A Visuospatial Dataset for Naturalistic Verb Learning

Dylan Ebert
Brown University
dylan_ebert@brown.edu

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

Abstract

We introduce a new dataset for training and evaluating grounded language models. Our data is collected within a virtual reality environment and is designed to emulate the quality of language data to which a pre-verbal child is likely to have access: That is, naturalistic, spontaneous speech paired with richly grounded visuospatial context. We use the collected data to compare several distributional semantics models for verb learning. We evaluate neural models based on 2D (pixel) features as well as feature-engineered models based on 3D (symbolic, spatial) features, and show that neither modeling approach achieves satisfactory performance. Our results are consistent with evidence from child language acquisition that emphasizes the difficulty of learning verbs from naive distributional data. We discuss avenues for future work on cognitively-inspired grounded language learning, and release our corpus with the intent of facilitating research on the topic.

1 Introduction

While distributional models of semantics have seen incredible success in recent years (Devlin et al., 2018), most current models lack "grounding", or a connection between words and their referents in the non-linguistic world. Grounding is an important aspect to representations of meaning and arguably lies at the core of language "understanding" (Bender and Koller, 2020). Work on grounded language learning has tended to make opportunistic use of large available corpora, e.g. by learning from web-scale corpora of image (Bruni et al., 2012) or video captions (Sun et al., 2019), or has been driven by particular downstream applications such as robot navigation (Anderson et al., 2018).

In this work, we take an aspirational look at grounded distributional semantics models, based on the type of situated contexts and weak supervision from which children are able to learn much of their early vocabulary. Our approach is motivated by the assumption that building computational models which emulate human language processing is in itself a worthwhile endeavor, which can yield both scientific (Potts, 2019) and engineering (Linzen, 2020) advances in NLP. Thus, we aim to develop a dataset that better reflects both the advantages and the challenges of humans' naturalistic learning environments. For example, unlike most vision-and-language models, children likely have the advantage of access to symbolic representations of objects and their physics prior to beginning word learning (Spelke and Kinzler, 2007). However, also unlike NLP models, which are typically trained on image or video captions with strong signal, children's language input is highly unstructured and the content is often hard to predict given only the grounded context (Gillette et al., 1999).

We make two main contributions. First (§2), using a virtual reality kitchen environment, we collect and release[1] the New Brown Corpus[2]: A dataset containing 18K words of spontaneous speech alongside rich visual and spatial information about the context in which the language occurs. Our protocol is designed to solicit naturalistic speech and to have good coverage of vocabulary items with low average ages of acquisition according to data on child language development (Frank et al., 2017). Second (§3), we use our corpus to compare several distributional semantics models, specifically comparing mod-

[1] https://github.com/dylanebert/nbc

[2] Our university namesake, plus paying homage to important Brown corpora in both NLP (Francis and Kucera, 1979) and Child Language Acquisition (Brown, 1973).

*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 143–153
Barcelona, Spain (Online), December 12–13, 2020

els which represent the environment in terms of objects and their physics to models which represent the environment in terms of pixels. We focus on verbs, which have received considerably less attention in work on grounded language learning than have nouns and adjectives (Forbes et al., 2019). More so than nouns, verb learning is believed to rely on subtle combinations of both syntactic and grounded contextual signals (Piccin and Waxman, 2007) and thus progress on verb learning is likely to require new approaches to modeling and supervision. In our experiments, we find that strong baseline models, both feature-engineered and neural network models, perform only marginally above chance. However, comparing models reveals intuitive differences in error patterns, and points to directions for future research.

2 Data

The goal of our data collection is to enable research on grounded distributional semantics models using data that better resembles the type of input young children receive on a regular basis during language development. Doing this fully is ambitious if not impossible. Thus, we focus on a few aspects of children's language learning environment that are lacking from typical grounded language datasets and that can be emulated well given current technology: 1) spontaneous speech (i.e. as opposed to contrived image or video captions) and 2) rich information about the 3D world (i.e. physical models of the environment as opposed to flat pixels).

We develop a virtual reality (VR) environment within which we collect this data in a controlled way. Our environment data is described in Section 2.1 and our language data is described in Section 2.2. Our collection process results in a corpus of 152 minutes of concurrent video, audio, and ground-truth environment information, totaling 18K words across 18 unique speakers performing six distinct tasks each. The current data is available for download in json format at `https://github.com/dylanebert/nbc`. The code needed to implement the described environment and data recording is available at `https://github.com/dylanebert/nbc_unity_scripts`.

2.1 Environment Data Collection

2.1.1 Environment Construction

Our environment is a simple kitchen environment, implemented in Unity with SteamVR and our experiments are conducted using an HTC Vive headset. We choose to use VR as opposed to alternative interfaces for simulated interactions (e.g. keyboard or mouse control) since VR enables participants to use their usual hand and arm motions and to narrate in real time, leading to more natural speech and more faithful simulations of the actions they are asked to perform.

We design six different kitchen environments, using two different visual aesthetics (Fig. 1) with three floorplans each. This variation is so that we can test, for example, that learned representations are not overfit to specific pixel configurations or to exact hand positions that are dependent on the training environment(s) (e.g. "being in the northwest corner of the kitchen" as opposed to "being near the sink"). Each kitchen contains at least 20 common objects (not every kitchen contains every object). These objects were selected because they represent words with low average ages of acquisition (described in detail in §2.2) and were available in different Unity packages and thus could be included in the environment with different appearances. Across all kitchens, the movable objects used are: `Apple`, `Ball`, `Banana`, `Book`, `Bowl`, `Cup`, `Fork`, `Knife`, `Lamp`, `Plant`, `Spoon`, `Toy1:Bear|Bunny`, `Toy2:Doll|Dinosaur`, `Toy3:Truck|Plane`. The participant's hands and head are also included as movable objects. We also include the following immovable objects: `Cabinets`, `Ceiling`, `Chair`, `Clock`, `Counter`, `Dishwasher`, `Door`, `Floor`, `Fridge`, `Microwave`, `Oven`, `Pillar`, `Rug`, `Sink`, `Stove`, `Table`, `Trash Bin`, `Wall`, `Window`.

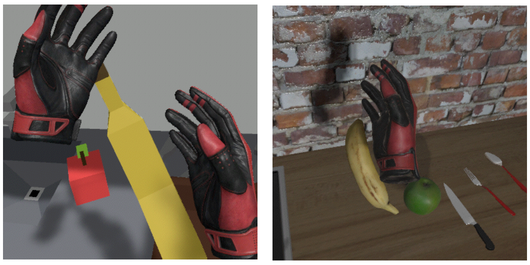


Figure 1: Screenshots of a person picking up a banana in each of our two kitchen aesthetics.

Our environments are constructed using a combination of Unity Asset Store assets and custom models. All paid assets (most objects we used) come from two packs: 3DEverything Kitchen Collection 2 and Synty Studios Simple House Interiors, from the Unity asset store[3]. These packs account for the two visual styles. VR interaction is enabled using the SteamVR Unity plugin, available for free on the Unity asset store.

2.1.2 Data Recording

During data collection, we record the physical state of each object in the environment, according to the ground-truth in-game data, at a rate of 90fps (frames per second). The Vive provides accurate motion capture, allowing us to record the physical state of the user's head and hands (Borges et al., 2018) as well. For each object, we record the physical features described in Table 1. Audio data is also collected in parallel to spatial data, using the built-in microphone. We later transcribe the audio using the Google Cloud Speech-to-Text API[4]. Word-level timestamps from the API allow us to match words to visuospatial frames. While spatial and audio data are recorded in real-time, video recording is not, since this would introduce high computational overhead and drop frames. Instead, we iterate back over the spatial data, and reconstruct/rerender the playback frame-by-frame. This approach makes it possible to render from any perspective if needed, though our provided image data is only from the original first-person perspective.

2.2 Language Data Collection

We design our protocol so as to solicit the use of vocabulary items that are known to be common among children's early-acquired words. To do this, we first select 20 nouns, 20 verbs, and 20 prepositions/adjectives which have low average ages of acquisition according to Frank et al. (2017) and which can be easily operationalized within our VR environment (e.g. *"apple"*, *"put (down)"*, *"red"*, see Appendix A for full word list). We then choose six basic tasks which the participants will be instructed to carry out within the environment. These tasks are: set the table, eat lunch, wash dishes, play with toys, describe a given object, and clean up toys. The tasks are intended to solicit use of many of the target vocabulary items without explicitly instructing participants to use specific words, since we want to avoid coached or stilted speech as much as possible. One exception is the "describe a given object" task, in which we ask participants to describe specific objects as though a child has just asked what the object is, e.g. *"What's a spoon?"*. We use this task to ensure uniform coverage of vocabulary items across environments, so that we can construct good train/test splits across differently appearing environments. See Appendix B for details on distributing vocabulary items.

We recruited 18 participants for our data collection. Participants were students and faculty members from multiple departments involved with language research. We asked each participant to perform each of our tasks, one by one, and to narrate their actions as they went, as though they were a parent or babysitter speaking to a young child. The exact instructions given to participants before each task are shown in Appendix C. An illustrative example of the language in our corpus is the following: *"okay let's pick up the ball and play with that will it bounce let's see if we can bounce it exactly let's let it drop off the edge yes it bounced the ball bounced pick it up again..."*. The full data can be browsed at `https://github.com/dylanebert/nbc`.

Our study design was determined not to be human subjects research by the university IRB. All participants were informed of the purpose of the study and provided signatures consenting to the recording and release of their anonymized data for research purposes (consent form in Appendix D).

2.3 Comparison to Child Directed Speech

Since our stated goal was to collect data that better mirrors the distribution of language input a young child is likely to receive, we run several corpus analyses to assess whether this goal was met.

2.3.1 Vocabulary Distribution

First, we compare the distribution of vocabulary in our collected data to that observed in the Brent-Siskind Corpus (Brent and Siskind, 2001), a corpus of child-directed speech consisting of 16 English-speaking mothers speaking to their preverbal children. For reference, we also compare with the vocabulary distributions of three existing corpora which could be used for training distributional semantics models: 1) MSR-VTT (Xu et al.,

[3] `https://assetstore.unity.com/`

[4] `https://cloud.google.com/text-to-speech/`

Name (Type)	Description
pos (xyz)	Absolute position of object center, computed using the `transform.position` property; equivalent to position relative to an arbitrary world origin, approximately in the center of the floor.
rot (xyzw)	Absolute rotation of object, computed using the `transform.rotation` property.
vel (xyz)	Absolute velocity of object center, computed using the `VelocityEstimator` class included with SteamVR.
relPos (xyz)	Position of object's center relative to the person's head, computed using Unity's built-in `head.transform.TransformPoint(objectPosition)`.
relRot (xyzw)	Rotation of object relative to the person's head, computed by applying the inverse of the head rotation to the object rotation.
relVel (xyz)	Velocity of the object's center, from the frame of reference of the person's head
bound (xyz)	Distance from object's center to the edge of bounding box
inView (bool)	Whether or not the object was in the person's field of view, computed using Unity's `GeometryUtility` to compute if an object is in the Camera renderer bounds. This is based on the default camera's 60 degree FOV, not the wide headset FOV. The head and hands are always considered *inView*.
img_url (img)	Snapshot of the person's entire field of view as a 2D image. We compute this once per frame (as opposed to the above features which are computed once per object per frame).

Table 1: Object features recorded during data collection. Object appearance does not vary across frames; img_url does not vary across objects. All other features vary across object and frame.

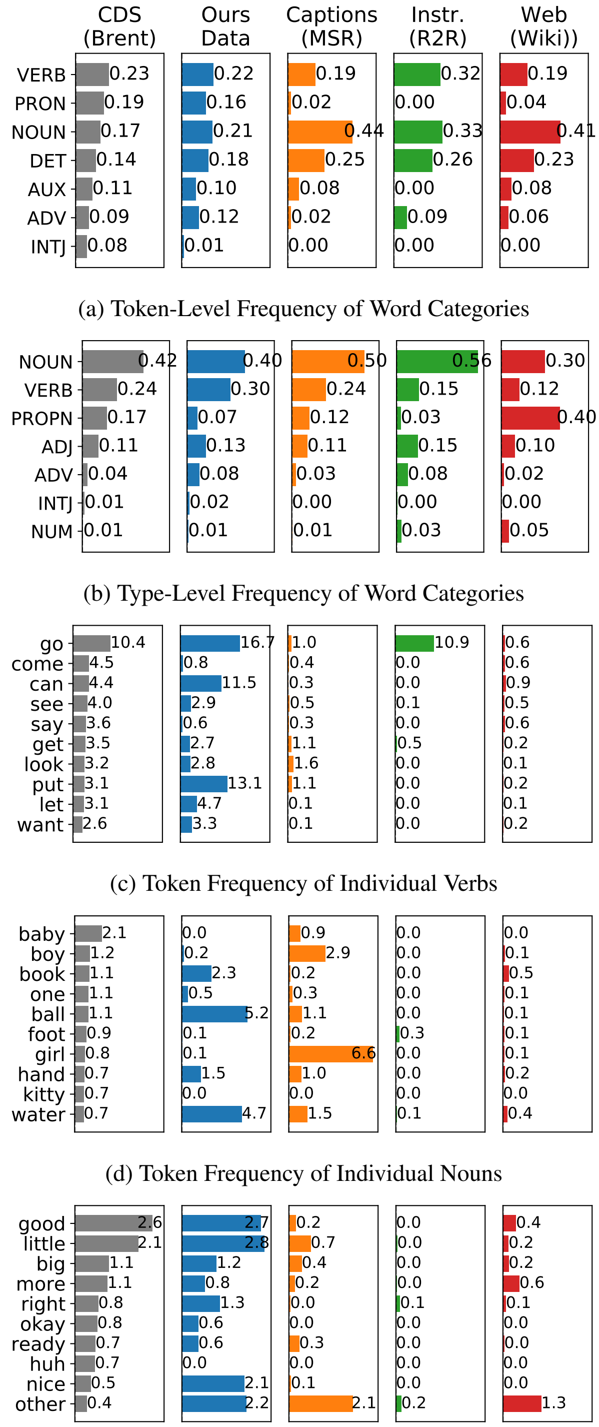
2016), a large dataset of YouTube videos labeled with captions, 2) Room2Room (R2R) (Anderson et al., 2018), a dataset for instruction following within a 3D virtual world, and 3) a random sample of sentences drawn from Wikipedia. Since our primary focus is on grounded language, MSR and R2R offer the more relevant points of comparison, since each contains language aligned with some kind of grounded semantic information (raw RGB video feed for MSR and video+structured navigation map for R2R). We include Wikipedia to exemplify the type of web corpora that are ubiquitous in work on representation learning for NLP.

Figure 2 shows, for each of the five corpora, the token- and type-level frequency distributions over major word categories[5] and of individual lexical items. In terms of word categories, we see that our data most closely mirrors the distribution of child-directed speech: Both our corpus and the Brent corpus contain primarily verbs ($\sim$23% when computed at the token level) followed by pronouns ($\sim$19%) followed by nouns at around 17%. In contrast, the MSR video caption corpus and Wikipedia both contain predominantly nouns ($\sim$40%) and the R2R instruction dataset contains nouns and verbs in equal proportions ($\sim$33% each). None of the baseline corpora contain significant counts of pronouns. Additionally, in terms of specific vocabulary items, our corpus contains decent coverage for many of the most frequent verbs observed in CDS, while the baseline corpora are dominated by a single verb each (*"go"* for R2R and *"be"* for MSR and Wikipedia). For nouns and adjectives, we also see better coverage of top-CDS words in our data compared to the other corpora analyzed, though we note that the difference is less obvious and that the lexical items in these categories are much more topically determined.

2.3.2 Word-Context Alignment

We next look at how well the language corresponds to the the salient objects and events in the context of its use. This property is important as it relates to how strong the "training signal" would be for a model that is attempting to learn linguistic meaning from distributional signal. It is hard to directly estimate the quality of the "training signal" available to children. However, experiments in psychology using the Human Simulation Paradigm (HSP) (Gillette et al., 1999; Piccin and Waxman, 2007) come close. In the HSP design, experimenters collect audio and video recordings of a child's normal activities (i.e. via head-mounted cameras). Given this data, adults are asked to view segments of videos and predict which words are said at given points in time. This technique is used to estimate how "predictable" language is given only the grounded (non-linguistic) input to which a child has access. Using this technique, Gillette et al. (1999) estimates that nouns can be predicted at 45% accuracy and verbs at 15% accuracy.

[5] We preprocess all corpora using the SpaCy 2.3.2 preprocessing pipeline with the `en_core_web_lg` model. For our data and Brent, we process the entire corpus. Since MSR, R2R, and Wikipedia are much larger, we process a random sample of 5K sentences from each.



(a) Token-Level Frequency of Word Categories

(b) Type-Level Frequency of Word Categories

(c) Token Frequency of Individual Verbs

(d) Token Frequency of Individual Nouns

(e) Token Frequency of Individual Adjectives

Figure 2: Comparison of word category and lexical distributions. Lexical item frequencies labels are ×1000. Distributions are over the most frequent categories/words according to the Brent-Siskind corpus of child-directed speech.

While not directly comparable to our setting, this provides us with an approximate point of comparison against which to benchmark the word-to-context alignment of our collected data. Rather than try to guess the word given a video clip, we instead view a short (5 second) video clip alongside an uttered word and make a binary judgement for whether or not the clip depicts an instance of the word: e.g., *yes or no, does the clip depict an instance of "pick up"?* We chose this design over the HSP design since it provides a more interpretable measure of the quality of the training signal from the perspective of NLP and ML researchers using the data. We expect this variant of the task to yield higher numbers than the HSP design, since it does not require guessing from the entire vocabulary. We take a sample of (up to) five instances for each of our target nouns and verbs (fewer if the word occurs less often in our data) and label them in this way. We find inner annotator agreement on this task to be very high (91% when computed between two researchers on the project) and thus have a single annotator label all instances.

Table 2 shows the results of this analysis. We see the expected trend, in which grounded context is a considerably better signal of noun use than verb use. We also note there is substantial variation in training signal across verbs. For example, while some verbs (e.g. *"pick"*, *"take"*, *"hold"*) have strong signal, other verbs (*"eat"*) tend to be used in contexts sufficiently detached from the activities themselves. The noisiness of this signal is one of the biggest challenges of learning from such naturalistic data, as we will discuss further in §3.4.

3 Experiments

Using the above data, we now compare several grounded distributional semantics models (DSM) in terms of how well they encode verb meanings, focusing in particular on differences in how the environment is represented when put in to the DSM. Our hypothesis is that models will perform better if they represent the environment in terms of 3D objects and their physics rather than pixels, since work in psychology has shown that children learn to parse the physical world into objects and agents very early in life (Spelke and Kinzler, 2007), long before they show evidence of language understanding. We also explore how models vary when they have access to linguistic supervision early in the pipeline, during environment encoding, in addition to later, during language learning. We note that the models explored are intended as sim-

Nouns			Verbs		
w	N	P	w	N	P
table	81	1.0	go	238	0.0
spoon	76	1.0	put	193	0.4
banana	75	0.8	pick	162	0.8
apple	68	1.0	eat	77	0.0
cup	57	1.0	take	63	0.8
ball	54	0.6	get	43	0.4
toy	48	1.0	wash	38	0.6
fork	47	0.8	play	37	0.8
bowl	42	1.0	walk	25	0.4
knife	40	0.8	throw	25	0.6
book	25	1.0	hold	21	1.0
plant	22	1.0	drop	17	0.4
bear	18	1.0	stop	13	0.0
chair	16	0.4	give	13	0.0
doll	13	0.8	open	3	0.3
clock	12	0.6			
lamp	2	1.0			
door	2	0.0			
window	1	1.0			
Avg.	37	0.8	Avg.	64	0.4

Table 2: Estimates of training signal quality for nouns and verbs. N is the number of times the word occurs in the training data. P is the precision–given a 5 second clip in which the word is used, how often does the clip depict an instance of the word? Note that the verb *"go"* is an outlier, since it appears most often as *"going to"*.

ple instantiations to test the parameters of interest given our (small) dataset. Future work on more advanced models should no doubt yield improvements.

3.1 Preprocessing

Our raw data consists of continuous video and game-engine recordings of the environment, and parallel transcriptions of the natural language narration. To convert this into a format usable by our DSM, we perform the following preprocessing steps. This preprocessing phase is common to all the models evaluated. First, we segment the environment data into "clips". Each clip is five seconds long[6] and thus consists of 450 frames (since the VR environment recording is at 90fps), which we subsample to 50 frames (10fps). Since our grounded DSMs require associating a word w with its grounded context c, we consider the clip immediately following the utterance of w to be the context c. See earlier discussion (§2.3.2) for estimates of the signal-to-noise ratio produced by this labeling method. Training clips that are not the context of any word are discarded. We hold out two subjects' sessions (one from each visual aesthetic) for test, and use the remaining 16 subjects' sessions for training.

Finally, since this verb-learning problem proves quite challenging, we scope down our analysis to the following 14 verbs, which come from the 20 verbs specified in our initial target vocabulary (§2.2) less 6 which did not ultimately occur in our data: *"walk", "throw", "put (down)", "get", "go", "give", "wash", "open", "hold", "eat", "play", "take", "drop", "pick (up)"*. Again, these words all have low average ages of acquisition (19 to 28 months) and thus should represent reasonable targets for evaluation. Nonetheless, we will see in §3.3 that models struggle to perform well on this task; we elaborate on this discussion in §4.

3.2 Models

We train and evaluate four different DSMs, each of which represent a word w in terms of its grounded context c. The parameters we vary are 1) the feature representation of c ($3.2.1) and 2) the type of supervision provided to the DSM (§3.2.2). All models share the same simple pipeline. First, we build a word-context matrix M which maps each token-level instance of w to a featurized representation of c. We then run dimensionality reduction on M. Finally, we take the type-level representation of w to be the average row vector of M, across all instances of w. All of our model code is available at `http://github.com/dylanebert/nbc_starsem`.

3.2.1 Context Encoders

Object-Based. In our Object-Based encoder, we take a feature-engineered approach intended to provide the model with a knowledge of the basic object physics likely to be relevant to the semantics of the verbs we target. Specifically, we represent each clip using four feature templates (`trajectory`, `vel`, `dist_to_head`, `relPos`), defined as follows. First, we find the "most moving object", i.e., the object with the highest average velocity over the clip. We then compute our four sets of features for this most moving object. Our `velocity` and

[6] The length of 5 seconds was chosen heuristically prior to model development.

relPos features are simply the mean, min, max, start, end, and variance of the object's velocity and relative position, respectively, over the clip. For our dist_to_head feature, for each position dimension (xyz), we compute the following values of the distance from the object's center to the participant's head: start, end, mean, var, min, max, min_idx, max_idx, where min/max index is the point at which min/max value was reached (recorded as a % of the way through the clip). Finally, our trajectory features are intended to capture the shape of the objects trajectory over the clip. To compute this, for each of position dimension (xyz), we compute four points during the clip: start, peak (max), trough (min), end. Then, if max happens before min, we consider the max to be "key point 1" (kp1) and the min to be "key point 2" (kp2), and vice-versa if the min happens before the max. We then compute the following features: kp1-start, kp2-kp1, end-kp2, end-start.

Pretrained CNN. To contrast with the above featured-engineered approach, we also implement an encoder based on the features extracted by a pretained CNN. Our CNN encoder has an advantage over the Object-Based encoder in that it has been trained on far more image data, but has a disadvantage in that it lacks domain-specific feature engineering. We use pretrained VGG16 (Simonyan and Zisserman, 2014), which is a 16-layer CNN trained on ImageNet that produces a 4096-dimensional vector for each image. We compute this vector for each frame in the clip, and then compute the following features along each dimension in order to get a vector representation of the full clip: start_value, end_value, min, max, mean.

3.2.2 Dimensionality Reduction

Given a matrix M that maps each word instance to a feature vector using one of the encoders above, we run dimensionality reduction to get a 10d vector[7] for each word instance. We consider two settings. In the unsupervised setting, we run vanilla SVD. In the supervised setting, we run supervised LDA in which the "labels" are the words uttered at the start of the clip as described in §3.1.

3.3 Evaluation

We evaluate our models in terms of their precision when assigning verbs to unseen clips. Specifically, for our two heldout subjects, we partition the full session into consecutive 5-second clips, resulting in 189 clips total. For testing, unlike in training, we include all clips, even those in which the subject is not speaking. Then, for each model, we encode each clip using the model's encoder and then find the verb with the highest cosine similarity to the encoded clip. The authors then view each clip alongside the predicted verb and make a binary judgement for whether or not the verb accurately depicts the action in the clip, e.g. *yes or no, does the clip depict an instance of "pick up"?* To avoid annotation bias, all four models plus a random baseline are shuffled and evaluated together, and annotators do not know which prediction comes from which model. Annotator agreement was high (91%).

3.4 Results and Analysis

Table 3 reports our main results for each model. We compute both "strict" precision, in which a prediction is only considered correct if both annotators deemed it correct, as well as "soft" precision, in which a prediction is correct as long as one annotator deemed it correct. As the results show, no model performs especially well. Random guessing achieves 32% (soft) precision on average. The supervised Object-Based model and the unsupervised CNN model both perform a bit better (40% on average), but we note that the samples are small and we cannot call these differences significant (see 95% bootstrapped confidence intervals given in Table 3). Only the unsupervised Object-Based model stands out in that it performs significantly worse than all other models (20% soft precision). For the CNN models, we do not see a significant difference with the supervised dimensionality reduction. Figure 3 shows example clips for each encoder.

Table 4 shows a breakdown of model performance by verb. We see a few intuitive differences between the CNN-based model and the Object-Based model, discussed below. We note these observations are based on a small number of predictions, and thus should be taken only as suggestive.

Low-level actions. The Object-Based models achieve higher precision on low-level verbs like *"pick"*, *"take"*, and *"hold"*. This makes intuitive

[7] 10d is chosen since we are only attempting to differentiate between 14 words, and thus our supervised LDA cannot use more than 13d.

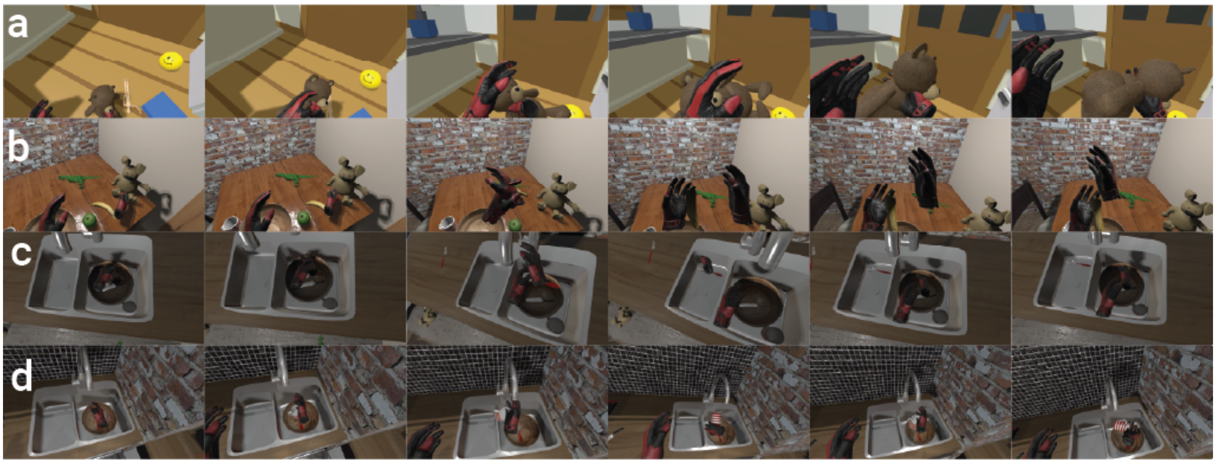


Figure 3: Example clips, subsampled to 6 frames. **(b)** is **(a)**'s nearest-neighbor using the Object-Based model. In each of these clips, the participant picks up an object with their right hand. **(d)** is **(c)'s** nearest-neighbor using the CNN. In each, the participant is washing dishes in a similar looking sink.

	Soft	Strict
Random	0.32 (0.25–0.39)	0.23 (0.17–0.29)
Obj.	0.20 (0.14–0.25)	0.13 (0.08–0.19)
CNN	0.40 (0.33–0.47)	0.29 (0.22–0.36)
Obj+Sup.	0.40 (0.33–0.47)	0.28 (0.22–0.34)
CNN+Sup.	0.35 (0.28–0.42)	0.25 (0.19–0.31)

Table 3: Precision of each method with 95% bootstrapped CI. "Soft" means a prediction is correct as long as one annotator considers it to be so; "strict" means prediction is only considered correct if both annotators agree that it is correct.

	CNN		Obj		Obj+Sup.	
	N	Prec.	N	Prec.	N	Prec.
pick	0	0.00	1	1.00	4	1.00
take	0	0.00	3	1.00	12	0.67
hold	11	0.64	5	0.80	17	0.65
get	32	0.56	5	0.00	13	0.54
go	29	0.21	1	0.00	17	0.47
put	4	0.00	11	0.27	31	0.29
play	7	0.29	17	0.18	6	0.17
walk	16	0.44	33	0.30	26	0.15
throw	36	0.08	25	0.00	16	0.06
drop	4	0.25	16	0.06	2	0.00
eat	2	0.00	2	0.00	19	0.00
give	17	0.00	35	0.00	5	0.00
open	8	0.00	30	0.00	10	0.00
wash	23	0.48	5	0.00	11	0.00

Table 4: Analysis of model precision broken down by verb. Top-level columns are the unsupervised CNN, unsupervised obj model, and supervised obj model.[8] For each, N is the number of times the model predicts that verb. Precision is the proportion of the time that prediction was correct.

sense, since the 3D spatial features are designed to capture these types of mechanical actions, independent of the objects with which they co-occur. The 2D visual data, on the other hand, may struggle to ground a visually diverse set of objects-in-motion to these low-level mechanical actions.

Visual cues. Some actions are strongly predicted by specific objects, which are well captured by visual cues. This is most obvious in the case of *"wash"*, on which the CNN achieves higher precision than the Object-Based models. This is again intuitive as *wash* tends to co-occur with a clear view of the sink, which is a large, visually-distinct part of the field of view.

Vague actions. Actions like *"go"*, *"walk"*, and *"hold"* occur frequently, even when the language signal does not reflect it. That is, in any given clip, there is a high chance that the participant walks, goes somewhere, or holds something. Thus, models which happen to predict these verbs frequently may have artificially high accuracy. For example, the unsupervised Object-Based model only predicts *"go"* once and *"hold"* 5 times , which may contribute to the unsupervised Object-Based model performing significantly worse than random, despite seeming to capture low-level actions well.

Special cases. We note that some verbs are very difficult or impossible to detect given limitations of our data. In particular, *"give"*, *"eat"*, and *"open"* have a precision of 0 across all models, as well as in the training signal (§2.3.2). For example, *"give"* only occurs twice in our data (*"fluffy teddy bear going to give it a little hug"* and *"turn on the water give it a little sore[sic] and we can let it dry there"*), but cannot occur in its prototypical sense since there is no clear second agent to be a recipient. During instances of *"eat"* and *"open"*, participants tended to mime the actions, but the in-game physics data does not faithfully capture the semantics of these verbs (e.g., containers do not actually open). These words highlight limitations of the environment which may be addressed in future work.

4 Discussion

We compare two types of models for grounded verb learning, one based on 2D visual features and one based on 3D symbolic and spatial features. Our analysis suggests that these approaches favor in different aspects of verb semantics. One open question is how to combine these differing signals, and how to design training objectives that encourage models to chose the right sensory inputs and time scale to which to ground each verb.

We evaluated on a small set of verbs that are acquired comparably early by children. Nonetheless, our models perform only marginally better than random. This disconnect highlights an important challenge to be addressed by work on computational models of grounded language learning: Can statistical associations between words and contexts result in more than simple noun-centric image or video captioning, eventually forming general-purpose language models? While that question is still wide open, research from psychology could better inform work on grounded NLP. For example, Piccin and Waxman (2007) argues that verb learning in particular is not learned from purely grounded signal, but rather is "scaffolded" by earlier-acquired knowledge of nouns and of syntax. From this perspective, the models we explored here, which are similar to what is used for noun-learning, are far too simplistic for verb learning. More research is needed on ways to combine linguistic and grounded signal in order to learn more abstract semantic concepts.

5 Related Work

We contribute to a large body of research on learning grounded representations of language. Grounded representations have been shown to improve performance on intrinsic semantic similary metrics (Hill et al., 2017; Vulić et al., 2017) as well as to be better predictors of human brain activity (Anderson et al., 2015; Bulat et al., 2017). Much prior work has explored the augmentation of standard language modeling objectives with 2D image (Bruni et al., 2011; Kiela et al., 2017; Lazaridou et al., 2015; Silberer and Lapata, 2012; Divvala et al., 2014) and video (Sun et al., 2019) data. Recent work on detecting fine-grained events in videos is particularly relevant (Hendricks et al., 2018; Zhukov et al., 2019; Fried et al., 2020, among others). Especially relevant is the data collected by Gaspers et al. (2014), in which human subjects were asked to play simple games with a physical robot and narrate while doing so. Our data and work differs primarily in that we focus on the ability to ground to symbolic objects and physics rather than only to pixel data. Past work on "situated language learning", inspired by emergence theories of language acquisition (MacWhinney, 2013), has trained AI agents to learn language from scratch by interacting with humans and/or each other in simulated environments or games (Wang et al., 2016; Mirowski et al., 2016; Urbanek et al., 2019; Beattie et al., 2016; Hill et al., 2018; Mirowski et al., 2016),

6 Conclusion

We introduce the New Brown Corpus, a dataset of spontaneous speech aligned with rich environment data, collected in a VR kitchen environment. We show that, compared to existing corpora, the distribution of vocabulary collected is more comparable to that found in child-directed speech. We analyze several baseline distributional models for verb learning. Our results highlight the challenges of learning from naturalistic data, and outlines directions for future research.

7 Acknowledgements

This work was supported by DARPA under award number HR00111990064. Thanks to George Konidaris, Roman Feiman, Mike Hughes, members of the Language Understanding and Representation (LUNAR) Lab at Brown, and the reviewers for their help and feedback on this work.

References

Andrew James Anderson, Elia Bruni, Alessandro Lopopolo, Massimo Poesio, and Marco Baroni. 2015. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, 120:309–322.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. 2016. Deepmind lab. *arXiv preprint arXiv:1612.03801*.

Emily M. Bender and Alexander Koller. 2020. 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. Association for Computational Linguistics.

Miguel Borges, Andrew Symington, Brian Coltin, Trey Smith, and Rodrigo Ventura. 2018. Htc vive: Analysis and accuracy improvement. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2610–2615. IEEE.

Michael R Brent and Jeffrey Mark Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–B44.

Roger Brown. 1973. *A first language: The early stages.* Harvard U. Press.

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 22–32, Edinburgh, UK. Association for Computational Linguistics.

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1219–1228. ACM.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1091, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277.

Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural naturalist: Generating fine-grained image comparisons. 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)*, pages 708–717, Hong Kong, China. Association for Computational Linguistics.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor*, 5(2):7.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2017. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677–694.

Daniel Fried, Jean-Baptiste Alayrac, Phil Blunsom, Chris Dyer, Stephen Clark, and Aida Nematzadeh. 2020. Learning to segment actions from observation and narration. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2569–2588, Online. Association for Computational Linguistics.

Judith Gaspers, Maximilian Panzner, Andre Lemme, Philipp Cimiano, Katharina J. Rohlfing, and Sebastian Wrede. 2014. A multimodal corpus for the evaluation of computational models for (grounded) language acquisition. In *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pages 30–37, Gothenburg, Sweden. Association for Computational Linguistics.

Jane Gillette, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition*, 73(2):135–176.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1380–1390, Brussels, Belgium. Association for Computational Linguistics.

Felix Hill, Karl Moritz Hermann, Phil Blunsom, and Stephen Clark. 2017. Understanding grounded language learning agents. *arXiv preprint arXiv:1710.09867*.

Felix Hill, Karl Moritz Hermann, Phil Blunsom, and Stephen Clark. 2018. Understanding grounded language learning agents.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2017. Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. pages 153–163.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Brian MacWhinney. 2013. The emergence of language from embodiment. In *The emergence of language*, pages 231–274. Psychology Press.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. 2016. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*.

Thomas B Piccin and Sandra R Waxman. 2007. Why nouns trump verbs in word learning: New evidence from children and adults in the human simulation paradigm. *Language Learning and Development*, 3(4):295–323.

Christopher Potts. 2019. A case for deep learning in semantics: Response to pater. *Language*, 95(1):e115–e124.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Elizabeth S Spelke and Katherine D Kinzler. 2007. Core knowledge. *Developmental science*, 10(1):89–96.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. 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)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2368–2378, Berlin, Germany. Association for Computational Linguistics.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3537–3545.