Phonetic and Lexical Discovery of Canine Vocalization

Theron S. Wang[1*], **Xingyuan Li**[2*], **Chunhao Zhang**[3], **Mengyue Wu**[4†], **Kenny Q. Zhu**[5†]
[1,5]University of Texas at Arlington, Arlington, Texas, USA
[2,3,4]Shanghai Jiao Tong University, Shanghai, China
{[1]sinong.wang,[5]kenny.zhu}@uta.edu,
{[2]xingyuan,[3]forest_zch,[4]mengyuewu}@sjtu.edu.cn

Abstract

This paper attempts to discover communication patterns automatically within dog vocalizations using a data-driven approach, which breaks the barrier that exists in previous methods that heavily rely on human prior knowledge of limited data. We present a self-supervised approach with HuBERT, enabling the accurate classification of phones, and an adaptive grammar induction method that identifies phone sequence patterns suggesting a preliminary vocabulary within dog vocalizations. Our results show that a subset of this vocabulary has substantial causal relations with certain canine activities, suggesting signs of stable semantics associated with these "words."

1 Introduction

The concept of "animal language" hinges on the intricate ways through which non-human species communicate, revealing a spectrum of vocalizations, gestures, and behavioral cues that resemble humans' capabilities to convey information and emotions. The study of animal communication has captivated researchers across numerous disciplines, from biology to linguistics (Rutz et al., 2023; Paladini, 2020; Robbins, 2000; Pardo et al., 2024).

While it is commonly accepted that most animals lack a language system comparable to human languages, recent advancements in natural language processing have opened up new avenues for investigating the patterns and structures embedded within animal vocalizations (Huang et al., 2023; Wang et al., 2023; Sharma et al., 2024). The study of animal language abilities, therefore, not only broadens our understanding of the animal kingdom but also deepens our insights into the evolution and functions of communication systems across species.

Despite observable acoustic variations in canine vocalizations that hint at potential patterns of communication, asserting that dogs possess a language is fraught with limitations. The absence of identifiable phones and structured syntax in their vocalizations challenges traditional linguistic frameworks (Holdcroft, 1991). While dogs exhibit a range of sounds that vary in pitch, duration, and intensity, as illustrated in Figure 1 (sound clips extracted from Audioset (Gemmeke et al., 2017)), six distinct dog barking sounds are identified. However, these variations alone do not constitute a language. The lack of a structured, consistent phonetic system and the inability to form complex ideas through their sounds significantly limit the comparison of their vocalizations to human language.

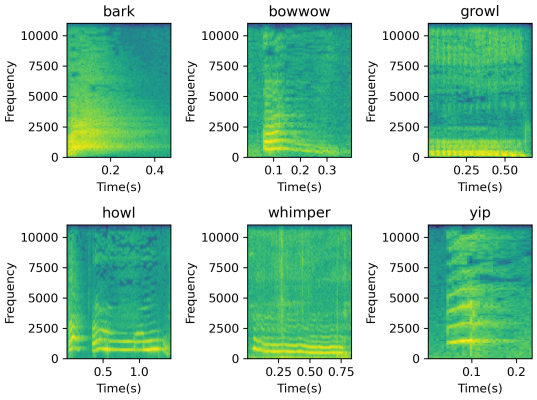

Figure 1: Spectragrams of six different dog barking sounds from AudioSet (Gemmeke et al., 2017).

Nevertheless, undertaking the challenge of exploring the concept of animal language is a daunting task. Unlike well-researched human language, the frequency range and phonetic variations remain underexplored, rendering the classification approach based on sound amplitude inadequate for discerning the fundamental phones of dog vocalizations. The difficulty lies not only in interpreting the acoustic variations but also in identifying meaningful patterns within the vast array of animal sounds. This complexity is compounded by

[*] These authors made equal contribution.
[†] Corresponding authors.

the need to distinguish between mere noise and significant vocalizations that could indicate some form of structured communication. The endeavor requires innovative methodologies and a departure from traditional linguistic analysis.

Our approach to navigating these challenges involves leveraging advanced signal processing techniques and self-supervised learning models. By focusing on the acoustic features of dog vocalizations, we aim to uncover underlying patterns that suggest an elementary form of phonetics. This involves a systematic process of audio cleanup, sentence extraction, phone recognition, and combination across vocalizations from different individual dogs. Given the lack of prior knowledge of dog vocalizations, we apply a self-supervised approach, HuBERT (Hsu et al., 2021), for representation pretraining and phone identification. HuBERT can effectively reference the contextual information of the audio and generate vector representations, which provides robustness when faced with vocalizations that have slight variations in context.

Utilizing the precise sequences of phones acquired from various dog vocalizations, we explored the feasibility of creating a vocabulary of frequently occurring subsequences of phones, which we call "words" here [1]. We identify these words by applying an adaptor grammar induction algorithm (Zhai et al., 2014) on a simple but universal syntax. We assess the validity of a word by evaluating the statistical correlation between its occurrence and the dog activities before, during, and after the utterance of the word, which serves as an indication of consistent lexical semantics. Our analysis revealed that a sizable portion of the vocabulary possesses semantics.

Our contributions lie in three aspects:

1. We developed a first-of-its-kind automatic pipeline for transcribing dog vocalizations into a sequence of phones (with unique labels) and further parsing it into a sequence of words with semantics;

2. Utilizing a shallow grammar, we were able to discover a vocabulary of words without supervision that covers 91.89% of phones in more than 34,000 dog "sentences";

3. By classifying the dog activity events surrounding the dog vocalizations and computing the causal strength between words and surrounding activities, we validated that 87.1% of the discovered words carry meanings.

2 Data Collection and Segmentation

In this section, we detail our workflow[2] (Figure 2), including audio cleanup by AudioSep, sentence extraction, dataset preparation, phone recognition, word discovery, and semantics discovery. The first three parts are primarily intended to gather large amounts of high-quality data on dog barking, while the last three parts aim to investigate potential phones and words in dog vocalizations.

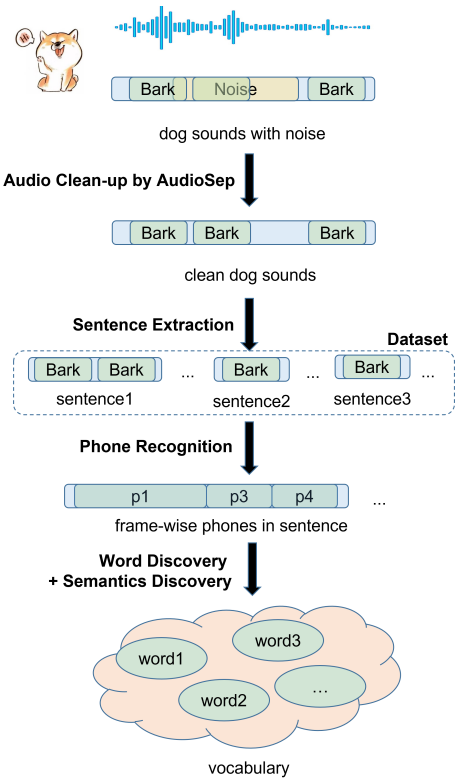

Figure 2: A workflow to discovery phones and words in canine vocals.

2.1 Audio Clean-up by AudioSep

A mixture of dog sounds and noises will inevitably occur due to the use of videos from public social media. These noises can include background music edited in by the video uploader, human speech, toy noises, etc. We expect a cleaner dataset, so we need to separate dog sounds from all mixed audio. In this work, we use AudioSep (Liu et al.,

[1]For ease of discussion, we are reusing the terms "phone", word", and "sentence" in a similar way as we do for human languages, despite the lack of universal agreement on the linguistic capability of dogs. We provide the definitions of "phone", "word", and "sentence" in Sec. 3.1, Sec. 3.2, and Sec. 2.2, respectively.

[2]The code and dataset are available at https://github.com/UTA-ACL2/canineLexical

2023), a foundation model for open-domain audio source separation with natural language queries. AudioSep is pre-trained on large-scale multimodal datasets, including the AudioSet (Gemmeke et al., 2017), VGGSound (Chen et al., 2020), and AudioCaps (Kim et al., 2019) datasets, etc. We apply AudioSep, using "Dog" as the input text query, to separate dog sounds from all audio. After this step, we will be working with higher-quality dog vocals.

2.2 Sentence Extraction

After the separation of dog sounds, the audio will contain mostly barking, silence, and a small amount of noise that cannot be removed. Next, we segment the dog vocalization audio clips into "sentences," each containing a sequence of dog vocals using an approach similar to Huang et al. (2023), which defines a sentence as *a continuous sequence of dog barks with no more than 0.5 seconds of silence in between*. We apply PANNs (Kong et al., 2020) pretrained on the large-scale AudioSet dataset to extract dog sentences. In order to acquire higher quality data, we initially manually labeled some dog barking data from our dataset. The data consisted of 1,483 dog barking audio clips with a total duration of more than 9,597 seconds. We utilized this data to fine-tune the pre-trained PANNs model and achieved an F1 score of 0.69. Then, consecutive clips less than 0.5 seconds apart are combined to form a dog "sentence."

To further reduce noise, sentences that meet one of the following conditions will be removed: (1) "Dog" label score less than 0.1; (2) One of the top 10 sound event tags is not related to dogs and has a score greater than the score for the dog tag, or the difference between the score and the score for the dog tag is less than 0.7. With this process, we obtain more and cleaner sentences than Huang et al. (2023).

To verify the effectiveness of AudioSep and to confirm that it did not have a significant impact on sound quality, we manually labeled 1,467 seconds (1,137 seconds for training and 330 seconds for testing) of dog barking audio to fine-tune PANNs. The results (Table 1) show that AudioSep can effectively reduce noise interference without significantly impacting the quality of dog sounds. The setting with the highest $F1$ score is adopted for all subsequent experiments and analyses.

Train Data	Test Data	F1 Score
-	-	0.6916
✓	-	0.6797
-	✓	**0.7755**
✓	✓	0.7709

Table 1: The result of PANNs. Those with a ✓ mark use AudioSep.

2.3 Dataset

Our raw dataset contains more than 6.8k YouTube videos of various types of dogs, including those from Huang et al. (2023) and AudioSet (Gemmeke et al., 2017). After the sentence extraction in our workflow, we obtained 37,919 dog "sentences"(more than 23 hours) from more than 1,300 users. Compared to previous studies(Huang et al., 2023; Abzaliev et al., 2024; Yin and McCowan, 2004), our dataset significantly exceeds them in terms of dog barking duration and the number of dogs.

Obviously, mispronunciation problems can arise in human speech datasets due to various reasons, such as non-native speakers, speech disorders, or simply errors in articulation. This might also be the case with our dog dataset or other dog datasets (Huang et al., 2023; Abzaliev et al., 2024; Yin and McCowan, 2004). With the largest amount of high-quality dog data compared to previous studies, we can mitigate the impact of mispronunciation problems. Furthermore, our dataset may be easily extended.

3 Phonetic and Lexical Discovery

In this section, we detail our approach in training a model to identify canine phones from the above segmented data and discover possible semantic units.

3.1 Phone Recognition

After extracting clean dog vocaliztions, we endeavor to discover *minimal sound units* (referred as phones) from these clips. In our study, we have meticulously isolated distinct vocalizations from dogs, aiming to identify the fundamental sound components, which we refer to as phones, within these recordings. To achieve this, we have employed HuBERT (Hsu et al., 2021), a self-supervised learning technique that assimilates acoustic and linguistic data from ongoing audio streams. Since there is no established phone set for canine vocalizations, it is difficult to manually label each audio clip with designated phone labels.

Such a self-supervised method has been instrumental in our analysis of canine vocalizations. The research by Hsu et al.'s (2021) has already established HuBERT's proficiency in delineating the nuances of human speech, where it is posited that the output classes from HuBERT could be analogous to phones (or sub-phones) in the context of human languages, serving as carriers of phonetic information.

Specifically, we pretrain a Dog-HuBERT using all sequences of dog vocalizations, which total more than 20 hours of dog sentences. We used 100 clusters at the first stage and 150 clusters at the second stage, a learning rate of 0.0001, and 80k training steps at the first stage and 60k training steps at the second stage. The other settings are the same as those in Hsu et al. (2021). Then, we used features from the 11th transformer layer of the second-stage model to train a K-Means model with 140 clusters. Figure 3 illustrates the clusters, indicating dog phones and noise labels, respectively. The clusters are determined based on the sum of distances from each sample to its cluster center for different clusters (Figure 4). It is a good choice when the curve flattens out, allowing for enough clusters without adding too many noise labels. Clearly, the cluster centers are evenly distributed, and the noise phones are mainly concentrated in one corner of the image, indirectly indicating the reliability of the phone discovery results.

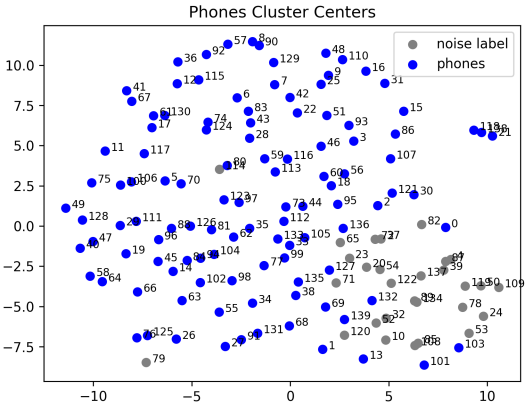

Figure 3: t-SNE plot of 140 different phones from Dog-HuBERT.

We regard the distinct classes produced by Dog-HuBERT as the basic units of canine vocal expression. Consequently, we define these output classes as the fundamental components within a dog's bark, dubbing them the dog's "phones." This terminology aligns with the concept of phones in linguistics, signifying the smallest units of sound that carry meaning.

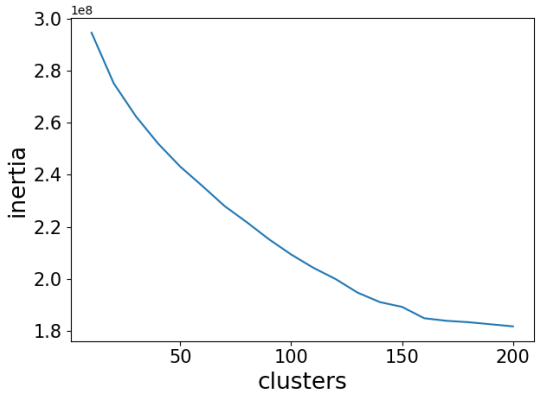

Figure 4: Inertia under different number of clusters K. 140 is picked as the optimal number of clusters due to the "Elbow Method."

3.2 Word Discovery

We define a "word" as *the smallest sequence of phones that consistently appears in one or more specific situations*. In our proposed pipeline, after the phone recognition step, we acquire a set of phones and can transcribe each sentence into a phone sequence. We continue to explore the potential words from these sentences.

The lack of prior knowledge in animal language prevents us from using discriminative deep learning methods (Baevski et al., 2021), while brute-force methods fail to capture the language structure. Adaptor Grammar (Johnson et al., 2006) can statistically learn recurrent sequence segments and build context-free grammar (see Appendix **??** for more discussion about Probabilistic Context-Free Grammars (PCFGs)). Following a previous well-performing method, we adopt Hybrid Variational-MCMC Inference (Zhai et al., 2014) to train the parameters of the Adaptor Grammar, ultimately obtaining a candidate vocabulary.

We use $\mathbf{S}_i = \{p_j | 1 \leq j \leq J_i\}$ to denote a sentence, or a series of phones, where a phone $p_j \in \mathbb{P}$, in which $\mathbb{P}$ is a set of integers, and J_i stands for the length of sentence i.

To discover the latent hierarchical linguistic structures in canine sentences, we use underlines to indicate adapted non-terminals and use $^+$ to indicate right-branching recursive rules for non-terminals.

$$\text{Sentence} \rightarrow \underline{\text{Word}}^{+}$$

$$\underline{\text{Word}} \rightarrow \text{Phone}^{+}$$

$$\text{Phone} \rightarrow p \quad \text{for } p \in \mathbb{P}$$

We treat the nonterminal *Word* as an adapted nonterminal, learning the relationship between *Word* and observation segments during training. We parse the entire training data using the trained model, as shown in Figure 5, obtaining the parse results for each utterance. By calculating the occurrence counts of each word across all sentences, we ultimately obtain a ranked list of candidate words.

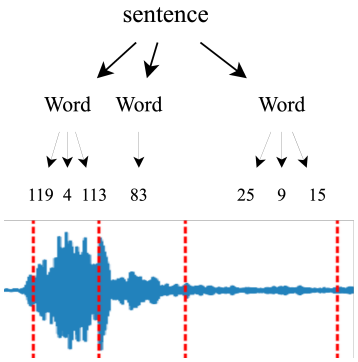

Figure 5: Parsing a sentence.

3.3 Semantics Discovery

To thoroughly understand dog vocalization, we need to obtain complete contextual information such as weather, environment, and the vocalization target, since these factors could be reactions to or causes of changes in the context. It is very challenging to obtain complete contextual information; however, we can obtain partial context to explore the coarse-grained semantics behind that dog barking sound (Berthet et al., 2023).

We believe the meaning of a word is determined by a causal relation between the context and the word. Specifically, if a contextual event causes a word to be uttered, this usually means that the word is a **reaction** to the event. On the other hand, if a word causes an event to occur, this means that the word is a **request** for the event. We broadly categorize the dog events into 14 different dog activities (Table 2), following Wang et al. (2023). We implemented a method for recognizing canine activities in three distinct phases: before, during, and after the utterance of a word. For each isolated segment of a dog's vocalization, which we refer to as a "sentence," we expanded the analysis to include a 5-second window both before and after the vocalization. To ascertain the dog's activities within these time frames, we employed a video understanding model called Temporal Segment Network (TSN) (Wang et al., 2016), one of the state-of-the-art models designed for video analysis.

To enhance the model's performance, we supplemented it with manually labeled 2,534 clips with a balanced number in each category, which allowed us to fine-tune the pre-trained TSN model. This process resulted in an improvement in accuracy, achieving 0.61 for top-1 accuracy and 0.92 for top-5 accuracy. Subsequently, the identified canine "words" were categorized into specific time segments using an activity position classification algorithm (Algorithm 1). This approach ensures a comprehensive understanding of the dog's activity *before, during, and after* its vocalizations, providing insights into the context and triggers of barking incidents.

Type	**Categories**
Activity	Standing, Walking, Sitting, Laying down, Eating, Sleeping, Running, No dog, Taking a shower, Sniffing, Playing with human, Playing with a toy, Swimming, Begging for food

Table 2: The categories of dog's activities.

Algorithm 1 Activity Position Classification Algorithm

Input: $activity\_start, activity\_end, word\_start, word\_end$
Output: $activity\ position$
$len1 \leftarrow word\_start - activity\_start$;
$len2 \leftarrow activity\_end - activity\_start$;
$len3 \leftarrow activity\_end - word\_end$;
$len4 \leftarrow word\_end - word\_end$;
if $len3 \leq 0 \parallel len1 \leq 0$ **then**
  if $len1/len2 \geq 0.5$ **then return** "before";
  else if $len3/len2 \geq 0.5$ **then return** "after";
  end if
  return "during";
end if
if $len4 \geq len1$ && $len4 \geq len3$ **then return** "during";
end if
if $len1/len2 \geq 0.3$ && $len3/len2 \geq 0.3$ **then**
  return ("before", "during", "after");
end if
if $len1/len2 \geq 0.3$ && $len3/len2 \leq 0.3$ **then return** "before";
end if
if $len1/len2 \leq 0.3$ && $len3/len2 \geq 0.3$ **then return** "after";
end if

To compute the causality strength (CS) between an activity and a word, we adopt the following approach.

$$CS(a \rightsquigarrow w) = \frac{count(a \rightarrow w)}{count(w)} - \frac{count(a)}{\sum_a count(a)}$$

where $a \rightarrow w$ denotes activity a is *before* word w.

$$CS(w \rightsquigarrow a) = \frac{count(w \rightarrow a)}{count(w)} - \frac{count(a)}{\sum_a count(a)}$$

where $w \rightarrow a$ denotes activity a is *after* word w.

If any candidate word has a strong enough causal relationship with an activity, we say that this word is genuine.

4 Evaluation

In order to investigate: (1) whether the phones after Dog-HuBERT are distinct and accurate, and (2) whether the words discovered are complete and semantically consistent, we evaluate our models' performances on phone recognition accuracy and vocabulary discovery.

4.1 Phone Evaluation

Setup A successfully classified phone should possess the following two properties: (1) the same phones should sound very similar, and (2) distinct phones should sound different (Twaddell, 1935). We verify the reliability of the phones we obtained by comparing consecutive identical phone audio samples. To assess the similarity of phones identified as the same, we randomly sampled 2 audio clips from different users for each phone, forming a test set of 140 pairs. To verify the differences between different phones, we selected 50 pairs of different phones with the closest cluster centers and randomly sampled 3 audio segments from different users for each phone, with each segment consisting solely of that phone, forming a test set of 150 pairs. In total, there are 290 pairs.

The consistency among testers and the results of distinguishing identical or distinct phones are indicators for measuring the reliability of the phones.

The testers are two college students majoring in engineering who love small animals and participated in the experiment as volunteers.

Results The phone evaluation results are shown in Table 3. Under the condition where the testers' agreement rate is greater than 71%, they can be considered capable of accurately distinguishing the same or different phones from an acoustic perspective.

For the internal consistency of each phone, an accuracy of at least 62% indicates that the instances of the same phone are indeed similar. Additionally, testers reported that the audio of canine calls was generally similar, whereas the differences between noises belonging to the same noise label were relatively large.

For the external differences between different phones, we selected the 50 pairs of phones with the smallest Euclidean distance between their cluster centers, which significantly increased the difficulty of distinguishing these phones. An accuracy rate exceeding 50% would indicate that these phones are distinct from each other.

Table 4 shows the average duration, median duration, and standard deviation of the dog phones and noise labels. The shorter average duration of noise labels and the longer average duration of dog phones further demonstrate the accuracy of phone recognition.

Tester	AP	SPP	DPP
Tester 1	58.28%	62.86%	54.00%
Tester 2	60.00%	66.43%	54.00%
Agreement	71.38%	75.00%	68.00%

Table 3: Accuracy and agreement result on testing the reliability of phonem discovery. AP: all pairs, SPP: same phone pairs, DPP: different phone pairs.

	Dog	Noise	All
Mean	60.5 ms	37.3 ms	52.5 ms
Median	40 ms	20 ms	40 ms
Std	65.5 ms	35.3 ms	58.0 ms

Table 4: The durations of dog phones and noise labels.

Overall, we have obtained reasonably accurate phone discovery, both in terms of the internal consistency of instances within the same phone and the differentiation between phones. These phones provide a foundation for further exploration into "words" in canine language.

4.2 Lexical Evaluation

Setup To measure the potential semantics within the candidate words we discovered, we select the 200 words with the highest occurrence counts in the sentences after excluding words that contain more than half noise phones. We also calculated the sentence coverage rate and word length using the top 100 and top 200 candidate words to measure the importance of the latter 100 candidate words.

In addition to the properties of the candidate words themselves, we also calculate the **reaction CS** score and **request CS** score for the top 200 candidate words in relation to each activity. The higher the CS score, the stronger the association between the word and the activity. Conversely,

the closer the CS score is to zero, the weaker the association between the word and the activity. A negative CS score indicates a negative correlation between the word and the activity.

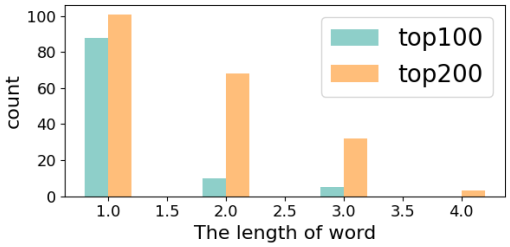

Figure 6: The distribution of word lengths in our dictionary.

Results We first look at the top 100 and 200 candidate words obtained through adaptor grammar induction. Figure 6 shows the distribution of discovered candidate words over the word length (i.e., number of phones in a word). We can see that most candidate words are unigrams, but there is a significant number of bigrams and trigrams in the discovered vocabulary. The average duration of the top 100 candidate words is 0.064 seconds, while that of the top 200 candidate words is 0.072 seconds. The variance in duration for the top 100 candidate words is 0.0037 seconds, and for the top 200 candidate words, it is 0.0047 seconds.

	Sentences (%)	Phones (%)
Top 100	89.57	80.18
Top 200	90.75	91.89

Table 5: The coverage of sentences and phones by top words in our vocabulary.

Next, we look at the coverage of our entire corpus by these top-ranked candidate words. Table 5 shows that they cover 90% of the sentences and also the majority of the phone sequences in the corpus.

Table 6 shows a set of words that have a causality strength greater than or equal to 0.07 with each activity. These words are considered genuine by the framework. Most of these words are bigrams or trigrams. One interesting word is 116-46-3, which is a reaction to "eating" but requests "laying down." After checking against the raw videos, we realize that the eat - bark - lay down procedure occurs with a number of dogs, and from the videos, we speculate that 116-46-3 expresses something like "I'm full."

Activity	Word	React/Request
Standing	59-124-11	Both
Walking	125	Both
Sitting	92-36	Both
Laying down	7-42-22	React
	116-46-3	Request
Eating	116-46-3	React
	31-105	Request
Sleeping	126-104	Both
Running	25	React
	59-74-139	Request
Taking a shower	81-4-81	Both
Sniffing	124-11-104	Both
Playing(Human)	34	Both
Playing(Toy)	38-135	Both
Swimming	113-4-59	React
	43-25-9	Request
Begging for food	92-36	React
	128	Request

Table 6: Activity and their most associated words

Then, they look at the flip side of the experiment, and Table 7 shows the top activities associated with the top 35 words. It is interesting to see that there is a strong correlation between sitting and begging for food, which suggests that dogs tend to beg for food while sitting, rather than standing. We also find that there are many bigrams, trigrams, and even 4-grams among the top 35 words.

Finally, we assess the accuracy of the top 35 words through human evaluation. Appendix B shows the causal strength of all activities with respect to these words. We employ three human judges to examine these graphs and label each word as either plausible or not plausible based on their reasoning regarding the relationship between the peak activities in each graph. Specifically, if a graph contains a single significant peak, or if the graph contains multiple peaks that have a strong semantic connection, then the word is considered plausible. The annotated results indicate that 87.1% of the top 35 words are plausible.

5 Related Work

To decode a dog's language, it is necessary to analyze its basic sound units, linguistic structure, lexicon, meaning, and more. The past few years have seen a surge of interest in using machine learning (ML) methods for studying the behavior of nonhuman animals (Rutz et al., 2023). Much of the past work has primarily focused on dog behavior (Abzaliev et al., 2024; Ide et al., 2021; Ehsani et al., 2018) and the meaning of dog sounds (Molnár et al., 2008; Hantke et al., 2018; Larranaga et al., 2015; Hantke et al., 2018; Pongrácz et al., 2006). Most of these

Word	Activities
59-124-11	standing
113-59-124	standing
124-11-104	standing
27-91	standing, sitting
110-31	standing
124-11	standing, walking
59-124-74	standing
11-104	standing
7-42-22	laying down
91-131	standing
59-124	standing
27	standing
46-3	laying down
116-46-3	laying down
128	sitting, begging for food
92-36	sitting, begging for food
7-42-51	laying down
34	playing with human
81-4-81	sitting, taking shower
61-70	begging for food
125	walking
36	sitting, begging for food
5-70	sitting, begging for food
38-135	laying down
113-59-83-17	laying down
59-74-130	standing
57-92	sitting, begging for food
56	sitting, playing with human
51-22	laying down
12-36	sitting
98	playing with human
136	standing
126-4-126	playing with human
66	standing
83-9-16	begging for food

Table 7: Top 35 words ($CS > 0.07$) and their associated activities

studies only classify the audio of dog sounds into multiple categories, including activities, contexts, emotions, and ages. They did not study the sound units of the dog's language.

Several researchers (Hagiwara, 2023; Abzaliev et al., 2024) have shown that self-supervised methods are equally adept at analyzing and characterizing the vocalizations of animals. Specifically, the work by Hagiwara et al.'s (2024) has utilized HuBERT to establish a phonetic alphabet that transcends species, facilitating the transcription of animal sounds.

The above work illustrates the existence of multiple distinct sound units in dog language. Many species that appear to use only a handful of basic call types may possess rich vocal repertoires (Rutz et al., 2023). Numerous studies demonstrate the diversity of animal sounds (Paladini, 2020; Robbins, 2000; Bermant et al., 2019). Huang et al. (2023) and Wang et al. (2023) conducted fine-grained studies of dog sound units. However, using a priori knowledge of human language directly may not be applicable. Hagiwara et al. (2024) and Hagiwara (2023) also conducted fine-grained studies of dog sound units, but they did not explore the possible meanings of these units.

Conducting unsupervised lexicon discovery on speech without any prior knowledge is a very challenging task (Park and Glass, 2007). Lee et al. (2015) achieved end-to-end human lexicon discovery from audio by jointly training Hidden Markov Models (HMMs)(Schwartz et al., 1984) for phone discovery and Adaptor Grammar(Johnson et al., 2006) for lexicon discovery. However, the accuracy of their lexicon discovery remains low. In this paper, we use an improved approach to train the parameters for Adaptor Grammar.

6 Conclusion

In this paper, we present a method to parse and understand canine vocalizations. In contrast to previous work, this approach uses a self-supervised method instead of relying on human language knowledge to explore sound units in canine language. This is better suited for examining fine-grained phonetics and semantics in a language with no prior linguistic knowledge. We then use Hybrid Variational-MCMC Inference to train the parameters of the Adaptor Grammar, ultimately obtaining a candidate vocabulary. By simultaneously considering dog activity events surrounding the vocalizations, we find that some words in the vocabulary have strong correlations with certain activities.

Limitation

The discovery of dog sound units heavily relies on the quality of the dataset. Even though we have implemented multiple measures to enhance the dataset's quality, noise may still be present due to various factors, including the recording equipment, background noise, and added noise from video uploaders. Discovering words in an unfamiliar human language presents a challenging problem due to noise, mispronunciation, and other factors. The same difficulties exist with dogs. One possible solution to achieve better results is to acquire more datasets and consider the broader context of the dog barks.

Acknowledgment

Kenny Q. Zhu is partially supported by NSF Award No. 2349713.

References

Artem Abzaliev, Humberto Pérez Espinosa, and Rada Mihalcea. 2024. Towards dog bark decoding: Leveraging human speech processing for automated bark classification. *arXiv preprint arXiv:2404.18739*.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839.

Peter C Bermant, Michael M Bronstein, Robert J Wood, Shane Gero, and David F Gruber. 2019. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific reports*, 9(1):12588.

Mélissa Berthet, Camille Coye, Guillaume Dezecache, and Jeremy Kuhn. 2023. Animal linguistics: a primer. *Biological reviews*, 98(1):81–98.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, and Ali Farhadi. 2018. Who let the dogs out? modeling dog behavior from visual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4051–4060.

Thomas S Ferguson. 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Masato Hagiwara. 2023. Aves: Animal vocalization encoder based on self-supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Masato Hagiwara, Marius Miron, and Jen-Yu Liu. 2024. Ispa: Inter-species phonetic alphabet for transcribing animal sounds. *arXiv preprint arXiv:2402.03269*.

Simone Hantke, Nicholas Cummins, and Bjorn Schuller. 2018. What is my dog trying to tell me? the automatic recognition of the context and perceived emotion of dog barks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5134–5138. IEEE.

David Holdcroft. 1991. *Saussure: signs, system and arbitrariness*. Cambridge University Press.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jieyi Huang, Chunhao Zhang, Mengyue Wu, and Kenny Zhu. 2023. Transcribing vocal communications of domestic shiba lnu dogs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13819–13832.

Yuta Ide, Tsuyohito Araki, Ryunosuke Hamada, Kazunori Ohno, and Keiji Yanai. 2021. Rescue dog action recognition by integrating ego-centric video, sound and sensor information. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, pages 321–333. Springer.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Ana Larranaga, Concha Bielza, Péter Pongrácz, Tamás Faragó, Anna Bálint, and Pedro Larranaga. 2015. Comparing supervised learning methods for classifying sex, age, context and individual mudi dogs from barking. *Animal cognition*, 18(2):405–421.

Chia-ying Lee, Timothy J O'donnell, and James Glass. 2015. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. 2023. Separate anything you describe. *arXiv preprint arXiv:2308.05037*.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Csaba Molnár, Frédéric Kaplan, Pierre Roy, François Pachet, Péter Pongrácz, Antal Dóka, and Ádám Miklósi. 2008. Classification of dog barks: a machine learning approach. *Animal Cognition*, 11:389–400.

Aleida Paladini. 2020. The bark and its meanings in inter and intra-specific language. *Dog behavior*, 6(1):21–30.

Michael A Pardo, Kurt Fristrup, David S Lolchuragi, Joyce H Poole, Petter Granli, Cynthia Moss, Iain Douglas-Hamilton, and George Wittemyer. 2024. African elephants address one another with individually specific name-like calls. *Nature Ecology & Evolution*, pages 1–12.

Alex S Park and James R Glass. 2007. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197.

Jim Pitman and Marc Yor. 1997. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

Péter Pongrácz, Csaba Molnár, and Adam Miklosi. 2006. Acoustic parameters of dog barks carry emotional information for humans. *Applied Animal Behaviour Science*, 100(3-4):228–240.

Robert L Robbins. 2000. Vocal communication in free-ranging african wild dogs (lycaon pictus). *Behaviour*, pages 1271–1298.

Christian Rutz, Michael Bronstein, Aza Raskin, Sonja C Vernes, Katherine Zacarian, and Damián E Blasi. 2023. Using machine learning to decode animal communication. *Science*, 381(6654):152–155.

Rich Schwartz, Y Chow, S Roucos, M Krasner, and J Makhoul. 1984. Improved hidden markov modeling of phonemes for continuous speech recognition. In *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 21–24. IEEE.

Pratyusha Sharma, Shane Gero, Roger Payne, David F Gruber, Daniela Rus, Antonio Torralba, and Jacob Andreas. 2024. Contextual and combinatorial structure in sperm whale vocalisations. *Nature Communications*, 15(1):3617.

Erik Sudderth and Michael Jordan. 2008. Shared segmentation of natural scenes using dependent pitman-yor processes. *Advances in neural information processing systems*, 21.

W. Freeman Twaddell. 1935. On defining the phoneme. *Language*, 11(1):5–62.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer.

Yufei Wang, Chunhao Zhang, Jieyi Huang, Mengyue Wu, and Kenny Zhu. 2023. Towards lexical analysis of dog vocalizations via online videos. *arXiv preprint arXiv:2309.13086*.

Sophia Yin and Brenda McCowan. 2004. Barking in domestic dogs: context specificity and individual identification. *Animal Behaviour*, 68:343–355.

Ke Zhai, Jordan Boyd-Graber, and Shay B Cohen. 2014. Online adaptor grammars with hybrid inference. *Transactions of the Association for Computational Linguistics*, 2:465–476.

A Probabilistic Context-Free Grammars and Adaptor Grammars

Probabilistic Context-free Grammars (PCFGs) (Manning and Schutze, 1999) are an extension of context-free grammars that assign a probability to each production rule, providing a probabilistic framework for generating strings. A PCFG is defined as a tuple (N, Σ, R, S, P) where N is a set of non-terminal symbols, Σ is a set of terminal symbols, R is a set of production rules of the form $A \rightarrow \beta$, S is the start symbol, and $P(A \rightarrow \beta)$ is the probability associated with the production rule $A \rightarrow \beta$ such that $\sum_{\beta} P(A \rightarrow \beta) = 1$ for all $A \in N$.

The adapted tree distributions H_n is generated by using a *Pitman-Yor process* (Pitman and Yor, 1997), a generalization of Dirichlet process (Ferguson, 1973). A draw $H_n \equiv (\pi_n, z_n)$ is formed by the stick breaking process (Sudderth and Jordan, 2008) parametrized by scale parameter a, discount factor b, and base distribution G_n:

$$\pi\prime_k \sim \text{Beta}(1-b, a+kb),\ z_k \sim G_n$$
$$\pi_k \equiv \pi\prime_k \prod_{k-1}^{j=1}(1-\pi\prime_j),\ H \equiv \sum_k \pi_k \delta_{z_k} \quad (1)$$

Algorithm 2 Parse Algorithm

1: For nonterminals $e \in N$, draw rule probabilities $p_e \sim Dir(\alpha_n)$ for PCFG G
2: **for** adapted nonterminal n **do**
3: Draw $H_n \sim$ PYGEM(a_n, b_n, G_n) according to Equation 1, where G_n is defined by the PCFG rules R.
4: **end for**
5: For $i \in 1, ..., D$, generate a hierarchical structure tree $t_{s,i}$ using the PCFG rules $R(e)$ at non-adapted nonterminal e and the H_c at adapted nonterminals n.
6: The yields of trees $t1, ..., t_D$ are observations $x_1, ..., x_D$.

B Causal strength of top 35 words

Here are the causal strength for the top 35 words and activities.

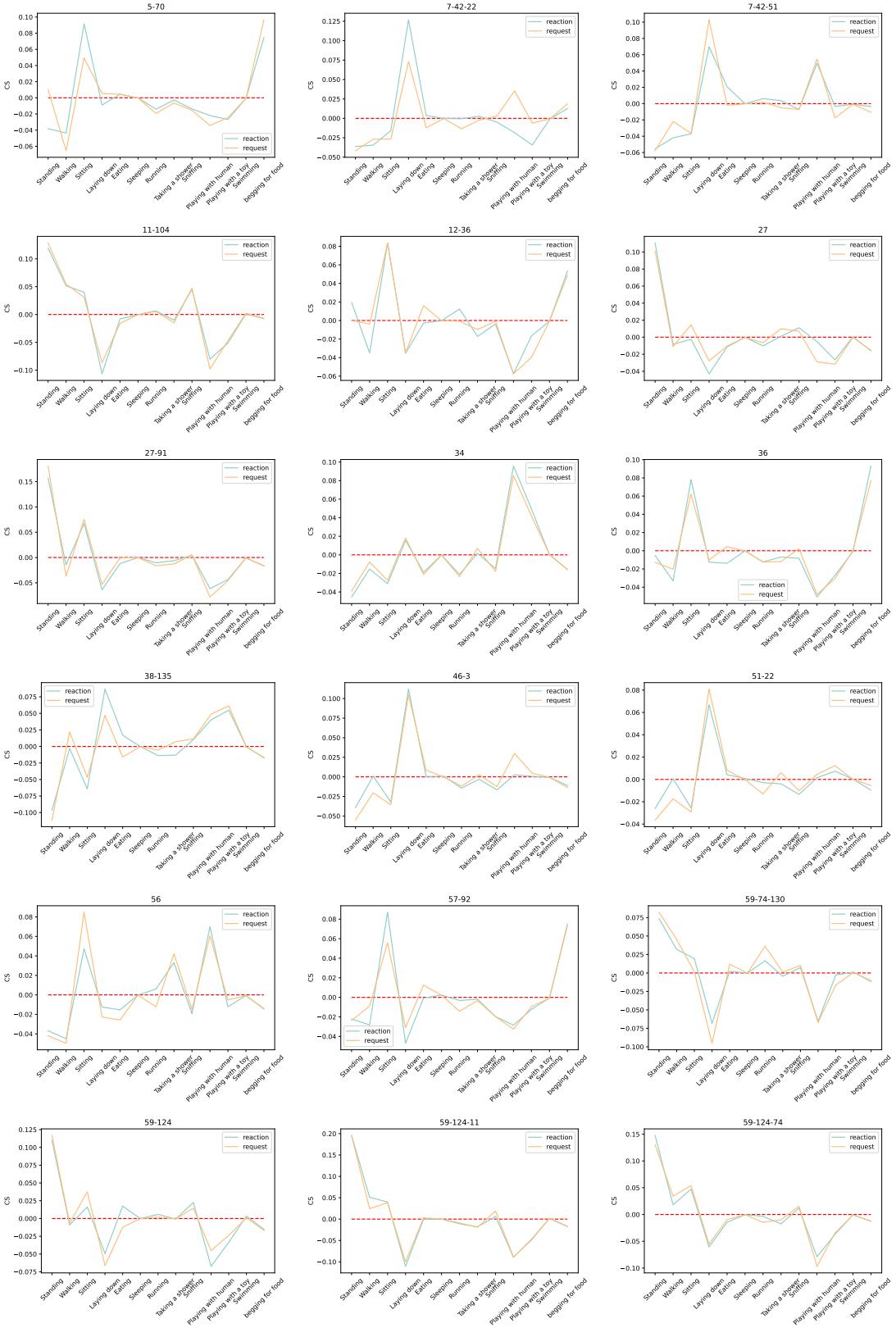

Figure 7: Causal strength of top 35 words (part 1).

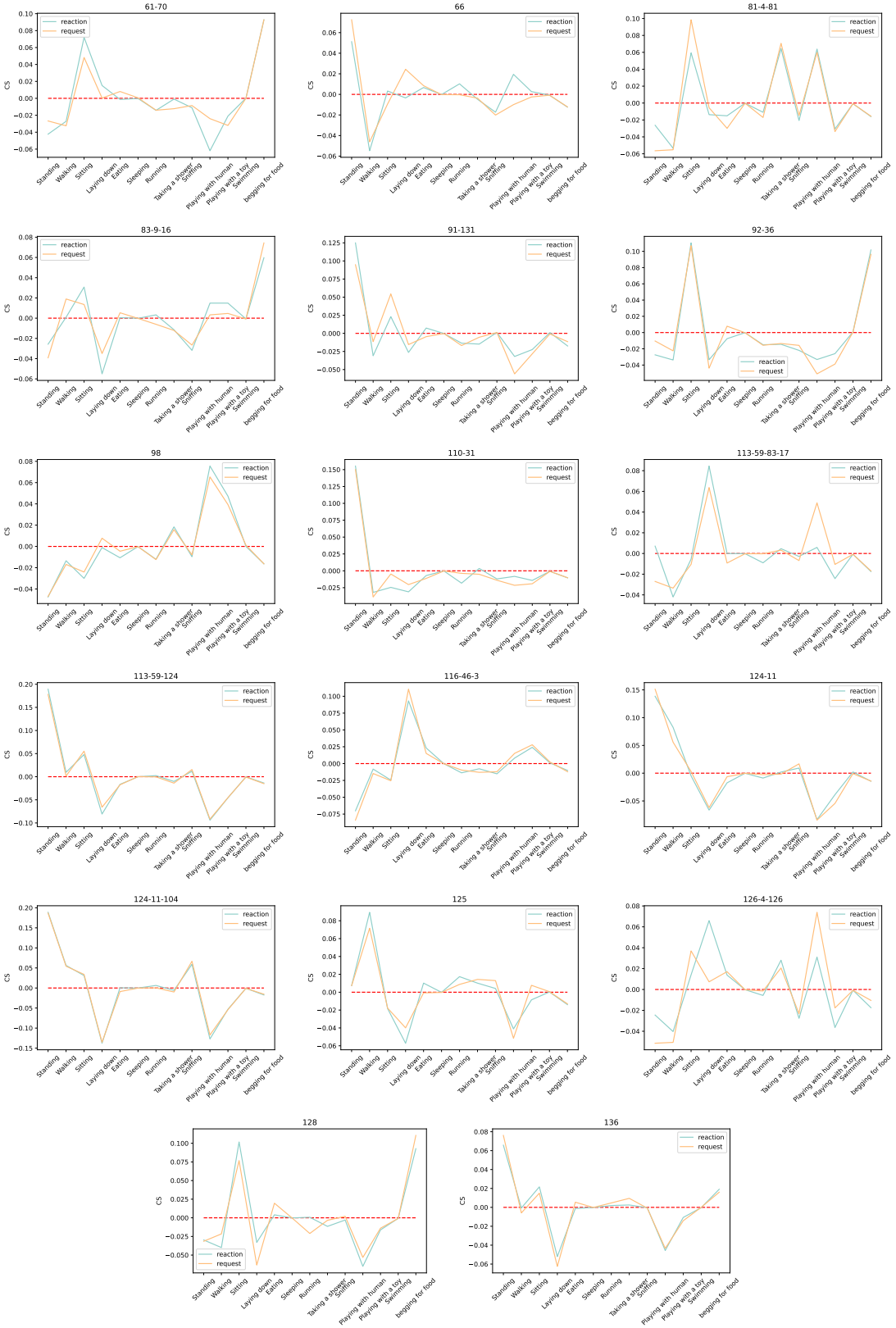

Figure 8: Causal strength of top 35 words (part 2).