Towards Lexical Analysis of Dog Vocalizations via Online Videos

Anonymous submission

Abstract

Deciphering the semantics of animal language has been a grand challenge. This study presents a data-driven investigation into the semantics of dog vocalizations via correlating different sound types with consistent semantics. We first present a new dataset of Shiba Inu sounds, along with contextual information such as location and activity, collected from YouTube with a well-constructed pipeline. The framework is also applicable to other animal species. Based on the analysis of conditioned probability between dog vocalizations and corresponding location and activity, we discover supporting evidence for previous heuristic research on the semantic meaning of various dog sounds. For instance, growls can signify interactions. Furthermore, our study yields new insights that existing word types can be subdivided into finer-grained subtypes and minimal semantic unit for Shiba Inu is word-related. For example, whimper can be subdivided into two types, attention-seeking and discomfort.

1 Introduction

Animal languages have captured the curiosity of scientists for years and animals use vocal expressions to communicate (Garcia and Favaro 2017). Despite various attempts from diverse perspectives, deciphering the intricately complex semantic meanings within animal communication systems remains a challenge (Andreas et al. 2022; Scott-Phillips and Heintz 2023). Acquiring a deeper comprehension of animal language holds significant implications for unraveling their social structures, and intelligence, and facilitating human-animal interactions.

Within the expansive realm of animal languages, the study of **dog language** holds particular interest. Dogs, as one of the most popular and widely kept pets, engage in constant interaction with human beings through their vocal expressions. Given the massive interactions between dogs and people during the domestication process, dogs' vocal behavior undergoes considerable changes (Huang et al. 2023; Feddersen-Petersen 2000) and it is reasonable to infer that diverse sounds emitted by dogs in varying scenes carry distinct significances. Previous works on dog language have largely relied upon experimental knowledge and heuristic subjective observations, which depend on long-term experiences and costly data-collection and can be limited and prone to biases (Yin 2002; Pongrácz, Molnár, and Miklósi 2010; Faragó et al. 2017). Only coarse-grained meanings can be drawn given limited data and corresponding scenes from these studies. Our web-data-driven exploration leverages a more comprehensive methodology to uncover finer-grained semantics with a broader context. The utilization of **web data** offers a wealth of information, introducing numerous possible variables and semantic clues, thus enabling a more comprehensive analysis.

There is a wide range of vocalizations dogs can produce (Yeon 2007), which are affected by various engaged objects, the emotion of the dog, the surrounding environment the dog is located in, the activity the dog is doing, the object the dog interacts with, and even the age and gender of the dog may play a key factor (Pongrácz et al. 2005; Molnár et al. 2009). Given the benefits of using web data, we opt to focus on the Shiba Inu breed for our research as Shiba Inu is a widely adopted dog at home and plenty of video data is available on YouTube. Hereby, we investigate the semantic meaning of the Shiba Inu dog sound according to two important factors of the context, the **location** and the **activity** of the dog as these two factors are currently available.

To understand the semantic meaning of dog language, previous works always record dog sounds in different scenarios and then analyze them. As we have online videos, there are several challenges: extract the meaningful dog sound and context from videos and ascertain if they show consistent patterns with context. In this work, we implement the first data-driven, evidence-based research sourced from social media to give fine-grained semantics to understand the vocal language of Shiba Inu. We construct data as Figure 1 for our dataset to map dog sound with context which enables us to take a closer look into the behaviors of animals and some hidden **semantic patterns**. For example, one kind of dog sound, bow-wow, is usually mixed with bark by previous researchers. However, in our work, we find that it shows the curiosity of dogs for the surroundings while bark does not exhibit such meaning. On the other hand, our method has more detailed contexts for analyzing animal languages, which will benefit future works. We identified as many as 11 locations and 14 activities, in which dogs might produce 6 different vocal sounds to signify various meanings. To associate vocal expression patterns with possible lexical meanings, we need to respectively extract vocal sounds and their transcriptions as well as the activity and location that might give rise to a change of meanings. With these fine-grained labels, we

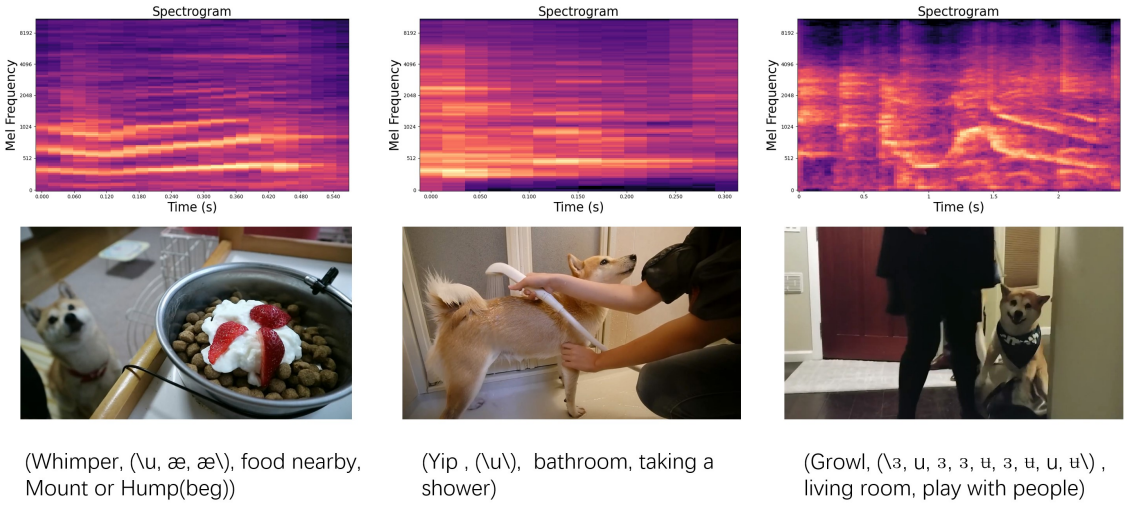


Figure 1: Introducing scenario: Spectrograms of dog sounds in the corresponding context. When the dog is begging for food, taking a shower, or playing with people, its vocal spectrograms differ a lot. We map possible words and subwords extracted from dog barking audio with location and activity, forming our quadruplet dataset. `<word, IPA symbols, location, activity>`.

can capture subtle semantic differences and constant vocal patterns under these contexts to explore the semantics of vocalizations for Shiba Inu dogs. Our main contributions can be summarized as follows:

- We propose a universal pipeline to process and analyze dog-related videos on YouTube to understand the lexical semantics of dog language. The framework is reusable to other animal species for which videos are available.
- We are the first to implement data-driven research to study dog semantic language from web data. We build a dataset of 10,779 quadruplets that contains 6 distinct words, subwords, and corresponding context for exploring dog language. We define fine-grained 14 activities and 11 locations that could imply different semantic meanings which can be extracted from videos.
- Through our investigation of dog sound patterns, we have uncovered several conclusions that align with existing human knowledge and previous research. Additionally, we have gained some unique insights that have been underexplored.

2 Problem

Our goal is to understand the lexical semantics of dogs and explore the minimal semantic unit. We seek to address the following technical problems:

1. Do dogs use consistent vocal patterns to signify certain meanings?
2. How to compute the correlation between vocal expressions with possible factors that give rise to different certain meanings?

To answer these questions, we need to classify distinct sound types, which are defined as "words" and we further phonetically transcribe these words, which are signified as subwords in Section 3.2. Regarding contextual information to uncover the semantics, we define a diversed and comprehensive list for location and activity and utilize respective extraction methods in Section 3.3.

3 Approach

In this section, we present our pipeline (as shown in Figure 2) including data collection, vocalization processing procedures (word segmentation, subword extraction, and phonetic transcription), as well as contextual information extraction methods (location and activity recognition). Implementation details can be found in Appendix A.

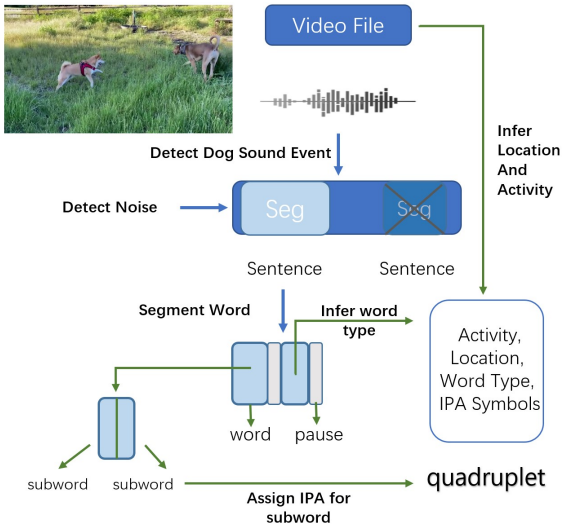


Figure 2: The overall view of pipeline.

3.1 Collect Shiba Inu Online Videos

We first use the keyword "Shiba Inu" to identify users who own such dogs and only upload shiba inu videos with dog sounds then we collect up to 13,164 Shiba Inu related videos posted by these users within days which can be easily expanded in the future.

3.2 Sentence Split, Word Segmentation and Subword Extraction

Once videos are downloaded, we begin to process the audio tracks. Similar to human language, where words serve as the fundamental units for constructing sentences, we hypothesize that a similar concept can be applied to dog language. We define a **"word"** as an independent and contiguous dog vocal sound that typically lasts around 1 second, and it is bounded by some noticeable pauses. A **"sentence"** is a sequence of consecutive words. A **"subword"** is a subpart of "word" and is represented by an IPA symbol.

Algorithm 1: Sentence Extraction

```
Data: Audio tracks
Result: "sentences" of dogs
while pass accross segments of audio track do
    feed the audio segment feature into PANNs;
    if PANNs infers as dog sound without
      accompanying noise then
        audio segment belongs to a "sentence";
    else
        pass to next segment;
    end
end
```

Sentence Extraction To extract the word clips, we adopt a similar methodology as Huang et al. (2023). The initial step is detecting dog vocalizations and splitting the audio tracks into "sentences", continuous sequences of dog vocal sounds. We pass through the frames of audio and determine whether they contain dog sounds. If so, we extract these audio clips and then we remove noise other than dog vocalizations to ensure that there is no accompanying music or human speech in the background. We apply PANNs (Kong et al. 2020), a sound event detection model that is pre-trained on the large-scale Audioset (Gemmeke et al. 2017) dataset with 527 sound classes.

Word Segmentation These "sentences" may contain short pauses in the middle that can be used to separate the words and the next step is to segment "words" in "sentences". We finetune PANNs to determine each start and end of a single "word" by detecting a frame in the audio clip which transits from a silence frame to a dog frame with a gap of 0.1 seconds between frames. We manually create labels for the event "dog" with a total data length of 715 seconds. This finetuned model is capable of detecting the small pauses within "sentences" and extracting the singular "words". Lastly, we follow the definition of 6 different dog vocalization patterns defined in Audioset, which are distinct from each other. We labeled 240 audio clips and trained a classification model.

Algorithm 2: Word Segmentation

```
Data: "Sentences"
Result: "Words"
while pass through segments of "sentences" do
    Feed segment feature into PANNs;
    if PANNs infers as "dog" other than "silence"
      then
        segment belongs to a "word";
    else
        Use segment to split "words";
    end
end
```

$$\text{Wordtype} = \arg\max_i P_{PANNs}(\text{soundtype}_i|\text{word}). \quad (1)$$

Subword Segmentation and Phonetic Transcription We first split "words" into "subwords" with the method as (Räsänen, Doyle, and Frank 2018), which uses sonority fluctuation in audio to segment words into smaller units. Thus, a word can be further split into smaller parts according to the sonority changes. In order to present the vocal characteristics of each subpart, we phonetically transcribe each subword with international phonetic alphabet (IPA) symbols. We compute the acoustic feature distances with standard pronunciations (Peter 2023; Bruce 2023). In this way, each subword is represented by an IPA vowel symbol, and a "word" is represented by a sequence of IPA symbols.

3.3 Surrounding Context Extraction

Since images and sound are naturally aligned in a video, we can extract the *location* and *activity* of the dog while it utters a particular "word", from the image frames that synchronize with the audio frames. The end product of this phase will be a sequence of quadruplets consisting of `<word, subwords, location, activity>`, extracted from the videos.

Location Given the begin and end timestamp of a "word", five image frames in that time range are sampled, then sent into an image classification model to determine the location of the dog when the "word" occurs. Here we finetune the pre-trained model from Zhou et al. (2017) trained on the scene-centric datasets Places365 based on thousands of pictures sampled from Shiba Inu videos to predict the dog's location. The fine-tuned model achieves 77.99% on Top-1 accuracy and 97.13% on Top-5 accuracy.

We denote the following variables: t_{word}: Timestamp of the "word" occurrence between begin and end, I_{i}: An image from a batch of images constructed from five frames sampled around the timestamp t_{word}, M_{class}: The class predicted by the model given I_{i}, L_{dog}: Location of the dog when the "word" occurs, Majority Vote is a function that selects the class with the highest frequency among the predicted class of the individual images in the batch. The formula for location

inference is as:

$$L_{\text{dog}} = \text{Majority Vote}\left(\bigcup_{i=1}^{n} M_{\text{class}}(I_i)\right) \quad (2)$$

Activity To get the activity information about the dog, we sample a five-second video clip based on the timestamp of the "word" that we chooce a range of (begin timestamp - 1s, end timestamp + 1s), and then a video understanding model is applied to decide what the dog is doing. The pre-trained model Temporal Segment Networks (TSN) (Wang et al. 2018) performs well on the task of video understanding. In this study, we annotated 2534 video clips sampled from the Shiba Inu videos and finetune the pre-trained model. The fine-tuned model achieves 61.40% on Top-1 accuracy and 92.40% on Top-5 accuracy. Based on these two fine-tuned models, given the timestamp of a "word", we can explore the location and activity from the video.

4 Dataset

We obtain a large-scale timestamp-aligned dataset of `<word, subwords, location, activity>`. We present the details of our dataset and quality of it in the following part.

4.1 Statistics of the dataset

The dataset contains 10,779 quadruplets from corresponding 3,068 videos and 5,834 sentences, including 6 different types of dog sounds, 11 location categories, 14 activity categories, and 20 IPA vowels for subwords as shown in Table 1. We follow the definitions of dog vocalization types defined in Audioset and we adopt the most frequent locations and activities tailored for Shiba Inu by combining commonsense and data manual checking for location and activity. More details of the dataset can be seen in Appendix C.

Context	Possible values
Word Type	Bow-wow, Bark, Whimper, Growl, Howl, Yip
Phonetic Symbol	\u,æ,ɜ,ʉ,ɯ,ɐ,a,ɵ,œ,ə,ʊ,ɪ,ɒ,ɛ,ʌ,ɑ,œ,ɔ,i,e\
Location	Living Room, Food Nearby, Grass, Cage, Road, Bathroom, Snowfield, Beach, Square, Vehical Cabin, Others
Activity	Mount Or Hump (beg), Play With People, Sit, Lay Down, Walk, Sniff, Eat, Stand, Take a Shower, NoDog, Run, Be Touched, Unknown, Fight With Dogs, Show Teeth or Bit

Table 1: Respective categories for dog vocalizations and contextual information in our dataset.

Figure 3 shows the prior distribution of the items in the dataset across different word types, subword IPA symbols, location types, and activity types. This data imbalance reflects real-world data distribution, that pet dogs kept at home are more prone to stay in the living room and exhibit activities like standing and fighting with dogs. We consider this imbalanced prior distribution when computing the correlations between them.

4.2 Quality of the dataset

We present the accuracy of each step to ensure the high quality of our dataset. For word segmentation and word classification, we randomly sample 200 segmented words to listen to. Three annotators have to label whether this is a singular dog word and whether the word is correctly classified. Their respective accuracy is 0.95 and 0.84. We achieve an accuracy of 0.78 for location classification on our manually labeled pictures and an accuracy of 0.59 for the top 1 and 0.9416 for the top 5 for activity classification.

5 Results and Analysis

To answer the first question, we analyze the relationship between words and context. We then explore the relationship between words and subwords. We present a summarization of findings that uncovers dog vocalization patterns. To answer the second question, we adopt conditioned probability.

5.1 Correlation between word type and location

The vocalization of a dog varies depending on its location. Based on the dataset, we compute the correlation between location and word as shown in Figure 4a:

$$\frac{P(\text{location}|\text{word})}{P(\text{location})} \quad (3)$$

By dividing each item in the formula by its prior probability, we can offset the influence of the original umbalanced frequencies of the word types and locations.

"Bow-wow" can be used to express curiosity. Bow-wows are usually used when they are outside and in unfamiliar surroundings like snowfields, grass, and other places. This word may implicate an exploration of the environment.

"Whimper" can be seen as attention-seeking when food nearby. "Whimper" can be used to elicit people attention (Handelman 2012). When there is food nearby, they may beg for food.

5.2 Correlation between word type and activity

We use the same method to analyze the relationship between word types and activities in Figure 4b as:

$$\frac{P(\text{activity}|\text{word})}{P(\text{activity})} \quad (4)$$

"Bow-wow" may indicate movements or food. Dogs exhibit a preference for using the "bow-wow" sound when engaging in activities that involve movement, such as walking and sniffing. "Bow-wow" may also be a signal for food as begging for food or eating.

"Growl" expresses interation with outside. "Growl" appears to be used to express interactions (Handelman 2012). It is prevalent when be touched, fight with dogs, play with people and show teeth or bite.

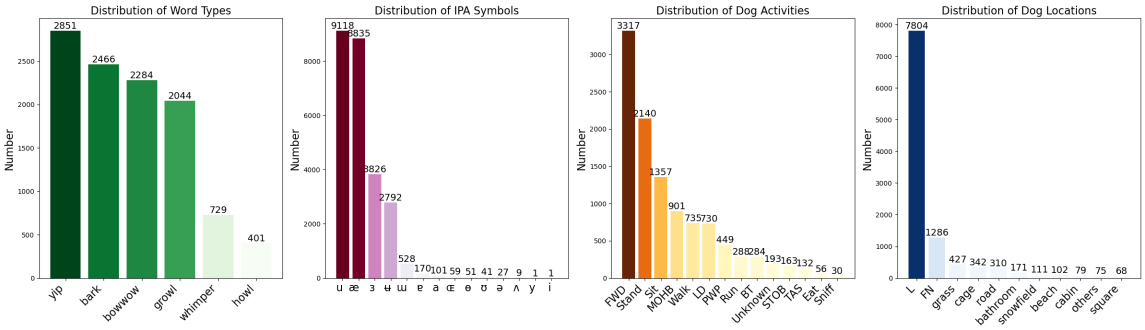


Figure 3: Prior distributions of word types(green), IPA symbols(pink), locations(blue) and activities(red) in the quadruplet sequences. The numbers shown are the number of times for different word types, locations, and activities. "L", "FN", "FWD", "MOHB", "LD", "PWP", "BT", "STOB", "TAS" represent "Livingroom", "Food Nearby", "Fight With Dogs", "Mount Or Hump (Beg)", "LayDown", "Play With People", Be Touched", "Show Teeth Or Bit", "Take A Shower" respectively.

"Whimper" and "howl" are used when relatively steady. Dogs are usually howling when sitting down and standing. We find that Shiba Inu tends to "howl" when they are begging for food which is a new observation. Dogs often "whimper" when engaged in activities like sitting and begging for food. It may indicate contact seeking and a kind of submission (Pongrácz, Molnár, and Miklósi 2010).

5.3 Correlation between word type and context

When putting location and activity together as in Figure 4c:

$$\frac{P((\text{location, activity})|\text{word})}{P((\text{location, activity}))} \tag{5}$$

Because of the numerous combination possibilities for context, we define a threshold value of 100, indicating that a particular combination is considered statistically significant only if it occurs more than 100 times. Figure 4c depicts this relationship between word type and context.

"Howl" and "whimper" come up frequently with food. "Whimper"represent attention-seeking, thus it can be interpreted as attention for food when "food nearby" and the dog "begs". "Howl" is used when sitting and food nearby.

"Yip" and "whimper" can express discomfort. They can express pain or discomfort (Yeon 2007; Pat 2011). When the dog is located in the cage and laying down, it tends to yip and whimper.

"Bark" can express discomfort. Shiba Inu barks when it is taking a shower in the bathroom. This can express its discomfort and warning (Yeon 2007) because dogs usually don't like baths.

5.4 Analysis of the sequence of quadruplets

To analyze quadruplet sequences, we start by picking words in the same sentence where the word type changes. We hypothesize that a sequence of different word types in a similar context carries a specific meaning. We only consider combinations that appear frequently, setting a threshold of 10 to distinguish patterns from chance occurrences. We explore the formula as shown in Figure 4d:

$$\frac{P((\text{location, activity})|(w_1; w_2))}{P((\text{location, activity}))} \tag{6}$$

while $(w_1; w_2)$ represents two consecutive words with different types, which is defined as a *bi-gram* here. We highlight several insights. Our findings suggest that the sequence of words might not strongly indicate distinct semantic meanings. For instance, comparing "Bark Bow-wow" and "Bow-wow Bark" reveals differing context distributions. However, in the cases of "Whimper Bow-wow" and "Bow-wow Whimper", we observe similar distributions despite the word order variation.

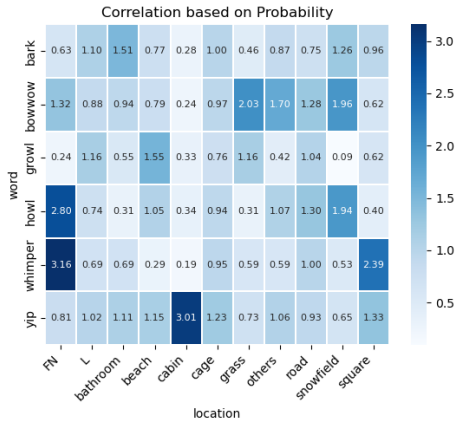
To provide additional evidence, we examine the bi-gram probability of two-word sequences as w_1 followed by w_2, where w_1 and w_2 represent two words. To enhance clarity, we normalize each item by dividing it by the prior distribution of w_2, revealing relative changes. The result for:

$$P(w_1|w_2) \tag{7}$$

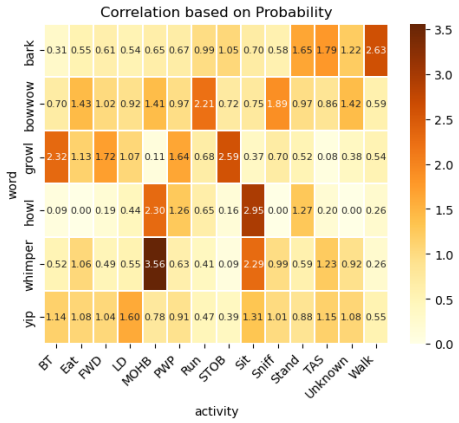
is shown in Figure 5a. As observed, "howl" and "whimper" are highly probable of following each other and this implies their semantics are kind of overlapping. Same for "bark" and "bow-wow". This shows the words contain multiple semantics.

5.5 Analysis of subwords semantics

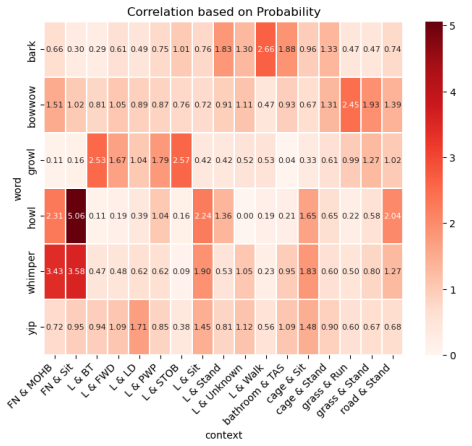
Previous works show for one word type, it conveys multiple meanings. Each word contains one or multiple different subwords. By combining analyzing the multiple meanings for one word type and its possible subwords, we prove that one word type can be further divided into finer-grained types, which is the first data-driven experimental proof. For subwords in the same word, we first collect the most frequent subwords that can be one or a sequence of IPA vowel symbols, then we illustrate their distribution on the context and they vary a lot, which means that they convey different semantics. As shown in Figure 5b, we present the activity distribution of the most frequent subwords for yip. As we can



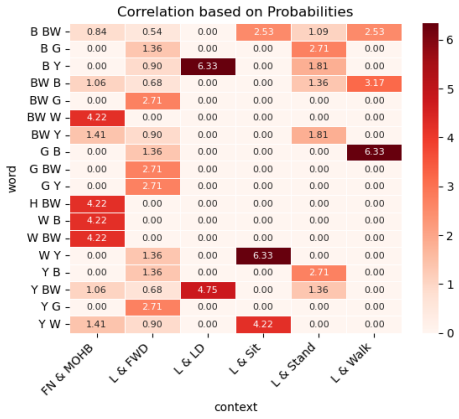
(a) Correlation between word types and locations.



(b) Correlation between word types and activities.



(c) Correlation between word types and contexts.



(d) Correlation between bi-gram of words and contexts. "B", "BW", "G", "H", "Y", "W" represent "bark", "bow-wow", "growl", "howl", "yip" and "whimper" respectively.

Figure 4: Correlation to explore semantics of words.

see, for the same scene fight with dogs, different IPA symbol sequence shows an dissimilar distribution.

We further observe that these frequently different subwords are not restricted to one dog, which means this semantic difference is not by accident or caused by specific characters of one dog. A summary table is appended in Appendix C. By combining analysis of words and subwords, we explore that the distribution of the IPA symbol is mainly influenced by the distribution of the word which contains this symbol and this illustrates that the minimal semantic unit is word-related instead of subword. A detailed explanation is in Appendix D.

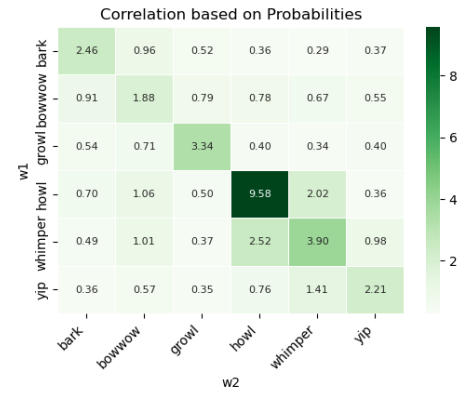
5.6 Comparison with Previous Works

We present our findings including the evidence to support previous theoretical studies and our new observations in Table 2. Through the analysis, we found most of the patterns we mentioned are consistent with the previous qualitative research. We also discover more detailed interpretations of those existing patterns and provide possible new findings. We are the first web-data-driven approach to analyze the minimal semantic unit of the Shiba Inu dog language.

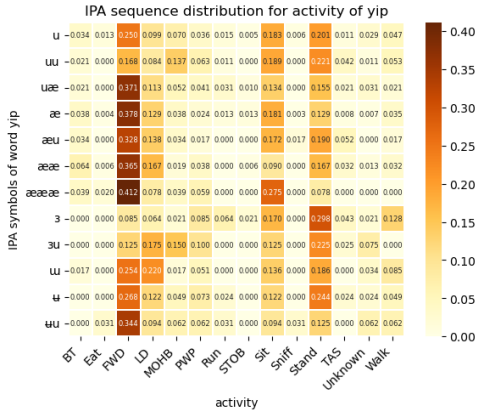
6 Related Work

It has long been difficult to understand what animals are trying to express. Since we can't speak to them directly, just trying to understand their language becomes an explorable target. Early studies have contributed to our understanding of dogs: dog vocalizations in different context (Molnár et al. 2009; Robbins 2000), emotion recognition through vocal cues (Pongrácz, Molnár, and Miklósi 2006), and the development of image and video analysis techniques for pet understanding (Mao and Liu 2023). These studies give us a basic understanding of dog sounds. However, they either only conduct testing experiments, or only studied images and sound signals without mining the relationship between them. Our study does lexical analysis and connects it with the goal of understanding dogs.

In our communication with dogs, visual signals complement our understanding through sound signals. There are some interesting datasets that encourage visual tasks



(a) Correlation based on probability for P($w_2|w_1$).



(b) Activity distribution of most frequent subword combinations for yip.

Figure 5: Exploring sequence of words and activity distribution of subwords for yip

which help us understand dogs, such as first-person videos from a camera on the back of dogs for activity classification (Iwashita et al. 2014), videos with skeletons labeled which help to detect poses (Cao et al. 2019) and a collection of videos about different animal behavious (Ng et al. 2022). These studies have greatly broadened the methods for animal action recognition, but there is a lack of a dataset for dog activities in domestic scenes.

In addition to dogs, some animals are social animals and have a lot of interspecies interactions and it is an interesting topic about how they communicate with each other. Cetaceans (Bermant et al. 2019), elephants (Rossman et al. 2020) have a high degree of social complexity, and acoustic features can be used for detection or analyzing responses to other signals. Voice of birds (Koh et al. 2019; Salamon et al. 2017; Adavanne et al. 2017) also brings information that can be used to detect birds or classify breeds. However, it is expensive to record sound and the restricted experiment environment restricts the diversity of data. Our data-driven research pipeline takes advantage of the vast amount of data available on the Internet to build a scalable and diverse dataset.

Also, it is of vital importance to make sure the boundary

Lexical Symbols	**Previous Meaning**	**Our Meaning**
Whimper	attention-seeking (Handelman 2012), discomfort (Chelsea 2018)	attention-seeking, beg for food, steady, discomfort
Yip	discomfort (Pat 2011)	loneliness
Bow-wow	NA	show curiosity, movement
Growl	playing interaction (Handelman 2012)	interation with outside
Bark	warning (Handelman 2012)	discomfort
Howl	warning, play, group cohesion (Siniscalchi et al. 2018)	signal for food, steady

Table 2: Summary of findings with comparison to previously-discovered meanings.

of the words are precise and the background is clear. AudioSet (Gemmeke et al. 2017) consists of hundreds of audio event classes with human-labeled sound clips and further study (Hershey et al. 2021) collected frame-wise labels for a portion of the AudioSet to improve the detection performance. Based on that large-scale audio event dataset, PANNs (Kong et al. 2020), including several models for sound event detection are pre-trained. Previous research also tries to decipher different dog sounds (Hennifer 2023; Pat 2011). Our work proposes the definition of words and develops word segment methods based on these foundation works.

By further splitting the words, we also explore the semantics of subwords and the minimal semantic unit for Shiba Inu dog language. The results show that subwords expressed by IPA vowels do not show a special meaning. This could be attributed to the possibility that IPA vowels are tailored for human language rather than animal communication, yielding the potential need for a broader range of symbols to accurately capture the nuances in vocalization differences.

7 Conclusion

In this paper, we introduce a data-driven approach for exploring the semantics of Shiba Inu vocalization and constructing a dataset including dog words and corresponding context. Compared to the former approaches, it is cost-saving and extensible for datasets. Due to the large amount of data, it provides a probability to explore new contexts and find fine-grained semantics. The approach can be transferred to other animals easily.

We also make some preliminary observations and analyses on the dataset. The analysis shows that the different dog words are used in various contexts. Most of our findings are consistent with previous research and we also explore new semantics for words. We present the word type conveys multiple meanings and can be further divided into more fine-grained types. By given evidence, we declare the minimal semantic unit for the Shiba Inu dog language is word-related. For future work, we can further classify dog sounds into more fine-grained types because we realize a word conveys multiple meanings.

References

Adavanne, S.; Drossos, K.; Çakir, E.; and Virtanen, T. 2017. Stacked convolutional and recurrent neural networks for bird audio detection. In *2017 25th European signal processing conference (EUSIPCO)*, 1729–1733. IEEE.

Andreas, J.; Beguš, G.; Bronstein, M. M.; Diamant, R.; Delaney, D.; Gero, S.; Goldwasser, S.; Gruber, D. F.; de Haas, S.; Malkin, P.; et al. 2022. Towards Understanding the Communication in Sperm Whales. *Iscience*, 104393.

Bermant, P. C.; Bronstein, M. M.; Wood, R. J.; Gero, S.; and Gruber, D. F. 2019. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific reports*, 9(1): 12588.

Bruce, H. 2023. Vowel Chart with Sound Files. Website. https://linguistics.ucla.edu/people/hayes/103/Charts/VChart/.

Cao, J.; Tang, H.; Fang, H.-S.; Shen, X.; Lu, C.; and Tai, Y.-W. 2019. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9498–9507.

Chelsea, G. 2018. 10 sounds your dog makes and what they mean. Website. https://www.insider.com/what-dog-sounds-mean-2018-11.

Faragó, T.; Takács, N.; Miklósi, Á.; and Pongrácz, P. 2017. Dog growls express various contextual and affective content for human listeners. *Royal Society open science*, 4(5): 170134.

Feddersen-Petersen, D. U. 2000. Vocalization of European wolves (Canis lupus lupus L.) and various dog breeds (Canis lupus f. fam.). *Archives Animal Breeding*, 43(4): 387–398.

Garcia, M.; and Favaro, L. 2017. Animal vocal communication: function, structures, and production mechanisms.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.

Handelman, B. 2012. *Canine behavior: A photo illustrated handbook*. Dogwise Publishing.

Hennifer, N. 2023. Canine Communication: Deciphering Different Dog Sounds. Website. https://www.akc.org/expert-advice/advice/canine-communication-deciphering-different-dog-sounds/#:~:text=Your%20dog%20can%20be%20afraid,or%20another%20dog%20in%20play.

Hershey, S.; Ellis, D. P.; Fonseca, E.; Jansen, A.; Liu, C.; Moore, R. C.; and Plakal, M. 2021. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 366–370. IEEE.

Huang, J.; Zhang, C.; Wu, M.; and Zhu, K. 2023. Transcribing Vocal Communications of Domestic Shiba lnu Dogs. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Iwashita, Y.; Takamine, A.; Kurazume, R.; and Ryoo, M. S. 2014. First-person animal activity recognition from egocentric videos. In *2014 22nd International Conference on Pattern Recognition*, 4310–4315. IEEE.

Koh, C.-Y.; Chang, J.-Y.; Tai, C.-L.; Huang, D.-Y.; Hsieh, H.-H.; and Liu, Y.-W. 2019. Bird Sound Classification Using Convolutional Neural Networks. In *CLEF (Working Notes)*.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.

Mao, Y.; and Liu, Y. 2023. Pet dog facial expression recognition based on convolutional neural network and improved whale optimization algorithm. *Scientific Reports*, 13(1): 3314.

Molnár, C.; Pongrácz, P.; Faragó, T.; Dóka, A.; and Miklósi, Á. 2009. Dogs discriminate between barks: The effect of context and identity of the caller. *Behavioural processes*, 82(2): 198–201.

Ng, X. L.; Ong, K. E.; Zheng, Q.; Ni, Y.; Yeo, S. Y.; and Liu, J. 2022. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19023–19034.

Pat, M. 2011. The Meanings Behind Different Dog Sounds. Website. https://www.whole-dog-journal.com/behavior/body-language/the-meanings-behind-different-dog-noises/.

Peter, L. 2023. A Course in Phonetics. Website. http://www.phonetics.ucla.edu/course/chapter1/chapter1.html.

Pongrácz, P.; Molnár, C.; and Miklósi, Á. 2006. Acoustic parameters of dog barks carry emotional information for humans. *Applied Animal Behaviour Science*, 100(3-4): 228–240.

Pongrácz, P.; Molnár, C.; and Miklósi, Á. 2010. Barking in family dogs: an ethological approach. *The Veterinary Journal*, 183(2): 141–147.

Pongrácz, P.; Molnár, C.; Miklósi, Á.; and Csányi, V. 2005. Human listeners are able to classify dog (Canis familiaris) barks recorded in different situations. *Journal of comparative psychology*, 119(2): 136.

Räsänen, O.; Doyle, G.; and Frank, M. C. 2018. Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171: 130–150.

Robbins, R. L. 2000. Vocal communication in free-ranging African wild dogs (Lycaon pictus). *Behaviour*, 1271–1298.

Rossman, Z. T.; Padfield, C.; Young, D.; Hart, B. L.; and Hart, L. A. 2020. Contagious yawning in African elephants (Loxodonta africana): Responses to other elephants and familiar humans. *Frontiers in Veterinary Science*, 252.

Salamon, J.; Bello, J. P.; Farnsworth, A.; and Kelling, S. 2017. Fusing shallow and deep learning for bioacoustic bird species classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 141–145. IEEE.

Scott-Phillips, T.; and Heintz, C. 2023. Animal communication in linguistic and cognitive perspective. *Annual Review of Linguistics*, 9: 93–111.

Siniscalchi, M.; D'Ingeo, S.; Minunno, M.; and Quaranta, A. 2018. Communication in Dogs. *Animals*, 8(8).

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2018. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11): 2740–2755.

Yeon, S. C. 2007. The vocal communication of canines. *Journal of Veterinary Behavior*, 2(4): 141–144.

Yin, S. 2002. A new perspective on barking in dogs (Canis familaris.). *Journal of Comparative Psychology*, 116(2): 189.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.