Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Indonesian idiomatic zoo: Profiling the distribution of animal idiomatic expressions in an Indonesian corpus

Prihantoro Prihantoro

prihantoro@live.undip.ac.id

Diponegoro University

Elizabeth Ika Hesti Aprilia Nindia Rini

Diponegoro University

Dian Annisa Nur Ridha

Diponegoro University

Tanjun Liu

Xi'an Jiaotong-Liverpool University

Article

Keywords: idiomatic expression, semantics, corpus, Indonesian, sentiment

Posted Date: May 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8458469/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Humanities and Social Sciences Communications on October 10th, 2026. See the published version at https://doi.org/10.1057/s41599-026-09311-4.

An Indonesian idiomatic zoo: Profiling the distribution of animal idiomatic expressions in an Indonesian corpus

AUTHOR1[i♠], AUTHOR2[♠], AUTHOR3[♠], AUTHOR4[♥]

AFFILIATION1[♠], AFFILIATION2[♥]

Abstract

Idiomatic expressions in a language may include animals, for instance, 'scapegoat', as one of the compositional elements. In this project, we study animal idiomatic expressions in Indonesian, such as *cinta monyet* (lit. monkey love) 'puppy love', *tikus kantor* (lit. office rat) 'corruptor', and *kupu-kupu malam* (lit. night butterfly) 'prostitutes'. We aim to understand their empirical distributions, degrees of ambiguity, and sentiments. Corpus approaches are used in the data collection and analysis. Combining data-driven and introspective methods, we create a corpus query that we apply to the LCC Indonesian 2023 CQPweb version, a large Indonesian corpus. Most expressions are highly idiomatic with moderate frequencies; negative sentiments are overwhelming. No meaningful associations are present between sentiment and specific animal types: the animal imagery seems present by random distribution. The corpus-based insights we offer contribute to understanding how animal idiomatic expressions function in Indonesian may support cross-linguistic idiom studies and semantic analysis.

Keywords: idiomatic expression, semantics, corpus, Indonesian, sentiment

1. Introduction

The composition of idiomatic expressions may include lexical items from various semantic domains such as WEATHER, TEMPERATURE, PLANT, and ANIMAL. As for ANIMAL, phrases like 'puppy love', 'cash cow', or 'scapegoat' are known English idiomatic expressions with animal as one of their lexical constituents. This study focuses on animal idiomatic expressions in Indonesian, a standardised Malay variety (Eberhard et al. 2022) used as Indonesia's national and official language with more than 250 million speakers.

In Indonesian, idiomatic expressions that include animals in their compositions are widespread, as in *cinta monyet* (lit. monkey love) 'puppy love', *tikus kantor* (lit. office rat) 'corruptor', and *kupu-kupu malam* (lit. night butterfly) 'prostitutes'. Some animals are used in more than one idiomatic expression. The word *buaya* 'crocodile', for instance, is used in several idiomatic expressions, such as in *lidah buaya* (lit. crocodile tongue) 'aloe vera', *air mata buaya* (lit. crocodile tears) 'fake tears', *buaya darat* 'playboy/womaniser'.

While there are some regarding specific animal idiomatic expressions in Indonesian or their comparison with animal idiomatic expressions in other languages, a comprehensive and corpus-based research regarding animal idiomatic expressions in Indonesian is rare; most are based on introspection (elaborated in "literature review section"). Aspects such as distribution, degrees of

ambiguity, and sentiments of the animal idiomatic expressions in Indonesian are underexplored. We translate this background into three main research questions. Each is complemented by subsidiary research questions, which are presented below:

1. How are animal idiomatic expressions distributed in the corpus?
    a. What animal idiomatic expressions are present/absent in the corpus?
    b. Are the idiomatic expressions present in the corpus distributed evenly?
    c. Does the frequency distribution exhibit a pattern consistent with Zipf's Law?
2. How can animal idiomatic expressions be empirically categorised based on the degree of ambiguity?
    a. Which expressions are more frequently observed as idiomatic?
    b. Are there any expressions having more than one idiomatic meaning?
    c. How are the idiomatic expressions distributed based on idiomaticity and frequency?
3. To what extent are the animals associated with the domain and sentiment of the idiomatic expressions?
    a. What animals are present in the idiomatic expressions?
    b. What are the target domains of these animal idiomatic expressions?
    c. How are sentiments for these idiomatic expressions distributed?
    d. Is there any association between the animals and the sentiment of the idiomatic expressions?

2. Literature Review

Qiu and Shi (2022) studied Indonesian idiomatic expressions to build an educational application. Even though they rely on computational methods to create the applications, they include an idiomatic expression lexicon as a part of their database. The database has 247 Indonesian idiomatic expressions, but only four animal idiomatic expressions are found, namely *cinta monyet* (lit. monkey love) 'puppy love'*, ekor kuda* (lit. horse tail) 'braided hair'*,* and *tahi lalat* (lit. fly feces) 'mole/beauty spot'. Most studies regarding animal idiomatic expressions in Indonesian compare the Indonesian expressions to English (Ardianto 2022; Ningsih 2018; Syahri et al. 2023; Mirahayuni and Garnida. 2017). Few studies have compared Indonesian to non-English languages such as Chinese (Waru et al. 2022) or Japanese (Suryadimulya et al. 2023). These studies share a focus on contrastive analysis and translation studies.

Some studies focus on animal idiomatic expressions in Indonesian beyond the translation framework. One example is from Bachrun and Suhandano (2023). While the discussed animal idiomatic expressions vary, they focus on only one animal: dog. Unlike Bachrun and Suhandano, Mirahyuni and Garnida (2017) studied more animals: horses, cows, buffalo, donkeys. However, the number of unique idiomatic expressions is still very limited. Arthanti (2024) is one study in which the coverage of animal idiomatic expressions is quite high: more than 80 expressions. However, her study relies heavily on introspective methods. Conversely, Rusmana et al. (2023) use a corpus approach, drawing on data from a corpus of research articles. Even though they present a frequency distribution of the sentiments (negative, positive, neutral) of the animal idiomatic

expressions, no frequency distribution is given for the animal idiomatic expressions found in the corpus. While they acknowledge that some expressions are ambiguous, no salient analysis is presented. Therefore, it is impossible to understand which animal idiomatic expressions are more frequent or ambiguous than others. This is a handicap, because frequency distribution is one means to discover how linguistic items, in this case animal idiomatic expressions, are distributed in a corpus.

In this study, we aim to offer more rigorous methods, more representative data, more accessible tools, and more comprehensive analysis relative to the relevant previous studies. We combine data-driven and introspective methods, which we present in greater detail in the "Methodology" section. Instead of using a specialised corpus as in Rusmana et al. (2023), we target LCC Indonesian 2023[ii] CQPweb Lancaster version, a general open-access corpus with more than 500 million tokens, as our corpus data. Indeed, a larger general Indonesian corpus, Indonesian Web 2024[iii], exists in Sketch Engine (Kilgariff et al., 2014), but unlike CQPweb (Hardie 2012), Sketch Engine is a proprietary corpus query system. This stands in contrast to CQPweb, which is an open-source software. In terms of analysis, we combine qualitative and quantitative approaches to determine the distribution of Indonesian animal idiomatic expressions, to discover to what extent they are ambiguous, and to understand if meaningful association is present between animals and sentiments incorporated into the idiomatic expressions, which are understudied in the previous works we mentioned.

3. Methodology

3.1. Procedures to Find the Distribution

The following steps were implemented to answer the first RQ: 'How are animal idiomatic expressions distributed in the corpus?' We created a list of idiomatic expressions derived from Arthanti (2024) and Rusmana et al. (2023), two studies in which a considerably large animal idiomatic expression database can be found. Subsequently, we request three Indonesian language specialists to review and improve our list. They amplify the list with some additional idiomatic expressions by introspection. This overall gives us 96 animal idiomatic expressions. We then transform the list into a CQPweb-readable query for use in the target corpus.

The concordance lines that were output from applying the query to the target corpus were carefully examined to verify whether the node in an output concordance line is idiomatic. See **Table 1**, its spreadsheet version. For an expression to qualify as idiomatic, there has to be a domain transfer. We excluded (x) concordance lines where the candidates proved non-idiomatic. For instance, in line 1, Table 1, *sapi* 'cow' refers to a cow (+ANIMAL), whereas in line 2, it refers to a company being exploited for money (+INSTITUTION). This domain transfer confirms that the expression is indeed idiomatic. Therefore, we obtained only lines in which animal idiomatic expressions were used (v).

Table 1. Sample of a concordance lines matrix to validate the use of idiomatic expressions in contexts

No	Idiomatic	Context before	Query item	Context After
1	x	*Nestle juga membantu peternak* Nestle also helped farmers of	*sapi perah* cows for milk (dairy cows)	*di Jawa Timur* at East Java
2	v	*PLN akan menjadi* PLN will be a	*Sapi perah* cash cow	*Untuk mendapatkan dana bagi partai politik* To fund political parties

Subsequently, we filter out the concordances only for the idiomatic expressions we have identified as above. We then tabulate each instance with its respective frequency distribution and proportion. This way, we can examine and rank their occurrence in the corpus. By comparing this to the idiomatic expressions in the query, we can identify which idiomatic expressions are present or absent in the corpus. We transform our findings into a frequency distribution table, which lists each unique idiomatic expression, its corresponding raw frequency, proportion, and Z-score (to know to what extent it deviates from the mean). By observing the proportion and Z-score, we can examine to what extent these expressions are evenly distributed. The animal idiomatic expressions are ranked from highest to lowest frequency, allowing us to build a frequency distribution chart to understand whether or not their distribution follows Zipf's law. These steps help answer RQ1 and its subsidiary RQs.

3.2. Procedures to Categorise the Degree of Ambiguity

The following steps are implemented to answer the 2nd RQ: 'How can animal idiomatic expressions be empirically categorised based on the degree of ambiguity?'. We first extract the frequency distribution of the overall instances from the concordance lines, including the non-idiomatic ones. Then, we measure how many of the expressions are idiomatic and non-idiomatic. This way, we can observe to what degree an expression is idiomatic and to what degree it is ambiguous between idiomatic and literal senses. As seen in **Table 2**, *tikus kantor* 'corruptor' is empirically never used in the literal sense 'office rat' even though its composition allows it to do so, and thus has the highest degree of unambiguity. However, *kupu-kupu malam* is ambiguous, even though it is more frequently used non-idiomatically. Subsequently, frequency and idiomaticity levels are added following the guides given in **Table 3** and in **Table 4**. We created these descriptors as there is no universally accepted intervals for idiomatic expressions. There are indeed frequency intervals for vocabularies for general or pedagogic purposes (Brezina and Gablasova 2015; Coxhead 2000; Dang et al. 2017; Nation 2001), but they cannot be used for Indonesian idiomatic expressions as they target English.

Table 2. Sample of a frequency distribution table comparing idiomatic and non-idiomatic instances

No	Instances	Freq	Idiomatic (%)	Non-Idiomatic (%)	Freq. Level	Idiom. Level
1	*Kupu-kupu malam*	114	50 'prostitute' (43%)	64 'night butterfly' (57%)	5	4
2	*Tikus kantor*	14	14 'corruptor' (100%)	0 'rat in the office' (0%)	3	6

Table 3. Frequency level descriptor

Level	Abs. Freq	Rel. Freq	Remarks
1	0–3	0 – 5.23	Never used/Extremely Rare
2	4–9	6.97 – 15.69	Rare
3	10–24	17.44 – 41.84	Uncommon
4	25–69	43.59 – 120.29	Low Frequency
5	70–199	122.03 – 347.09	Moderate Frequency
6	200–399	348.84 – 695.88	Fairly Frequent
7	400–699	697.63 – 1,219.56	Frequent
8	700–999	1,221.31 – 1,742.24	Very Frequent
9	1000–1499	1,744.00 – 2,615.99	Highly Frequent
10	≥1500	≥2,617.74	Extremely Frequent / Most Common

Table 4. Idiomaticity level descriptor

Level	Grade	Description
6	100% idiomatic	The expression is exclusively idiomatic. No literal usage found in any observed context
5	70–99% idiomatic	Strongly idiomatic in most contexts, though rare literal readings exist.
4	40–69% idiomatic	Balanced usage between idiomatic and literal. Context is key to interpretation.
3	10–39% idiomatic	Predominantly literal; idiomatic usage is secondary and context-dependent.
2	1–9% idiomatic	Idiomatic use is extremely rare. Literal meaning is overwhelmingly dominant.
1	0% idiomatic	No idiomatic usage at all. The expression is always used literally.

If an idiomatic expression has more than one idiomatic meaning, we expand the frequency distribution columns depending on how many different idiomatic meanings it has. This will correspond to the gloss. For instance, under the idiomatic gloss in **Table 5**, there are two idiomatic meanings for *ekor kuda* (either a kind of hairstyle or plant). The frequency of the first meaning refers to M1, and the second meaning refers to M2. We also provide its percentage count to understand the productivity of one idiomatic meaning as compared to another. Our procedure here helps answer RQ2 and its subsidiary RQs.

Table 5. Sample output of an expression with poly-idiomatic meanings

Expression	Gloss:non-idiomatic	Gloss: idiomatic	SUM	M1	(%)	M2	(%)

ekor kuda	horse's tail	1) horse's tail (ponytail hairstyle), 2) a kind of ferns in equisetaceae (plant) family	73	68	93.15	5	6.85

3.3. PROCEDURES TO ASSOCIATE DOMAIN AND SENTIMENT

As for RQ3: 'To what extent are the animals associated with the domain and sentiment of the idiomatic expressions?', we implement the following procedures, whose output sample can be observed in **Table 6**. First, we add an animal label to each idiomatic expression, depending on the animal present. For instance (see line 1), we assign FLY as in *tahi lalat,* as the animal word *lalat* refers to a fly. This allows us to measure how many unique idiomatic expressions an animal may be represented. This is different from frequency of occurrence. For instance, PIG is only used once in *membabi buta* 'run amok even though the frequency is very high. This counts for one unique idiomatic expression. However, CAT may have multiple unique idomatic expressions as in *kumis kucing* 'cat whisker plant'*, lidah kucing* 'a kind of cake'*,* or *nasi kucing* 'small portion rice'.

We then label sentiment for each instance, one among three: NEGATIVE, NEUTRAL, and POSITIVE. We then add DOMESTICATED, SEMI-DOMESTICATED, or WILD labels to inform whether the animal in the idiomatic expressions is domesticated. We also add the target domain, i.e. the semantic domain of the entity referenced by the idiomatic expressions (e.g. HUMAN or more specific HUMAN BODY PARTS, PLANT, etc). For instance, even though the semantic domain of the 'goat' in 'scapegoat' is ANIMAL, the actual semantic domain of the overall term is HUMAN. In the case of ambiguous idiomatic meanings, the cell will be split to accommodate both possible target domains. To measure associations between the type of animals/domestication level and sentiments, we will use chi-squared test. These steps help answer RQ3 and its subsidiary RQs.

Table 6. Sample distribution of animal idiomatic expressions by animal word, domestication status, sentiment, and target domains

No	ANIMAL	Instances	Meaning	Sentiment	Domesticated	Domain
1	FLY	*Tahi lalat*	'mole/beauty spot'	NEUTRAL	WILD	HUMAN BODY PARTS
2	CAT	*Kumis kucing*	'Cat whisker plant'	NEUTRAL	DOMESTICATED	PLANT
3	CAT	*Nasi kucing*	'Small portion rice'	NEUTRAL	DOMESTICATED	FOOD
4	CHICKEN	*Ayam sayur*	'weakling'	NEGATIVE	DOMESTICATED	HUMAN
5	…	…	…	…	…	…

4. THE ZOO OF INDONESIAN ANIMAL IDIOMATIC EXPRESSIONS

4.1. Distribution

In this sub-section, I will answer RQ1: *How are Indonesian animal idiomatic expressions distributed in the corpus?* Here, we will present a list of animal idiomatic expressions that we find from our corpus query and those that are absent. These expressions are not evenly distributed. We map the frequency distribution of these items and discover that the patterns follow Zipf's law (Linders and Louwerse 2023; Zipf 1949). Even though the law is typically applied to individual lexical items, it also seems to hold for phrases or multi-word formats (these animal idiomatic expressions). These observations are elaborated upon in the detailed analysis that follows.

4.1.1. The absence of monkey grip

RQ1.1 is: 'What animal idiomatic expressions are present and absent in the corpus?'. We empirically observed 82 unique idiomatic expressions present in the corpus. Due to its length, we here display the top-ten animal idiomatic expressions (by frequency of occurrence), as shown in **Table 7**. Even from the top-ten list only, we see various animals used in the idiomatic expressions, such as crocodile (1), tiger (2), pig (3), goat (4), snapper (5), and many others. We confirmed a specific kind of animal, snapper, a kind of fish, which confirms Sokolowska's (2018) finding that animals at different taxonomic ranks may be used to create idiomatic expressions. Note that we only saw two different levels, animal and its kind. We did not find a more specific categorisation, such as FISH > SNAPPER > RED SNAPPER. As compared to Bachrun and Suhandano (2023) or Waru et al. (2022), our study is more comprehensive, as they solely focus on specific animals such as dogs and ox. Unlike Damanik and Nasution (2022), who restrict their observation of animal idiomatic expressions to a specific animal group (animals in the zodiac) such limitation is absent in our study.

Table 7. Top-ten animal idiomatic expressions

No	Expression	Literal gloss	Idiomatic gloss
1	*lidah buaya*	crocodile's tongue	*aloe vera*
2	*macan kemayoran*	*kemayoran* tiger	*symbol of a Jakarta football club*
3	*membabi buta*	acting blindly like a wild boar	*run amok*
4	*kambing hitam*	black goat	*scapegoat*
5	*kelas kakap*	large fish class	*high-class, elite group*
6	*gajah putih*	white elephant	*Thailand*
7	*adu domba*	sheep fight	*playing both sides*
8	*kuda lumping*	flat horse	*dancing horse (traditional javanese dance)*
9	*negeri kangguru*	kangaroo country	*Australia*
10	*tahi lalat*	fly feces	*mole (beauty spot)*

As for the absent animal idiomatic expressions, they can be divided into two categories. See **Table 8**. The first category covers the expressions not found in the corpus, such as *ayam abg* 'young sex worker', or *sawan babi* 'epilepsy'. This means the query with these strings yields no result (0). Ten animal idiomatic expressions fall into this category. Most of these expressions are from Arthanti (2024) , whose study is not corpus-based. This suggests that while certain expressions are present in one's mental dictionary, to what extent they are actually used (or perhaps never used empirically), can be reconfirmed by corpus data.

Table 8. Two categories of idiomatic expressions that are absent in the corpus

Status in the target corpus	No	Expression	Literal gloss	Idiomatic gloss
Not found	1	*ayam abg*	teenage chicken	young sex worker (derogatory slang)
	2	*ayam jalanan*	street chicken	prostitute
	3	*ayam plat merah*	red-plate chicken	sex worker connected to government officials (corruption context)
	4	*ayam tambatan*	tethered chicken	old employee
	5	*babi duyung*	mermaid pig	dugong
	6	*membabi jalang*	to pig-bitch	to commit adultery
	7	*gajah menyusu*	elephant breastfeeding	roof of a house
	8	*genggam monyet*	monkey grip	strong grip
	9	*nyamuk gajah*	elephant mosquito	big mosquito
	10	*pintu monyet*	monkey door	1) batwing door 2) small door or escape hatch (often in public transport)
	11	*sawan babi*	pig seizure	epilepsy (colloquial, now considered insensitive)
Found, but never idiomatic	12	*bintang babi*	pig star	morning star
	13	*anjing galak*	fierce dog	a person who is fiercely protective
	14	*buaya buntung*	crocodile without a tail	a man who flirts women but not more than tempting

The second category covers, also presented in Table 8, expressions found in the corpus, but the meaning is never idiomatic. For example, the instances of *buaya buntung* are found in the corpus. Still, the meaning never idiomatically refers to 'a man who flirts with women but is not more than tempting'. Instead, it refers to 'a crocodile without tail' following the compositional meaning of each word in the animal idiomatic expression. Our findings are methodologically new because a list of absent idiomatic expressions is not discussed in previous studies (Ardianto, 2022; Rusmana et al., 2023; Syahri et al., 2023). This may be obtained by our approach, which combines data-driven and introspective methods.

4.1.2. Crocodile, Tiger, and Pig on top

In this subsection, we will answer RQ1.2: 'Are the animal idiomatic expressions present in the corpus distributed evenly?' This will complement RQ1.1, which identifies the animal idiomatic expressions absent in the corpus. The list of animal idiomatic expressions present in the corpus can be observed in **Table 9**. Some expressions are redacted because the list is too long (82 items). In the table, these animal idiomatic expressions are ranked by the frequency of occurrences found in the corpus and complemented by the proportion relative to the sum of all animal idiomatic expressions in the corpus. We can observe that *lidah buaya* 'aloe vera' comes on top, while *ayam abu-abu* 'senior high school prostitute' is ranked at the bottom. We could not compare our findings with findings from Arthanti (2024) , Ardianto (2022), Syachri et al. (2023), or Waru et al. (2022), as in these studies, no frequency count or coverage is offered. We could not find the frequency distribution even with Rusmana et al. (2023), who report that their research is corpus-based.

Table 9. List of top-15 animal idiomatic expressions ranked by frequency of instances (and coverage %)

No	Expressions	Literal	Idiomatic	Count	%
1	*lidah buaya*	crocodile's tongue	aloe vera	1793	12.72
2	*macan kemayoran*	kemayoran tiger	symbol of a jakarta football club	1365	9.68
3	*membabi buta*	acting blindly like a wild boar	run amok	1142	8.10
4	*kambing hitam*	black goat	scapegoat	951	6.75
5	*kelas kakap*	large fish class	high-class, elite group	665	4.72
6	*gajah putih*	white elephant	Thailand	661	4.69
7	*adu domba*	sheep fight	playing both sides	606	4.30
8	*kuda lumping*	flat horse	dancing horse (traditional Javanese dance)	566	4.02
9	*negeri kangguru*	kangaroo country	Australia	549	3.90
10	*tahi lalat*	fly feces	mole (beauty spot)	475	3.37
11	*kabar burung*	bird news	rumor	385	2.73
12	*kutu buku*	book louse/flea	bookworm (avid reader)	299	2.12
13	*kutu loncat*	jumping flea	a person who frequently changes jobs/alliances	285	2.02
14	*kumis kucing*	cat's mustache	cat whisker plant	271	1.92
15	*jalan tikus*	rat path	secret shortcut or illegal route)	268	1.90

We argue that animal idiomatic expressions are not distributed evenly in the target corpus. While this is obvious from the frequency and percentage, it is reconfirmed by the Z-score of each animal's idiomatic expression. Some major and noticeable drops (in Z-score) are observed from the top-9 animal idiomatic expressions, as shown in **Table 10**. Starting from the tenth most frequent animal idiomatic expression, the drop is considered normal. The trend prolongs to the bottom of the idiomatic expression list.

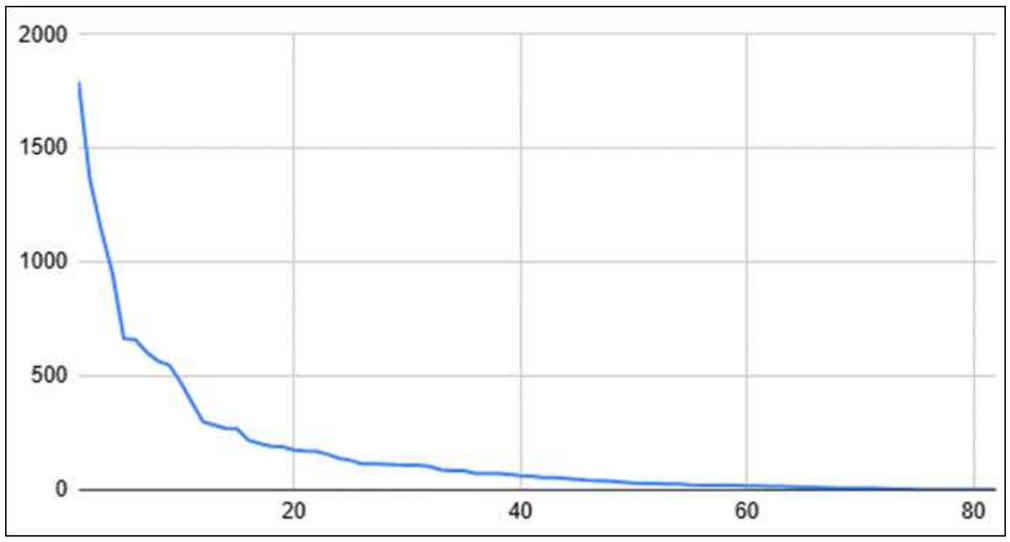
Table 10. List of animal idiomatic expressions ranked by frequency and Z-Scores

Rank	Expressions	Count	Percentage Drop	Z-score	
1	*lidah buaya*	1793	0	4.75	MAJOR DROP
2	*macan kemayoran*	1365	3.04	3.46	MAJOR DROP
3	*membabi buta*	1142	1.58	2.79	MAJOR DROP
4	*kambing hitam*	951	1.36	2.22	MAJOR DROP
5	*kelas kakap*	665	2.03	1.36	Noticeable Drop
6	*gajah putih*	661	0.03	1.34	Noticeable Drop
7	*adu domba*	606	0.39	1.18	Noticeable Drop
8	*kuda lumping*	566	0.28	1.06	Noticeable Drop
9	*negeri kangguru*	549	0.12	1.01	Noticeable Drop
10	*tahi lalat*	475	0.53	0.78	Normal
11	*kabar burung*	385	0.64	0.51	Normal
12	*kutu buku*	299	0.61	0.25	Normal
13	*kutu loncat*	285	0.10	0.21	Normal
14	*kumis kucing*	271	0.10	0.17	Normal
15	*jalan tikus*	268	0.02	0.16	Normal
...	...	...	...	...	...
72	*rabun ayam*	6	0.01	-0.63	Normal
73	*baju monyet*	3	0.02	-0.64	Normal
74	*kuda pacu*	3	0.00	-0.64	Normal
75	*mati ayam*	2	0.01	-0.64	Normal
76	*otak udang*	2	0.00	-0.64	Normal
77	*ayam biru*	2	0.00	-0.64	Normal
78	*celana monyet*	2	0.00	-0.64	Normal
79	*anjing penjilat*	1	0.01	-0.64	Normal
80	*buta ayam*	1	0.00	-0.64	Normal
81	*ayam abu-abu*	1	0.00	-0.64	Normal

4.1.3. Adherence to Zipf Law

In this subsection, we will present our answer to RQ1.3: 'Does the frequency distribution exhibit a pattern consistent with Zipf's Law?'. We convert the complete list of the frequency distribution into a smooth line chart in which the x-axis corresponds to idiomatic expressions by rank, and the y-axis corresponds to the frequency. As observed in **Figure 1**, frequencies of the idiomatic expressions follow an inverse power law by rank. This distribution aligns with Zipf's Law. In addition, the chart resembles frequency distribution charts generated using frequency distribution data from major corpora such as BNC, Brown, and LOB.

Figure 1. Frequency distribution chart of the animal idiomatic expressions in the corpus (x=frequency, y=rank)



In many studies (Hordijk 2022; Linders and Louwerse 2023; Takamichi et al. 2023), Zipf's Law is applied to individual word tokens. However, in this study, we have demonstrated that its application to idiomatic expressions, specifically animal idiomatic expressions, exhibits similar results. This may align with findings from Williams et al.(2015) or Ha et al. (2009), who studied the application of Zipf law in phrases and n-grams.

4.2. Ambiguity

In this sub-section, we will answer RQ2: 'How can animal idiomatic expressions be empirically categorised based on the degree of ambiguity?'. We here have devised a 6-grade measure to categorise the idiomatic expressions, ranging from those that are always idiomatic to those that are rarely so, on the basis of the target corpus data. Our gradation is different from Rusmana et al. (2023), in which the expressions are divided into 3 grades: literal, semi-opaque, and full idiom.

With the 6-grade scale, we can empirically pinpoint expressions ranging from always idiomatic to rarely or never idiomatic. Some idioms, in theory, may be polysemous, but our corpus data observations only show a small number of instances of expressions with multiple idiomatic meanings. Regarding distribution, we discover two groups in which many idiomatic expressions congregate by frequency and idiomaticity axes: highly idiomatic with moderate frequency and minimally idiomatic with low frequency.

4.2.1. 'Crocodile Tongue' is Always Idiomatic but not 'Crocodile Tail'

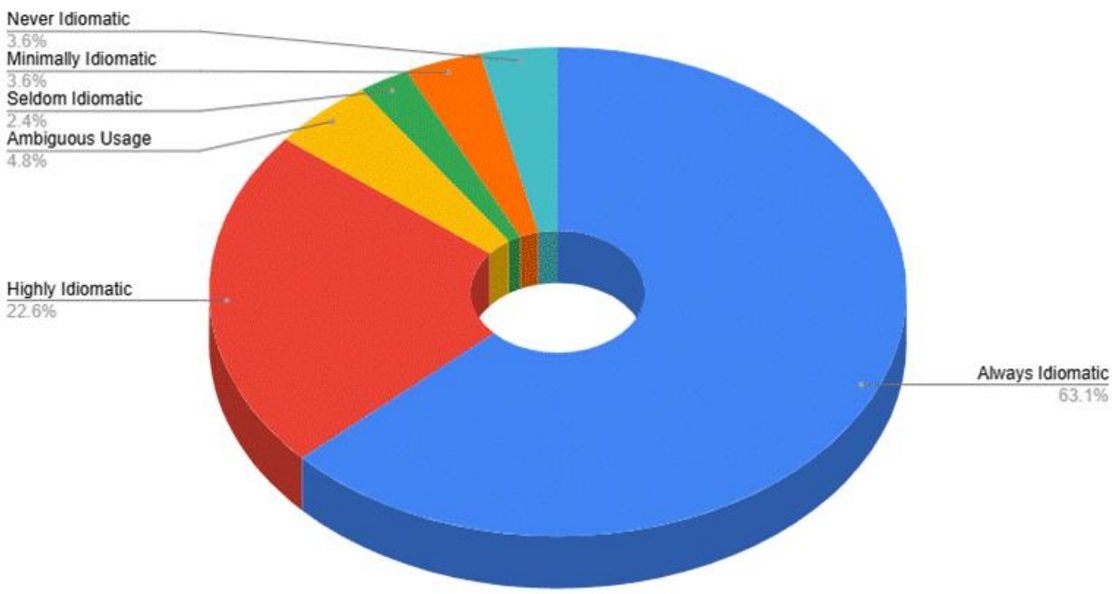
The first step to answer RQ2 is by answering RQ2.1: 'Which expressions are more frequently observed as idiomatic?'. In this question, we examine to what extent animal expressions are present idiomatically compared to their literal meanings. We present this in **Table 11** (redacted for space limitations), which shows some selected expressions which correspond to their gradations. Figure 2 summarises the table based on the proportion.

Table 11. Frequency distribution of the animal idiomatic expressions (graded)

Expressions	Literal	Idiomatic	Idiomatic (%)	Degree of idiomaticity
lidah buaya	crocodile's tongue	aloe vera	100.00	6
macan kemayoran	kemayoran tiger	symbol of a jakarta football club	100.00	6
membabi buta	acting blindly like a wild boar	run amok	100.00	6
kambing hitam	black goat	scapegoat	100.00	6
kelas kakap	large fish class	high-class, elite group	100.00	6
...	...	...	...	...
kutu loncat	jumping flea	a person who frequently changes jobs/alliances	98.28	5
kumis kucing	cat's mustache	cat whisker plant	98.91	5
buaya darat	land crocodile	womanizer, playboy	98.57	5
...	...	...	...	...
kupu-kupu malam	night butterfly	prostitute	43.86	4
otak udang	shrimp brain	foolish person	40.00	4
...	...	...	...	...
anjing penjaga	guard dog	security officer	14.15	3
belalai gajah	elephant's trunk	A portable tunnel that connects a plane door to an airport door	20.00	3
...	...	...	...	...
sapi perah	dairy cow	exploited person/group	2.31	2
...	...	...	...	...
buaya buntung	crocodile without a tail	a man who flirts women but not more than tempting	0.00	1

As observed, many expressions are always idiomatically used (grade 6) as in *kambing hitam* 'scapegoat', *membabi buta* 'run amok', and *lidah buaya* 'aloe vera plant'. They account for 62% of the total idiomatic expressions. Grade 5 idioms are highly idiomatic, covering 22% of the overall expressions. These include animal idiomatic expressions such as *ayam kampus* 'college student prostitute', *jalan tikus* 'less-known alternative route (often illegal), and *tikus berdasi* 'corrupt businessman/official'. Altogether, 84% of the overall data is composed of grade 5 and 6, which is a very significant amount. While grade 6 and grade 5 animal idiomatic expressions are the most frequent, we observe a contrasting trend with the remaining expressions. The animal idiomatic expressions in grades 4,3,2, and 1 only range between 2-5% as shown by **Figure 2**. While Rusmana et al. (2023) give some examples of idiomatic expressions with different gradations, as indicated earlier, it seems that gradation is not transparently applied as it is in our study. A more elaborate description may be found in Suryadimulya et al. (2023), but they aim to compare body-part idiomatic expressions from Indonesian and Japanese.

Figure 2. Distribution of animal idiomatic expressions (graded)



4.2.2. Poly-idiomaticity

In this subsection, we address RQ 2.2: 'Are there any expressions having more than one idiomatic meaning?'. As shown in **Table 12**, only three animal idiomatic expressions from our list which exhibit poly-idiomaticity (more than one idiomatic meaning across different contexts). These expressions are *cakar ayam, ekor kuda,* and *kuping gajah.* The first two expressions have two idiomatic senses (M1,M2), while the last has 3 idiomatic senses (M1, M2, M3).

Table 12. Sample output of an expression with poly-idiomatic meanings

Expression	Gloss: non-idiomatic	Gloss: idiomatic	SUM	M1	(%)	M2	(%)	M3	(%)
cakar ayam	chicken's claw	1) a kind of foundation with multiple branching downward the earth, 2) Selaginella doederleinii, a kind of plant, 3) messy handwriting	109	108	99.08	1	0.92	0	0.00

ekor kuda	horse's tail	1) horse's tail (ponytail hairstyle), 2) a kind of ferns in equisetaceae (plant) family	73	68	93.15	5	6.85	0	0.00
kuping gajah	elephant's ears	1) a kind of sweet-savory snack shaped like ear, 2) a kind of orchid with realtively big leafs, 3) a trophy with 2 handles (loving cup)	62	46	74.19	15	24.19	1	1.61

However, not all idiomatic meanings presumed to be present by introspection are attested empirically in the corpus. Our corpus analysis reveals that for *cakar ayam*, only 2 out of 3 meanings have instances in authentic usage. The third idiomatic meaning of *cakar ayam,* as messy handwriting (M3), is not found. As for *ekor kuda* (2 meanings) and *kuping gajah* (3 meanings), all idiomatic meanings have instances. Also, the frequency distribution for empirically present meanings is uneven. For *kuping gajah,* for example, the instance of CAKE (74.19 %) meaning ranks first, followed by PLANT (24.19 %), and the last one is CUP (1.61%). The same trend applies to the other two animal idiomatic expressions.

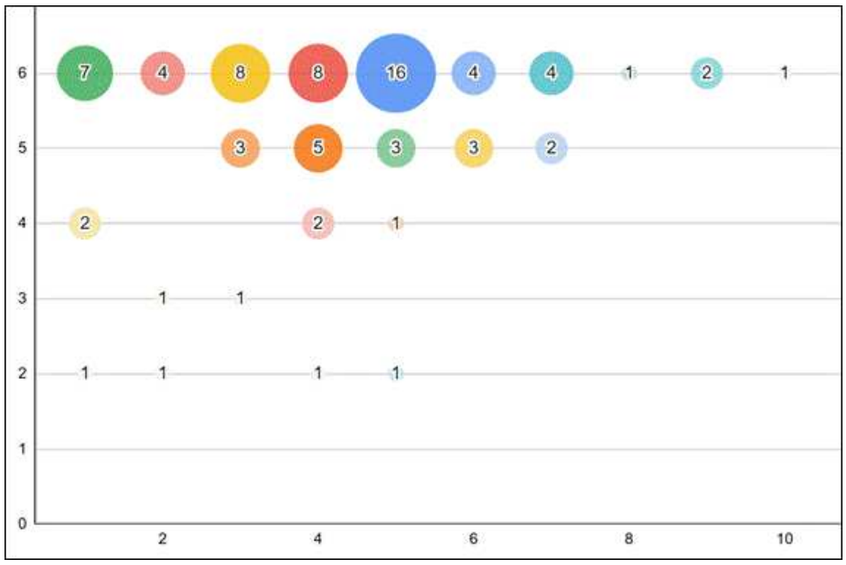
Rusmana et al. (2023) only list some animal idiomatic expressions that may have more than one meaning by adding numbers next to the expressions. However, it is not very clear which idiomatic meanings are present or absent. For instance, in the case of *cakar ayam,* they only wrote *cakar ayam1* and *cakar ayam2.* In this way, confirming which two of the three meanings (as shown in **Table 12**) are present is difficult. As for our case, we can quickly pinpoint which meanings are present or absent. In addition, they also do not correlate the findings with frequency, and thus, it is impossible to compare which is more productive. Conversely, we can observe the productivity as we supply frequency counts for all the animal idiomatic expressions. For instance, in the case of *kuping gajah,* it is mostly used under the domain of CAKE, as compared to PLANT or CUP.

4.3.3. Frequency vs Idiomaticity

In this subsection, we will answer RQ2.3: 'How are the idiomatic expressions distributed based on idiomaticity and frequency?'. So far, we have found no study comparing these two scales in the same way as ours. In Alavi and Rajabpoor(2014), and Kovaliuk (2023), frequencies are transparently given, even though the study is on idiomatic expressions in general. Note that idiomaticity is not analysed. Popiel and McRae (1988) compare idiomaticity and frequency, but based on introspection rather than corpus data.

We created a bubble chart with frequency (x) and idiomaticity (y) axes, as shown in **Figure 3,** allowing us to create coordinates. There are 60 grids by raw counts, but empirically, only 24 grids were observed as the coordinates where the frequency and idiomaticity expressions are anchored. The size of each bubble in the chart represents the number of unique animal idiomatic expressions clustered within the same coordinate. Therefore, the bigger the bubble, the more idiomatic expressions cluster in the same grid

Figure 3. Frequency (x) and idiomaticity (y) grids of animal idiomatic expressions



The chart shows most animal idiomatic expressions clustering at grid 5-6 (column 5, row 6) with 16 expressions, such as *cinta monyet* 'puppy love'*, kelinci percobaan* 'guinea pig'*, nasi kucing* 'small portion rice' . These are expressions that are always idiomatic and with moderate frequency. Grid 3-6 and 4-6 (rare/low frequency and highly idiomatic) rank second with 8 expressions such as *malu-malu kucing, air mata buaya, tikus kantor* (half of the first rank grid). This suggests that most animal idiomatic expressions are highly or even always idiomatic with very low to fairly frequent frequency.

4.3. Animals and Their Association

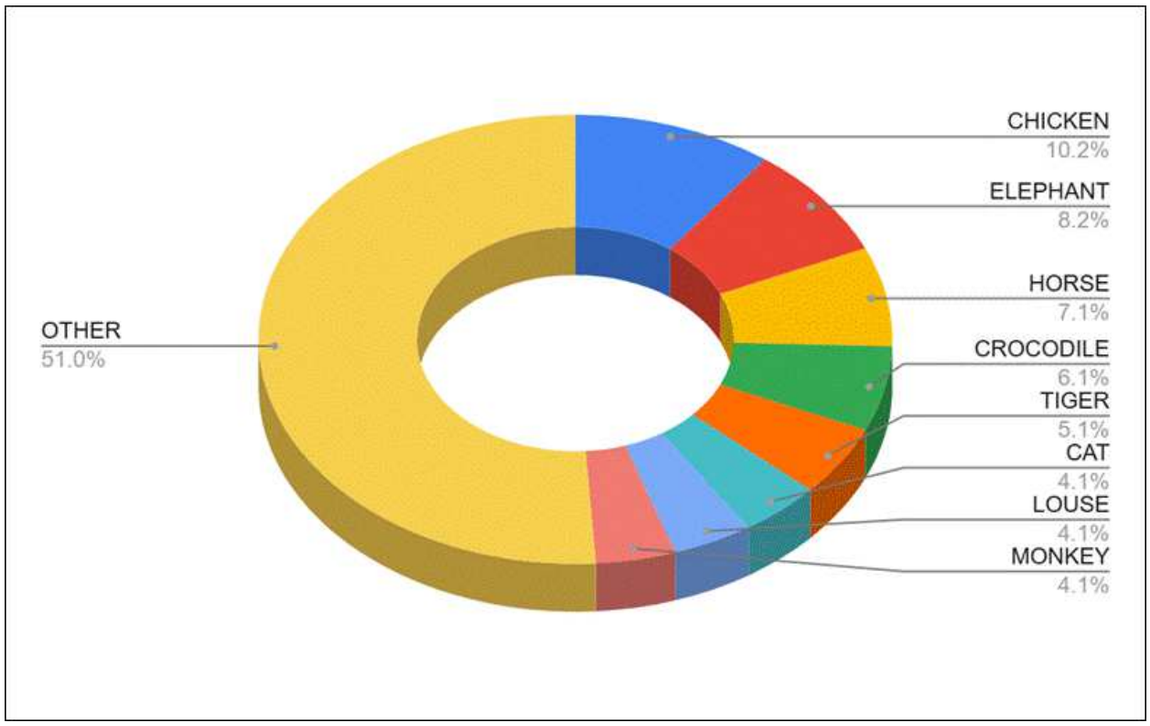
In this sub-section, I will answer RQ3: 'To what extent are the animals associated with the domain and sentiment of the idiomatic expressions?'. The majority of the idiomatic expressions analysed in this corpus carry negative sentiments. However, no meaningful association emerges between the animal reference and the sentiment expressed when examined across different animal types. Although cultural associations may suggest certain emotional connotations, the data indicates that sentiment is distributed independently of the specific animal invoked. These findings and the sentiment classifications and examples are discussed in greater detail in the following section.

4.3.1. Animal Types

This subsection will answer RQ3.1: 'What animals are present in the idiomatic expressions?'. Our empirical findings show a wide range of animals involved in idiomatic usage. In total, there are 32 distinct animals namely: CHICKEN (ten unique idiomatic expressions), ELEPHANT (8), HORSE (7), CROCODILE (6), TIGER (5), CAT (4), LOUSE (4), MONKEY (4), GOAT (3), PIG (3), RAT (3), COW (2), DOG (2), DUCK (2), FISH (2), BIRD (1), BULL (1), BUTTERFLY (1), FLY (1), FROG (1), KANGAROO (1), LEECH (1), LION (1), OCTOPUS (1), RABBIT (1), RHINO (1), SHEEP (1), SHRIMP (1), SQUIRREL (1), TURTLE (1), WOLF (1), SHEEP (1), WORM (1).

The frequencies of the animals being used in the idiomatic expressions range from only one unique idiomatic expression (*burung* 'bird' in *kabar burung* 'rumor') to ten (e.g. *ayam* 'chicken' in *ayam kampus* 'college student whose side job is a prostitute', *ayam sayur* 'weakling' etc). In terms of their distribution, the three most frequent animals are chicken (10), elephant (8), and horse (7). I merged animals whose number of unique idiomatic expressions is three or less into OTHER. This includes *kupu-kupu* 'butterfly' as in *kupu-kupu malam* 'prostitute (general)', *udang* 'shrimp' in *otak udang* 'foolish', or *kelinci* 'rabbit' in *kelinci percobaan* 'test subject'. Altogether, they form more than 50% of the animals (types). The presence of these animals confirms findings studied in Bachrun and Suhandanu (2023), Rusmana et al. (2023), Waru et al., (2022), Arthanti (2024) and Ardianto (2022). Altogether, we reconfirm the diversity of animals in the idiomatic expressions, even though some animals appear in more unique idiomatic expressions than others. For their distribution, see **Figure 4**.

Figure 4. Distribution of unique animals in the idiomatic expressions



If we consider the frequency of occurrence instead of uniqueness, the results are shown in Table 13. Impressionistically, when an idiomatic expression has an extremely high frequency, it will remain near the top position. Conversely, even though the animal is present in multiple unique idiomatic expressions, if the combined frequency is low, it will not be able to surpass the previously explained case. CHICKEN, for example, has ten unique idiomatic expressions. However, their combinations of frequency of occurrences cannot outrank extremely frequent idiomatic expressions, for instance, PIG, even though it has only 3 unique idiomatic expressions. This is because one idiomatic expression, in this case, *membabi buta* 'run amok', has an extremely high frequency (1142 times). **Table 13** lists the ten animals each with the highest combined frequency of occurrences across idiomatic expressions.

Table 13. Top ten animal idiomatic expressions (merged by animal type)

ANIMAL	FREQUENCY
CROCODILE	1936
TIGER	1878
PIG	1275
ELEPHANT	1098
GOAT	979
HORSE	796
LOUSE	778
FISH	693
SHEEP	606
KANGAROO	549

4.3.2. Target Domains

In this subsection, we will answer RQ3.2: 'What are the target domains of these animal idiomatic expressions?'. Our sample data of idiomatic expressions and their target domains is shown in **Table 14**. The overall target domains (and number of unique idiomatic expressions) are: HUMAN (29), FOOD (5), PLANT (4), COUNTRY (3), HUMAN DISEASE (4), INSTITUTION (2), HUMAN ACTION (2), HUMAN BEHAVIOUR (2), HUMAN MIND (2), HUMAN STATUS (2), HUMAN TEAR (2), HUMAN/INSTITUTION (2), THING (2), ART (1), CLASS/GRADE (1), CLOTHING (1), CONSTRUCTION (1), DANGER (1), DISEASE (1), ECONOMY (1), FOOD/PLANT (1), HAIR/PLANT (1), HUMAN BODY PART (1), HUMAN FEATURE (1), HUMAN FEELING (1), HUMAN STANCE (1), HUMAN STATE (1), HUMAN TRAIT (1), HUMAN TRICK (1), HUMAN/ANIMAL (1), LOCATION (1), NATURE (1), NEWS (1), SITUATION (1), STATE (1), THING/FOOD (1) The procedure for mapping the target domain of each idiomatic has been discussed at length in the "Methodology" section.

Table 14. Some animal idiomatic expressions with their target domains

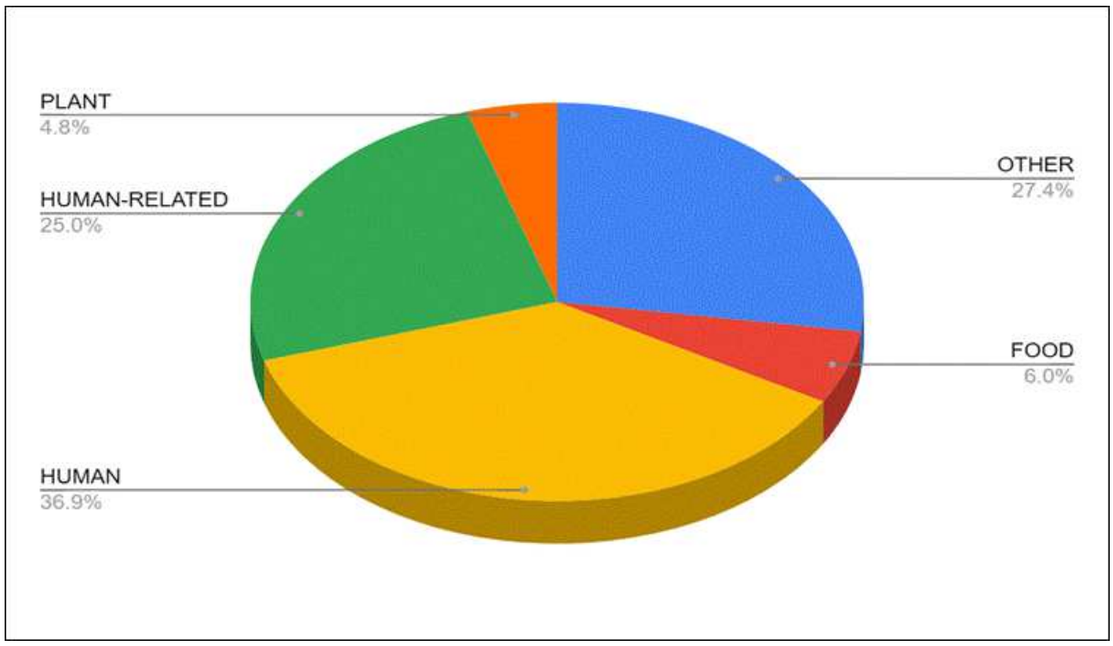
TARGET DOMAIN	Expression	Gloss: idiomatic
HUMAN ACTION	*membabi buta*	run amok
HUMAN	*kambing hitam*	scapegoat
HUMAN STATUS	*kelas kakap*	high-class, elite group
COUNTRY	*gajah putih*	Thailand
HUMAN	*adu domba*	playing both sides
ART	*kuda lumping*	dancing horse (traditional javanese dance)
COUNTRY	*negeri kangguru*	Australia
HUMAN BODY PART	*tahi lalat*	mole (beauty spot)
NEWS	*kabar burung*	rumor
HUMAN	*kutu buku*	bookworm (avid reader)
HUMAN	*kutu loncat*	a person who frequently changes jobs/alliances

See **Figure 5**. Our findings show that there is an intense domination of HUMAN aspect (HUMAN and other HUMAN-related target domains such as HUMAN BODY PART, HUMAN

ACTION, HUMAN STATUS) as the references in the idioms involving animals, as in *kutu* 'louse' in *kutu buku* 'an avid reader' or *domba* 'sheep' in *adu domba* 'to divide and conquer'. This means animal imagery often refers to human or human-related traits, actions, or features. Other target domains, such as PLANT or FOOD, while present, occupy very few slots, ranging from 1% to 6%. For chart visibility purposes, I merged target domains with frequency of occurrence 3 or less into OTHER categories (COUNTRY, NATURE, DISEASE, ECONOMY etc).

Our categorisation here is more fine-grained than Rusmana et al. (2023), which map the domains into HUMAN and NON-HUMAN. However, we share findings that the domain mapping to HUMAN is more dominant than others. While not explicitly written so, the same trend is also observed in Waru et al. (2022) Arthanti (2024) and Ardianto (2022). HUMAN remains the dominant target domain even in non-Indonesian languages such as English or Chinese (Damanik & Nasution, 2022; Liu, 2019).

Figure 5. Distribution of target domains



4.3.3. Sentiments

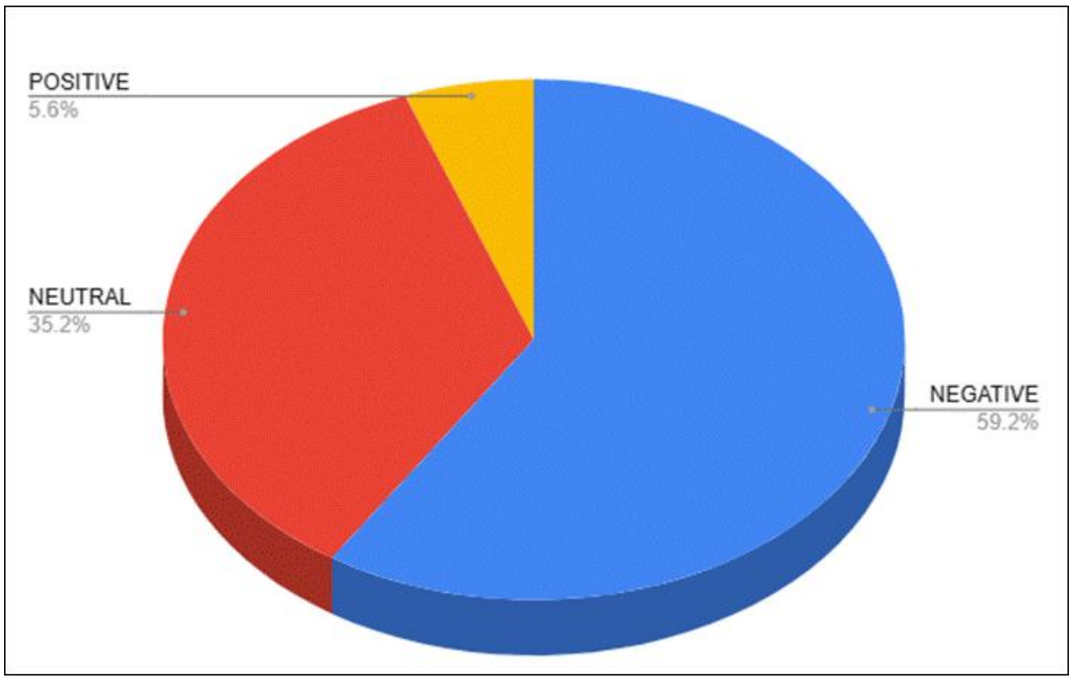
In this subsection, we examine the overall distribution of Indonesian animal idiomatic expressions' sentiments to answer RQ3.3. Some sample data is presented in **Table 15**. All animal idiomatic expressions are assigned sentiment labels, as the "Methodology" section explains. We discovered that the majority (67.1%) of the animal idiomatic expressions are negative, followed by neutral expressions at 28.2%, with very few animal idiomatic expressions with positive (4.7%) sentiment, as shown in **Figure 6**.

Table 15. Some animal idiomatic expressions with their sentiments

Sentiment	Expressions	Literal	Idiomatic
NEGATIVE	*kabar burung*	bird news	rumor

NEUTRAL	*kutu buku*	book louse/flea	bookworm (avid reader)
NEGATIVE	*kutu loncat*	jumping flea	a person who frequently changes jobs/alliances
NEUTRAL	*kumis kucing*	cat's mustache	cat whisker plant
NEGATIVE	*jalan tikus*	rat path	secret shortcut or illegal route)
NEGATIVE	*kelas teri*	anchovy class	low-class, insignificant group
NEUTRAL	*bajing loncat*	jumping squirrel	refers to highway robbers

Figure 6. A frequency distribution chart of the sentiments of animal idiomatic expressions



The domination of negative sentiments suggests that many animal-based idiomatic expressions are used to imply undesirable circumstances, characters, or features, as shown in *adu domba* (lit. sheep fight), which denotes a 'divide and conquer' tactic, *buaya darat* (lit. land crocodile), which refers to a womaniser or *kambing hitam* (lit. black goat), which corresponds to a scapegoat. These empirical findings support the argument that animal imagery is often used to express negative sentiments. Conversely, animal idiomatic expressions with positive sentiments are extremely rare (only 3 found, which constitutes only around 3%), namely *macan Asia* (lit. Asian tiger) 'strong Asian economies', *kuda pacu* (lit. racing horse) 'front runner', and *buaya keroncong* 'experts in *keroncong* songs'.

Approximately one-third of the expressions are neutral. In some cases, this refers to the animal idiomatic expressions that are inherently neutral, as in *kuping gajah* 'a kind of cake' and *baju monyet* 'overall (clothing)'. They are mainly attributed to entities other than humans, mostly plants. However, in some cases, it is also assigned to an expression whose sentiment may be positive or negative based on the context.

We share the same findings with Damanik and Nasution (2022) and Rusmana et al. (2023) in terms of the order of the sentiments (NEGATIVE, NEUTRAL, POSITIVE). As for other studies

(Ardianto 2022; Arthanti 2024; Syahri et al. 2023; Waru et al. 2022), such a phenomenon is also observed, even though no explicit frequency distribution is offered.

4.3.4. No Association Found

In this subsection, we will answer RQ3.4: Is there any association between the animals and the sentiment of the idiomatic expressions? We implemented a Chi-Square Test of Independence to find the answer to this question. We used the animal and sentiment data from **Table 16**. The first test aims to discover whether sentiment polarity (negative, neutral, or positive) is associated with the type of animal. The results showed no statistically significant association between these variables, $\chi^2(62) = 49.99$, $p = 0.863$. We argue that different sentiments associated with varying types of animals are more likely to be caused by a random distribution instead of a systematic association. Indeed, some animals, such as pigs, crocodiles, or chickens, are more frequently associated with negative sentiment, but the results show no statistically meaningful association.

Table 16. Redacted data on sentiment and domestication type of animal idiomatic expressions

Domestication	Sentiment	Expressions	Literal	Idiomatic
Semi-Domesticated	NEGATIVE	*kabar burung*	bird news	rumor
Wild	NEUTRAL	*kutu buku*	book louse/flea	bookworm (avid reader)
Wild	NEGATIVE	*kutu loncat*	jumping flea	a person who frequently changes jobs/alliances
Domesticated	NEUTRAL	*kumis kucing*	cat's mustache	cat whisker plant
Wild	NEGATIVE	*jalan tikus*	rat path	secret shortcut or illegal route)
Semi-Domesticated	NEGATIVE	*kelas teri*	anchovy class	low-class, insignificant group
Wild	NEUTRAL	*bajing loncat*	jumping squirrel	refers to highway robbers

The second test observed whether sentiment is related to the level of domestication: wild, domesticated, or semi-domesticated. Here, we use the domestication and sentiment data whose sample is shown in **Table 16** earlier. We could not find any similar study on Indonesian idiomatic expressions that applies the same procedure. Cortellazo (2021) focuses on the idiomatic expressions of wild animals in English. However, no semi-domesticated or domesticated animals are observed. She also did not apply the same measure as we did.

As for our experiment, with this more generic categorisation (compared to individual animal type), we still fail to conclude a meaningful association: $\chi^2(4) = 2.958$, $p = 0.564$. The reading shows that the sentiments of idiomatic expressions are not correlated with the status of the animals (wild, domesticated, or somewhere in between). For example, *macan* ‘tiger’, a wild animal, can express a negative sentiment, as in *macan ompong* ‘someone who looks scary but weak’, or a positive sentiment, as in *macan asia* ‘economically strong country in Asia’. This shows that idiomatic expressions are highly idiosyncratic.

5. Conclusion

We here have answered all the RQs established in the introduction. Our findings contribute to the understanding of animal idiomatic expressions in Indonesian by systematic frequency distribution of ambiguity, idiomaticity, sentiment, and semantic domains. We have confirmed the uneven distribution of idioms and their alignment with Zipf's Law. Our findings suggest that intuition-based assumptions of idiomatic expressions should be confirmed by corpus data. We here argue that these findings may contribute to machine translation, cross-linguistic or cognitive analyses, among others. For example, a direct pattern matching to dictionary items may be safe to use for idiomatic expressions in grade 6 (always idiomatic). Directions for further studies, may include an investigation, or compare idiomatic patterns across corpus sections with different metadata (demographic information, time, among others). For instance, by studying the data diachronically, an evolution of animal idiomatic expressions remains to be studied.

REFERENCES

Alavi, Sepideh, and Aboozar Rajabpoor (2014) Analyzing Idioms and Their Frequency in Three Advanced ILI Textbooks: A Corpus-Based Study. English Language Teaching 8(1): 170. https://doi.org/10.5539/elt.v8n1p170

Ardianto, Ardik (2022) Animal Metaphors in Indonesian and English. Jurnal Humaya: Jurnal Hukum, Humaniora, Masyarakat, Dan Budaya 2(2): 84–91. https://doi.org/10.33830/humayafhisip.v2i2.3949

Arthanti, D (2024) Pembentukan Ekspresi Dengan Nama Hewan Dalam Bahasa Indonesia. Thesis, Universitas Airlangga. https://repository.unair.ac.id/129204/

Bachrun, Nitis Asmara, and Suhandano Suhandano (2023) Metaphorical Conceptualization Employed by Animal Lexicon 'Dog' in English and Indonesian Proverbs: A Comparative Study. Journal of Social Research 2(3): 892–900. https://doi.org/10.55324/josr.v2i3.741

Brezina, Vaclav, and Dana Gablasova (2015) Is There a Core General Vocabulary? Introducing the New General Service List. Applied Linguistics 36 (1): 1–22. https://doi.org/10.1093/applin/amt018

Cortelazzo, Guidlao Maria (2021) Wild Animals in English Idiom Translation. Zoology and Animal Physiology 2(4): 25-37 https://doi.org/10.38007/ZAP.2021.020403.

Coxhead, Averil (2000) A New Academic Word List. TESOL Quarterly 34(2): 213. https://doi.org/10.2307/3587951.

Dang, Thi Ngoc Yen, Averil Coxhead, and Stuart Webb (2017) The Academic Spoken Word List. Language Learning 67(4): 959–97. https://doi.org/10.1111/lang.12253

Eberhard, David M., Gary Francis Simons, and Charles D. Fennig (2022) Ethnologue®: Languages of Asia. Twenty-Fifth edition. SIL International. Available via https://www.ethnologue.com/. Accessed 2024-05-27

Ha, Le Quan, Philip Hanna, Ji Ming, and F. J. Smith (2009) Extending Zipf's Law to n-Grams for Large Corpora. Artificial Intelligence Review 32(1-4): 101–13. https://doi.org/10.1007/s10462-009-9135-4.

Hardie, Andrew (2012) CQPweb Combining Power, Flexibility and Usability in a Corpus Analysis Tool. International Journal of Corpus Linguistics 17(3): 3. https://doi.org/10.1075/ijcl.17.3.04har.

Hordijk, Wim (2022) Snooker Statistics and Zipf's Law. Stats 5(4): 985–92. https://doi.org/10.3390/stats5040058.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, et al (2014) The Sketch Engine: Ten Years On. Lexicography 1(1): 1. https://doi.org/10.1007/s40607-014-0009-9.

Kovaliuk, Yurii (2023) Idioms in General English Corpora: On Frequency, Register, and Cross-Variety Variation. Respectus Philologicus 44(49): 11–24. https://doi.org/10.15388/RESPECTUS.2023.44.49.104

Linders, Guido M., and Max M. Louwerse (2023) Zipf's Law Revisited: Spoken Dialog, Linguistic Units, Parameters, and the Principle of Least Effort. Psychonomic Bulletin & Review 30(1): 77–101. https://doi.org/10.3758/s13423-022-02142-9

Mirahayuni, Ni Ketut, and Susie Chrismalia Garnida (2017) Sapi Perah' or 'Kuda Beban': Cultural Values in Indonesian Proverbs Associated with Horse, Cow, Buffalo and Donkey. KnE Social Sciences 1(3): 175. https://doi.org/10.18502/kss.v1i3.737.

Nation, I. S. P (1st eds) (2001) Learning Vocabulary in Another Language. Cambridge University Press, Cambridge https://doi.org/10.1017/CBO9781139524759.

Ningsih, Siti Ayu (2018) Penggunaan Istilah Binatang Dalam Metafora Bahasa Indonesia Dan Bahasa Inggris. Arkhais: Jurnal Ilmu Bahasa Dan Sastra Indonesia 9(1): 49–58. https://doi.org/10.21009/ARKHAIS.091

Popiel, Stephen J., and Ken McRae (1988) The Figurative and Literal Senses of Idioms, or All Idioms Are Not Used Equally. Journal of Psycholinguistic Research 17(6): 475–87. https://doi.org/10.1007/BF01067912

Qiu, Xinying, and Guofeng Shi (2022) InDEX: Indonesian Idiom and Expression Dataset for Cloze Test. Paper presented at the 2022 International Conference on Asian Language Processing (IALP), Chinese University of Hong Kong (CUHK), Singapore, 27-28 October 2022, https://doi.org/10.1109/IALP57159.2022.9961288

Rusmana, Kadek Bagus, I Putu Kelvin Meiadiaksa Permana, and Gerdarus Kevin Sebastian Darsana (2023) Analysis of Corpus Research Articles on Indonesian Idioms Based on Animal Names. Paper presented at the 2nd International Student Conference on Linguistics (ISCL 2022), Universitas Warmadewa Bali, Bali, 20 October 2022, https://doi.org/10.2991/978-2-38476-014-5_3

Sokołowska, Olga (2018) Levels of Categorization in Animal Idiomatic Expressions. Beyond Philology An International Journal of Linguistics, Literary Studies and English Language Teaching 5(1): 69–86. https://doi.org/10.26881/bp.2018.1.04

Suryadimulya, Agus Suherman, Reiza D. Dienaputra, Susi Machdalena, Awaludin Nugraha, and Nyai Kartika (2023) Unveiling Cultural Intelligence: A Comparative Study of Japanese and Indonesian Idiomatic Expressions. Indonesian Journal of Applied Linguistics 13(1): 1–12. https://doi.org/10.17509/ijal.v13i1.58248

Syahri, Indawan, Masagus Sulaiman, and Muhammad Faisal (2023) English and Indonesian Idioms of Animals. Darussalam English Journal (DEJ) 3(2): 188–224. https://doi.org/10.30739/dej.v3i2.2475

Takamichi, Shinnosuke, Hiroki Maeda, Joonyong Park, Daisuke Saito, and Hiroshi Saruwatari. (2023) Do Learned Speech Symbols Follow Zipf's Law?" Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2309.09690.

Waru, Dian Sari Unga, Ria Riski Marsuki, and Firman Saleh (2022) The Comparative Analysis of Cultural Aspect in the Idioms of Zodiac Animals in Indonesian and Chinese Language. Paper presented at the 9th Asbam International Conference (Archeology, History, & Culture In The Nature of Malay) (ASBAM 2021), Universitas Hasanuddin, Makassar, 27-28 October 2021, https://doi.org/10.2991/assehr.k.220408.037

Williams, Jake Ryland, Paul R. Lessard, Suma Desu, et al (2015) Zipf's Law Holds for Phrases, Not Words. Scientific Reports 5(1): 12209. https://doi.org/10.1038/srep12209.

Zipf, George Kingsley (1949) Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley, Cambridge. https://pure.mpg.de/rest/items/item_2407822_4/component/file_2562959/content.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

The data used for this research are also available via CQPweb Lancaster and can be accessed at https://cqpweb.lancs.ac.uk/lccindonesianv3/. Concordance and frequency list are accessible at http://tinyurl.com/animal-idiom-id

All analytical tools available in CQPweb Lancaster were used in this study.

Ethical Approval

Ethical approval was not required for this study.

Informed Consent

No human or animal participants were involved in this research. The data used are open-access and licensed under CC BY-NC 4.0 (Creative Commons Attribution–NonCommercial 4.0).

Author Contributions

Prihantoro is the corresponding author (correspondence to: prihantoro@live.undip.ac.id). He was responsible for the research design, methodological framework, data analysis, project supervision, manuscript drafting, and revision. He wrote Section 3 (overall) and Section 4.1.

Elizabeth Ika Hesti Aprilia Nindia Rini was responsible for data analysis and manuscript writing. She wrote the Abstract (overall), Section 1 (overall), and Section 4.2.

Dian Annisa Nur Ridha was responsible for data analysis and manuscript writing. She wrote Section 2 (overall), Section 4.3, and Section 5.

Tanjun Liu was responsible for data extraction, technical support, data analysis, and manuscript writing. She wrote Section 4.4.

---

[i] Corresponding author: ***(***@live.*****.ac.id)
[ii] https://cqpweb.lancs.ac.uk/lccindonesianv3
[iii] https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fidtenten24_tt2