

Presentation and Analysis of a Multimodal Dataset for Grounded Language Learning

Patrick Jenkins, Rishabh Sachdeva, Gaoussou Youssouf Kebe, Padraig Higgins, Kasra Darvish, Edward Raff,¹ Don Engel, John Winder,² Francis Ferraro, Cynthia Matuszek

University of Maryland, Baltimore County, ¹Booz Allen Hamilton, ²Johns Hopkins Applied Physics Laboratory
1000 Hilltop Circle, Baltimore, MD, 21250

pjenk1,rishabs1,mb88814,phiggin1,kasradarvish,edraff1,donengel,jwinder1,ferraro,cmat@umbc.edu

Abstract

Grounded language acquisition — learning how language-based interactions refer to the world around them — is a major area of research in robotics, NLP, and HCI. In practice the data used for learning consists almost entirely of textual descriptions, which tend to be cleaner, clearer, and more grammatical than actual human interactions. In this work, we present the Grounded Language Dataset (GoLD), a multimodal dataset of common household objects described by people using either spoken or written language. We analyze the differences and present an experiment showing how the different modalities affect language learning from human input. This will enable researchers studying the intersection of robotics, NLP, and HCI to better investigate how the multiple modalities of image, text, and speech interact, as well as how differences in the vernacular of these modalities impact results.

1 Introduction

Grounded language acquisition is the process of learning language as it relates to the world—how concepts in language refer to objects, tasks, and environments (Mooney 2008). *Embodied* language learning specifically is a significant field of research in NLP, machine learning, and robotics. There are many ways in which robots learn grounded language (Chen and Mooney 2008; Branavan et al. 2009; Tellex et al. 2011; Thomason et al. 2015; 2018; Matuszek 2018; Anderson et al. 2018; Das et al. 2018; Chevalier-Boisvert et al. 2019; Hu et al. 2019; Vanzo et al. 2020), but they all require either multimodal data or natural language data—usually both.

A significant goal of modern robotics research is the development of robots that can operate in human-centric environments. Examples include domestic service robots (DSRs) that handle common household tasks such as cooking, cleaning, and caretaking (Beckerle et al. 2017), robots for elder care (Bedaf, Gelderblom, and Witte 2014), assistive robotics for providing support to people with disabilities (Chen et al. 2013), and rehabilitation robotics (Kubota et al. 2020). To be useful for non-specialists, such robots

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

	Typed description	Spoken description	Google ASR
	 it's a		 Arizona white coffee mug
	it's a coffee mug	“There is a white coffee mug.”	Arizona white coffee mug

Figure 1: GoLD is comprised of RGB and depth point cloud images of 47 classes of objects in five high-level categories. It includes 8250 text and 4059 speech descriptions gathered with Amazon Mechanical Turk (AMT).

will require easy-to-use interfaces (Beer et al. 2012). Spoken natural language is an appropriate interface for such systems: it is natural, expressive, and widely understood, as shown by the proliferation of natural language-based home devices (Haeb-Umbach et al. 2019). To have a robotic system flexibly understand language in dynamic settings and realize it in physical, goal-oriented behaviors, it is necessary to ground linguistic and perceptual inputs to a learned representation of knowledge tied to actions.

Current approaches to grounded language learning require data in both the perceptual (“grounded”) and linguistic domains. While existing datasets have been used for this purpose (Johnson et al. 2016; Krishna et al. 2017; Das et al. 2018; Nguyen and Daumé III 2019; Thomason et al. 2019), the language component is almost always derived from either textual input or manually transcribed speech (Matuszek et al. 2014; Tellex et al. 2011). In practice, robots are likely to need to operate on imperfectly understood spoken language. To that end, we present the **Grounded Language Dataset (GoLD)**, which contains images of common household objects and their description in multiple formats: text, speech (audio), and automatically recognized speech derived from the audio files. We present experiments that demonstrate the utility of this dataset for grounded language learning.

The primary contributions of this paper are as follows: 1. We provide a freely available, multimodal, multi-labelled dataset of common household objects, with paired image

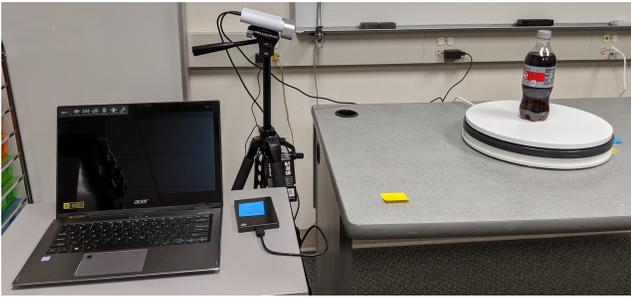


Figure 2: The data collection setup, inspired by Lai et al. (2011). An Azure Kinect (Kinect 3) is mounted on a tripod, pointed at the target object (in this case a soda bottle) on a white turntable. Image and depth data is collected as the object rotates on the turntable.

and depth data and textual and spoken descriptions. 2. We demonstrate this dataset’s utility by analyzing the result of known grounded language acquisition approaches applied to transfer and domain adaptation tasks.

2 The GoLD Dataset

GoLD is a collection of visual and natural language data in five high-level groupings: food, home, medical, office, and tools. These were chosen to reflect and provide data for domains in which dynamic human-robot teaming is a near-term interest area. Perceptual data is collected as both images and depth while natural language is collected in both text and speech. There are 47 object classes spread across these high-level categories, each containing four to five instances of the object for a total of 207 object instances. During imaging, the objects are rotated on a turntable, allowing us to select four representative frames from different angles for a total of 825 views. For example, within the food category there is an apple class with five instances of apples, each with four distinct frames. Complete depth video of the objects from all angles is also available, but only these frames are labeled with multiple natural language descriptions in both text and speech. GoLD is available on GitHub https://github.com/iral-lab/UMBC_GLD.

Accuracy of Speech Transcriptions

Obtaining accurate transcriptions of speech in sometimes noisy environments is a significant obstacle to speech-based interfaces (Li, Tsao, and Sim 2013). In creating GoLD we used the popular Google Speech to Text API, chosen because it is widely available, easy to use, and not tied to a specific domain or hardware setup. However, the resulting transcriptions are therefore not tuned for optimal performance. For a particular use case, a more focused automatic speech recognition (ASR) system could be used on the sound files included in the dataset. In order to understand the degree to which learning is affected by ASR errors, 100 randomly selected transcriptions were evaluated on a 4-point scale (see table 2). These descriptions were also manually transcribed (see table 5 for examples). Of those, 77% are high quality,

Table 1: Classes of objects in GoLD.

Topic	Classes of Objects
food	<i>potato, soda bottle, water bottle, apple, banana, bell pepper, food can, food jar, lemon, lime, onion</i>
home	<i>book, can opener, eye glasses, fork, shampoo, sponge, spoon, toothbrush, toothpaste, bowl, cap, cell phone, coffee mug, hand towel, tissue box, plate</i>
medical	<i>band aid, gauze, medicine bottle, pill cutter, prescription medicine bottle, syringe</i>
office	<i>mouse, pencil, picture frame, scissors, stapler, marker, notebook</i>
tool	<i>Allen wrench, hammer, measuring tape, pliers, screwdriver, lightbulb</i>

Table 2: Human ratings of 100 automatic transcriptions. These ratings are designed strictly to assess the accuracy of the transcription, not the correctness of the spoken description with respect to the described object.

Rating	Transcription Quality Guidelines	#
1	wrong or gibberish / unusable sound file	13
2	slightly wrong (missing keywords / concepts)	10
3	pretty good (main object correctly defined)	14
4	perfect (accurate transcription and no errors)	63

i.e., ‘perfect’ or ‘pretty good,’ while 13% are rated ‘unusable.’

In order to evaluate the replicability of the human-provided ratings in table 2, two subsets of these ratings was evaluated using Fleiss’ kappa (Fleiss 1971) (κ) to measure inter-annotator agreement across three raters. In both trials, $\kappa \approx 0.6$, representing moderate/substantial agreement among the raters. Although a higher agreement would be preferable, we observe that disagreement was never more than one unit between the raters, most commonly two ratings of 1 and one rating of 2. As Fleiss’ kappa does not incorporate concepts such as “near agreement,” for larger datasets, a weighted kappa statistic may be more appropriate.

To get a more detailed understanding of transcription accuracy, we compare the ASR transcriptions and the human-provided transcriptions using the standard NLP metrics of Word Error Rate (WER) and Bilingual Evaluation Understudy (BLEU) score. Word Error Rate is recognized as the *de facto* metric for automatic speech recognition systems, as WER strongly influences the performance of speech (Cavazza 2001; Park et al. 2008; Saon, Ramabhadran, and Zweig 2006). WER measures the minimum-edit distance between the system’s results, the *hypothesis*, and manually transcribed text, the *reference*. WER is typically calculated as the ratio of word insertion, substitution, and deletion errors in a hypothesis to the total number of words in reference (McCowan et al. 2004). We evaluate WER for

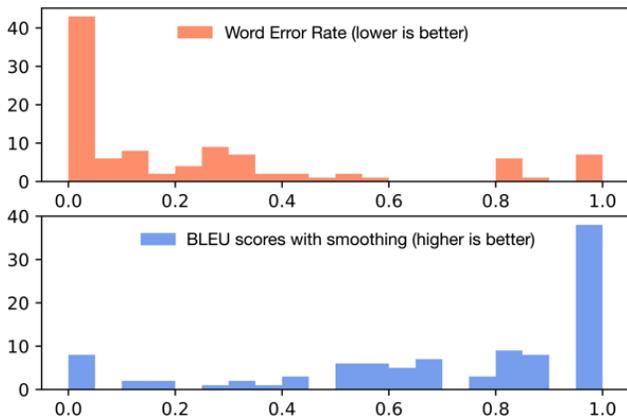


Figure 3: Distribution of Word Error Rate and BLEU scores of automated compared to manual transcriptions. For higher-quality transcriptions (rated 3 or 4), the mean WER drops to 0.09, and the mean BLEU score rises to 0.83.

the same subset, and the mean WER per transcription is approximately 21.3%. Out of 100 transcriptions, 42 had a zero error rate, meaning the reference and hypothesis match exactly.

BLEU is a closeness metric inspired by the WER metric. The main idea is to use a weighted average of variable length phrase matches against the references (Papineni et al. 2002). BLEU scores are widely used to measure the accuracy of language translations based on string similarity; we adopt this system to evaluate the goodness of transcriptions. BLEU scores are calculated by finding n -gram overlaps between machine translation and reference translations. In our analysis, we use $N = 4$.

$$BLEU = \min \left(1, \frac{output - length}{reference - length} \right) \left(\prod_{i=1}^N precision_i \right)^{\frac{1}{N}} \quad (1)$$

To mitigate the tendency of the BLEU score to penalize longer sentences, we apply a smoothing function while calculating scores and add 1 to both numerator and denominator while calculating precision (Lin and Och 2004). We find that the mean BLEU score of the same subset of 100 transcriptions is 0.71, and Figure 3 shows the distribution.

Comparative Analysis

We analyze both the text and speech descriptions for the number of words used as well as mentions of color, shape, and object name. Since both modes are used to interface with robots, we wish to find any similarities or differences that might inform system design techniques for grounded language models. One of the learning targets in grounded language acquisition is to learn attributes of physical objects (as identified by natural language) such as color, shape, and object type (Pillai and Matuszek 2018). These categories are limited because they are expert-specified and prescribed; the GoLD dataset is intended to support learning of unconstrained, “category-free” (Richards and Matuszek 2019)

linguistic concepts. This would allow learning of attribute terms such as “white” or “cylindrical,” but also unexpected concepts such as materials (e.g., “ceramic coffee mug”).

To that end, we analyze the natural language descriptions and find that color, shape, and object names often appear in natural language descriptions of images. We apply a list of 30 common color terms from large language corpora and compared each description to see if it included one of these terms (Mylonas et al. 2015). Similarly, we use a vocabulary list of shape terms to count how many descriptions included shape descriptions. It is worth noting that shape descriptions are less well defined than colors and that a better vocabulary of shape descriptions would be helpful towards this kind of analysis. Finally, we consider how often descriptions contain object labels, which would allow them to be linked to external models.

Our initial hypothesis was that people would use more words when describing objects verbally than when typing, as it is lower effort to talk than to type. When comparing description length, we balance the number of speech and text descriptions, using 4059 of each. However, we found no significant differences in the average length of descriptions between speech and text ($p \geq 0.35$ using a Welch test or t-test) and in the distributions of mentions of color, shape, or object name between the two. While speech has slightly more average words per description, 8.72, compared to text at 8.38, when stop words are removed the averages are 4.52 and 4.38 respectively (see fig. 4).

The larger mean drop in the speech descriptions is likely due to the tendency of ASR systems to interpret noise or murmur utterances as filler words, the inclusion of which has been shown to detract from meaning (Engel 2001; Engel, Charniak, and Johnson 2002; Stolcke and Droppo 2017). Text descriptions are a more consistent length than speech, with a standard deviation of $\sigma_{textlength} = 5.14$ words, versus $\sigma_{speechlength} = 8.0$. When we remove stop words, the standard deviation is 2.58 for text and 3.96 for speech.

As the GoLD dataset contains 47 different classes, it is useful to note the class-wise differences in length of descriptions, in terms of the number of words. Table 3 is a sample of some classes with interesting length differentials. In general, people tend to speak more, compared to what they type, when describing relatively complex objects, like “syringe”, “measuring tape” or “cell phone.” On the other hand, speech descriptions for more basic objects such as “banana”, “spoon”, or “eye glasses” tend to be shorter even than their text descriptions. However, when taken all together the differential between text and speech length per object is not significant with an average of 0.14 more words in speech descriptions.

We use the Stanford Part-of-Speech (POS) Tagger¹ (Toutanova et al. 2003) to count the number of nouns, adjectives, and verbs in the descriptions. We are interested in evaluating these occurrences because they play a central role in defining groundings associated with any object. We find that the mean number of noun tokens per description is

¹<https://nlp.stanford.edu/software/tagger.shtml>

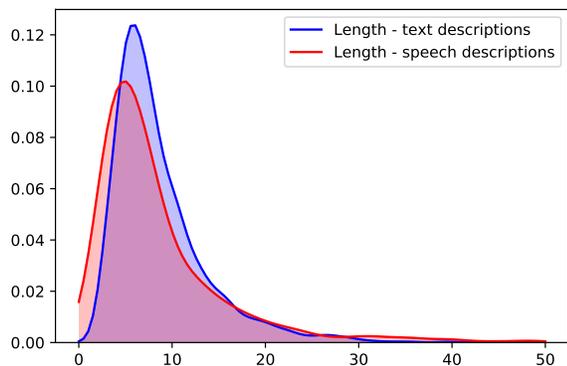


Figure 4: Density Estimate plots of sentence length (the number of words) for the natural language descriptions with stop words removed. We found no significant difference in length of description between text and speech. $\mu_{textlength} = 8.38$, $\sigma_{textlength} = 5.14$; $\mu_{speechlength} = 8.72$, $\sigma_{speechlength} = 8.0$.

slightly higher in the text data (2.59) than the speech data (2.49). Similarly, the average number of adjectives per description is marginally higher for text data (1.25) compared to speech (1.17). Whereas the mean verb occurrence for text and speech are 0.52 and 0.62, respectively demonstrating the reverse trend.

Table 4 shows the top twenty most frequent tokens in both categories. There is substantial overlap, as expected, since the same objects are being described. Words related to color are most commonly used to describe objects. People use more filler words when describing the objects using speech; for example, the word ‘like’ appears 166 times in speech data whereas it was not significant in the text data. The word ‘used’ appears frequently, typically used to describe the functionality of certain objects. Developing grounded language models around functionality for the analysis of affordances in objects is an important research avenue that our dataset enables, which is not possible in prior datasets that do not contain the requisite modalities.

Table 3: Mean length (number of words with stop words removed) in descriptions for selected categories, by description modality.

Category	Length (Text)	Length (Speech)
measuring tape	4.72	6.12
syringe	3.8	4.75
cell phone	4.17	4.98
sponge	4.025	3.925
food can	4.06	3.899
scissors	4.98	4.79
spoon	3.98	3.23
eye glass	4.86	4.03
banana	4.14	3.21

Table 4: Most frequent words in text (left) and speech (right).

Token	% Frequency	Token	% Frequency
black	13.96	black	13.42
object	12.66	white	12.31
white	10.95	red	10.29
blue	10.49	blue	9.87
red	11.76	bottle	8.86
bottle	10.03	yellow	8.37
yellow	9.61	object	6.87
small	6.37	handle	6.62
used	6.19	color	5.93
green	6.01	green	5.56
pair	5.56	used	5.46
plastic	4.53	small	4.67
box	3.90	silver	4.06
silver	3.61	light	3.89
metal	3.30	box	3.81
pink	2.94	pair	3.74
picture	2.54	like	3.59
orange	2.42	plastic	3.57
large	2.53	looks	3.07
jar	2.08	pink	2.58

Related Datasets

Grounded language acquisition is in the unique position of requiring a dataset that combines sensory perception with language. These combined datasets are frequently hand-crafted for the specific task that the research seeks to accomplish (?; Roy 2002), often leading to datasets with narrow applications. For example, CLEVR (Johnson et al. 2016) was designed as a benchmark for question answering tasks. The dataset itself consists of scenes of colored toy blocks arranged in various positions. These scenes are annotated with the color, shape, size, and spatial relation to other objects within the scene. The simplicity of the scene along with the narrow scope of annotations in turn limits the type and complexity of questions that can be asked. As question answering and grounded language systems become more advanced, there is a need for our datasets to reflect real world scenes both in their composition and annotation. GoLD achieves this by including real world objects of varying types and natural language.

A barrier to creating a dataset that includes speech is the high cost of collecting audio or transcribing it into a form that is usable by the intended system. Roy (2002) presents a grounded language system that can generate descriptions for targets within a scene of colored rectangles. The visual data for this task is easily generated, but for the speech descriptions, an undergraduate recorded 518 utterances over three hours. The audio from this collection was then manually transcribed into text. The manual audio transcription task can take anywhere between four and ten hours per hour of audio depending on the quality of the audio being transcribed and the final quality of the transcription (Evers ; 2011; Weizs 2019). We overcome this challenge by utilizing Speech-to-Text technology and evaluating the transcriptions for their quality as described in section 2.

While not a grounded language dataset itself, it should be

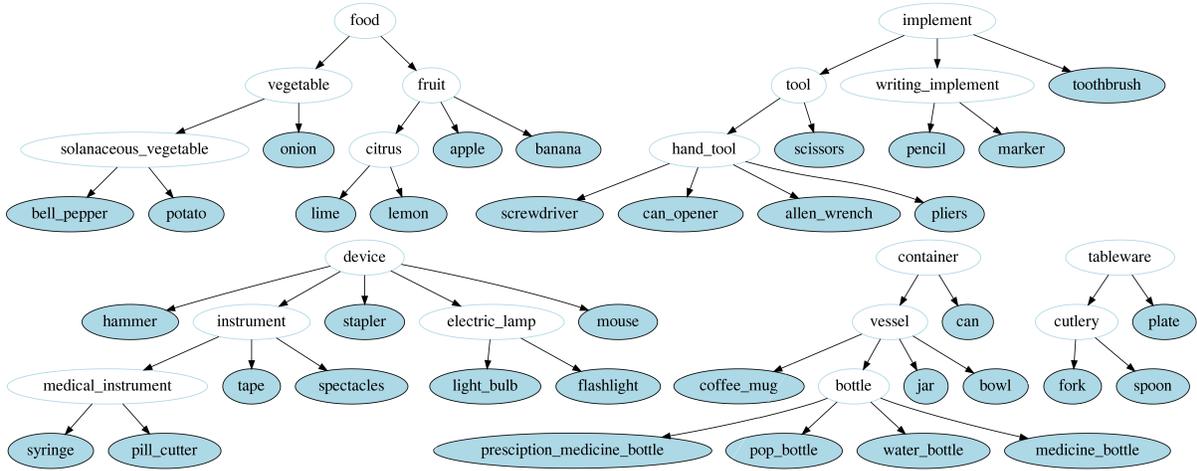


Figure 5: Subtrees highlighting object classes that appear in our dataset (light blue nodes), and the hierarchical structure of their related concepts as derived from WordNet (Miller 1995) (white nodes). These object class subtrees mirror similar category hierarchies reported in ImageNet (Deng et al. 2009) and UW RGB-D (Lai et al. 2011) datasets.

noted that the image collection of this work is heavily influenced by the University of Washington RGB-D dataset (Lai et al. 2011). Both datasets contain large numbers of everyday objects from multiple angles. Our dataset is collected with a now-state of the art sensor which enables us to capture smaller objects at a finer level of detail (such as an Allen key, which is nearly flat against the surface). Additionally, we select objects based on their potential utility for specific human-robot interaction scenarios, such as things a person might find in a medicine cabinet or first aid kit, allowing learning research relevant to eldercare and emergency situations (Beckerle et al. 2017).

3 Dataset Creation

In this section, we discuss the steps involved in collecting images, depth images, speech, and typed descriptions for objects in the GoLD dataset. This includes the tools used and the crowd-sourcing activities required.

RGB-D Collection

Visual perception data is collected using a Microsoft Azure Kinect, colloquially known as a Kinect 3, using Microsoft’s Azure Kinect drivers for Robot Operating System (ROS)(Koubâa 2017).² The Kinect 3 is an RGB-D camera consisting of both a Time-of-Flight (ToF) depth camera and a color camera which enables it to capture high-fidelity point cloud data. We collected raw image and point cloud data from 47 classes of objects across the five categories of objects. Approximately five instances of each class are imaged for a total of 207 instances. Table 1 shows the high-level topics as well as the specific classes that are collected within each topic. The dataset contains 207 90-second depth videos, one per instance, showing the object performing one complete rotation on a turntable. It also contains **825 pairs**

of image and depth point cloud from 207 objects, consisting of manually selected representative frames showing different angles of each object (an average of 3.98 frames per instance).

The Kinect Azure depth camera uses the Amplitude Modulated Continuous Wave (AMCW) ToF principle. Near infrared (NIR) emitters on the camera illuminate the scene with modulated NIR light and the camera calculates the time of flight for the light to return to the camera. From this a depth image can be built converting the time of flight to distance and then encoded into a monochromatic depth image. ROS allows for the registration of the color and depth images, matching pixels in the color to pixels in the depth image, to build a colored point cloud of the scene. Point Cloud Library (PCL) (Rusu and Cousins 2011) passthrough filters are used to crop the raw point cloud to only include the object being collected and the turntable.

Text Description Collection

We collect our text descriptions using the popular crowd-sourcing platform Amazon Mechanical TURK, or AMT. As described above, for each object instance, a subset of representative frames was manually chosen. For each task on AMT, these frames were shown for five randomly-chosen objects, each paired with a textbox. The AMT worker is asked to describe the object on the turntable in one or two short, complete sentences; they are specifically asked to not mention the turntable, table, other extraneous objects in the background. Each task was performed ten times, for a total of 40 text descriptions per instance. Removing objects for which there were representative frame errors, this allowed for the collection of **8250 total text descriptions**.

The purpose of taking images of objects from a variety of angles is to diversify what workers see. It is a known problem in vision systems that pictures tend to be taken from ‘typical’ angles that most completely show the object; for example, it is rare for a picture of a banana to be taken end-

²<https://docs.microsoft.com/en-us/azure/kinect-dk/>

on. This aligns with our motivation of creating a dataset of household objects to support research on grounded language learning in a human-centric environment: a robot talking to a person may have a partial view or understanding of an object, or vice versa. Thus we consider it essential to capture multiple views of objects in our dataset, and have those perhaps atypical views reflected in natural language descriptions.

Speech Description Collection

As speech interaction is becoming more common with current technologies, our dataset will allow researchers to design and test grounded learning solutions using this popular input modality. We collect audio data to capture the nuances between spoken and written natural language. It is common for people to restructure sentences before writing them, but while speaking, we do not have the liberty to reframe or restructure them. Therefore, spoken sentences tend to be unplanned, less well framed or grammatically incorrect (Redeker 1984). Humans support speech with body gestures, eye gaze, expressions or pitch of the voice, details that are missing in writing (Jaimes and Sebe 2007). Experienced writers may be able to overcome these differences while communicating but these people usually hold formal education (Miller 2006).

To collect spoken natural language data we develop a user interface utilizing the `MediaStream` recording API.³ The audio clips are stored in an Amazon S3 bucket⁴, which is a cloud storage service. Workers can play the recorded audio and if not satisfied can record it again. A similar approach is reported in recent work (Lane et al. 2010; Lee et al. 2015) to collect data using web-based and mobile application-based systems. We embed the interface into AMT and the recorded audio files are collected from these tasks.

The task on Amazon Mechanical Turk had a simple interface, showing a single image with “Record”, “Play”, and “Submit” buttons. Each ‘task’ had five such images, shown sequentially. In order to make the audio files compatible with ASR systems, missing metadata was added. The audio files were converted to text using Google’s Speech to Text API⁵. A subset of these transcriptions was evaluated for quality, as explained in section 2 (some examples are shown in table 5). This process resulted in a spoken-language dataset of **4059 verbal descriptions of 207 objects**.

4 Applications

Grounded language is useful for many robotic tasks. Grounded language acquisition (Alomari et al. 2017) is the general task of learning the structure and meanings of words based on natural language and perceptual inputs, usually visual but sometimes including other modalities such as haptic feedback or sound (Thomason et al. 2018). Other tasks include navigation (Ünal, Can, and Yemez 2019) where a

robot needs to either understand directions to get to a destination, or generate directions for someone or something else. Teaching and understanding tasks (Chai, Cakmak, and Sidner 2017) as well as asking for help on tasks (Tellex et al. 2014) are important areas of language that make interactions with robotic systems capable and productive. We focus on *manifold alignment for grounded language acquisition* as our example use of GoLD, because it is a grounded language acquisition task that highlights the multimodal nature of the data and the challenges unique to that setting. Since grounded language acquisition is the most general and used to some degree in all of these applications, we choose to focus on an example that highlights the use of GoLD towards a grounded language acquisition task.

Example: Manifold Alignment

We conduct a learning experiment to show how GoLD might be used as a means to learn grounded language. We use manifold alignment (Richards et al. 2019; Wang and Mahadevan 2011; 2008; Andrew et al. 2013) with triplet loss (Baltas et al. 2016; Schroff, Kalenichenko, and Philbin 2015) to embed the perceptual and language data from GoLD into a shared lower dimensional space. Within this space, a distance metric is applied to embedded feature vectors in order to tell how well a particular utterance describes an image. Novel pairs can then be embedded to determine correspondence. Alternatively, inputs from either domain can be embedded in the intermediate space to find associated instances from the other domain.

For example, a picture of a lemon and the description “The object is small and round. It is bright yellow and edible.” should be closer together in the embedded space than the same picture of a lemon and the incorrect description “This tool is used to drive nails into wood,” since the latter description was used to describe a hammer. Through this technique, even novel vision or language inputs should be aligned, meaning that a new description of a lemon should still be closely aligned in the embedded space. We would additionally expect other similar objects, such as an orange, to be described in a somewhat similar way, allowing for potential future learning of categorical information.

Vision. The vision feature vectors are created following the work of Eitel et al (Eitel et al. 2015). The color and depth images are each passed through a convolutional neural network that has been pretrained on ImageNet (Deng et al. 2009; Russakovsky et al. 2015) with the last layer (which is used for predictions) removed so that the final layer is a learned feature vector. The two vectors, one from color and one from depth, are then concatenated into a 4096-dimensional visual feature vector.

Language. The language features are extracted using BERT (Devlin et al. 2019). Each natural language description is fed to a “BERT-base-uncased” pretrained model which gives us the individual embeddings of all the tokens in the description. We obtain the description embedding by performing average pooling over the word embeddings. Due

³https://developer.mozilla.org/en-US/docs/Web/API/MediaStream_Recording_API

⁴<https://aws.amazon.com/s3/>

⁵<https://cloud.google.com/speech-to-text>

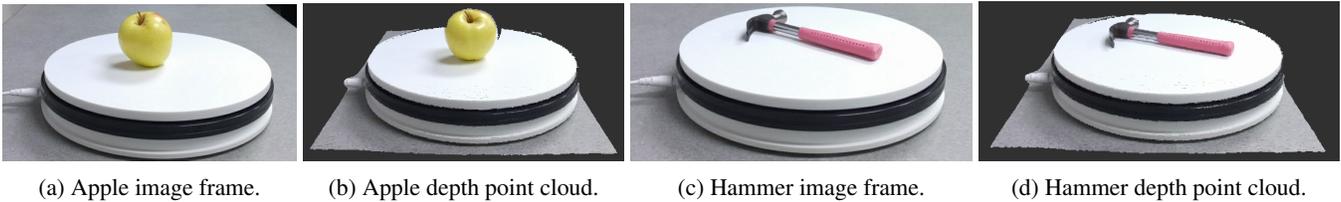


Figure 6: Samples showing the alignment of the visual data in GoLD. Each instance contains a stream of RGB image frames (taken as the object rotates on the turntable), as well as an aligned 3D point cloud capturing depth information. Note: the RGB images have been cropped from the full size for display here.

Table 5: Some examples of transcription-quality ratings. The transcriptions with the exact match are rated as 4.

Rating	Class of Object Described	Google Speech to Text Transcription	Manual Transcription
1	toothpaste	Institute best	It's a toothpaste
1	spoon	did Persephone used to serving before	this is spoon made up with wood used for serving food
1	soda_bottle	lovesick 100 African Buffalo	it is a plastic one and half liter bottle of coke
2	stapler	this is the stuff inside mechanical device which joins Legends of paper	this is a stapler it is a mechanical device which joins pages of paper
2	can_opener	emmanuel 10 opener with a blue handle	A manual tin opener with a blue handle
2	hand_towel	its a folded great owl	it's a folded gray towel
3	shampoo_bottle	what is a bottle of shampoo	that is a bottle of shampoo
3	mouse	Addison black color Mouse can be used in laptop or system	it is a black color Mouse can be used in laptop or system
3	coffee_mug	Arizona white coffee mug	There is a white coffee mug

to the contextual nature of its embeddings, BERT can differentiate between different meanings of the same word in different contexts. This results in semantically richer language features and a more meaningful embedding space. The resulting 3072-dimensional vector is taken as the description's language feature vector and associated to the visual feature vector of the frame it describes. Since the dataset contains ten natural language descriptions for each frame of an object, each visual feature vector is paired with ten different language feature vectors. The same process is repeated for the speech transcriptions.

Triplet Loss. The basic triplet loss function (Balntas et al. 2016; Schroff, Kalenichenko, and Philbin 2015) uses one training example as an “anchor” and two more points, one of which is in the same class as the anchor (the positive), and one which is not (the negative). For example, while classifying images of dogs and cats the anchor might be a cat image, the positive would be a different cat image, and the negative would be an image of a dog. The loss function then encourages the network to align the anchor and positive in the embedded space while repelling the anchor and the negative. Typically the positive and negative instances are from the same domain as one another. However, they may also be from the same domain as the anchor in order for the the network to be internally consistent, or from the other domain to align the two networks. Each of the four cases is chosen uniformly at random during training for each training instance.

Negative Sampling. In our case there is no obvious conceptualization of positive and negative examples of language. Because language is not exhaustive, the fact that a description omits particular concepts does not mean that the omitted language would not describe the object (for example, describing a lemon as a “yellow lemon” does not make it a good counterexample for the concept ‘round’). Similarly, a description can include language that is accurate for a description of a different object or even a different class; even in our dataset, which focuses on deep coverage of a small number of classes, “a round yellow thing” can be a lemon, a light bulb, or an onion. This underpins the widespread difficulty of finding true negative examples for natural language processing.

One approach to solving this problem for natural language processing relies on a different use of feature embedding (Pillai and Matuszek 2018). We calculate the cosine similarity between a language feature vector and all other language feature vectors within the training set. Vectors that are semantically similar will have a distance close to 1 while those further apart will be closer to 0. Therefore, we take the feature vector with the smallest cosine similarity as the negative and the largest as the positive. To get positive and negatives of images, we find the positive and negative of the image's associated language and then take the associated images of those instances. For anchors (A), positive instances (P), and negative instances (N), we compute embeddings of these points, then compute triplet loss in the standard fashion

with a margin $\alpha = 0.4$ (Schroff, Kalenichenko, and Philbin 2015) with Equation 2:

$$\mathcal{L}(A, P, N) = \max(\|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + \alpha, 0), \quad (2)$$

where f is the relevant model for the domain of the input points.

Training. Two models are trained from the data, a text-based language model and a transcribed speech-based language model. The text model, T, is trained for 50 epochs on 6600 paired visual/text feature vectors and evaluated on a held out set of 1650 examples from GoLD. A speech model, S, is trained from 3232 language vectors and their associated images and evaluated on a held out set of 828 transcriptions. A third model, T+S, is trained from both text and speech transcriptions to see how the combination of domains affects learning. We are interested in how training may be affected by differences in the way people describe objects through their word choice or structure. The automated transcription process also introduces noise into the speech descriptions, which has an effect on downstream performance.

Evaluation. We evaluate the network using the Mean Reciprocal Rank (MRR). The MRR is calculated by finding the distance of an embedding of a vector in one domain to all of the embeddings of the other domain, ordering them by Euclidean distance, and finding the rank of the testing instance in the ordered list. The reciprocals of these ranks are summed over the testing set and then averaged by the number of testing examples. When the number of testing examples is very high, the MRR can quickly approach zero even when the rank of the instance is in the top half of examples, rendering the metric difficult to interpret. To counteract this and to evaluate our model on a scenario that is more realistic to what it might be used for, instead of ranking the entire testing set, we rank a select few instances. The first ranking is only on the target, positive, and negative instances and the second is on the target and four other randomly selected instances. The first evaluation tests that the triplet loss is having an effect on the final model, while the second mimics what the model might do when faced with identifying objects in a cluttered scene.

In both cases, we test (1) identifying objects from language descriptions, and (2) choosing a description given an object. We use $x \rightarrow y$ to denote a test query from domain x , which identifies something to be returned from domain y . $L \rightarrow V$ therefore denotes the test case in which language is provided and an object must be chosen from its perceptual data, and vice versa.

The combined “T + S” model is evaluated three separate times. First, it is tested individually on held-out sets where V is drawn first from text, then from speech. It is then evaluated on the combination of the two held-out sets. Because we expect the speech and text to be similar, testing on the combination of them should perform better than testing on either evaluation set in isolation, and in fact, this is what we see. Our results show that a grounded language model can

be learned from the GoLD data. In particular, when ranking the distances between an embedded target instance from one domain with a selection of embedded instances from the other domain, we expect the target to appear in the top half of the rankings, which we consistently see.

Discussion. We are investigating, first, whether using the GoLD data to train these models in a manifold alignment experiment yields better performance than a random baseline, and second, how the performance of a language model trained on speech performs *vs.* typed text. Table 6 shows the results from our experiments. In all cases our experiments outperform the random baseline, where the target is expected to have a rank halfway down the ordering. For the Triplet MRR and Subset MRR, we would thus like our models to perform better than $1/2$ and $1/3$ respectively since there are three objects in the triplet evaluation and five in the subset. Table 6 shows, then, that our model has effectively achieved manifold alignment, aligning similar examples while repelling dissimilar ones. The fact that the Subset MRR is above $1/3$, which is true in all cases (vision to language, language to vision, whether speech or text), tells us that our model is not just randomly selecting a target instance. So, when given a subset of instances, we can say that our model is able to select the target or rank the target highly.

The speech model performs marginally worse than the text model. This could be due to the smaller training dataset, but is probably fully explainable by the noise generated by the transcription process. Speech transcriptions allow for a one to one comparison of the two domains, but in future work, we will train a model over the raw audio. In a system that utilized one of these models, this would eliminate the need for a transcription step and may provide more accurate results since there may be tonal or inflection data that is lost in the transcription process.

When the combined model is evaluated against each individual domain, we find the performance drop that we expected. The minor drop in performance when evaluating on the combined held out set implies that the two domains are different enough that a model trained on both modalities together has more difficulty reaching the level of performance of a learned model trained on uniform data. While this is not unexpected, it is heartening that the performance drop is relatively small. Perhaps more importantly, the improvement on the combined test set demonstrates that the model is being trained to effectively interpret either spoken or typed input.

Another interesting aspect of these results is the relative difference in the direction of the mapping, that is, which domain is chosen as the target. In particular, the $V \rightarrow L$ case outperforms $L \rightarrow V$. That is, selecting the associated language given a visual input is easier than selecting the associated vision given a language input. We suspect this observed behavior could be due to differences in the manifolds of the feature vector spaces. BERT is a highly pre-trained model (Devlin et al. 2019). The visual domain, however, is much more complex. Raw images, like an apple and a red ball, may be very close together while it is unlikely that their

Table 6: Experimental results. Mean Rank Reciprocal for models trained on Text and Speech descriptions. For Triplet, MRR values above 0.5 demonstrate a successfully learned alignment between language and perceptual data; for Subset, MRR values above 0.33 demonstrate success. Triplet MRR is calculated from the target and a positively and negatively associated test data point, while Subset MRR is calculated from the target and a subset of four random test data points.

Model	Domain	Triplet MRR	Subset MRR
Text	L → V	0.6658	0.4560
	V → L	0.7342	0.4669
Speech	L → V	0.6661	0.4391
	V → L	0.7289	0.4562
T + S	L → V	0.5954	0.4255
	L → V (Test on T)	0.4670	0.4547
	L → V (Test on S)	0.6651	0.4520
	V → L	0.6762	0.4519
	V → L (Test on T)	0.4389	0.4605
	V → L (Test on S)	0.4587	0.4594

descriptions would be close together. BERT uses the context of a word to generate the embeddings so in this example even if the word “red” were used in both descriptions, the language embedding would be different. The raw images also all contain similar background scenery, the turntable, and the table.

5 Discussion and Future Work

In this paper we present GoLD, a grounded language dataset of images in color and depth paired with natural language descriptions of everyday household objects in text and speech. We aim to make this resource a useful starting point for downstream grounded language learning tasks such as spoken natural language interfaces for personal assistants and domestic service robots.

To demonstrate a potential use of GoLD, we use the data to train models that perform heterogeneous manifold alignment. We hope this dataset serves researchers as a resourceful starting point from which to explore many more techniques, model architectures, and algorithms that further our understanding of grounded language. In particular, the inclusion of speech alongside written textual descriptions allows for side-by-side comparisons of the two domains grounded to physical objects, or for novel multimodal techniques involving all three domains of vision, text, and speech.

The idiosyncratic properties of GoLD suggest many research questions for future study. For example, we remark that GoLD includes descriptions of the same object as observed from multiple angles. One interesting question to explore, then, would be how to identify objects from a different perspective or when information is missing. Another property of GoLD is that some descriptions focus on the use of the object (mostly agnostic to perspective), while others report perceptual qualities of the objects such as logos and other identifying features uniquely visible from the annota-

tor’s current perspective on that object. This aspect could be incorporated into a human-robot interaction study that examines grounding language to objects in a physical as viewed by embodied agents from different vantage points.

In the near term, we are interested in leveraging this dataset to train robots to understand natural language in order to perform tasks in a domestic context. The inclusion of medical and kitchen supplies is critical to training a robot for tasks such as cooking, cleaning, and administering care. As we work toward this goal, we anticipate creating an expanded catalog of items including the diverse ways in which people describe and talk about the wide variety of items they encounter every day.

References

- Alomari, M.; Duckworth, P.; Hogg, D. C.; and Cohn, A. G. 2017. Natural language acquisition and grounding for embodied robotic systems. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I. D.; Gould, S.; and van den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3674–3683.
- Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In Dasgupta, S., and McAllester, D., eds., *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, 1247–1255. Atlanta, Georgia, USA: PMLR.
- Balntas, V.; Riba, E.; Ponsa, D.; and Mikolajczyk, K. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. 119.1–119.11.
- Beckerle, P.; Salviotti, G.; Unal, R.; Prattichizzo, D.; Rossi, S.; Castellini, C.; Hirche, S.; Endo, S.; Amor, H. B.; Ciocarlie, M.; et al. 2017. A human–robot interaction perspective on assistive and rehabilitation robotics. *Frontiers in Neuro-robotics* 11(24):1.
- Bedaf, S.; Gelderblom, G. J.; and Witte, L. D. 2014. Overview and categorization of robots supporting independent living of elderly people: What activities do they support and how far have they developed. *Assistive Technology* 27(2):88–100.
- Beer, J. M.; Smarr, C.-A.; Chen, T. L.; Prakash, A.; Mitzner, T. L.; Kemp, C. C.; and Rogers, W. A. 2012. The domesticated robot: design guidelines for assisting older adults to age in place. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 335–342. ACM.
- Branavan, S. R.; Chen, H.; Zettlemoyer, L. S.; and Barzilay, R. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, 82–90. Association for Computational Linguistics.

- Cavazza, M. 2001. An empirical study of speech recognition errors in a task-oriented dialogue system. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*, 1–8. Association for Computational Linguistics.
- Chai, J. Y.; Cakmak, M.; and Sidner, C. 2017. Teaching robots new tasks through natural interaction. *Interactive Task Learning: Agents, Robots, and Humans Acquiring New Tasks through Natural Interactions*.
- Chen, D. L., and Mooney, R. J. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, 128–135. ACM.
- Chen, T. L.; Ciocarlie, M. T.; Cousins, S. B.; Grice, P. M.; Hawkins, K. P.; Hsiao, K.; Kemp, C. C.; King, C.-H.; Lazewatsky, D. A.; Leeper, A.; Nguyen, H.; Paepcke, A.; Pantofaru, C.; Smart, W. D.; and Takayama, L. 2013. Robots for humanity: using assistive robotics to empower people with disabilities. *IEEE Robotics & Automation Magazine* 20:30–39.
- Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T. H.; and Bengio, Y. 2019. BabyAI: First steps towards grounded language learning with a human in the loop. In *International Conference on Learning Representations*.
- Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; and Batra, D. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2054–2063.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- Eitel, A.; Springenberg, J. T.; Spinello, L.; Riedmiller, M. A.; and Burgard, W. 2015. Multimodal deep learning for robust rgb-d object recognition. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 681–687.
- Engel, D.; Charniak, E.; and Johnson, M. 2002. Parsing and disfluency placement. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 49–54.
- Engel, D. 2001. The utility of filled pauses, interjections, and parentheticals in parsing conversational language. Master's thesis, Brown University.
- Evers, J. *Kwalitatief interviewen: kunst én kunde*.
- Evers, J. C. 2011. From the past into the future. how technological developments change our ways of data collection, transcription and analysis. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 12.
- Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378.
- Haeb-Umbach, R.; Watanabe, S.; Nakatani, T.; Bacchiani, M.; Hoffmeister, B.; Seltzer, M. L.; Zen, H.; and Souden, M. 2019. Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal Processing Magazine* 36(6):111–124.
- Hu, R.; Fried, D.; Rohrbach, A.; Klein, D.; Darrell, T.; and Saenko, K. 2019. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6551–6557.
- Jaimes, A., and Sebe, N. 2007. Multimodal human-computer interaction: A survey. *Computer vision and image understanding* 108(1-2):116–134.
- Johnson, J. E.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. B. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1988–1997.
- Koubâa, A. 2017. *Robot Operating System (ROS)*. Springer.
- Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.
- Kubota, A.; Peterson, E. I. C.; Rajendren, V.; Kress-Gazit, H.; and Riek, L. D. 2020. Jessie: Synthesizing social robot behaviors for personalized neurorehabilitation and beyond. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI '20*, 121–130. New York, NY, USA: Association for Computing Machinery.
- Lai, K.; Bo, L.; Ren, X.; and Fox, D. 2011. A large-scale hierarchical multi-view rgb-d object dataset. 1817–1824.
- Lane, I.; Waibel, A.; Eck, M.; and Rottmann, K. 2010. Tools for collecting speech corpora via mechanical-turk. In *Proceedings of the NAACL HLT*, 184–187.
- Lee, K. A.; Larcher, A.; Wang, G.; Kenny, P.; Brummer, N.; van Leeuwen, D.; Aronowitz, H.; Kockmann, M.; Vaquero, C.; Ma, B.; Li, H.; Stafylakis, T.; Alam, J.; Swart, A.; and Perez, J. 2015. The reddots data collection for speaker recognition. In *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*.
- Li, B.; Tsao, Y.; and Sim, K. C. 2013. An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition. In *Interspeech*, 3002–3006.
- Lin, C.-Y., and Och, F. J. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 605. Association for Computational Linguistics.
- Matuszek, C.; Bo, L.; Zettlemoyer, L.; and Fox, D. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.
- Matuszek, C. 2018. Grounded language learning: Where robotics and nlp meet. 5687–5691.

- McCowan, I. A.; Moore, D.; Dines, J.; Gatica-Perez, D.; Flynn, M.; Wellner, P.; and Bourlard, H. 2004. On the use of information retrieval measures for speech recognition evaluation. Technical report, IDIAP.
- Miller, G. A. 1995. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM* 38:39–41.
- Miller, J. 2006. In Aarts, B., and McMahan, A., eds., *The Handbook of English Linguistics*. Blackwell Publishing Ltd. chapter 28, 673–675.
- Mooney, R. J. 2008. Learning to connect language and perception.
- Mylonas, D.; Purver, M.; Sadrzadeh, M.; Macdonald, L.; and Griffin, L. 2015. The use of english colour terms in big data.
- Nguyen, K., and Daumé III, H. 2019. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. 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)*, 684–695.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.
- Park, Y.; Patwardhan, S.; Visweswariah, K.; and Gates, S. C. 2008. An empirical analysis of word error rate and keyword error rate. In *Ninth Annual Conference of the International Speech Communication Association*.
- Pillai, N., and Matuszek, C. 2018. Unsupervised selection of negative examples for grounded language learning. In *Proc. of the Thirty-second AAAI Conference on Artificial Intelligence (AAAI)*.
- Redeker, G. 1984. On differences between spoken and written language. *Discourse processes* 7(1):43–55.
- Richards, L. E., and Matuszek, C. 2019. Learning to understand non-categorical physical language for human-robot interactions. In *Proceedings of the RSS 2019 workshop on AI and Its Alternatives in Assistive and Collaborative Robotics (RSS: AI+ACR)*.
- Richards, L. E.; Nguyen, A.; Darvish, K.; Raff, E.; and Matuszek, C. 2019. A manifold alignment approach to grounded language learning. In *Unpublished Proceedings of the 8th Northeast Robotics Colloquium*.
- Roy, D. K. 2002. Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language* 16:353–385.
- Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3):211–252.
- Rusu, R. B., and Cousins, S. 2011. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*.
- Saon, G.; Ramabhadran, B.; and Zweig, G. 2006. On the effect of word error rate on automated quality monitoring. In *2006 IEEE Spoken Language Technology Workshop*, 106–109. IEEE.
- Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. 815–823.
- Stolcke, A., and Droppo, J. 2017. Comparing human and machine errors in conversational speech transcription. *arXiv preprint arXiv:1708.08615*.
- Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A. G.; Teller, S.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.
- Tellex, S.; Knepper, R.; Li, A.; Rus, D.; and Roy, N. 2014. Asking for help using inverse semantics.
- Thomason, J.; Zhang, S.; Mooney, R. J.; and Stone, P. 2015. Learning to interpret natural language commands through human-robot dialog. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
- Thomason, J.; Sinapov, J.; Mooney, R. J.; and Stone, P. 2018. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Thomason, J.; Murray, M.; Cakmak, M.; and Zettlemoyer, L. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.
- Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, 173–180. Association for Computational Linguistics.
- Ünal, E.; Can, O. A.; and Yemez, Y. 2019. Visually grounded language learning for robot navigation. In *1st International Workshop on Multimodal Understanding and Learning for Embodied Applications*, 27–32.
- Vanzo, A.; Croce, D.; Bastianelli, E.; Basili, R.; and Nardi, D. 2020. Grounded language interpretation of robotic commands through structured learning. *Artificial Intelligence* 278:103181.
- Wang, C., and Mahadevan, S. 2008. Manifold alignment using procrustes analysis. 1120–1127.
- Wang, C., and Mahadevan, S. 2011. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI'11*, 1541–1546. AAAI Press.
- Weizs, A. 2019. How long it really takes to transcribe (accurate) audio.