Journal of Japanese Language Education and Linguistics
Volume 9 No. 1, 2025, 49-64
e-ISSN: 2615-0840 p-ISSN: 2597-5277
DOI: https://doi.org/10.18196/jjlel.v9i1.25371
https://journal.umy.ac.id/index.php/jjlel/issue/view/1142

Proposing a Statistical Approach to Idiomatics Phrase Research

Eishi Hirose*, Vinky Irawan, Robby Yussac Tallar
Japan Center, Universitas Kristen Maranatha, Indonesia

***Corresponding author email:** japan.center@maranatha.edu

Submitted: 22 January 2025, Revised: 19 February 2025, Accepted: 26 February 2025

==

Abstract

This study explores the limitations of traditional idiom studies and proposes a new approach based on statistical methods. Conventional research has often linked idioms simplistically to society and culture, but quantitative linguistics can identify flaws in such arguments and conclusions. This study employed quantitative lexical analysis to clarify the meanings associated with idioms, as well as the underlying linguistic systems and statistical trends governing their usage. By adopting a macro-level and dynamic perspective, this approach provides new insights into the relationship between language culture and idioms, yielding significant findings compared to previous studies.

Keywords: Idiom studies, Quantitative linguistics, Quantitative lexical analysis, Linguistic systems, Language culture

INTRODUCTION

Discussions on what language represents have a long history that dates back to Ferdinand de Saussure’s concept of linguistic arbitrariness and the segmentation (boundaries) created by language. On the other hand, debates about how language interacts with elements of human experience, such as thought and culture—referred to here as the extralinguistic world—have ancient roots. Notably, the Sapir-Whorf hypothesis posits that the influence of language on cognition varies depending on cultural and linguistic backgrounds.

Saussure’s theory has faced criticism, often citing examples like onomatopoeia and affixes. However, counterarguments based on the concept of motivated relationships also exist. In Japan, Tokieda Motoki's "Theory of Language Processes" argues that psychological processes in humans influence language comprehension and generation.

Conversely, the Sapir-Whorf hypothesis has been criticized by scholars such as Popper, K., Avram Noam Chomsky, Steven Arthur Pinker, Malotki Ekkehart and Berlin Brent & Kay Paul (Mizuno, 2022), who largely rejected the strong hypothesis, though the weaker version (Linguistic Relativity Theory) remains plausible.

The ongoing debate surrounding these philosophical theories of language has highlighted that the relationship between language and the extralinguistic world remains unresolved. Although a total disconnect has not been definitively established, some relationship—albeit undefined—exists, especially when considering interpretative frameworks such as motivated relationships. However, this connection should not be overstated to imply an excessive influence on human thought, society, or culture. Above all, it is important to emphasize the perspective of linguistics: linguistic discussions should address linguistic problems and seek linguistic causes.

Language is often described as a communication tool; however, this term can be misleading. Unlike a physical "tool," language plays a unique role in the extralinguistic world. It connects to values, identity, emotions, and creativity. Therefore, it is more accurate to refer to language as a medium of communication rather than a simple tool. Because we communicate through this linguistic medium, it inevitably influences us. For instance, the frustration of not being able to fully convey emotions or thoughts is an effect of this medium.

This study addresses certain assumptions and hasty conclusions about idiom research. While idioms are undeniably connected to the extralinguistic world, it is necessary to consider the relationship between language and the extralinguistic world. At a minimum, the strong Sapir-Whorf hypothesis—which suggests that the extralinguistic world strongly determines language—should be avoided in both premises and conclusions. This will be further discussed in the research background.

The objectives of this research are as follows: (1) Clarifying problematic aspects of idiom research, (2) Proposing new methodologies and approaches, (3) Introduce new perspectives and insights from these methodologies.

By achieving these objectives, this paper aims to deepen our understanding of how idioms relate to the extralinguistic world from a linguistic perspective, particularly to caution against oversimplified conclusions drawn in previous research.

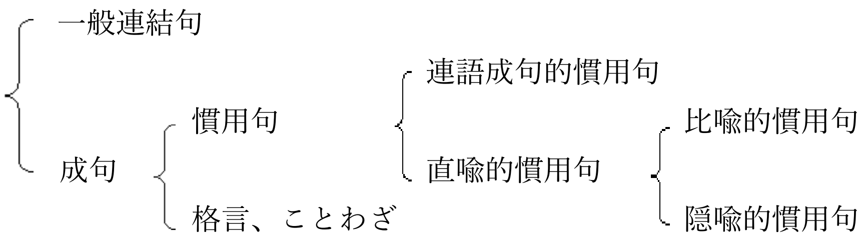
Background and Issues in Idiom Research

Problem of Defining Idiom

In Indonesia, Chaer (1984) defines idioms as "language units whose meanings cannot be derived from general grammatical rules or predicted from the lexical meanings of their constituent elements." Other definitions, such as those by Chaedar (1985), Kridalaksana (1993), and Pratiwi (2018), highlight the distinction between the combined meaning of an idiom and the literal meanings of its components. They generally describe idioms as cohesive language units that convey special meanings that are distinct from the meanings of their constituent elements.

In Japan, idioms are "expressions in which two or more words always combine in the same way to form a specific meaning. This is also referred to as idiomatic phrases or idioms" (Nihon Kokugo Daijiten, Second Edition). Scholars like Hiroshi Miyachi (1982) and Ryo Morita (1985) are frequently cited in this field.

From these definitions, it appears that the concept of idioms is consistent across languages, with little variation between Indonesian and Japanese. However, Miyachi's (1982) classification provides additional insight:



Although the details of these classifications are beyond the scope of this paper, subsequent studies have further refined the distinctions between idioms and other fixed expressions (Hirose, 2023). The term "idiom" in many countries often corresponds to Japanese “seiku” (set phrases). In Japanese, “kan'youku” (idioms) denote a subset of “seiku”. Hence, what is labeled as an idiom in Indonesian often aligns with “seiku” and does not strictly correspond to “kan'youku” in Japanese. However, a precise definition of “kan'youku” in Japanese remains elusive.

Simplistic Connections to the Extralinguistic World in Bilingual Comparative Studies

As noted in Section 1, many idiom studies relied on simplistic assumptions about connections between idioms and the extralinguistic world. This is particularly evident in comparative studies of idioms between two languages, which often focus on idioms involving body-related vocabulary.

These comparisons categorize idioms into three groups: Idioms consisting of identical lexical items with identical meanings; Idioms composed of identical lexical items but with different meanings; and Idioms consist of different lexical items with identical meanings.

While these studies have provided valuable insights and practical applications, especially in the context of language education, there is a pressing need to advance to the next level of research.

Idioms, with their unique forms of expression, are closely tied to specific cultures and languages, drawing interdisciplinary attention. In the context of language education, idioms have become a focal point because of factors such as linguistic habits, age-related cognitive differences, and the presence or absence of prior knowledge. Similarly, in academic research, idioms are examined from diverse perspectives, including literature, linguistics, and cultural studies. Linguistically, idiom research emphasizes aspects such as grammar, semantics, lexical composition, combinatorial strength, metaphorical characteristics, and language acquisition.

The following are the key areas of existing idiom research: (1) Metaphorical Studies on Idioms: Many idioms are derived from metaphorical meanings. Understanding the linguistic mechanisms underlying these metaphors is crucial. (2) Grammatical Studies on Idioms: These focus on the syntactic relationships among constituent elements. In Japanese, idioms often involve auxiliary words that provide unique insights, including their grammatical roles. (3) Degrees of Idiomaticity: Research in this area examines the variability of idioms from highly fixed expressions to those with greater flexibility. (4) Cognitive Linguistic Studies on Idioms: Cognitive linguistics explores how language comprehension and production relate to human cognition. Although the weak Sapir-Whorf hypothesis is weak, the two approaches differ in their methods and focal points. (5) Corpus-Based Idiom Studies: using large-scale data, this approach uncovers statistical tendencies and distinctive features of idioms, providing objective, evidence-based insights.

Other studies highlight the cultural dimensions of idioms, suggesting that factors such as geography, climate, lifestyle, traditional modes of transportation, religion, and cultural practices contribute to idiom formation in Indonesian. While these claims are valid, they also highlight the difficulty of analyzing the relationship between cultural context and linguistic expressions through linguistic features such as grammar and phonology. To explore this relationship more deeply, comparative analyses across languages and anthropological approaches are essential.

In summary, while the relationship between idioms and culture is undeniable, it is crucial to avoid simplistic conclusions that rely on predetermined assumptions. Researchers must critically reflect on and move beyond superficial interpretations.

Additionally, the researchers want to emphasize one more point in this paper. Qualitative research has long been prioritized in linguistic studies, and in the home country of Japan, quantitative research has often been criticized as simply "counting" (Maekawa, 2016). For instance, some scholars argue that vocabulary such as "earth," "soil," and "water" appears relatively frequently in Indonesian idioms because these words often appear in their holy texts (such as the Quran and the Bible). However, the mere presence of words from these texts in idioms does not necessarily establish causal relationships. The formation of idiomatic expressions is likely influenced by multiple factors, with the influence of the Quran being only one possible factor. Although words appearing in the Quran can sometimes be idiomatic expressions, this alone does not provide a sufficient explanation. If one argues that the Quran influences idiomatic expression formation, it is essential to demonstrate the extent and significance of this influence through research. It is crucial to consider not just the "source match" but also how these expressions are actually used within the language community, as well as their relation to other factors. Quantitative linguistic research, such as the approach taken in this paper, aims to clarify the relationship between idioms and specific influencing factors and to measure the extent of their influence.

Moreover, corpus-based quantitative linguistics is now widely used in fields such as English linguistics, and its significance is increasingly recognized. A notable achievement of quantitative research is the development of machine translation systems. Traditional rule-based machine translation (RBMT) involves designing grammar rules and transformation rules for human translation. However, this method struggled with handling exceptions and achieving fully accurate translations. With the advent of statistical machine translation (SMT) and the subsequent evolution to neural machine translation (NMT), AI has been used to automatically adjust translation patterns based on large datasets, leading to more accurate translations. In this way, quantitative linguistics has produced practical results through methods such as morphological analysis, large-scale text data analysis, modeling, language processing technologies, machine learning (ML), and deep learning (DL).

This paper does not criticize theoretical research. Theoretical studies, based on hypotheses and inferences, hold significant value. In contrast, quantitative research explores the truth through an understanding of the present situation and the meaning of numerical data, addressing real-world issues. "Quantification" is a scientific method, and when applied to linguistics, it requires analysis based on deep linguistic knowledge rather than merely focusing on the quantity of numbers. Thus, while quantitative linguistics deals with "quantity," it is not qualitatively inferior to theoretical research.

METHOD

This study investigates idiomatic expressions in Indonesian and Japanese that incorporate natural vocabulary, compiling a comparative list of how such vocabulary is utilized in idiomatic contexts. Unlike previous studies, this research employs statistical methods to provide scientific evidence, focusing on idiomatic expressions related to biological elements in both languages.

To clarify the patterns and characteristics of idiomatic expressions involving natural vocabulary in Indonesian and Japanese, the following steps were undertaken:

Data Collection

Utilization of Idiom Dictionaries:

Data were collected from the 2024 edition of the Weblio Japanese Idiom Dictionary and the 2010 edition of Kamus Peribahasa Indonesia*[7].

- Extraction of Natural Vocabulary:

Idiomatic expressions containing natural vocabulary were identified in both languages and compiled into a list.

Classification and Organization

Category Classification. Idiomatic expressions involving natural vocabulary were categorized into five main categories: “Celestial Bodies,” “Earth and Sky,” “Water,” “Air,” and “Others.”

Semantic Classification. Based on the Bunrui Goi Hyo (Japanese Word Classification Table) by the National Institute for Japanese Language and Linguistics, idiomatic expressions were classified semantically and organized in Excel sheets.

Statistical Analysis

Frequency Calculation. The frequency of idiomatic expressions was calculated for each category and semantic classification.

Comparative Analysis. The frequencies and patterns of idiomatic expressions involving natural vocabulary in Indonesian and Japanese were compared statistically.

Interpretation of Results

Clarification of Features. The characteristics of idiomatic expressions involving natural vocabulary in both languages were clarified. In addition, the similarities and differences were interpreted comprehensively.

Cultural Background Examination. The cultural backgrounds and contexts reflected in idiomatic expressions involving natural vocabulary were examined and correlated with the statistical findings.

Presentation of Scientific Evidence

Application of Statistical Methods. The collected data were processed statistically, and the comparative results of idiomatic expressions in Indonesian and Japanese were presented based on scientific evidence.

Through this research methodology and approach, the study aims to statistically elucidate the features and patterns of idiomatic expressions involving natural vocabulary in Indonesian and Japanese. By providing a fresh perspective, this research seeks to contribute to the comparative study of idiomatic expressions between the two languages, promoting a deeper understanding and setting a new direction for future studies.

RESULTS AND DISCUSSION

Table 1. Results of Natural Vocabulary Idiomatic Expressions Survey

Classification	Indonesian Idioms		Japanese Idioms	
1. Celestial Bodies	5(2)	4.6%	24(0)	6.7%
2. Earth and Land	39(9)	36.1%	128(22)	35.6%
3. Water	40(7)	37.0%	115(25)	31.9%
4. Air	20(3)	18.5%	69(15)	19.2%
5. Others	4(2)	3.7%	24(6)	6.7%
TOTAL	**108**	**100%**	**360(68)**	**100%**

Table 1 shows the results of the survey on idiomatic expressions incorporating natural vocabulary. Both Indonesian and Japanese idioms frequently utilize the categories "Earth and Land" and "Water," followed by "Air." The numbers in parentheses indicate overlapping idiomatic expressions, such as "Air dan Minyak" ("Water and Oil"), where "Air" (Water) is counted under "Water" and "Minyak" (Oil) under "Others."

Table 2 summarizes the classification results based on the intermediate categories in the Bunrui Goi Hyo (Classification Vocabulary Table). Categories such as "Mind," "Action," "Aspect," and "Life" are prominent, while there are noticeable instances with zero or only one example. To identify overarching trends or systems within natural vocabulary idioms, it is necessary to temporarily group into smaller categories for analysis, rather than ignoring them.

Numerical characteristics were used to perform cluster analysis for statistical classification. Subcategories under "2 Human Activity Actors," ."4 Products and Tools," and ."5 Natural Objects and Phenomena" were grouped together. Although the category "Substance" contained only two examples in Japanese idioms compared to 13 in Indonesian, these were temporarily consolidated for analysis.

The elbow method was applied to determine the optimal number of clusters for ".1 Abstract Relationships" and ".3 Human Activities," resulting in four clusters for the former and three clusters for the latter. The results of non-hierarchical clustering are shown in Table 3 and Table 4.

Table 2. Results of Intermediate Classification for Natural Vocabulary Idioms

Classification	Indonesian Idioms	Japanese Idioms
.1 Abstract Relationships	**48(10)**	**142(31)**
Space	4(0)	0(0)
Form	0(0)	1(0)
Action	14(1)	56(14)
Matter	0(0)	2(0)
Time	2(1)	19(5)
Existence	4(1)	15(2)
Aspects	12(3)	11(2)
Quantity	2(1)	6(1)
Force	1(0)	5(2)
Category	9(3)	27(5)
.2 Agents of Human Activities	**5(2)**	**5(0)**
Public and Private	2(1)	0(0)
Society	0(0)	3(0)
Humans	2(1)	0(0)
Individuals	1(0)	2(0)
.3 Human Activities	**43(10)**	**177(34)**
Economy	3(0)	10(1)
Arts	0(0)	1(0)
Language	1(0)	12(2)
Interaction	1(0)	17(3)
Action	11(4)	7(2)
Enterprise	0(0)	6(2)
Mind	17(4)	81(14)
Life	5(1)	25(5)
Treatment	5(1)	18(5)
.4 Products and Tools	**7(1)**	**2(0)**
Materials	2(1)	0(0)
Food	3(0)	2(0)
Land Use	1(0)	0(0)
Tools	1(0)	0(0)
.5 Natural Objects and Phenomena	**5(0)**	**34(3)**
Composition	0(0)	1(0)
Nature	0(0)	3(0)
Plants	0(0)	1(0)
Body	1(0)	7(1)
Life	1(0)	5(0)
Heaven and Earth	1(0)	4(1)
Substance	2(0)	13(1)
TOTAL	**108(23)**	**360(68)**

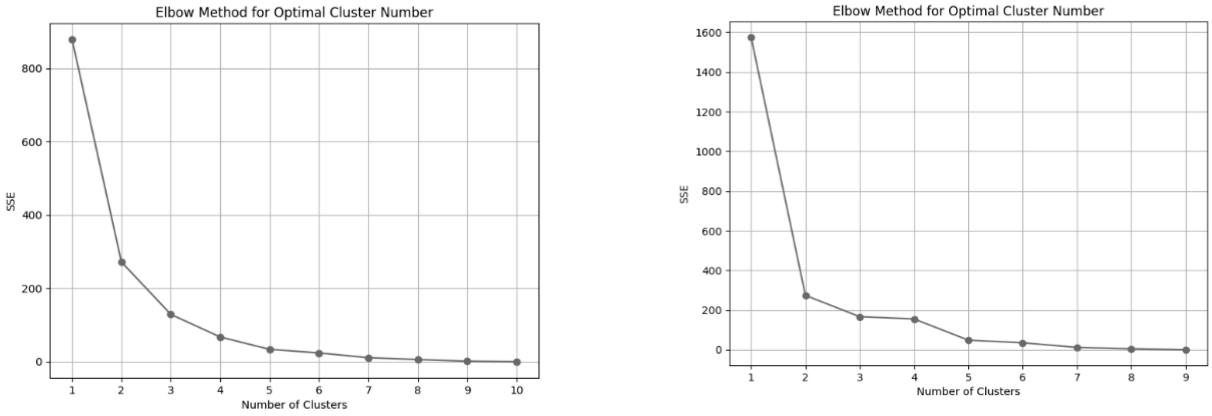


Figure 1. Left: Elbow Method for Abstract Relationships; Right: Elbow Method for Human Activities

Table 3. Abstract Relationships Clusters

Category	**Action**	**Time and Existence**	**Other Abstract Relations**
Category	Action	Time	Space
		Existence	Form
			Matter
			Aspect
			Quantity
			Force

Table 4. Human Activities Clusters

Mind	**Life and Interaction**	**Other Activities**
Mind	Interaction	Economy
	Life	Art
		Language
		Action
		Enterprise
		Treatment

Table 5 is a cross-tabulation of clusters and natural vocabulary derived from Table 3 and Table 4. It highlights the prominence of the "Mind" category across all clusters. Additionally, "Action" and "Life and Interaction" are frequently observed, differing slightly from the characteristics of idiomatic expressions involving physical vocabulary. These distinctions are clarified through statistical analysis.

Figure 2 presents the results of correspondence analysis on Table 5. The first axis explains 38.78% of the total variance, while the second axis accounts for 22.04%, with both axes contributing to 60.82% of the cumulative variance. The graph places "Action" and "Mind" at the center, indicating their strong association and their relationships with other concepts, such as "Category" and "Presence/Absence." On the other hand, "Life and Interaction" and "Celestial Bodies" are located farther apart, emphasizing their unique placement. "Products and Tools" are also distinctly isolated. Additionally, ".2 Human Activity Actors" is positioned near Indonesian categories "1. Celestial Bodies," "2. Earth and Land," and "3. Water."

Table 5. Cross-Tabulation of Clusters × Natural Vocabulary

Classification	**1. Benda langit**		**2. Land**		**3. Water**		**4. Air**		**5. Others**	
	Ind	**Jpn**	**Ind**	**Jpn**	**Ind**	**Jpn**	**Ind**	**Jpn**	**Ind**	**Jpn**
.1 Abstract Relationships	**2**	**12**	**18**	**44**	**14**	**52**	**10**	**22**	**4**	**12**
Action	0	3	4	17	6	21	4	12	0	3
Type	0	2	4	6	1	12	3	5	1	2
Time and Existence	0	6	4	12	2	10	0	2	0	4
Other Abstract Relations	2	1	6	9	5	9	3	3	3	3
.2 Agents of Human Activities	**1**	**0**	**1**	**3**	**2**	**2**	**1**	**0**	**0**	**0**
.3 Human Activities	**2**	**10**	**14**	**71**	**20**	**47**	**7**	**41**	**0**	**8**
Mind	2	6	5	34	7	21	3	17	0	3
Life and Interaction	0	11	1	12	2	18	3	0	1	0
Other Activities	0	3	8	26	11	14	1	6	0	5
.4 Products and Tools	**0**	**0**	**4**	**0**	**3**	**0**	**0**	**0**	**0**	**2**
.5 Natural Objects and Phenomena	**0**	**2**	**2**	**10**	**1**	**14**	**2**	**6**	**0**	**2**
TOTAL	**5**	**24**	**39**	**128**	**40**	**115**	**20**	**69**	**4**	**24**

In Table 1, we demonstrated that "Earth and Land" and "Water" were predominant in both Indonesian and Japanese idioms, followed by "Air." When examining why these categories dominate, one might wonder whether they can be clearly explained through cultural and social contexts. The issue lies in how humans conceptualize ideas and convey them through language. However, such considerations remain speculative. The table merely highlights objective facts.

To delve deeper, we conducted a more detailed analysis based on statistical evidence. By forming semantic categories grounded in numerical data, we objectively and comprehensively revealed the relationships with natural vocabulary.

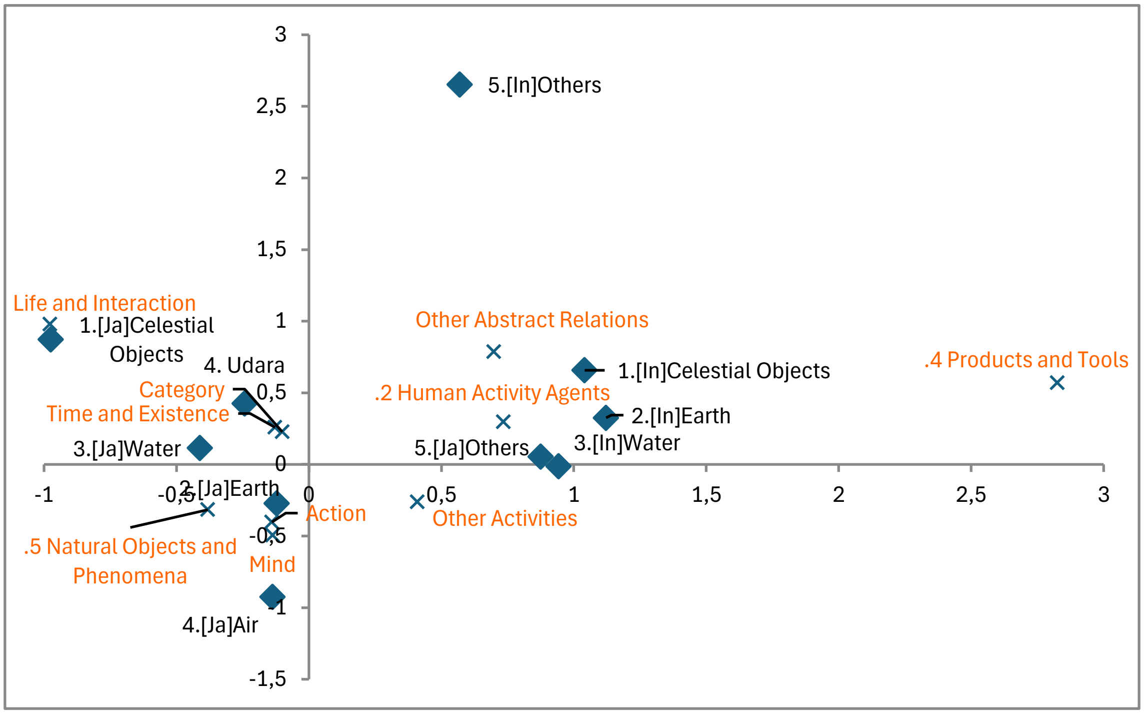


Figure 2. Correspondence Analysis of Clusters × Natural Vocabulary

The study found that the semantic category "Mind" was consistently prominent across all contexts. However, as Hirose (2024) stated, idioms fundamentally aim to express meanings related to human activities, particularly those within the "Mind" category. This aligns with the intrinsic nature of idioms and was confirmed in this study. Conversely, the high occurrence of "Action" and "Life and Interaction" differed slightly from the characteristics of idioms involving physical vocabulary, showcasing unique traits of natural vocabulary idioms.

The results of the correspondence analysis effectively illustrate the characteristics of the natural vocabulary idioms examined in this study. The central placement of "Action," "Mind," and other concepts (e.g., "Category" and "Presence/Absence") highlighted their interrelations. Categories with a high frequency of examples were concentrated, representing the core semantic features of natural vocabulary idioms. On the other hand, "Life and Interaction" and "Celestial Bodies" were positioned farther apart, indicating a potential connection between them. Similarly, the distant placement of "Products and Tools" reflected the scarcity of examples in this category.

Interestingly, "Human Activity Actors" was slightly distanced from the center and grouped with Indonesian idioms such as "1. Celestial Bodies," "2. Earth and Land," and "3. Water." This suggests a unique relationship between this semantic category and Indonesian idioms, a trait not observed in Japanese idioms.

Further exploration of these aspects is needed in future studies; however, due to space limitations, this paper concludes with the observations discussed so far.

CONCLUSION

This study aimed to compare Indonesian and Japanese natural vocabulary idioms based on statistical analysis, identifying their characteristics and patterns. We believe that this research has provided a fresh perspective on idiom studies, emphasizing the significance of numerical insights and holistic viewpoints. Unlike traditional methods that focus on individual analyses, our approach introduced a broader methodology.

Additionally, corpus-based collocation analysis provided insights into the connections between natural vocabulary and other words. This approach made it possible to explore the lexical relationships underlying the creation and use of idioms in greater depth.

The statistical approach to idiom research proposed in this paper, grounded in lexical analysis, introduced new perspectives to existing studies and allowed for a macro and dynamic understanding of idioms. However, while we succeeded in uncovering some relationships between meanings, further detailed examinations of idiomatic systems and trends remain as tasks for future research[9].

Future studies should focus on expanding data collection and analysis to strengthen statistical evidence. They should also employ corpus analysis to further unravel the systems and tendencies of natural vocabulary idioms. We hope this research serves as a step Towards new directions in idiom studies and contributes to the fields of linguistic and cultural studies.

*This paper is part of a collaborative research project with Mr. Robby titled "A Multidimensional Study of Water in Indonesia."

REFERENCES

Chaedar, A. (1985). *Beberapa mazhab dan dikotomi teori linguistic*. Bandung: Angkasa

Chaer, A. (1984). *Kamus idiom dalam Bahasa Indonesia*. Flores: Nusa Indah

Hirose, E. (2023). A comparative analysis of Japanese and Indonesian body idioms using structural analysis by mid-category and meaning field. *Goi Kenkyu, 20*.

Hirose, E. (2024). Towards new developments in comparative studies of Indonesian and Japanese idioms: Focusing on idioms related to living organisms. *ASJI International Symposium 2024 Proceedings*

Kridalaksana, H. (1993). *Dictionary of linguistic terms in Indonesian with English equivalent and Indonesian definitions*. Jakarta: Gramedia Pustaka Utama

Maekawa, K. (2016). Virtual lecture: Introduction to linguistic resource studies. *Japanese Linguistics, 35*(13), 2-11.

Miyachi, H. (1982). *The meaning and usage of idiomatic expressions*. Tokyo: Meiji Shoin

Mizuno, M. (2022). Reconsidering "language and thought" – Towards building a cognitive model of written Japanese. *Meiji University Journal of Liberal Arts, 562*. 141-164.

Morita, Y. (1985). Verb idiomatic expressions. *Nihongo-gaku, 4*(1), 37-44.

Pratiwi, H. A. (2018). Idiom pada rubrik berita nasional kategori pendidikan dalam CnnIndonesia.com. *Jurnal Pena Literasi, 1(*1), 1-16.

Strauss, S. (2013). *Discourse analysis: Putting our worlds into words*. Routledge