks where the goal is to follow high-level textual instructions (see section 2), but even these success stories may not generalize well to handling grounded language across domains or environments. LLMs and VLMs currently also have limited ability to handle multimodal sensor inputs of the sort that may occur in robotics settings, including auditory data.

Like many machine learning models, current approaches to grounded language learning tend to struggle with common sense reasoning [97], in which general, domain-agnostic background knowledge is key to understanding utterances. As an example, one description of an object might be "This is an apple, it's a kind of fruit." A robot with a good grounding system may learn the name and be able to identify apples subsequently, but will not be able to conclude that it is edible, or that it is similar to a banana. Another example has to do with a robot that has learned to hand someone a plate upon being instructed to do so, but would not from such a request conclude that the person is hungry or likely to engage in activities such as eating or setting the table. Efforts to combine common sense and language grounding exist [23, 15], but true common sense remains an elusive goal, as indeed it does in artificial intelligence generally.

There are also ethical questions and questions of bias and fairness associated with this problem area. There is a robust ongoing discussion in the machine learning community about discrimination and representation in machine learning technology, and the role of equitable development paradigms in addition to the deep-seated biases found in large data corpora (*inter alia*, [25, 22, 64]). These questions are highly relevant to the problem of making robots that learn from end users about their environment; a deployed system that works unevenly across different user demographics is inherently problematic, even if the system's average success is high. In considering this, it is necessary to bear in mind that discriminatory performance from machine learning models is not solely a product of unbalanced data [21, 11]. Model designs [73, 65], representational encoding choices [19], data collection methods [46], and learning paradigms all affect the inclusiveness of not only the results of machine learning, but the selection of the core questions being asked.

5 Conclusion

This chapter has discussed grounded language acquisition as a field where robotics and natural language understanding come together, and has discussed how learning to understand speech about the world plays a substantial role in human-robot interaction. A sampling of current work in this space has been described, with an emphasis on the challenges involved in going from understanding textual language to spoken language and in handling rich multimodal perception and communication. The chapter closes with an overview of some of the many outstanding challenges in the general space of understanding grounded language, including dealing with speech, learning-derived problems such as generalization, and classical artificial intelligence problems such as incorporating common sense reasoning.

Language is not synonymous with sound: speech is a carrier for linguistic content, but only one of several mechanisms by which communicative content can be conveyed. Nevertheless, speech is an obvious, intuitive mechanism for human-robot interaction, tightly coupled with questions of language understanding and understanding the world from complex perceptual context. Grounded language understanding, particularly from speech, represents a rich, promising research space that is tightly interwoven with questions of sound in a robotic environment. There is extensive work in this area and in the related areas of spoken language processing and human-robot interaction, and this chapter attempts to provide an overview of some of the ways in which these elements come together.

Acknowledgments

We would like to extend heartfelt thanks to Francis Ferraro and Luke E. Richards for their assistance in preparing this manuscript. The case study described in this chapter is based on the work of a number of colleagues in the Interactive Robotics and Language (IRAL) lab, especially including Gaoussou Youssouf Kebe and Kasra Darvish, as well as Luke E. Richards, Padraig Higgins, Patrick Jenkins, Rishabh Sachdeva, Ryan Barron, John Winder, Don Engel, and Edward Raff. This material is based in part upon work supported by the National Science Foundation under Grant Nos. 1813223, 1920079, 1940931, 2024878, and 2145642.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, 2022.

[2] Eiman Alsharhan and Allan Ramsay. Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition. *Language Resources and Evaluation*, 54(4):975–998, 2020.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[4] Dilip Arumugam, Siddharth Karamcheti, Nakul Gopalan, Lawson LS Wong, and Stefanie Tellex. Accurately and efficiently interpreting human-robot instructions of varying granularities. *arXiv preprint arXiv:1704.06616*, 2017.

[5] James Atwood, Yoni Halpern, Pallavi Baljekar, Eric Breck, D. Sculley, Pavel Ostyakov, Sergey I. Nikolenko, Igor Ivanov, Roman Solovyev, Weimin Wang, and Miha Skalic. The inclusive images competition. In Sergio Escalera and Ralf Herbrich, editors, *The NeurIPS '18 Competition*, pages 155–186, Cham, 2020. Springer International Publishing.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Monika Bansal, Munish Kumar, and Manish Kumar. 2d object recognition techniques: state-of-the-art work. *Archives of Computational Methods in Engineering*, 28:1147–1161, 2021.

[8] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pages 9453–9463, 2019.

[9] Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Luca Iocchi, Roberto Basili, and Daniele Nardi. Huric: a human robot interaction corpus. In *LREC*, pages 4519–4526, 2014.

[10] Emanuele Bastianelli, Danilo Croce, Andrea Vanzo, Roberto Basili, Daniele Nardi, et al. A discriminative approach to grounded spoken language understanding in interactive robotics. In *IJCAI*, pages 2747–2753, 2016.

[11] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[12] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online, July 2020. Association for Computational Linguistics.

[13] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786, 2007.

[14] 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. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, November 2020.

[15] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[16] Valts Blukis, Ross Knepper, and Yoav Artzi. Few-shot object grounding and mapping for natural language robot instruction following. In *Conference on Robot Learning*, pages 1829–1854. PMLR, 2021.

[17] Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482, 1996.

[18] Arthur Bucker, Luis Figueredo, Sami Haddadinl, Ashish Kapoor, Shuang Ma, and Rogerio Bonatti. Reshaping robot trajectories using natural language commands: A study of multi-modal data alignment using transformers. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 978–984, 2022.

[19] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[20] Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9, 2018.

[21] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? How? What to do? *arXiv preprint arXiv:2105.12195*, 2021.

[22] Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts, 2019.

[23] Haonan Chen, Hao Tan, Alan Kuntz, Mohit Bansal, and Ron Alterovitz. Enabling robots to understand incomplete natural language instructions using commonsense reasoning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1963–1969, 2020.

[24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[25] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[26] Grzegorz Chrupała. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*, 73:673–707, 2022.

[27] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[28] Yuchen Cui, Siddharth Karamcheti, Raj Palleti, Nidhya Shivakumar, Percy Liang, and Dorsa Sadigh. "No, to the right" – online language corrections for robotic manipulation via shared autonomy. In *18th ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 2023.

[29] Fabio De Ponte and Sarah Rauchas. Grounding words in visual perceptions: Experiments in spoken language acquisition. In *Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2022)*, 2022.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[31] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *European Conf. on Computer Vision*, September 2018.

[32] Felix Duvallet, Matthew R Walter, Thomas Howard, Sachithra Hemachandra, Jean Oh, Seth Teller, Nicholas Roy, and Anthony Stentz. Inferring maps and behaviors from natural language instructions. In *Experimental*

Robotics: The 14th International Symposium on Experimental Robotics, pages 373–388. Springer, 2016.

[33] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*, 2021.

[34] Michael Fleischman and Deb Roy. Grounded language modeling for automatic speech recognition of sports video. In *Proceedings of ACL-08: HLT*, pages 121–129, 2008.

[35] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Association for Computational Linguistics (ACL)*, 2022.

[36] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1), 1990.

[37] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244, 2015.

[38] William Havard, Laurent Besacier, and Olivier Rosec. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. *GLU 2017 International Workshop on Grounding Language Understanding*, Aug 2017.

[39] William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8618–8622, 2019.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 770–778. IEEE, Jun 2016.

[41] Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, et al. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*, 2017.

[42] Wei-Ning Hsu, David F. Harwath, C. Song, and J. Glass. Text-free image-to-speech synthesis using learned segmental units. *ArXiv*, abs/2012.15454, 2020.

[43] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*, 2022.

[44] Alyssa Ibarra and Michael K Tanenhaus. The flexibility of conceptual pacts: Referring expressions dynamically shift to accommodate new conceptualizations. *Frontiers in psychology*, 7:561, 2016.

[45] Peter Jansen. Visually-grounded planning without vision: Language models infer detailed plans from high-level instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4412–4417, Online, November 2020. Association for Computational Linguistics.

[46] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316, 2020.

[47] Johanna E. Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *CVPR*, pages 1988–1997, 2016.

[48] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. Visually grounded learning of keyword prediction from untranscribed speech. *arXiv preprint arXiv:1703.08136*, 2017.

[49] Gaoussou Youssouf Kebe, Padraig Higgins, Patrick Jenkins, Kasra Darvish, Rishabh Sachdeva, Ryan Barron, John Winder, Don Engel, Edward Raff, Franics Ferraro, and Cynthia Matuszek. A spoken language dataset of descriptions for speech-based grounded language learning. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2021.

[50] Gaoussou Youssouf Kebe, Luke E. Richards, Edward Raff, Francis Ferraro, and Cynthia Matuszek. Bridging the gap: Using deep acoustic representations to learn grounded language from percepts and raw speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2022.

[51] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14829–14838, June 2022.

[52] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020.

[53] Nicholas H Kirk, Daniel Nyga, and Michael Beetz. Controlled natural languages for language generation in artificial cognition. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6667–6672. IEEE, 2014.

[54] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 259–266. IEEE, 2010.

[55] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[56] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[57] Peter Lindes, Aaron Mininger, James R Kirk, and John E Laird. Grounding language for interactive task learning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 1–9, 2017.

[58] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[59] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *arXiv preprint arXiv:2210.06407*, 2022.

[60] Parsa Mahmoudieh, Deepak Pathak, and Trevor Darrell. Zero-shot reward specification via grounded natural language. In *International Conference on Machine Learning*, pages 14743–14752. PMLR, 2022.

[61] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[62] Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David Traum, and Zhou Yu. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, 71:101255, 2022.

[63] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423*, 2012.

[64] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[65] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[66] Raymond J. Mooney. Learning to connect language and perception. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1598–1601. AAAI Press, 2008.

[67] Andre T. Nguyen, Luke E. Richards, Gaoussou Youssouf Kebe, Edward Raff, Kasra Darvish, Frank Ferraro, and Cynthia Matuszek. Practical cross-modal manifold alignment for robotic grounded language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1613–1622, June 2021.

[68] Daniel Nyga, Subhro Roy, Rohan Paul, Daehyung Park, Mihai Pomarlan, Michael Beetz, and Nicholas Roy. Grounding robot plans from natural language instructions with incomplete world knowledge. In *Conference on Robot Learning*, pages 714–723, 2018.

[69] Sang-Min Park and Young-Gab Kim. Visual language navigation: A survey and open challenges. *Artificial Intelligence Review*, pages 1–63, 2022.

[70] Nisha Pillai and Cynthia Matuszek. Unsupervised selection of negative examples for grounded language learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2022.

[71] Shaohua Qi, Xin Ning, Guowei Yang, Liping Zhang, Peng Long, Weiwei Cai, and Weijun Li. Review of multi-view 3d object recognition methods based on deep learning. *Displays*, 69:102053, 2021.

[72] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021.

[73] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining

an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44, 2020.

[74] Gisela Redeker. On differences between spoken and written language. *Discourse processes*, 7(1):43–55, 1984.

[75] Luke E. Richards, Kasra Darvish, and Cynthia Matuszek. Learning Object Attributes with Category-Free Grounded Language from Deep Featurization. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8400–8407, October 2020.

[76] Deb Roy. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 5(2):197–209, 2003.

[77] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4694–4703, 2019.

[78] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[79] Lanbo She, Yu Cheng, Joyce Y Chai, Yunyi Jia, Shaohua Yang, and Ning Xi. Teaching robots new actions through natural language instructions. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 868–873. IEEE, 2014.

[80] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[81] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[82] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.

[83] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[84] Gabriel Skantze and Bram Willemsen. Collie: Continual learning of language grounding from language-image embeddings. *Journal of Artificial Intelligence Research*, 74:1201–1223, 2022.

[85] Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav Sukhatme. Embodied BERT: A transformer model for embodied, language-guided visual task completion. In *Novel Ideas in Learning-to-Learn through Interaction (NILLI) Workshop @ EMNLP*, 2021.

[86] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *ACL Workshop on Ethics in Natural Language Processing*, 2017.

[87] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.

[88] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond Mooney. Jointly improving parsing and perception for natural language commands through human-robot dialog. *Journal of Artificial Intelligence Research*, 67:327–374, 2020.

[89] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. In *Conference on Robot Learning*, pages 1691–1701. PMLR, 2022.

[90] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning multi-modal grounded linguistic semantics by playing "i spy". In *IJCAI*, pages 3477–3483, 2016.

[91] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[92] Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain. Automatic speech recognition of multiple accented english data. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[93] Wansen Wu, Tao Chang, and Xinmeng Li. Visual-and-language navigation: A survey and taxonomy. *arXiv preprint arXiv:2108.11544*, 2021.

[94] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[95] Chen Yu and Dana H Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)*, 1(1):57–80, 2004.

[96] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[97] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.