Learning English with Peppa Pig

Abstract

Attempts to computationally model or simulate the acquisition of spoken language via grounding in the visual modality have a long tradition but have gained momentum since around 2015 with the revival of neural networks. Current neural approaches are able to spot associations between the spoken and visual modality, and use these to represent speech and image/video data in a joint vector space. A major limitation of these works are the datasets used to train them. Most consist of static images or videos paired with spoken descriptions of what is depicted, and thus guarantee a strong correlation between speech and the visual world by construction. A child learning a language faces a very different and harder task: in the real world the coupling between the linguistic and the visual is much looser, and often contains confounds in the form of correlations with non-semantic aspects of the speech signal, such as voices of specific people and environmental sounds. The current study is a first step towards simulating such a naturalistic grounding scenario by using a dataset based on the children's cartoon *Peppa Pig*. We train a simple bi-modal architecture on the portion of the data consisting of naturalistic dialog between characters, and evaluate on segments containing descriptive narrations. Evaluation and analysis results indicate that despite the weak and confounded signal in this training data our model succeeds at learning aspects of the visual semantics of spoken language.

1 Introduction

Attempts to model or simulate the acquisition of spoken language via grounding in the visual modality date to the beginning of this century (Roy and Pentland, 2002) but have gained momentum since 2015. As noted by Chrupała (2021), most current approaches work well enough from an applied point of view but leave much to be desired as regards ecological validity. Most datasets consist of static images paired with their descriptions. Existing video datasets contain spoken descriptions of what happens in the video. The type of input that a child faces when learning a language is much more challenging. Firstly, speech is only loosely coupled with the visual modality. Secondly in addition to correlations between the visual scenes and the *meaning* of spoken utterances, there are also correlations with non-semantic aspects of the speech signal, such as the voice of specific characters and environmental noise. These non-semantic correlations make it harder for the learner to zoom in on those aspects of the audio signal most relevant to learning meanings of linguistic units.

In the current study we make a first step towards simulating such a naturalistic grounding scenario: we use the well-know children's cartoon *Peppa Pig* as a case study. Compared to commonly used video datasets, this data features a number of interesting characteristics. The visual modality is very schematic, and the language is also simple in terms of vocabulary size and syntactic complexity. Crucially, however, most of the speech in the videos consists of naturalistic dialogs between the characters. The utterances are only loosely and noisily correlated to the scenes and actions depicted in the videos. This choice of data thus allows us to directly address the ecological limitations of the current approaches.

We implement a model which learns to project visual scenes and spoken utterances into a join vector space, and train it on snippets of video containing dialog from the Peppa Pig cartoon, and carry out an in-depth evaluating of the nature of the learned representations using a variety of approaches.

2 Related work

3 Method

3.1 Dataset

The dataset consists of the set of videos of the English-language version of *Peppa Pig*. In addition to the raw videos we also use the annotation created by Papasarantopoulos and Cohen (2021).

These annotations feature written transcriptions aligned with the audio as well as segmentation into *dialog* and *narration*.[1] Dialogs are the parts spoken by the characters, while narrations are comments inserted by the narrator, which are more descriptive in nature. All the narration segments are uttered by the same actor. We use the dialogs for training the model, and set aside the narrations for evaluation purposes only. A small portion of the dialog data is also used for evaluation.

Specifically, we use dialog from episodes 1–196 for training, and 197–209 for validation. We set aside narrations from episodes 1–104 for validation and 105–209 for testing.

3.2 Preprocessing

For training, we do not use word or sentence level segmentation in order to make the setting more naturalistic. Instead we split the dialog sections into **XXX** second non-overlapping fragments. The video is subsampled to 10 frames per second, and to 180 × 100 resolution. The audio is converted to mono by averaging the two channels and the raw waveform is used as input.

For evaluation we have a number of different conditions and evaluation metrics described in detail in Section 3.3 and in some of these conditions we use the subtitles to guide segmentation. Table 1 shows the basic statistics of the training and validation splits.

[1] It should be noted that the quality of the alignment and segmentation in the original dataset is variable. In cases where exact alignment is needed, such as for word-level analyses, we re-align the transcriptions using `github.com/lowerquality/gentle`.

Split	Type	Triplet	Size (h)	Items	Mean length (s)
train	dialog	No	9.83	11058	3.20
val	dialog	No	0.65	729	3.20
val	narration	No	0.80	897	3.20
test	narration	No	0.51	570	3.20
val	dialog	Yes	0.65	828	2.81
val	narration	Yes	0.92	1492	2.21
test	narration	Yes	0.71	1052	2.44

Table 1: Dataset statistics. For the triplet condition, videos are split such that each segment corresponds to a line of subtitles. For the non-triplet condition, videos are split into 3.2s segments.

3.3 Evaluation

The most common approach to evaluation for visually grounded models trained on spoken image captions is caption-to-image retrieval (often combined with image-to-caption retrieval): in fact this technique is has been carried over from text-based image-caption modeling Chrupała (2021). With the standard spoken caption dataset this approach is unproblematic since the content of the captions is not correlated with extra-linguistic clues in the speech signal, such as speaker identity (since speakers are randomly assigned to captions) or non-speech environmental noise. Thus in this setting, a retrieval metric measures the ability of the model to match spoken utterances to images based on their semantic content. This not the case for the *Peppa Pig* dataset: here we can expect that when a video segment depicts a particular character (e.g. George) then the audio in this segment is more likely to contain utterances spoken by the voice actor playing George. George has a favorite toy dinosaur: when this toy appears in a video segment we can likewise expect higher than random chance of George's voice in the audio. Due to these factors, in a naive retrieval setting, a model could obtain a high score by mostly capturing these non-linguistic correlations.

In order to (partially) alleviate these concerns we leverage the narrator speech in the videos. These utterances are always spoken by the same actor, so speaker identity cannot be used as a clue for matching video and audio. Furthermore, the narration segments are akin to video captions in that they tend to describe what is happening in the video and thus their semantic content is more strongly correlated with the content of the video than in the case of the dialog, which is also a desirable feature for the purposes of system evaluation.

Retrieval For the retrieval evaluation, as for training, we use fixed segmentation into XXXXXXXs clips. We encode each audio clip in the validation (or test) data using the speech encoder part of the model; we encode each video clip using the video encoder. We then measure cosine similarity between the audio clip and all the video clips. If the video clip corresponding to the audio is among the n most similar video clips, we count that as a success. The proportion of successes across all audio clips gives us the retrieval metric known as recall@n: specifically in this paper we focus on $n = 10$.

GC: Add fixed vs jitter condition for retrieval too.

Triplets Retrieval metrics such as recall@10 have some disadvantages. Firstly the absolute value of this metric is hard to interpret as it depends crucially on the size of the candidate set (e.g. the size of the validation/test set). Thus these numbers cannot be directly between different datasets. Secondly, if we wanted to compare model performance with human performance, we could not feasibly ask human participants to provide the quadratic number of audio-video similarity judgments needed. For these reasons we evaluate model performance using the following simplified, controlled scenario: We extract clips aligned to a single subtitle line, group them by length, and for each pair of same-length video clips[2], we extract the audio from one of them (selected at random) – this is our *anchor*. The video clip from which the anchor was taken is the *positive* one, which the other video clip is the *negative* one. This triplet of stimuli form a single test item. We use the model's audio encoder to encode the anchor, and the video encoder to encode both video clips. We then check whether anchor is more similar to the positive or negative clips in terms of cosine similarity. More precisely, *triplet accuracy* is the mean over all triplets of the following quantity:

$$\frac{\text{signum}(\text{cosine}(A, P) - \text{cosine}(A, N)) + 1}{2} \tag{1}$$

with A being the anchor, P positive and N negative. The triplet accracy metric is inspired by the ABX score of Schatz (2016). For triplet accuracy, regardless of the specific set of test items, we expect random-guessing performance to be at 0.5, and perfect performance to be 1.0. To improve the reliability of this metric and provide information on its variance, triplets can be resampled N times from the dataset, and the mean accuracy and spread around the mean reported.

Targeted Triplets Inspired by 2-alternative forced choice (2AFC) paradigms in child language acquisition (Noble et al., 2011; Bergelson and Swingley, 2012), we design test trials that test the model's acquisition of grounded semantics under controlled circumstances.

The test can be seen as a special case of the triplets evaluation as described in the previous paragraph. Here, we aim to assess the model's acquisition of the semantics of commonly occurring words. Therefore, the *targeted* approach considers, in contrast to the general triplets evaluation, always pairs of triplets with *minimal differences* regarding one word in the transcripts of the anchor audios (e.g., "Peppa loves jumping" and "George loves jumping" can be used to test whether the model can discriminate the target word "Peppa" from the distractor word "George").

We search the transcripts of the validation data for phrases with minimal differences with respect to the most commonly occurring nouns and verbs.[3] Details on this search procedure can be found in Appendix A.1. Based on each pair of phrases, we create two counter-balanced test trials, an example and a corresponding counter-example as depicted in Figure 1. Here, the anchor A_x of the example triplet is the audio of "Peppa loves jumping" (a_p), the positive video P_x is the corresponding video (v_p) and the negative video N_x is the video corresponding to "George loves jumping" (v_g): $(A_x, P_x, N_x) = (a_p, v_p, v_g)$. In the

[2] To keep test items independent, the pairing of video clips is done such that each clip only occurs as a member of a single triplet.

[3] There were not enough adjectives in the dataset perform a comprehensive analysis of their acquisition.

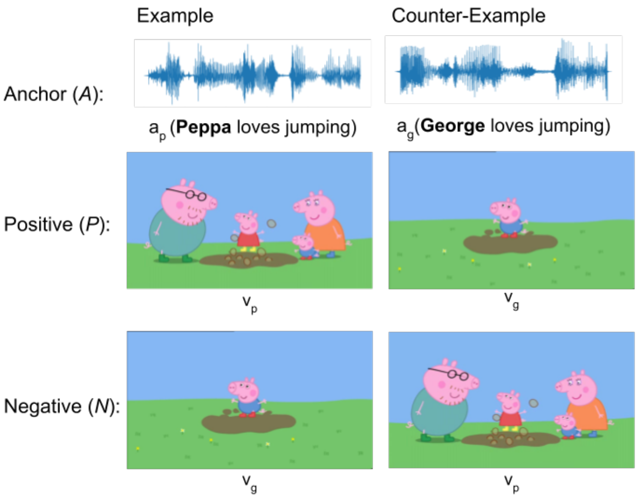

Figure 1: Targeted Triplets Evaluation

counter-example triplet, the anchor A_y is the audio of "George loves jumping", and the positive and negative video are flipped: $(A_y, P_y, N_y) = (a_g, v_g, v_p)$. In this way, we control the evaluation for linguistic biases in the dataset and ensure that a single-modality model that only considers the audio performs at chance (Nikolaus and Fourtassi, 2021).

We measure overall targetet triplet accuracy using Equation (1). Additionally, we report per-word accuracy by calculating the triplet accuracy for all triplets that contain a given word (e.g. "Peppa") either as target or distractor word, i.e. cases in which the model needs to succeed in either choosing a video containing the given word (the example triplet in Figure 1) or rejecting a video containing the given word (the counter-example triplet in Figure 1).[4] We report accuracy for all words for which we found at least 100 example-counterexample pairs of triplets. Results are reported both for nouns and verbs, however there were not enough examples for adjectives in the dataset to perform a targeted triplets evaluation on them.

3.4 Model

We adapt the high-level modeling approach from work on spoken image-caption data (Harwath et al., 2016; Chrupała et al., 2017): our objective function is based on a triplet-loss with margin which encourages the matching audio and video clip to be projected nearby in the embedding space, and mis-matching audio and video clips to be far away:

$$\ell = \sum_{av} \left[\sum_{a'} \max(0, S_{a'v} - S_{av} + \alpha) + \sum_{v'} \max(0, S_{av'} - S_{av} + \alpha) \right] \tag{2}$$

where α is a margin, S_{av} is a similarity score between a matching audio-video clip pair, and $S_{a'v}$ and $S_{av'}$ denote similarity scores between mismatched pairs,

[4] Both the example and the counter-example shown in Figure 1 are also used to assess the acquisition of the word "George".

ID	Static	Jitter	Pretraining	Resolution	R@10 (fixed)	R@10 (jitter)	Triplet Acc
68			AV	180x100	0.561	0.533	0.834
206974		Yes	AV	180x100	0.557	0.539	0.830
206975		Yes	A	180x100	0.391	0.393	0.744
206976		Yes	V	180x100	0.220	0.205	0.630
206977		Yes	None	180x100	0.194	0.195	0.612
206978	Yes	Yes	AV	180x100	0.442	0.430	0.781
206964		Yes	AV	360x200	0.594	0.589	0.852

Table 2: Retrieval and triplet scores on dialog validation data.

i.e. negative examples from the current batch. Our heuristic to generate positive and negative examples is very simple: namely we consider the example positive if the audio is exactly aligned with a video clip in our data. All other pairs of audio-video clips are considered negative.

The audio encoder portion of the model consists of a `small wav2vec2` model (Baevski et al., 2020) pretrained in an self-supervised fashion, with supervised fine tuning.[5] During training, we keep the feature extractor and the bottom $K = 3$ transformer layers of this encoder frozen. Its output is pooled across time using an attention mechanism with dimensionwise weights (Merkx et al., 2019):

$$\begin{aligned}\mathbf{A} =& \text{softmax}_t \left(\text{MLP}(\mathbf{X})\right) \\ \mathbf{z} =& \sum_t \left(\mathbf{A}_t \odot \mathbf{X}_t\right),\end{aligned} \tag{3}$$

where $\mathbf{X}$ is the tensor with the encoder output vectors for each time-step: an MLP followed by a time-wise softmax is used to compute an attention weight for each time step and for each dimension. The pooling is followed by a linear projection and L_2 normalization.

As a video encoder we use the 18-layer ResNet (2+1)D architecture (Tran et al., 2018) pretrained on the action recognition dataset Kinetics-400 (Kay et al., 2017). The pretrained model is available via Pytorch.[6] The output of this module is aggregated using the attention mechanism with the same architecture as for the audio module, linearly projected to the same dimensionality as the audio (512) and L_2 normalized.

4 Results

Performance metrics Table 2 and Table 3 show the performance of several model configurations on the retrieval and triplet tasks on the dialog and narration datasets respectively.

In the case of the narration data this scores is not confounded by speaker-based clues, which is an indication that the model possibly learned to detect some aspects of utterance meaning. We investigate this hypothesis further using multiple representational similarity analysis.

[5] Available from `https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt`.

[6] See `https://pytorch.org/vision/0.8/models.html#resnet-3d`.

ID	Static	Jitter	Pretraining	Resolution	R@10 (fixed)	R@10 (jitter)	Triplet Acc
68			AV	180x100	0.644	0.626	0.881
206974		Yes	AV	180x100	0.645	0.623	0.882
206975		Yes	A	180x100	0.585	0.574	0.869
206976		Yes	V	180x100	0.316	0.318	0.774
206977		Yes	None	180x100	0.362	0.346	0.770
206978	Yes	Yes	AV	180x100	0.556	0.555	0.850
206964		Yes	AV	360x200	0.691	0.685	0.904

Table 3: Retrieval and triplet scores on narration validation data.

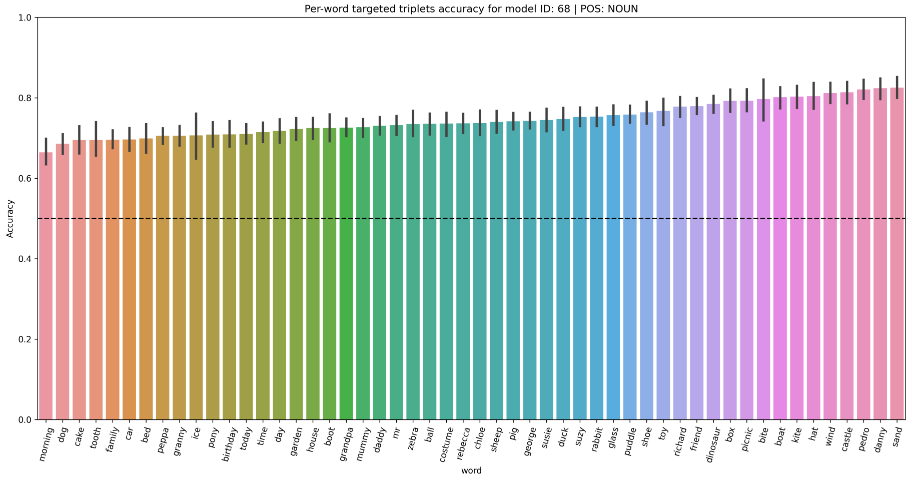

Figure 2: Per-word targeted triplets accuracy for nouns.

Targeted Triplets As a first baseline, we evaluate a model that has been pretrained but not fine-tuned on our dataset. The resulting performance is, as expected, close to chance level: 0.538 . Additionally, we evaluate a model that is trained using static (image) data instead of video. The average accuracy is 0.705 . Finally, the best performing model according to the performance metrics (ID 68, audio and video pretraining) achieves an average targeted triplets accuracy of 0.745 .

Figure 2 and Figure 3 show per-word accuracy for nouns and verbs, respectively. The vertical bars indicate the standard deviation as estimated using bootstrapping tests.

We further compute correlations between the per-word accuracy and two possible predictors of age of acquisition: frequency and concreteness. The resulting correlations are presented in Appendix A.1. We do not find any significant correlation between the model's per-word accuracy and word concreteness or input frequency of a word in the training data.

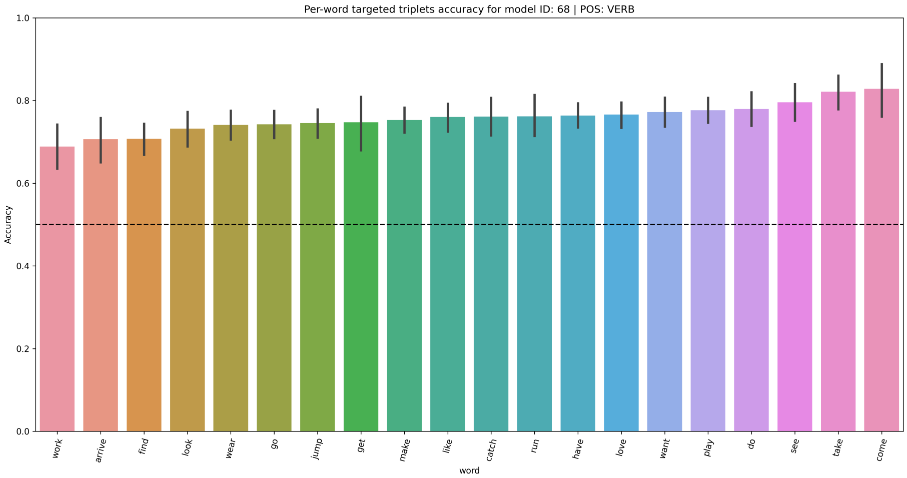

Figure 3: Per-word targeted triplets accuracy for verbs.

5 Conclusion

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Bergelson, E. and Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

Chrupała, G. (2021). Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. Preprint: `https://arxiv.org/abs/2104.13225`.

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

Harwath, D. F., Torralba, A., and Glass, J. R. (2016). Unsupervised learning of spoken language with visual context. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1858–1866.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python. *Zenodo*.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. (2017). The kinetics human action video dataset. *CoRR*, abs/1705.06950.

Merkx, D., Frank, S. L., and Ernestus, M. (2019). Language Learning Using Speech to Image Retrieval. In *Proc. Interspeech 2019*, pages 1841–1845.

Nikolaus, M. and Fourtassi, A. (2021). Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 200–210, Online. Association for Computational Linguistics.

Noble, C. H., Rowland, C. F., and Pine, J. M. (2011). Comprehension of argument structure and semantic roles: Evidence from english-learning children and the forced-choice pointing paradigm. *Cognitive science*, 35(5):963–982.

Papasarantopoulos, N. and Cohen, S. B. (2021). Narration generation for cartoon videos. Preprint: `https://arxiv.org/abs/2101.06803`.

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

Schatz, T. (2016). *ABX-discriminability measures and applications.* PhD thesis, Université Paris 6 (UPMC).

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

A Supplementary material

A.1 Targeted Triplets Evaluation Sets

To find commonly occurring nouns, adjectives, and verbs, we lemmatize and POS-tag all words in the transcripts of the validation dataset using spacy (Honnibal et al., 2020). Afterwards, we identify sets of all nouns $\{n_1, ..., n_n\}$, verbs $\{v_1, ..., v_o\}$ and adjectives $\{a_1, ..., a_p\}$ that occur at least 10 times in the validation data. Given these sets, we create sets of tuples $\{(n_1, n_2), (n_1, n_3), ..., (n_1, n_n), ..., (n_{n-1}, n_n)\}$ for all combinations of nouns and verbs, respectively. For each of these tuples, we search the validation data for pairs of phrases $(p_k = [w_1, ..., w_x], p_l = [w_1, ..., w_y])$ with same length $(x = y)$ and minimal difference regarding the tuple. That is, $n_1 \in p_1$, $n_2 \in p_2$, and if we replace n_1 with n_2 in p_1, it is equal to p_2.

For example, if n_1 = "peppa" and n_2 = "george", the phrases p_1 = ["peppa", "loves", "jumping"] and p_2 = ["george", "loves", "jumping"] are phrases with minimal differences. A phrase can also be a single word.

We set the minimum phrase duration to 0.3 seconds (for shorter sequences, we do not expect that the video data contains enough semantic information for a model to distinguish between target and distractor). For each phrase p_1 we look for the *longest* possible phrase p_2. Figure 4 shows the distribution of samples per duration, Figure 5 per number of tokens.

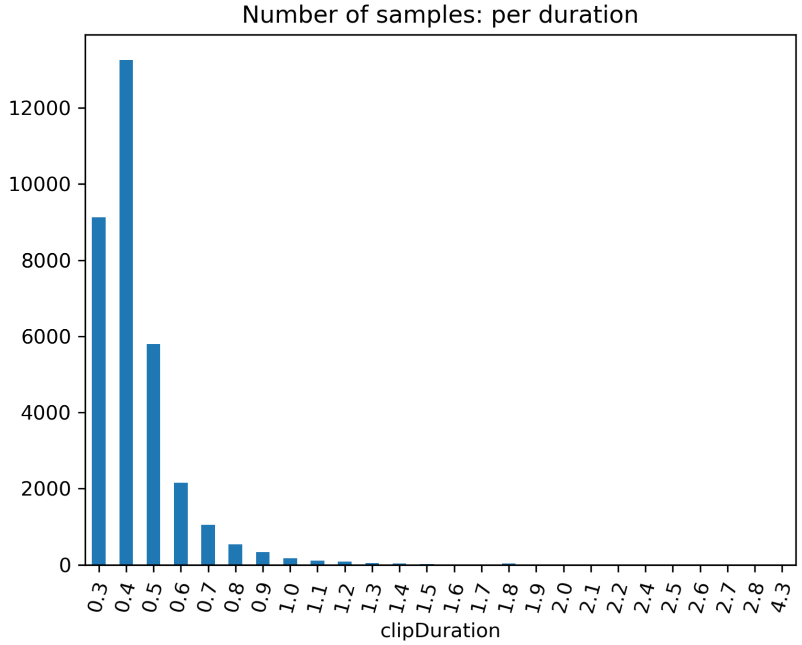

Figure 4: Number of samples per duration

Based on each minimal pair, we construct two counter-balanced test triplets as described in the main text.

Figures 6, and 7 show the number of samples for each noun and verb for which at least 100 sets of test triplets were available (for no adjective there were enough samples found).

A.2 Targeted Triplets Correlations

Figure 8 shows the correlation between per-word accuracy and frequency of this word in the training data. Figure 9 shows the correlation between per-word accuracy and concreteness scores.

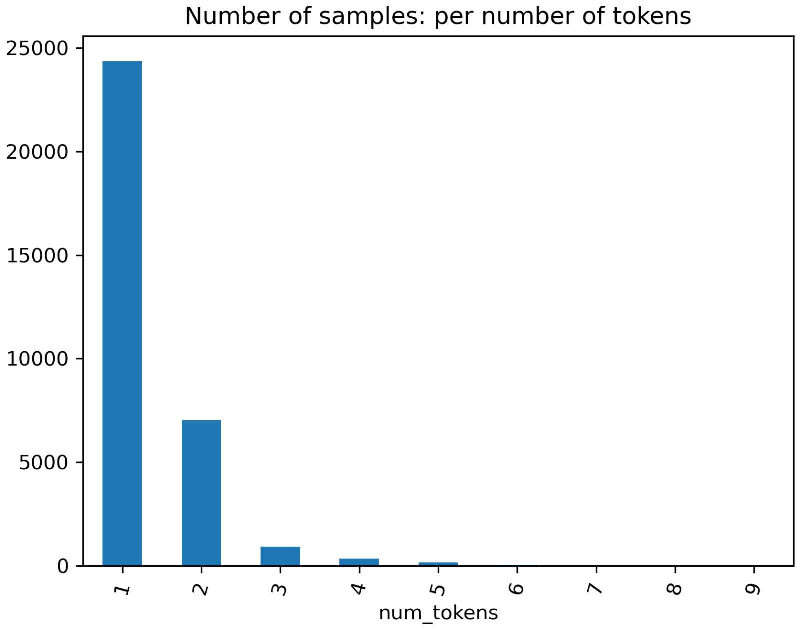

Figure 5: Number of samples per number of tokens

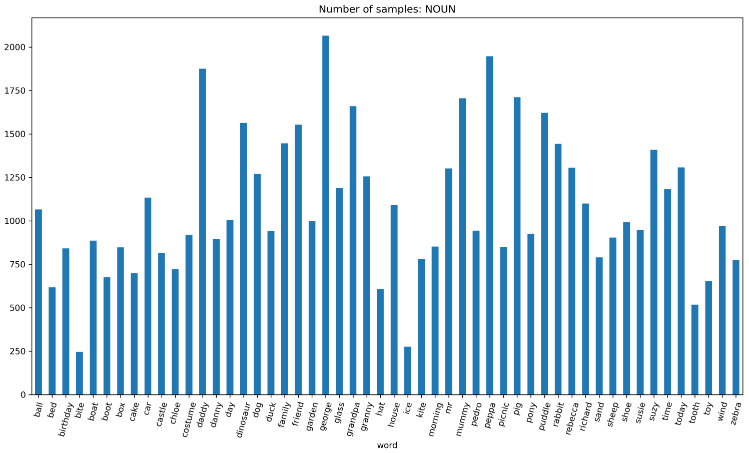

Figure 6: Number of samples: nouns

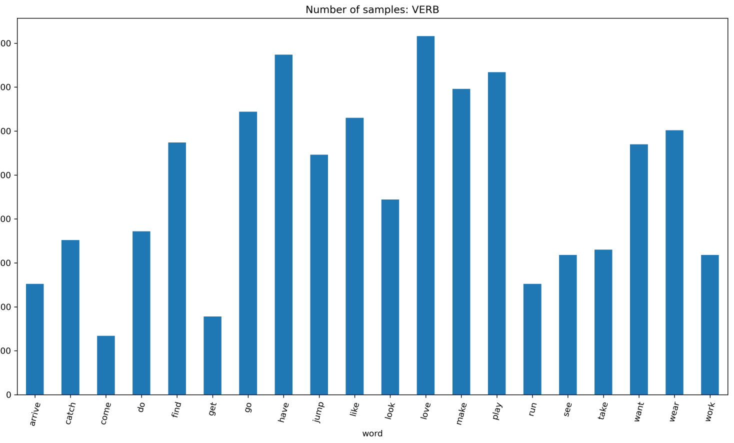

Figure 7: Number of samples: verbs

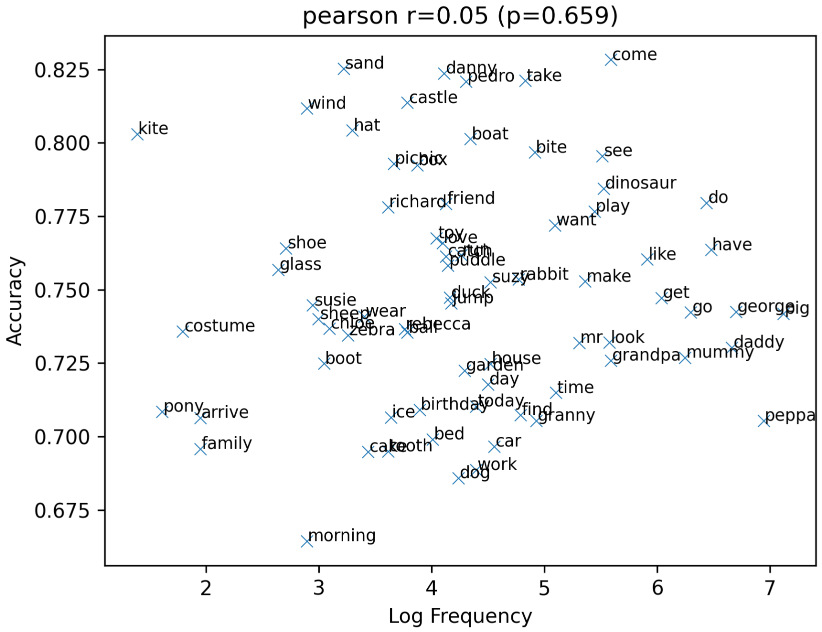

Figure 8: Correlation between accuracy and log frequency.

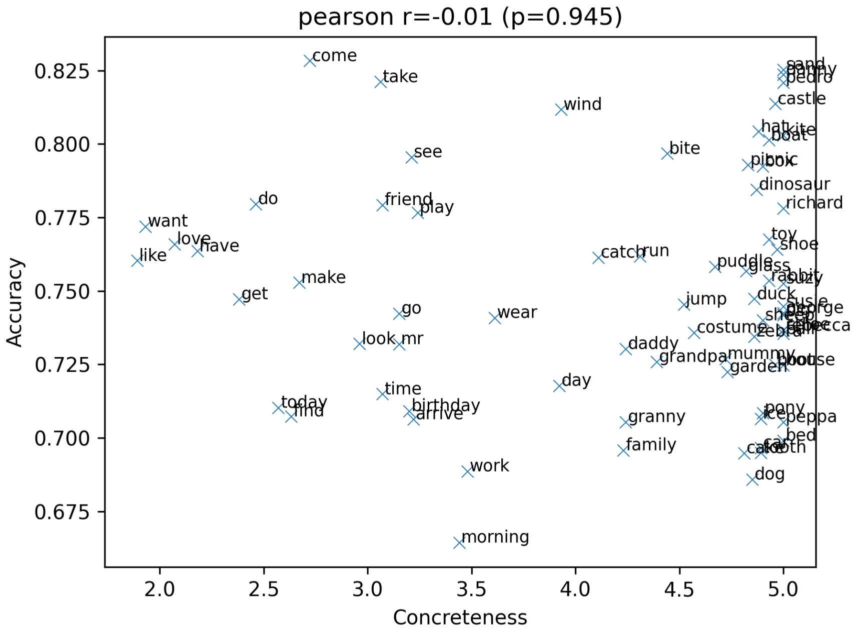

Figure 9: Correlation between accuracy and concreteness.